JOURNAL OF FORENSIC SCIENCES

AMERICAN ACADEMY OF FORENSIC SCIENCES
1948

VOLUME 65
NUMBER 6
NOVEMBER 2020

JFSCAS 65 (6) 1801–2202 (2020)

www.jfsjournal.com

American Academy of Forensic Sciences

410 North 21st Street
Colorado Springs, CO 80904-2798
Phone: 719-636-1100
Fax: 719-636-1993
Website: www.aafs.org

Journal of Forensic Sciences

Michael A. Peat, Ph.D., Editor-in-Chief
Journal of Forensic Sciences
6700 Woodlands Parkway, Suite 230–308
The Woodlands, TX 77381
Email: jfs.editor@att.net

Disclaimer: The Publisher, American Academy of Forensic Sciences and Editors cannot be held responsible for errors or any consequences arising from the use of information contained in this journal; the views and opinions expressed do not necessarily reflect those of the Publisher, American Academy of Forensic Sciences and Editors, neither does the publication of advertisements constitute any endorsement by the Publisher, American Academy of Forensic Sciences and Editors of the products advertised.

For submission instructions, subscription and all other information visit: www.jfsjournal.com.

The Journal of Forensic Sciences accepts articles for Open Access publication. Please visit https://authorservices.wiley.com/author-resources/Journal-Authors/open-access/onlineopen.html for further information about OnlineOpen.

Journal of Forensic Sciences

Vol. 65, No. 6
November 2020
CODEN JFSCA

Contents

Papers

Technical Notes

Case Reports

Book Review

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14528
Available online at: onlinelibrary.wiley.com

PAPER

ANTHROPOLOGY

Andrea Palmiotto,[1,2] Ph.D.; Traci L. Van Deest,[3] Ph.D.; Kyle McCormick,[2] Ph.D.; and Laurel Freas,[2] Ph.D.

Blast and Aircraft Crash Trauma: A Selection of WWII Cases from the Defense POW/MIA Accounting Agency Laboratory*'†

ABSTRACT: This study examines patterns of skeletal trauma in propeller-driven aircraft crashes and blast-related ground loss incidents from WWII. Specifically, descriptions and criteria used to characterize aircraft deceleration- versus blast-related skeletal injuries are examined from 35 recently identified forensic anthropology cases to determine possible diagnostic traits and characterize skeletal trauma associated with these events. Among these cases, blast trauma is more localized within the skeleton and is associated with one or few primary directions of force. It is recommended that analysts differentiate between secondary and nonspecific blast trauma categories. Conversely, aircraft crash deceleration trauma is more widespread throughout the skeleton, with torsional fractures and injuries occurring from multiple or indeterminate directions. These traits reflect factors such as more complex loading environments than is seen in blast events. Two case studies are presented in detail to further illustrate differences in aircraft crash and blast-related incidents. Both studies emphasize consideration of the body as a whole unit to facilitate interpretations. While the cases presented herein result from historic war-related casualties that characterize the Defense POW/MIA Accounting Agency's (DPAA) casework, these skeletal cases provide guidelines more appropriate than clinically derived criteria developed through assessments of soft tissue injuries. These guidelines can be used by anthropologists and pathologists working with skeletal remain from mass disasters and other complex contexts, as well as provide avenues for future research.

KEYWORDS: forensic anthropology, skeletal trauma, aircraft crash, blast injuries, historic aircraft, World War II, blunt force trauma, projectile trauma

Trauma assessment is a complicated endeavor that provides insight into the circumstances of death, perimortem interval, and life history of an individual. Because of their work with other forensic experts and in medico-legal situations, forensic anthropologists must demonstrate competency and accuracy in trauma analyses. As other researchers have noted, methods for trauma analyses require validation to ensure quality and scientific rigor of work (1,2).

Forensic anthropologists at the Defense POW/MIA Accounting Agency (DPAA) have the unique opportunity to conduct analyses on a wide array of skeletal cases associated with numerous past U.S. conflicts. While these types of cases may not be examined on such a large scale anywhere else in the world (3), the patterns associated with blast trauma and propeller-driven aircraft crashes are still applicable to many modern contexts of armed conflict and medico-legal casework. At DPAA, patterns of perimortem trauma in a skeletal case can later be compared to the circumstances of the documented historical loss incident for identified individuals, which permits a retrospective validation of the trauma analyses. In this study, we examine and assess patterns of bodily trauma observed in historic blast- and propeller-driven aircraft crash incidents associated with World War II (WWII). For consistency and clarity, all references to aircraft crash trauma in the remainder of the article specifically relate to propeller-driven aircraft. Often, these types of incidents result in more extensive or complicated forms of bodily trauma than other more limited blunt force trauma events seen in traditional forensic anthropology and medico-legal contexts (4,5).

Blast and aircraft crash incidents are influenced by various extrinsic and intrinsic forces, which lead to significant variation in the expression of skeletal trauma, and therefore can manifest in similar patterns of injury. However, limited research has comparatively examined the associated trauma patterning related to these events, and no well-defined criteria exist for analysis and interpretation.

Bone Biomechanics and Skeletal Trauma

Bone is a composite material consisting primarily of organic (collagen and noncollagenous proteins) and nonorganic

[1] Department of Anthropology, Indiana University of Pennsylvania, 441 North Walk, 15705, Indiana, Pennsylvania.

[2] Defense POW/MIA Accounting Agency Laboratory, 590 Moffet Street, Bldg 4077, 96853, JBPHH, Hawaii.

[3] Defense POW/MIA Accounting Agency Laboratory, 106 Peacekeeper Dr, Bldg 301, 68113, Offutt AFB, Nebraska.

Corresponding author: Andrea Palmiotto, Ph.D. E-mail: apalmiot@iup.edu

*Presented at the 71st Annual Scientific Meeting of the American Academy of Forensic Sciences, February 18–23, 2019, in Baltimore, MD.

†The views herein are those of the authors and do not represent those of the Defense POW/MIA Accounting Agency, Department of Defense, or United States Government.

[Correction added 11 August 2020. The Highlights section was removed.]

Received 23 Jan. 2020; and in revised form 8 Mar. 2020, 11 Apr. 2020, 8 June 2020; accepted 10 July 2020.

(hydroxyapatite mineral) components. Given its composition, bone is ideally suited to the normal stresses of daily biomechanical needs of stability, weight bearing, and flexion, to both support and move the human body. The two-phase structure creates viscoelasticity, a variable response to forces that exceed those acting on the skeleton during most normal daily activities (6). Multiple references extensively cover skeletal biomechanics and should be utilized for a greater understanding of this complex and essential foundation to skeletal trauma analysis (6–10).

At a basic level, skeletal trauma is the result of energy transfer that applies stress to bone beyond its ability to deform in strain, resulting in failure. Important extrinsic concepts for understanding osseous response to energy transfer, or applied force, include the velocity (speed and direction) and mass of the impacting object, as well as the surface area over which the force is applied. Furthermore, skeletal trauma is rarely the result of one force vector, but rather is an interplay of compression, tension, torsional, and shearing forces (5,11). Briefly, compression results in a reduction of bone dimensions while tension creates an elongation. Torsional forces create rotation around the long axis of the bone, and shear is the result of opposing forces acting parallel to the plane of the cross section of the bone. These factors are each distributed over a continuum and work in concert to fracture bone, making it difficult (though not impossible) to work from the skeletal trauma to identify the causal factors.

To simplify interpretation, forensic anthropologists typically bin skeletal trauma into the general categories of blunt-, projectile-, and sharp-force injuries. These categories of injury broadly reflect the bone's reaction to general ranges of energy transfer and the surface area of impact. Much research has been conducted regarding the expression of these events on both cranial and postcranial elements (1,2,5,11–16).

Perimortem fractures associated with blunt force trauma are typically identified and classified in terms of their orientation, using terms such as transverse, oblique, and spiral, among others (5). Although these terms may provide insight into the interplay of forces or directionality, they also allow for detailed descriptions and visualization of trauma appearances. The energy involved in blunt force trauma can be quite variable. Direct impacts from blunt instruments often reflect low-energy impacts, typically discussed in terms of low-velocity and slow-loading forces between a body and a blunt object (11). With slowly loaded blunt forces, bone has the ability to absorb stress and flex or deform in response. Generally, this results in a smaller number of fractures and the presence of plastic deformation wherein skeletal elements are permanently deformed due to excessive, sustained, slow-loading forces beyond the elastic limit of the bone. Other traumatic events often classified as blunt force injuries, such as a pressure wave from a blast, involve a high amount of energy, applied rapidly, but across a broad surface area. Here, the number of fractures may be few but can be quite severe, such as a complete fracture of the femoral diaphysis (5,17). Plastic deformation would be rare in such instances of more rapid loading and fracture responses.

A significant characteristic that differentiates blunt force from projectile trauma is the surface area of energy transfer (1,2). Blunt force trauma generally involves a larger surface area compared to the more focal impact of projectile trauma (5,11). In addition, projectile injuries are almost exclusively regarded as high-energy impacts, often referred to as high-velocity forces, where bone responds in a brittle fashion, with little flexion or permanent deformation at the site of impact (12). Projectile trauma is thus characterized by the combination of higher energy and more focused areas of impact, generally resulting in entrance and exit defects, radiating fractures, concentric and heaving fractures in the cranium, and usually less plastic deformation. These effects are illustrated in instances where the body is hit by bullets, pellets, or shrapnel (17).

Trauma Associated with Aircraft Crash and Blast Incidents

Understanding these basic biomechanical properties and differences in trauma mechanisms is important because skeletal injuries from aircraft and blast-related events reflect a complicated interplay of vector of force and the area over which that force is applied. Several researchers have discussed skeletal injuries due to aircraft crash or blast events alone, but rarely in comparison with one another (3,18–21).

As a point of clarification, we refer to high- and low-energy impacts when discussing aircraft crashes and blast events, rather than rapid- and slow-loading forces. The forces at work in both blast events and aircraft crashes can be considered rapid loading, particularly in comparison to traumatic forces seen in more typical forensic contexts. As such the terminology of high- and low-energy impacts is used throughout, as it is the most appropriate and informative in the context of this study.

Rapid deceleration events, such as aircraft crashes, vehicular accidents, and vertical falls, have been more extensively studied and reported in the literature than blast events. These deceleration events result in extensive blunt force trauma due to an object losing energy and inertia as it impacts a more stationary object. Vehicular accidents and vertical falls represent more common deceleration events and as such are more prevalent than aircraft crashes in anthropological and clinical medical literature (e.g., 5,15,22–30).

Blast trauma injuries related to high-energy explosions are often seen in association with explosive ordnance like grenades, landmines, mortars, and bombs. The explosions created by these devices are characterized by over-pressurized blast waves that radiate from the blast site upon detonation, resulting in air compression and a momentary, but significant, shift in atmospheric pressure. The resulting injuries arising from blast events are typically categorized as primary, secondary, tertiary, and quaternary trauma (17,18,31,32).

Primary blast trauma is associated with the blast wave and often is the cause of bodily amputation (17,18,31,32). Secondary blast trauma is associated with projectile trauma from shrapnel or fragmented debris thrown outward at high velocities by the ordnance. Tertiary blast trauma is associated with acceleration and deceleration trauma, from the pressure wave, resulting in impacts between the body and other objects or structures. This type of traumatic injury results in blunt force trauma and may appear similar to an aircraft crash event. Quaternary blast trauma is a catch-all term for other effects not described above, such as burns to the body, and is unlikely to result in identifiable skeletal trauma (31,32). Therefore, quaternary trauma is not considered relevant in this study.

It is important to recognize that not all of the four blast categories result in projectile trauma, and thus, not all blast victims will display projectile injuries. Additionally, bodies may be impacted by multiple categories of blast trauma, resulting in complicated trauma patterns on skeletal remains that may preclude identification of a single blast trauma category. Furthermore, clinical researchers typically distinguish blast trauma based on fracture patterns and proximity to shrapnel or blast waves (31,32). These studies often emphasize impacts to a fleshed body, rather than to skeletal remains. Therefore, the utility of these categories for skeletal trauma may not translate well for anthropological cases.

A volume dedicated to armed conflict and human rights situations by Kimmerle and Baraybar (17) provides context and case studies for blast trauma. In a case study of blast-related skeletal trauma associated with the Korean War, Willits et al. (21) rely on the presence of projectile defects or embedded metal as indicators for blast trauma. Christensen et al. (18) conducted blast experiments on porcine cadavers and suggested that inflicted trauma depends on the position and location of the body in relation to the blast location. Additionally, Christensen and Smith (20) showed that blast events can result in diagnostic inverted butterfly fractures to the ribs.

However, in general, limited anthropological research has been completed on comparisons between aircraft crash and blast-related trauma. In effort to distinguish between aircraft crash and blast trauma, Banks (33) assessed 52 DPAA cases where historical documents suggested that individuals were involved in these types of incidents. These cases spanned WWII, the Korean War, and the Vietnam Conflict and represented analytical reports completed during a 20-year range. Over this time period, analytical requirements and report-writing standards have changed significantly within the DPAA and its predecessor organizations. Therefore, these cases represent a wide variety of methods used to make biological determinations and trauma analyses.

Banks (33) reviewed the final anthropological reports. She assessed the DPAA blast cases using the blast trauma criteria presented in anthropological and clinical literature. She agreed that the presence of projectile trauma is one of the best ways to differentiate blast- from aircraft-related trauma. Additionally, she suggested that the directionality of impacts was not as useful in distinguishing trauma between these two events. Although she attempted to interpret additional information from photographs in the reports, Banks acknowledged that she did not have enough data to make substantial inferences. Furthermore, the different conflicts (and associated changes in ordnance and aircraft technologies) associated with the cases and differences in analytical methods and reporting standards may have impacted her ability to make comparable inferences.

Informal discussions with other DPAA analysts have suggested that certain regions of the body are more likely to be affected by blast-related or aircraft crash events. Analysts have acquired a suite of assumptions about trauma that has developed through years of anecdotal experience and independent casework. However, there is not uniform agreement among DPAA analysts regarding how these trauma events are expressed skeletally.

These expectations provide the foundation for this study. For example, analysts suggest that in some cases, the loss of lower limbs may reflect primary blast trauma amputation; however, incomplete recovery may also impact the presence of elements. There is disagreement whether fractured lower limbs reflect the bones of the feet absorbing impacts from landmines, deceleration events, or other diagnoses and if specific factors can be parsed from skeletal remains alone. Analysts also suggest that blast trauma will result in more predictable patterns of trauma than aircraft crash trauma, with the latter likely to result in more extensive, yet less patterned trauma throughout the skeleton. The wide variety of opinions and interpretations highlight both the complicated nature of trauma analysis, as well as the need to determine potential diagnostic traits and patterns associated with each type of incident.

All these studies and assumptions provide some criteria to assess aircraft crash and blast-related skeletal trauma but require additional research for validation. Based on this research, several hypotheses are explored. First, it is expected that aircraft crash trauma will display more widespread trauma throughout the skeleton than blast-related trauma. In aircraft crash trauma, associated bodily trauma may be characterized by extensive blunt force trauma throughout the skeletal remains, as well as complex and multiple loading vectors, with evidence of axial loading on the feet or other body parts that may have been braced for impact. On the other hand, blast-related skeletal trauma may be limited to fewer regions of the body than is seen in aircraft crash-related trauma due to proximity of the body (or certain body regions) to the explosives.

Next, it is expected that differential recovery and fragmentation of remains will be associated with aircraft crashes and blast events. Following the assumption that aircraft crashes will result in more diffuse skeletal trauma, increased fragmentation is also anticipated. Subsequently, it is assumed that a lower percentage of identifiable remains will be recovered and present for analysis. Following from an expectation of relatively limited regional impacts of blast trauma, comparatively less fragmentation and an overall higher recovery rate of remains in blast cases is anticipated. Furthermore, analytical capabilities, such as assessment of the biological profile, likely are impacted by differential patterns of skeletal fragmentation and recovery. It is expected that analysis of remains associated with aircraft crash incidents will be more limited than of those associated with blast events.

The amount of remains recovered and the degree of fragmentation reflects not only the incident type, but also a suite of factors such as the interval between the crash or blast incident and remains recovery, time since recovery, the manner of recovery, abiotic and biotic postmortem influences, intrinsic bone qualities, and the degree to which the remains have been handled. Despite these considerations, it is expected that the loss incident and associated trauma mechanisms will still have substantial influence on the condition of the remains at the time of analysis.

Additionally, it is expected that aircraft crash and blast-related trauma will display different mechanisms of trauma. Propeller-driven aircraft crashes may be associated with predominantly blunt force impacts. Blast-related injuries, on the other hand, represent a mixture of low- and high-energy events (e.g., blast waves, pressure waves, projectiles, acceleration, and deceleration) and occur due to interactions with explosives such as land mines or grenades. Therefore, blast incidents may be associated with a combination of blunt force impacts, projectile impacts, and embedded shrapnel and/or rust stains on the remains (21).

Previous research (33) suggests that assessment of the directionality of impacts does not facilitate interpretation of incident type. However, directionality was mentioned frequently in informal discussion among DPAA analysts. Therefore, we assess the utility of directionality and test the assumption that blast trauma will exhibit an overall pattern consistent with impact by forces from one primary direction based on proximity of the body to an explosive. In contrast, aircraft crash incidents likely reflect multiple or conflicting directions.

Regarding fractures, standard fracture terminology provides a way to describe and visualize common fracture types across multiple cases. These fracture types are sometimes associated with either blunt force or projectile trauma (5,12). For example, concentric cranial fractures are most common in projectile trauma (12), oblique fractures are common in blunt force trauma (5), while inverted butterfly fractures have been observed in experimental blast studies (20). Additionally, during informal conversation, analysts expressed ideas that shearing fractures are more common in aircraft crashes. Therefore, we expect to see possible correlations between incident type and observed fracture type(s).

In summary, blast- and aircraft-related incidents are influenced by various but similar extrinsic and intrinsic forces and may result in similar patterns of injury. However, limited research has been completed regarding differentiating between these types of trauma, and no well-defined criteria exist for identification and analysis. Therefore, this study breaks down perimortem trauma analyses from DPAA casework to examine patterns of trauma between aircraft crash and blast-related incidents.

Materials and Methods

Recently resolved WWII cases (n = 35), identified between 2015 and 2018, were examined that had established causes of death relating to blast (n = 11) or aircraft (n = 24) events. The reports were written by various anthropologists at DPAA who performed their analyses in the blind without knowledge of the specific details of the individual loss incident. Analysts were aware of the general context that the circumstances of death were believed to be within an armed conflict setting. Following forensic anthropological analyses, the causes of death were determined by a medical examiner based on the available information, including historical records and anthropological analyses. The cases are associated with European (n = 20) and Pacific Theater (n = 15) losses. The blast cases are from open-environment ground losses, while the aircraft cases involve propeller-driven aircraft.

Only the final anthropological reports were used to assess trauma, including written assessments and photographic evidence documented by the individual analyst. These reports were supplemented by the medical examiner summary and individual personnel files, which were used to confirm the type of traumatic incident for each case. All reports were previously subjected to a rigorous peer-review process and were completed using similar procedures. Most analyses were not confirmed through direct examination of the remains by the authors of this paper, though the sample does include cases that were originally analyzed and/or peer reviewed by one or more of the study authors.

One author recoded the data to minimize inter-observer error. All reports were recoded within a small window of time (several weeks) and double-checked after initial review to ensure consistent recoding among reports. The following information was recorded from each report: The estimated amount of recovered skeletal material, overall degree of fragmentation, analytical potential of the remains, and details of perimortem trauma. Some traits, such as recovery rate and fragmentation, were interpreted from the skeletal layout photographs.

Remains consisting of few elements and missing most major bones, such as the cranium and long bones were ranked as low recovery (25%); remains including some major bones, such as the cranium or long bones, but missing body regions, were ranked as moderate recovery (25–75%); and remains including most major bones and body regions were ranked as high recovery (>75%).

Fragmentation was estimated in a similar fashion and considers both perimortem and postmortem effects. Low fragmentation indicated nearly complete and intact bones; moderate fragmentation indicated the fragmentation of several major elements; and high fragmentation indicates that nearly all elements display fragmentation. When all four aspects of the biological profile were able to be assessed (sex, age, ancestry, and stature), analytical potential was described as complete. If any aspect could not be analyzed, then the analytical potential was described as limited.

Details of perimortem trauma were reported according to Kimmerle and Baraybar (17), to include the following: patterning of insults, directionality of insults, fracture types, and other perimortem traits. Additionally, because anthropological literature notes specific traits that may be observed in blast or aircraft crash trauma, the following characteristics were also documented when present: embedded metal, thermal alteration, and inverted butterfly fractures in the ribs (17,19–21,29). Identification of fracture types relied on the use of specific terminology (5,12,20) within the final forensic anthropology reports—that is, fracture types were not inferred based on descriptions or images within reports.

Tests of significances were performed to examine differences between aircraft crash and blast trauma for the following traits: recovery rate, relative fragmentation, analytical potential, trauma patterning, other perimortem traits, directionality, and trauma interpretation. Mann–Whitney U-tests were conducted for ordinal traits (e.g., relative fragmentation). Chi-square tests were conducted for nominal traits (e.g., trauma interpretation). Due to small sample sizes and the distribution of counts, chi-square p-values were computed via Monte Carlo simulation using 2000 replicates (34). All tests were evaluated against an alpha-level of 0.05, and all statistics were conducted in R (35).

Results

Among these cases, significant differences exist in the amount and overall condition of recovered remains and analytical potential (Tables 1–3). Half of aircraft (50%) and the majority of blast (90%) cases report at least 25% element recovery. Complete biological profiles were possible for some aircraft (29%) and most blast (90%) cases. These results support hypotheses regarding element recovery, fragmentation, and analytical potential.

Significant differences in the patterns of perimortem trauma in aircraft crash and blast cases (Tables 4–6) also support several of our hypotheses. Widespread trauma is found predominantly in aircraft cases (67%) and less frequently in blast cases (27%), in which trauma tends to be more localized. As for particular trauma mechanisms, blunt force trauma is observed in the majority of aircraft cases (66%) and a smaller percentage of blast cases (36%), while other blast cases (45%) display some evidence of projectile trauma. No definitive projectile trauma was observed in any of the aircraft crash cases. Indeterminate trauma was observed in some aircraft (29%) and blast (9%) cases, while the condition of one aircraft case precluded the definitive determination of perimortem trauma. Multiple or indeterminate directionality is evident in all aircraft (100%) and a number of blast (45%) cases, but one major directionality (and lack of conflicting directionality interpretations) is observed only in blast cases (55%).

TABLE 1—*Estimated recovery rate of skeletal materials per case.**

Estimated Recovery Rate	Aircraft (n)	Aircraft (%)	Blast (n)	Blast (%)
Low (<25%)	12	50.00	1	9.09
Moderate (25–75%)	9	37.50	4	36.36
High (>75%)	3	12.50	6	54.55
Total	24	100.00	11	100.00

*W = 57, p-value < 0.01.

TABLE 2—*Relative amount of fragmentation observed in each case.**

Relative Fragmentation	Aircraft (*n*)	Aircraft (%)	Blast (*n*)	Blast (%)
Low (<25%)	1	4.17	5	45.45
Moderate (25–75%)	3	12.50	4	36.36
High (>75%)	20	83.33	2	18.18
Total	24	100.00	11	100.00

*W = 223.5, p-value < 0.01.

TABLE 3—*Relative biological profile assessment capabilities per case.**

Analytical Capabilities	Aircraft (*n*)	Aircraft (%)	Blast (*n*)	Blast (%)
Limited	17	70.83	1	9.09
Complete	7	29.17	10	90.91
Total	24	100.00	11	100.00

*W = 50.5, p-value < 0.01.

TABLE 4—*General distribution of perimortem trauma per case.**

Trauma Distribution	Aircraft (*n*)	Aircraft (%)	Blast (*n*)	Blast (%)
Localized	7	29.17	8	72.73
Extensive	16	66.67	3	27.27
Total	24	100.00	11	100.00

*W = 180, p-value = 0.02.

TABLE 5—*Trauma mechanisms interpreted for each case.**

Trauma Mechanism	Aircraft (n=)	Aircraft (%)	Blast (n=)	Blast (%)
Blunt only	16	66.67	4	36.36
Projectile only	–	–	3	27.27
Blunt and Projectile	–	–	3	27.27
Indeterminate	7	29.17	1	9.09
No trauma	1	4.17	–	–
Total	24	100.00	11	100.00

*Chi-square = 16.09, DF = NA, p-value < 0.01.

TABLE 6—*Relative directionality of skeletal insults per case.**

Relative Directionality	Aircraft (*n*)	Aircraft (%)	Blast (*n*)	Blast (%)
One or few primary	–	–	6	54.55
Multiple or Indeterminate	24	100.00	5	45.45
Total	24	100.00	11	100.00

*W = 204, p-value < 0.01.

Regarding documented fracture types and characteristics (Tables 7–8), comminuted fractures were commonly encountered in both types of incidents (aircraft crash, 54%, and blast, 73%). However, spiral (33%) and butterfly (20%) fractures were only observed in aircraft crash cases. No inverted butterfly fractures were documented in either type of incident. Projectile defects (55%) and embedded metal (18%) were observed predominantly in blast cases. Thermal alteration of bone, though uncommon, was observed only in aircraft deceleration cases (8%). Documented fracture types could not be analyzed statistically; however, additional perimortem trauma characteristics are significantly different (Table 8).

The following case studies illustrate differences in typical patterns of trauma observed in aircraft crash incidents and blast incidents.

TABLE 7—*Documented fracture types* associated with each case.*

Fracture Type	Aircraft (*n*)	Aircraft (%)	Blast (*n*)	Blast (%)
Diastatic (cranium only)	3	12.50	3	27.27
Concentric (cranium only)	1	4.17	1	9.09
Linear	2	8.33	4	36.36
Oblique/Transverse	9	37.50	3	27.27
Spiral	8	33.33	–	–
Butterfly	5	20.83	–	–
Inverted Butterfly	–	–	–	–
Compression/Depressed	5	20.83	5	45.45
Shearing/Crushing	3	12.50	1	9.09
Comminuted	13	54.17	8	72.73
Other	5	20.83	2	18.18

*Standard fracture types as defined in the literature (5,12,20).

TABLE 8—*Additional perimortem trauma characteristics observed in each case.**

Perimortem Trauma Characteristics	Aircraft (*n*)	Aircraft (%)	Blast (*n*)	Blast (%)
Bone defect (projectile)	1	4.17	6	54.55
Thermal alteration	2	8.33	–	–
Embedded metal	1	4.17	2	18.18

*Chi-square = 15.83, DF = NA, p-value < 0.01.

Case Study 1: Aircraft Crash Trauma

This case involves the loss of a B-24J aircraft in January 1944. The B-24 had a takeoff speed of 110–130 mph (117–209 km/h) and a cruising speed of 140–160 mph (225–258 km/h). This aircraft crashed off Helen Island in the Tarawa Atoll shortly after takeoff. The remains of the aircrew were recovered from the crash site and reportedly buried in a temporary cemetery.

In 2017, a recovery team excavated a burial trench on the island and discovered human remains. The remains consist of a mostly complete skeleton (>75% of elements recovered) with a high degree of fragmentation (Figs 1 and 2). Based on historical analyses, the remains were associated with the January 1944 air loss. A complete biological profile was determined for the remains, which represent a male of European ancestry between 22 and 28 years of age with an estimated stature of 68.1–74.1 in.

Extensive, perimortem blunt force trauma consistent with a rapid deceleration event is observed throughout the skeleton (5,36). Definitive perimortem fractures are observed in the skull, rib cage, left forearm, right hand, both lower limbs, and both feet and are described as follows.

Extensive perimortem trauma is observed in the facial skeleton and basicranium. This includes fractures in the region of the left orbit (Fig. 3), involving the zygomatic process of the frontal bone and the greater wing of the sphenoid. A tripod fracture of the right zygomatic and a probable Le Fort I fracture of the maxillae are also present (Fig. 4). Additionally, diastatic fractures through the sphenotemporal and occipitotemporal fissures of the basicranium, and possibly through a number of the major and minor cranial vault sutures, are likely given the overall severity of cranial trauma. The mandible displays a left-side condylar fracture and a comminuted vertical fracture of the right mandibular body, immediately posterior to the mental foramen

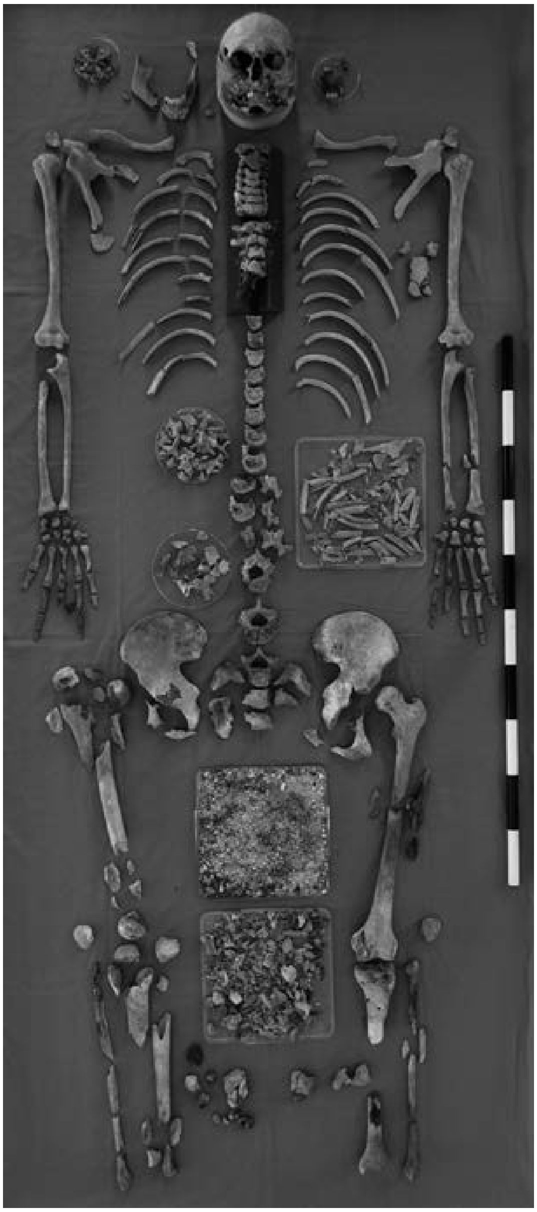


FIG. 1—*Aircraft crash case, skeletal layout. Petri dishes contain loose dentition, nondiagnostic osseous fragments, and sediment. Scale is in decimeters. Note the extensive fracturing and fragmentation of the remains, which is characteristic of aircraft crash cases. [Color figure can be viewed at wileyonlinelibrary.com]*

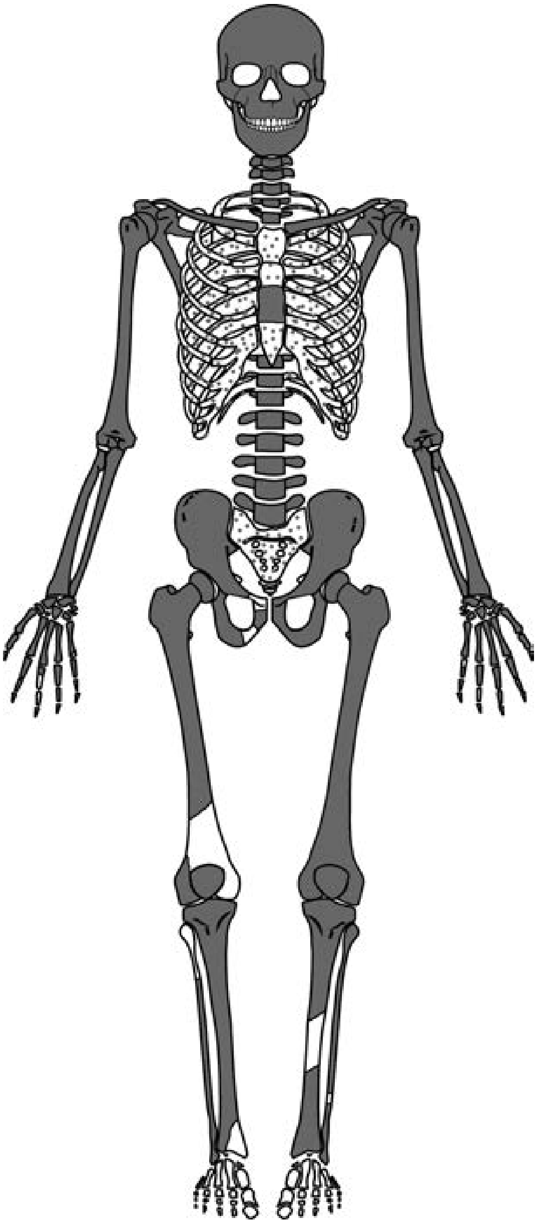


FIG. 2—*Aircraft crash case, skeletal diagram. Portions in gray are present. Stippled regions are present, fragmentary, and unreconstructed. Dentition and all damage are not depicted.*

(see Fig. 1). Such fractures on contralateral sides of the mandible are common in instances of significant impact force. Collectively, these skull fractures are indicative of widespread, high-energy blunt force impacts striking the head from multiple directions (5).

Transverse fractures of the costal angle are observed in at least seven right ribs (Fig. 5), and probable transverse and oblique shaft fractures are observed in at least two ribs. All ribs present for analysis are highly fragmented, and additional perimortem trauma to both sides of the ribcage may be present but obscured by subsequent postmortem damage.

The left radius and ulna display butterfly fractures in the distal thirds of their shafts (Fig. 6). The radial fracture is complete and comminuted, with a separate butterfly segment. The ulnar fracture is complete along its proximal arm, but incomplete along its distal arm, with the butterfly segment remaining attached to the distal ulna fragment. The radial and ulnar fractures are aligned when the elements are in anatomical position, indicating that both fractures can be attributed to a single impact; their orientation indicates that this impact was directed at the posteromedial aspect of the forearm (5,36).

The right third metacarpal displays comminuted transverse and longitudinal fractures, likely due to shearing or crushing forces (36). Similar fractures to the base of the articulating proximal phalanx are likely attributable to the same forces. The second, fourth, and fifth metacarpals are also fragmented and may represent additional instances of perimortem trauma (see Fig. 1); however, postmortem damage to these elements precludes a definitive assessment.

Numerous fractures are observed throughout the skeletal elements of both lower extremities. These include a comminuted

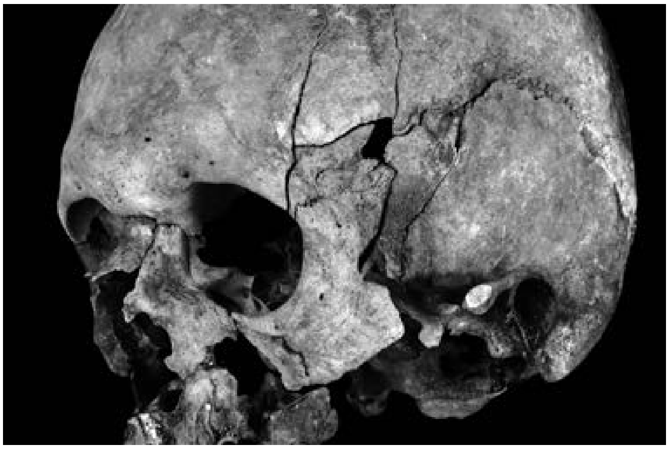


FIG. 3—*Aircraft crash case, detailed photograph of the cranium, depicting perimortem fractures to the left frontal, and greater wing of sphenoid. View is left oblique. [Color figure can be viewed at wileyonlinelibrary.com]*

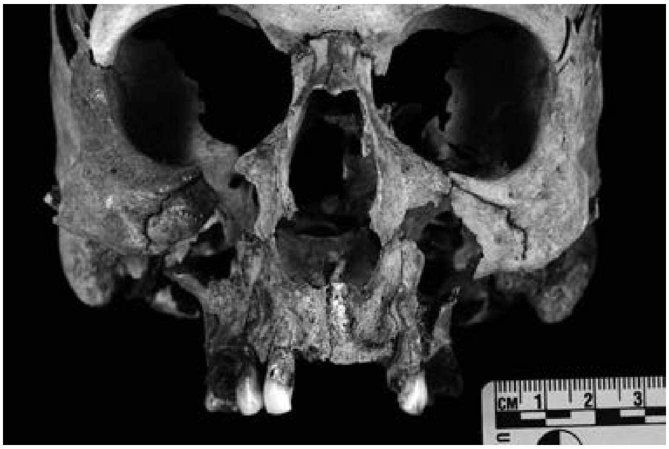


FIG. 4—*Aircraft crash case, detailed photograph of the cranium depicting perimortem tripod fracture to the right zygomatic and probable Le Fort I fracture to the maxillae. View is anterior. [Color figure can be viewed at wileyonlinelibrary.com]*

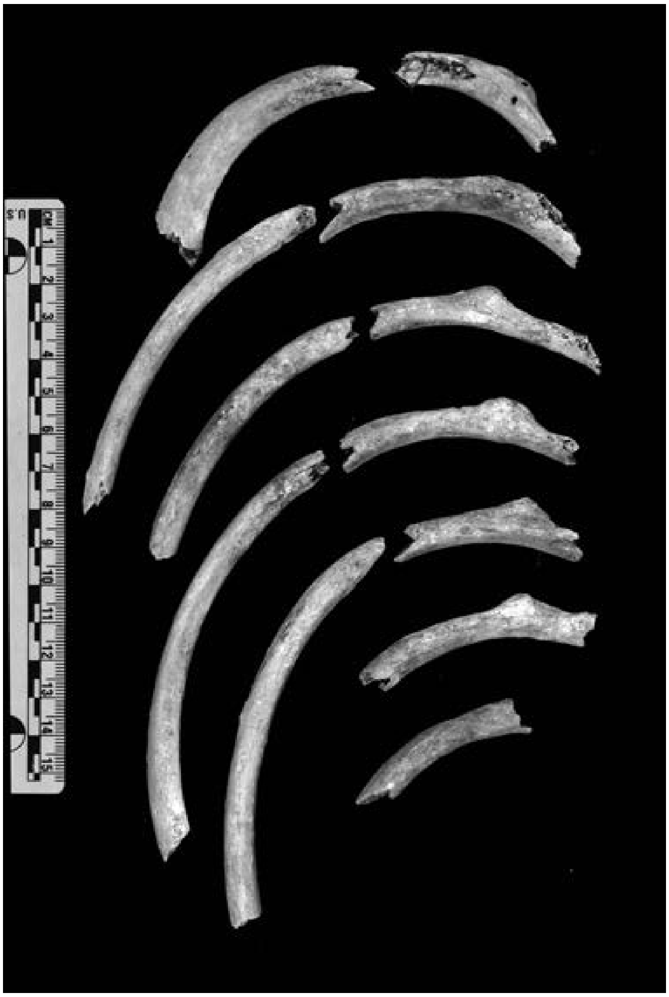


FIG. 5—*Aircraft crash case, detailed photograph of several right ribs, depicting perimortem fractures. [Color figure can be viewed at wileyonline library.com]*

spiral fracture near the midshaft of the left femur (Fig. 7), the extensive nature of which is attributed to a complex combination of torsional, bending, and axial load vectors. Other observed fractures in the femora include comminuted oblique or longitudinal fractures of the subtrochanteric region of the right femur; and comminuted oblique and transverse fractures of the distal third of the right femur (see Fig. 1). Oblique and transverse fractures can be attributed to direct impacts to skeletal elements, which create a bending force, while spiral fractures are the result of torsional, or twisting forces (5,36). Fractures evincing complex, multi-vector loading environments are characteristic of patterns of trauma observed in aircraft crash incidents.

Both tibiae display oblique, probably comminuted fractures of the distal thirds of their shafts (see Fig. 1). The right tibia also presents with a transverse/oblique fracture to the proximal shaft at the level of the nutrient foramen, and a probable comminuted shearing/crushing fracture of the distal tibial epiphysis. Distal tibia fractures of this type, also known as pilon fractures, are typically the result of high-energy compressive axial loads applied to the lower legs during rapid deceleration events, such as motor vehicle crashes and falls from great heights (5,36). The left tibia displays a probable perimortem oblique/transverse fracture to the proximal shaft; however, postmortem damage and adhering metal concretions derived from casket components preclude definitive assessment.

Correspondingly, extensive fractures are observed on both fibulae, with the right fibula fractured in four locations and the left fibula fractured in two locations (see Fig. 1). Extensive perimortem crushing and shearing fractures are observed in the left talus and calcaneus, and the right talus (Fig. 8), calcaneus, and navicular, resulting in high degrees of fragmentation of these elements. As with the pilon fracture noted above, such fractures to the bones of the feet are typically the result of high-energy compressive axial loads applied during rapid deceleration events (5,36).

Given the extent of perimortem blunt force trauma described above, additional perimortem fractures to other regions of the skeleton—including the vertebral column and pelvis—are highly likely, but they are obscured by subsequent postmortem damage and fragmentation. In summary, the observed fractures in these remains include oblique, transverse, butterfly, shearing, and longitudinal fractures, and many fractures are extensively comminuted. The overall nature and distribution of these fractures evince a variety of loading environments, including bending, torsion, compression, and shearing force vectors, indicating a high-

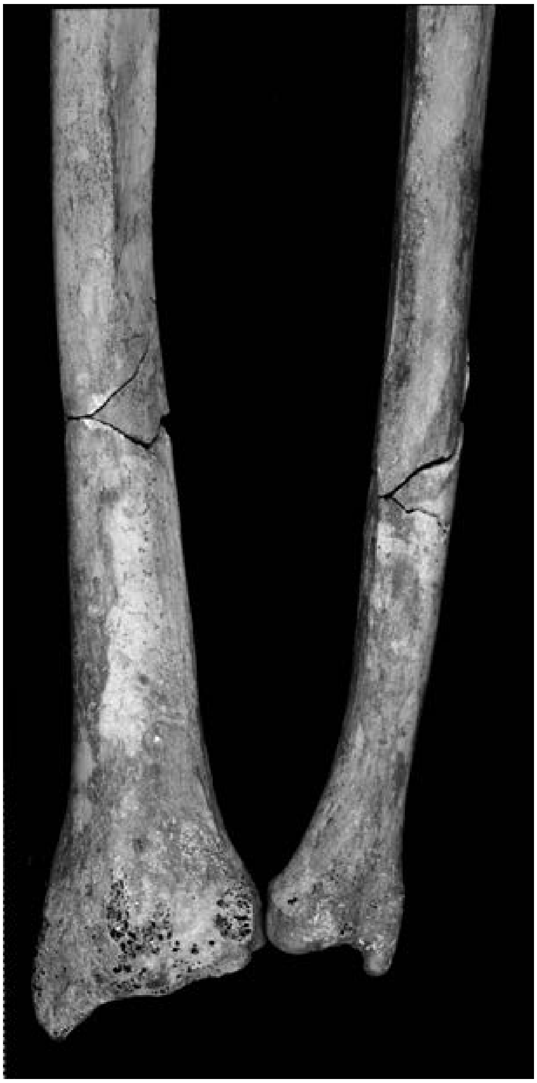


FIG. 6—*Aircraft crash case, detailed photograph of the left radius and ulna, displaying perimortem butterfly fractures to the distal thirds of both shafts. View is posterior. [Color figure can be viewed at wileyonlinelibrary.c om]*

energy, complex, chaotic impact incident, such as occurs in aircraft crashes (5).

Case Study 2: Blast Trauma

This case involves the November 1943 battle of Tarawa on Betio Island, Gilbert Islands (Republic of Kiribati). The individual was killed during the assault and reportedly buried in a temporary cemetery, but no remains were recovered from the site. In 2013, human remains were recovered from a burial trench on the island. In 2017, additional unidentified remains were disinterred from the National Memorial Cemetery of the Pacific. Based on original burial information, a shared mitochondrial DNA sequence, and anthropological analyses, these remains represent one individual.

The remains consist of an incomplete skeleton (25–75% complete) with a low degree of fragmentation (Figs 9 and 10). A complete biological profile was determined for the remains, which represent a male of European ancestry between 18 and 24 years of age with an estimated stature of 65.5–71.5 in.

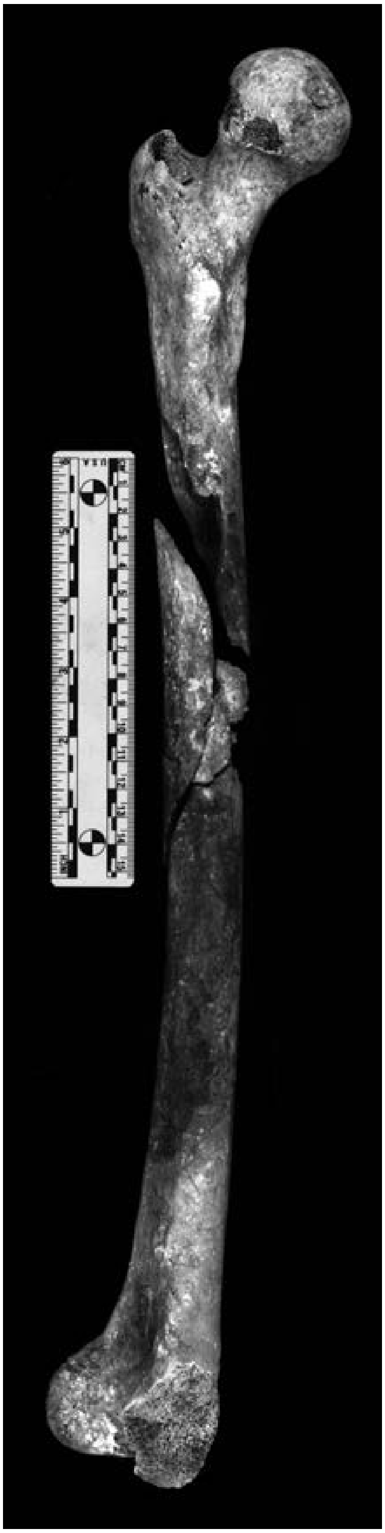


FIG. 7—*Aircraft crash case, detailed photograph of the left femur, depicting a perimortem comminuted spiral fracture. View is posteromedial. [Color figure can be viewed at wileyonlinelibrary.com]*

Projectile and extensive blunt force trauma were observed throughout the skeleton, including several ribs, the right maxilla, and both femora and tibiae. Details of these fractures are described as follows.

The right maxilla exhibits blunt force trauma (Fig. 11). There is an incomplete depressed fracture in the area of the canine

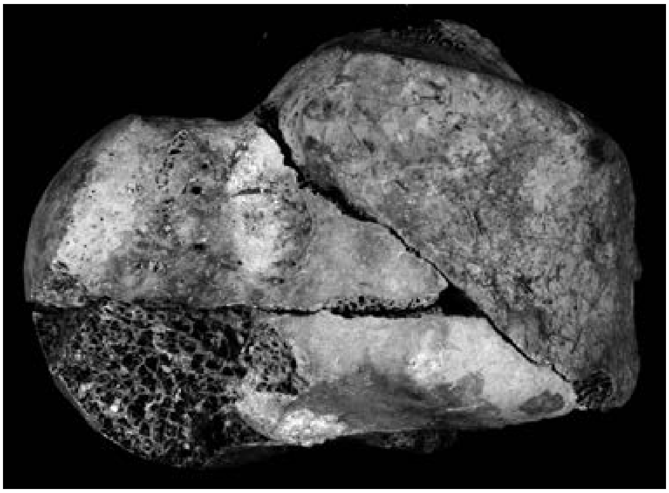


FIG. 8—*Aircraft crash case, detailed photograph of the right talus, depicting perimortem shearing/crushing fractures. View is superior. [Color figure can be viewed at wileyonlinelibrary.com]*

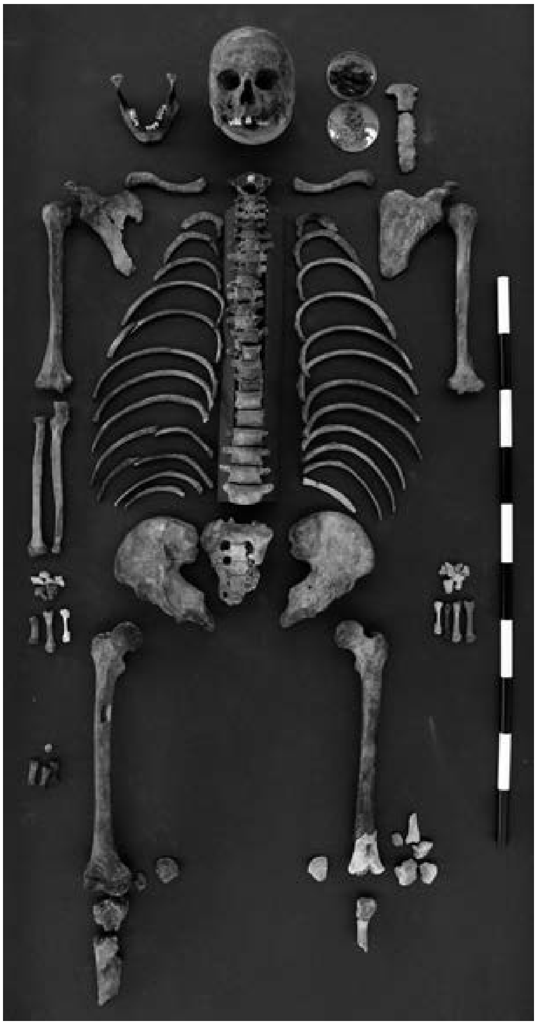


FIG. 9—*Blast trauma case, skeletal layout. Petri dishes contain hair, desiccated tissue, small osseous fragments, and dust. Scale is in decimeters. Note the localized fracturing and fragmentation of the remains, particularly the lower extremities in this case, which is characteristic of blast cases. [Color figure can be viewed at wileyonlinelibrary.com]*

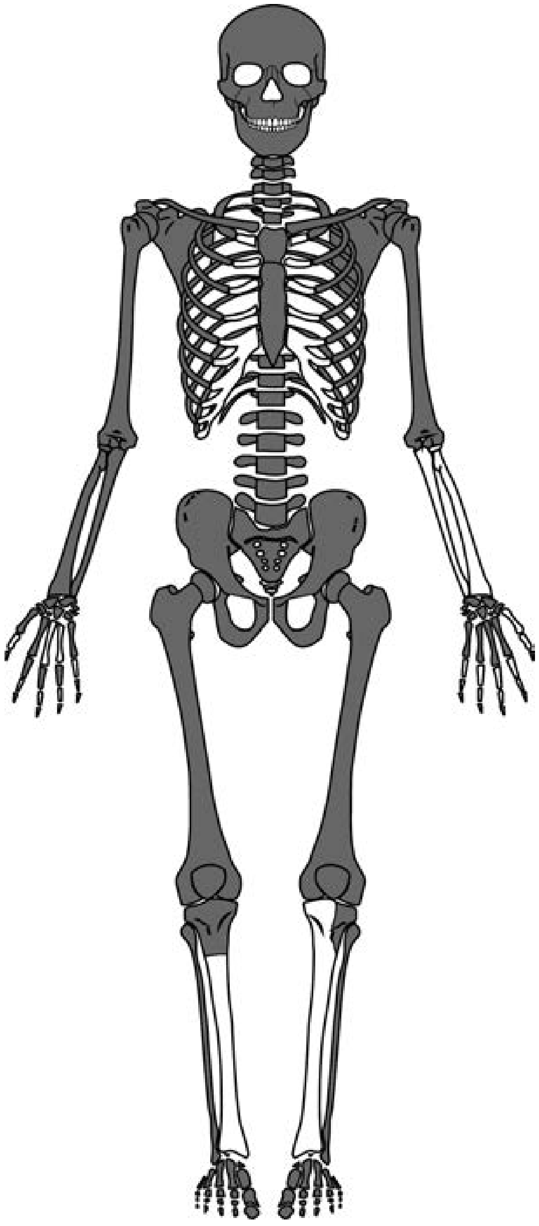


FIG. 10—*Blast trauma case, skeletal diagram. Portions in gray are present. Manual phalanges are arbitrarily sided. Dentition and all damage are not depicted.*

fossa, just inferior to the infraorbital foramen and an incomplete linear fracture extending inferiorly from the right nasal bone. The depressed fracture is the point of impact (5). A depressed fragment of bone perforates into the maxillary sinus, showing that the bone underwent plastic deformation prior to failure. This fracture morphology is consistent with a low-energy force (17). The linear fracture is likely associated with this trauma, occurring peripheral to the point of impact (5,17).

Projectile trauma is evident on the right 5th rib (Fig. 12). There is an externally beveled half-circular defect with a linear fracture radiating from this defect located on the superior border of the anterior third of the rib. This fracture morphology is consistent with a projectile entering the body posteriorly and exiting anteriorly. Other ribs have complete fractures that are also possible perimortem trauma, but postmortem damage limits further interpretation (see Fig. 9).

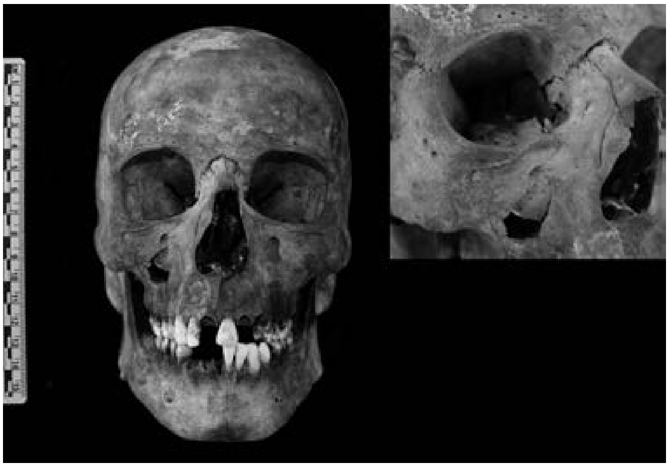


FIG. 11—*Blast trauma case, detailed photograph of the skull (anterior view) with inset of the right maxilla, depicting perimortem blunt force depressed and linear fractures. [Color figure can be viewed at wileyonlinelibrary.com]*

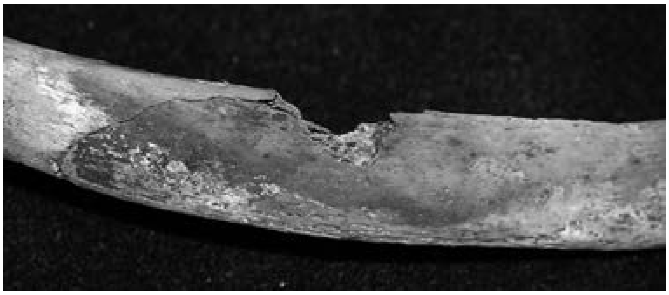


FIG. 12—*Blast trauma case, detailed photograph of the right 5th rib, depicting perimortem projectile trauma with associated external beveling and a radiating fracture. View is anterior. [Color figure can be viewed at wileyonlinelibrary.com]*

There is extensive perimortem trauma of the lower limbs consistent with blast trauma (see Fig. 9). The right femur exhibits a shearing fracture of the medial condyle (Fig. 13). The right tibia is highly comminuted through a combination of vertical, transverse, and oblique fractures (Fig. 14). The left distal femur (Fig. 15) and proximal tibia are also highly comminuted, with multiple transverse, oblique, and horizontal fractures. The pattern of trauma to the lower limb is consistent with an explosion and blastwave leading to a complex loading environment, predominately from a high-energy compressive force directed from distal to proximal (a blast occurring near the feet, with force traveling upwards; 37).

In summary, trauma is observed primarily on the lower extremities, as well as on the right 5th rib and maxilla. Additionally, possible trauma is observed on several ribs. The extent and patterning of the trauma is consistent with nonspecific blast injuries occurring near the feet and traveling superiorly.

Discussion

As illustrated in the case studies, the results of this analysis suggest that although there is considerable overlap in trauma expression, skeletal trauma resulting from historic aircraft crash and blast cases can be differentiated based on the type and pattern of skeletal trauma present. As expected, cases from aircraft crashes report a significantly lower recovery rate (see Table 1) and significantly higher degree of fragmentation (see Table 2) than blast-related cases, which likely translates to the limited analytical potential (see Table 3) of remains associated with aircraft crash events. Additionally, trauma is significantly more widespread throughout the skeleton in aircraft crash cases (see Table 4), likely reflecting the complex and multiple loading vectors associated with rapid deceleration of an aircraft. In contrast, cases from blast events report significantly higher recovery rates (see Table 1), less fragmentation (see Table 2), and increased analytical potential (see Table 3), likely related to the more localized distribution of trauma in these remains (see Table 4).

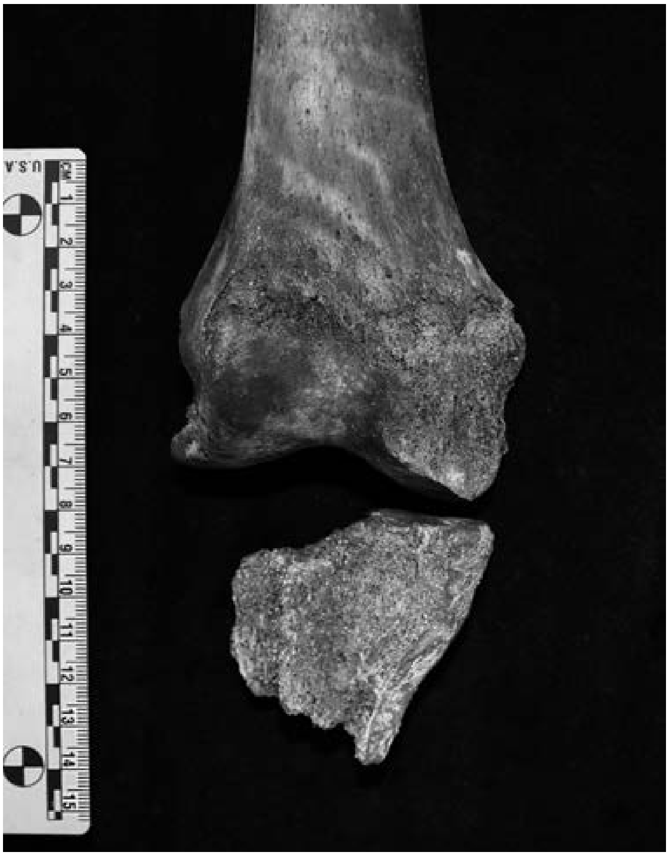


FIG. 13—*Blast trauma case, detailed photograph of the right femur, depicting a shearing fracture of the medial condyle. View is anterior. [Color figure can be viewed at wileyonlinelibrary.com]*

These findings are illustrated in the case studies. Although the aircraft crash case was selected in part for its high recovery rate to exemplify patterns observed in the analytical results, the skeletal layout depicts substantial fragmentation with widespread trauma distributed throughout the remains (see Fig. 1). In contrast, the individual from the blast case displays less overall fragmentation and more localized trauma distribution, predominantly to the lower legs (see Fig. 9).

It must be noted that recovery bias may be impacting the results of the analysis, such as the extended postmortem interval for cases within the study sample and the context from which the remains were recovered. Cases included in the study were received from both archeological field recoveries and cemetery disinterments from around the world. Field recoveries were undertaken recently using standard, contemporary archeological methods and techniques. Disinterments reflect the recovery of remains from battlefields and conflict areas within the weeks to

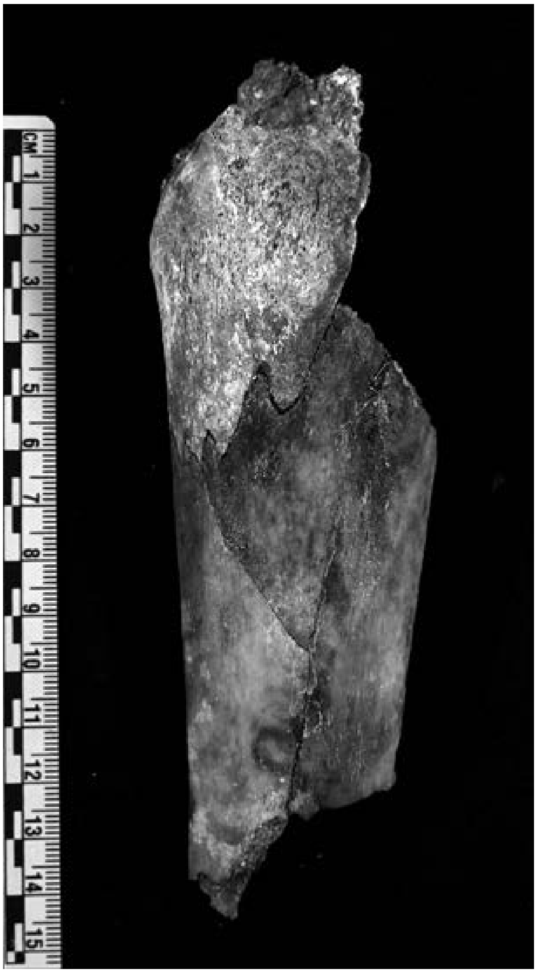


FIG. 14—*Blast trauma case, detailed photograph of the right tibia, displaying perimortem comminuted fractures with evidence of transverse, oblique, and vertical fractures. View is medial. [Color figure can be viewed at wileyonlinelibrary.com]*

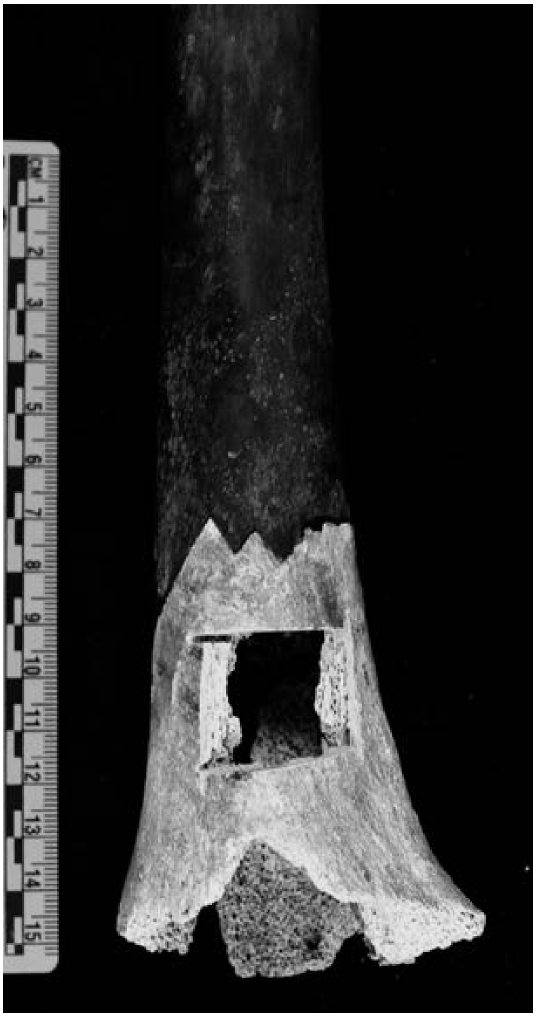


FIG. 15—*Blast trauma case, detailed photograph of the left femur, depicting perimortem trauma. Note the differential coloration of the femur, which is separated by a transverse fracture. The square cut represents the area sampled for DNA. View is anterior, distal is toward the bottom. [Color figure can be viewed at wileyonlinelibrary.com]*

years following the loss incident and have been processed prior to analysis. These remains were later buried as unknown individuals after it was determined that they could not be identified.

These factors impact the remains available for analysis in each case. While the type and proportion of the elements present for analysis in each case are influenced by the intrinsic properties of the elements themselves (i.e., larger, denser, and more recognizable bones are more likely to be recovered) (38), the condition of the remains also largely influences recovery rate. Percentage of remains present for analysis is a reflection of the recovery methods, but also the degree of fragmentation of the remains at the time of recovery. Increased fragmentation also increases the dispersion and surface area on which the postmortem environment can act as a destructive agent (39). As expected by the authors, this is reflected in the lower percentages of remains present for analysis seen in the aircraft crash cases and the higher percentages in the blast cases.

These recovery patterns may also be related to the mechanisms associated with each type of incident. As seen in Table 5 and illustrated in the case studies, hypotheses about trauma mechanisms are supported. Cases from propeller-driven aircraft crash events are associated only with blunt force trauma; however, it is extensive among the remains. Alternatively, cases from blast events often include both blunt force and projectile trauma, but it tends to be more localized within the remains. Unexpectedly, embedded metal was not commonly observed in blast cases (see Table 8), and, contrary to previous observations (21), few analysts made any mention of rust staining in association with perimortem trauma.

Not all cases displayed clearly identifiable trauma. Indeterminate trauma was predominantly interpreted for aircraft crash cases. The skeletal injuries could not be clearly associated with either blunt or projectile trauma. Additionally, one aircraft crash case, which consisted only of few, highly fragmented elements, did not display any apparent trauma. Indeterminate or no visible trauma can result from a myriad of factors, including postmortem damage, atypical instances of projectile or blunt force trauma, bias based on the paucity of remains, or conservative assessments on the part of the analyst. This interpretation likely reflects analytical difficulties owing to the high degree of fragmentation, the relatively low recovery rate, and the diffuse trauma distribution throughout the remains, all of which may inhibit an analyst's ability to draw clear interpretations.

Although previous research did not find a correlation between trauma event and the directionality of perimortem impacts (33), no aircraft crash cases in this study were associated with only one or few primary directions of insult (see Table 6). The majority of blast cases, however, are attributed to one or few primary directions of insults. These results suggest that directionality may be valuable for trauma assessment (e.g., related to the location of the landmine or grenade relative to the position of the body).

Moreover, interpretation of directionality must consider the position of the body when the insult occurred. Although it is tempting to review standard homunculi to interpret trauma directionality and alignment of insults, this can lead to analytical complacency. After all, it is unlikely that most individuals in these circumstances were simply standing in standard anatomical position at the time of the incident. Deliberation of body positioning can also facilitate interpretation of multiple or conflicting insults and help clarify apparent discrepancies.

Regarding fracture types, several appear to be more common to specific mechanisms or trauma events (see Table 7). For example, linear and compression/depressed fractures are more common in blast-related cases, while oblique/transverse, spiral, and butterfly fractures are more common in aircraft crash cases. In fact, no spiral or butterfly fractures were observed in any of the blast cases, again highlighting the differential loading environments associated with each type of event. The occurrence of spiral fractures reflects the increased complexity of the loading environments in aircraft compared to blast events, while the presence of butterfly fractures requires additional consideration due to their limited interpretive abilities (40).

No inverted butterfly fractures were described for the ribs in these cases, although they have been reported in other studies as characteristic injuries for blast trauma (20). Furthermore, no correlation was observed between deceleration trauma (in either aircraft or blast cases) and axial loading of the feet or other regions. Few cases explicitly mention axial loading to the feet or lower legs, regardless of loss type (but see *Case study 1: aircraft crash trauma*, above). The absence of clear inverted butterfly fractures or axial loading to the foot bones may reflect a number of factors. These fractures may not have been present in these historic conflict-related cases, analyses may have erred toward conservative assessments, identification may have been precluded due to the direction of force from multiple impacts, and interpretations may have been biased based on the paucity and condition of remains. Therefore, both traits should be assessed in future cases to further clarify their relationship with trauma.

Analysts must consider the sum total of analytical categories and diagnostic traits applied and observed throughout the skeleton. Interpretations may be hampered or precluded by consideration of only a single analytical category or if analysts emphasize individual elements rather than the body as a whole unit. For example, the presence of blunt versus projectile impacts alone is insufficient to determine the trauma mechanism. Blunt force trauma is observed in both aircraft crash and blast cases, while projectile trauma is primarily observed in blast cases. Although projectile trauma and embedded metal can be observed in aircraft crash cases, it is uncommon.

Furthermore, trauma does not occur in a vacuum nor are elements impacted in isolation. As observed in both case studies, numerous elements throughout the remains may display trauma that aligns and, in combination, provides valuable insight into the loss incident. We argue that analysts should critically consider how skeletal trauma is reported. Although it is tempting to itemize complicated trauma and/or leave interpretations more generalized, analysts should ensure they remain focused on a broader scale and emphasize meaningful descriptions that underscore the body as a whole unit. This may help decrease conclusions of indeterminate trauma mechanisms in complicated cases. In both case studies, a robust and informed interpretation is made through consideration of the body as a unit. Analyses that do not incorporate the condition and trauma pattern of the entire skeleton risk misinterpretation or oversimplified interpretations.

As for differentiating between blast trauma categories, it may not be possible to distinguish between primary (commonly associated with impacts resulting from the blast wave) and tertiary (commonly associated with impacts between the body and large objects or structures) blast trauma because both primary and tertiary blast trauma (e.g., blast waves and pressure waves, respectively) result in blunt force trauma. The amount of deformation that might be observed in primary and tertiary impacts is based on the amount of energy involved in each case. Additionally, while the amount of energy associated with each category differs, the location of these impacts on the body influences how force interacts with the body.

Although clinical literature identifies a correlation between primary blast trauma and amputation, the authors could not assess this expectation adequately due to the recovery contexts and lengthy postmortem interval associated with the skeletal remains. While potential amputation may be present in several cases (e.g., see *Case study 2: blast trauma*, above), numerous factors may account for this, including taphonomic and recovery factors. Incomplete recovery of fragmented remains may have impacted interpretations and led to false identification of amputation. Remains may not be fully recovered, or the postmortem burial environment may impact the condition and preservation of the remains. Either of these conditions can superficially mimic amputation.

Therefore, amputation is not a reliable criterion to distinguish blast trauma in skeletal remains, at least for the fully skeletonized cases typically analyzed by the DPAA, wherein the remains are decades old, may not have been recovered in a systematic manner, and may have been processed multiple times in the past. This cautionary consideration is important in the analysis of potential blast trauma in skeletal remains.

Although clinical literature may provide expectations and criteria for analysis, many of these studies and case reports are focused on examination of soft tissue remains that were recovered in a timely manner after the incident. Expected characteristics established in the clinical literature may be inappropriate for skeletonized remains because information may be lost in the postmortem interval or absent in the skeletal remains. For example, the tibia in the blast case study above displays a bending under compression fracture, in addition to the absence of the lower legs, which may be used to support interpretations of traumatic amputation. However, based on the available evidence, other causes cannot be ruled out. Further research is required to validate this hypothesis, and the potential impact of recovery bias in analytical conclusions cannot be ignored.

On the other hand, it may be possible to differentiate between secondary and other forms of blast trauma within skeletal remains by identifying evidence of projectile trauma, shrapnel, and/or metal stains. For example, the individual associated with the blast case study represents an example of nonspecific blast trauma. Trauma is observed to the lower extremities, but association with primary blast waves or tertiary pressure waves or deceleration cannot be confirmed from the skeletal remains. Although projectile trauma is identified on the remains, the projectile defect is on a rib and represents a posterior-to-anterior directionality. Therefore, the case study represents nonspecific blast trauma.

It is recommended that analysts clarify whether they are interpreting secondary/projectile or nonspecific blast trauma and consider the broader skeletal unit to interpret complicated insults appropriately. By assessing the skeletal trauma across the whole body as it was at the time of the traumatic insult, the clinical literature can be more readily used, although with caution, for overall pattern and distribution for expected injuries in blast events. For future studies and casework, analysts should critically assess the clinically derived blast trauma categories and their effectiveness when applied to skeletal remains. This may provide insight into considerations such as the type of blast trauma, the spatial relationship between the individual and the impact, or the type of projectile.

The results of this analysis indicate that blast and aircraft crash cases can be differentiated based on overall patterns in skeletal trauma. The condition of the remains, trauma mechanisms, trauma distribution, and fracture types together provide insight loss events. These WWII cases are valuable because DPAA analysts can retrospectively assess their results. Historic records are available to confirm skeletal interpretations, but many other agencies do not have access to this type of contextual information. Instead, many trauma-based analytical methods are based on experimental studies, often on nonhuman proxies, or isolated case studies (14,18). The DPAA cases present numerous examples by which to extrapolate representative patterns of skeletal trauma.

It is important to note that although numerous trauma features are presented here that may aid interpretations, they are not to be used as definitive criteria. Trauma analysis is complicated, and different types of trauma may display similar skeletal expressions, while the same type of trauma may display different expressions in different circumstances. Furthermore, this retrospective analysis emphasizes the need for standardized and clear terminology to ensure that specific perimortem trauma characteristics are described in an accurate manner to ensure the quality and scientific rigor of work (1,2). In addition to perimortem concerns, the condition of the remains, the amount of remains recovered, the postmortem environment, and even the experience and skill level of the analyst can impact the appearance and description of skeletal remains. Therefore, these case studies are intended to provide a comparative framework for differential diagnosis. Future research is required to better address these concerns.

Conclusions

This study provides guidance to support interpretation of aircraft crash and blast trauma in skeletal remains and lessen an analyst's reliance on their own individual experiences or on clinically based criteria developed through assessment of soft tissue remains. However, these case studies are meant to be illustrative, not definitive, for trauma interpretation. Trauma is variably expressed in the skeleton and different events can have similar skeletal affects, while similar events may result in different skeletal affects. Additionally, while blast trauma features prominently in clinical literature, anthropologists are still determining appropriate manners in which to operationalize these categories for skeletal remains.

Based on an analysis of recently resolved WWII cases, several criteria may be useful to differentiate historic aircraft crash from blast cases. Aircraft crash cases are associated with more limited biological profiles due to extensive fracturing throughout the skeleton, torsional fractures, and indeterminate or multiple directionalities of impacts. Blast cases are associated with higher element recovery, more complete biological profiles, more localized trauma patterns, and impacts from one or few directions of force. Because of the diverse categories of blast trauma, it can be characterized as secondary, when embedded metal or projectile defects are present, or nonspecific when particular blast-related mechanisms cannot be inferred. In cases where skeletal trauma cannot be more definitively attributed to a specific type of event (e.g., blast or aircraft deceleration), it is recommended that analysts extensively document the details, patterns, locations, and relationships of fractures and defects throughout the remains with an emphasis on the body as a whole unit. Future studies should explore the relationship between these types of extreme trauma events and axial loading, plastic deformation, bending, butterfly fractures, inverted butterfly fractures, and potential amputation.

Although the findings of this study are most directly applicable to the analysis of remains from similar historic military/armed conflict-related contexts, it can also be argued that these findings have broader relevance for forensic anthropologists working in other settings as well. While these practitioners may not have need to distinguish aircraft crash trauma from blast trauma in any one case, a greater understanding of the respective patterns of trauma observed in these incidents may facilitate trauma analyses in fatalities resulting from industrial, motor vehicle, or civil aviation accidents; violent incidents of civil unrest; and recent and historic instances of human rights violations. Further, in current-day military/armed conflict-related or human rights contexts, these patterns may be helpful for corroborating eyewitness testimony surrounding specific incidents, or as evidence of collateral civilian fatalities arising from military/paramilitary operations, for example. In other words, an aircraft crash is not the only circumstance that a forensic anthropologist may encounter involving complex, multi-vector loading environments; nor is explosive ordnance the only material capable of causing blast trauma. Thus, the broad patterns of skeletal trauma observed in blast and aircraft crash cases can be extrapolated and applied to other circumstances that generate similar forces, are characterized by similar loading environments, and result in similar mechanisms of injury.

.Acknowledgments

The authors would like to acknowledge DPAA anthropologists and Laboratory staff for the use of the numerous anthropological and identification reports that are included in this analysis. Additionally, the authors thank the anonymous reviewers for their comments.

References

1. Berryman HE, Lanfear AK, Shirley NR. The biomechanics of gunshot trauma to bone: research considerations within the present judicial climate. In: Dirkmaat DC, editor. A companion to forensic anthropology. Chichester, U.K.: Wiley-Blackwell, 2012;390–99.
2. Symes SA, L'Abbe EN, Chapman EN, Wolff I, Dirkmaat DC. Interpreting traumatic injury from bone in medicolegal investigations. In: Dirkmaat DC, editor. A companion to forensic anthropology. Chichester, U.K.: Wiley-Blackwell, 2012;340–89.
3. Emanovsky P. Low-velocity impact trauma: an illustrative selection of cases from the Joint POW/MIA Accounting Command – Central Identification Laboratory. In: Passalacqua NV, Rainwater CW, editors. Skeletal trauma analysis: case studies in context. Hoboken, NJ: John Wiley & Sons, 2015;156–66.
4. Di Maio VJM. Gunshot wounds: practical aspects of firearms, ballistics, and forensic techniques, 2nd edn. Boca Raton, FL: CRC Press, 1998;53–122.

5. Wedel VL, Galloway A, editors. Broken bones: anthropological analysis of blunt force trauma, 2nd edn. Springfield, IL: Charles C. Thomas, 2014;33–72.
6. Currey JD. Bones: structures and mechanics. Princeton, NJ: Princeton University Press, 2002;27–123.
7. Frost HM. Bone development during childhood: insights from a new paradigm. In: Schoenau E, editor. Paediatric osteology: new trends and developments in diagnostic and therapy. Amersterdam, the Netherlands: Elsevier Science, 1996;3–39.
8. Frost HM. Bone's mechanostat: a 2003 update. Anat Rec A Discov Mol Cell Evol Biol 2003;275(2):1081–101.
9. McGowan C. A practical guide to vertebrate mechanics. Cambridge, U.K.: Cambridge University Press, 1999;35–114.
10. Pearson OM, Lieberman DE. The aging of Wolff's "law": ontogeny and responses to mechanical loading in cortical bone. Am J Phys Anthropol 2004;125(S39):63–99. https://doi.org/10.1002/ajpa.20155
11. Passalacqua NV, Fenton TW. Developments in skeletal trauma: blunt-force trauma. In: Dirkmaat DC, editor. A companion to forensic anthropology. Chichester, U.K.: Wiley-Blackwell, 2012;400–12.
12. Berryman HE, Symes SA. Recognizing gunshot and blunt cranial trauma through fracture interpretation. In: Reichs K, editor. Forensic osteology II: advanced in the identification of human remains, 2nd edn. Springfield, IL: Charles C Thomas, 1998;333–52.
13. Loe L. Perimortem trauma. In: Blau S, Ubelaker DH, editors. Handbook of forensic anthropology and archaeology, 2nd edn. New York, NY: Routledge, 2009;263–83.
14. Willits NA, Hefner JT, Tersigni-Tarrant MA. Case studies in skeletal blast trauma. In: Passalacqua NV, Rainwater CW, editors. Skeletal trauma analysis: case studies in context. Hoboken, NJ: John Wiley & Sons, 2015;177–88.
15. Tomczak PD, Buikstra JE. Analysis of blunt trauma injuries: vertical deceleration versus horizontal deceleration injuries. J Forensic Sci 1999;44(2):253–62. https://doi.org/10.1520/JFS14449J
16. Wieberg DA, Wescott DJ. Estimating the timing of long bone fractures: correlation between the postmortem interval, bone moisture content, and blunt force trauma fracture characteristics. J Forensic Sci 2008;53 (5):1028–34. https://doi.org/10.1111/j.1556-4029.2008.00801.x
17. Seneviratne AB. Skeletal and soft tissue injuries resulting from a grenade. In: Kimmerle EH, Baraybar JP, editors. Skeletal trauma: identification of injuries resulting from human rights abuse and armed conflict. Boca Raton, FL: CRC Press, 2008;95–116.
18. Christensen AM, Smith VA. Rib butterfly fractures as a possible indicator of blast trauma. J Forensic Sci 2012;58(Suppl 1):S15–9. https://doi.org/10.1111/1556-4029.12019
19. Christensen AM, Smith VA, Ramos V, Shegogue C, Whitworth M. Primary and secondary skeletal blast trauma. J Forensic Sci 2012;57(1):6–11. https://doi.org/10.1111/j.1556-4029.2011.01938.x
20. Christensen AM, Smith VA. Blast trauma. In: Passalacqua NV, Rainwater CW, editors. Skeletal trauma analysis: case studies in context. Hoboken, NJ: John Wiley & Sons, 2015;167–76.
21. Willits NA, Hefner JT, Tersigni-Tarrant MA. Case studies in skeletal blast trauma. In: Passalacqua NV, Rainwater CW, editors. Skeletal trauma analysis: case studies in context. Hoboken, NJ: John Wiley & Sons, 2015;177–88.
22. Tersigni-Tarrant MA. Blunt force trauma associated with a fall from heights. In: Passalacqua NV, Rainwater CW, editors. Skeletal trauma analysis: case studies in context. Hoboken, NJ: John Wiley & Sons, 2015;147–55.
23. Rowbothan SK, Blau S. Skeletal fractures resulting from fatal falls: a review of the literature. Forensic Sci Int 2016;266:582.e1–e15. https://doi.org/10.1016/j.forsciint.2016.04.037
24. Abel SM, Ramsey S. Patterns of skeletal trauma in suicidal bridge jumpers: a retrospective study from the southeastern United States. Forensic Sci Int 2013;231(1–3):399.e1–e5. https://doi.org/10.1016/j.forsciint.2013.05.034
25. Pattimore D, Ward E, Thomas P, Bradford M. The nature and cause of lower limb injuries in car crashes. SAE Transactions 1991;100:1947–58. https://doi.org/10.4271/912901
26. Marella GL, Solinas M, Potenza S, Milano F, Manciocchi S, Perfetti E, et al. Identification of driver and front passenger in traffic accidents through skeletal injury pattern. EuroMediterranean Biomed J 2018;13 (1):1–4.
27. Jeffers RF, Tan HB, Nicolopoulos C, Kamath R, Giannoudis PV. Prevalence and patterns of foot injuries following motorcycle trauma. J Orthopaedic Trauma 2004;18(2):87–91. https://doi.org/10.1097/00005131-200402000-00005
28. Chalmers DJ, O'Hare DPA, McBride DI. The incidence, nature, and severity of injuries in New Zealand civil aviation. Aviat Space Environ Med 2000;71(4):388–95.
29. Byard RW, Tsokos M. Avulsion of the distal tibial shaft in aircraft crashes: a pathological feature of extreme decelerative injury. Am J Forensic Med Pathol 2006;27(4):337–9. https://doi.org/10.1097/01.paf.0000220928.77162.b6
30. Postma ILE, Winkelhagen J, Bloemers FW, Heetveld MJ, Bijlsma TS, Goslings JC. February 2009 airplane crash at Amsterdam Schiphol Airport: an overview of injuries and patient distribution. Prehosp Disaster Med 2011;26(4):299–304. https://doi.org/10.1017/S1049023X11006467
31. Dussault MC, Smith M, Osselton D. Blast injury and the human skeleton: an important emerging aspect of conflict-related trauma. J Forensic Sci 2014;59(3):606–12. https://doi.org/10.1111/1556-4029.12361
32. Dussault MC, Smith M, Hanson I. Evaluation of trauma patterns in blast injuries using multiple correspondence analysis. Forensic Sci Int 2016;267:66–72. https://doi.org/10.1016/j.forsciint.2016.08.004
33. Banks P. Skeletal blast trauma: an application of clinical literature and current methods in forensic anthropology to known blast trauma casualties [thesis]. Starkville, MS: Mississippi State University, 2017.
34. Hope ACA. A simple Monte Carlo significance test procedure. J R Stat Soc Series B (Methodol) 1968;30(3):582–98.
35. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2020. https://www.R-project.org/ (accessed May 31, 2020).
36. Salter RB. Textbook of disorders and injuries of the musculoskeletal system: an introduction to orthopaedics, fractures and joint injuries, rheumatology, metabolic bone disease, and rehabilitation, 3rd edn. Baltimore, MD: Lippincott, Williams, and Wilkins, 1999;561–655.
37. Covey DC. Blast and fragment injuries of the musculoskeletal System. J Bone Joint Surg Am 2002;84(7):1221–34. https://doi.org/10.2106/00004623-200207000-00022
38. Von Endt DW, Ortner DJ. Experimental effects of bone size and temperature on bone diagenesis. J Arch Sci 1984;11(3):247–53. https://doi.org/10.1016/0305-4403(84)90005-0
39. Conard NJ, Walker SJ, Kandel AW. How heating and cooling and wetting and drying can destroy dense faunal elements and lead to differential preservation. Paleogeogr Palaeoclimatol Palaeoecol 2008;266:236–45. https://doi.org/10.1016/j.palaeo.2008.03.036
40. L'Abbé EN, Symes SA, Raymond DE, Ubelaker DH. The Rorschach butterfly, understanding bone biomechanics prior to using nomenclature in bone trauma interpretations. Forensic Sci Int 2019;299:187–94. https://doi.org/10.1016/j.forsciint.2019.04.005

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14554
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Rebecca Campbell,[1] *Ph.D.; McKenzie Javorka,*[1] *M.A.; Dhruv B. Sharma,*[2] *Ph.D.; Katie Gregory,*[1] *Ph.D.; Matt Opsommer,*[3] *B.A.; Kristin Schelling,*[3] *M.S.; and Lauren Lu,*[3] *M.S.*

A State Census of Unsubmitted Sexual Assault Kits: Comparing Forensic DNA Testing Outcomes by Geographic and Population Density Characteristics*

ABSTRACT: A growing number of U.S. cities and states have large numbers of unsubmitted sexual assault kits (SAKs) in police property facilities. Prior research conducted in large urban cities has found that testing these kits yields a sizable number of DNA profiles that meet FBI eligibility for upload to the national criminal DNA database CODIS (Combined DNA Index System) and uploaded profiles return a substantial number of matches to existing criminal profiles in CODIS. It is unknown whether these findings are unique to large urban cities with high crime rates. The purpose of current study was to document forensic testing outcomes from a state census of previously unsubmitted SAKs, which included large urban–suburban centers, as well as smaller cities and rural counties. We inventoried all previously unsubmitted SAKs in Michigan (N = 3422 SAKs) and submitted all kits for forensic DNA testing. A total of n = 1239 SAKs had a DNA profile that met eligibility for upload into CODIS (36.2% unconditional, 56.5% conditional CODIS eligible rate) and n = 585 SAKs yielded a CODIS Hit (17.1% unconditional, 47.2% conditional CODIS hit rate). These rates are consistent with studies from urban areas suggesting approximately half of SAKs tested yield a CODIS profile and approximately half of those uploaded profiles yield a hit. We compared SAK forensic testing outcomes by geographic and population density characteristics, and although rates were often higher in larger metropolitan areas, the obtained rates in micropolitan and rural areas suggest testing is warranted in smaller jurisdictions as well.

KEYWORDS: forensic science, DNA, forensic testing, sexual assault kits, rape kits, CODIS, law enforcement

Over the past decade, the demand for forensic science services by the criminal justice system has increased substantially, especially for forensic biology DNA casework (1,2). As publicly funded crime laboratories have struggled to maintain timely forensic DNA testing throughput (3,4), law enforcement agencies (LEAs) have had to ration their requests for crime scene evidence testing (5). This has created substantial "unsubmitted" crime scene evidence, meaning law enforcement personnel have not made a request for testing and the evidence remains in police custody; by contrast, "untested" evidence has been submitted to a crime laboratory and is awaiting testing (5–8). Police have not been routinely submitting evidence into testing queues for decades, and for sexual assault cases, this has created sizable stockpiles of unsubmitted sexual assault kits (SAKs; also known as "rape kits") in police property storage facilities throughout the United States (5–10). SAKs contain semen, blood, saliva, hair, and/or fingernail scrapings collected from victims' bodies (11), but because these kits are not being submitted for testing, there is no opportunity for this evidence to inform investigations, support prosecutions, or exonerate those who have been wrongly accused (5,6). Estimates suggest there are at least 200,000 unsubmitted SAKs in U.S. LEAs, and large numbers (5000–10,000+) of unsubmitted rape kits have been uncovered in multiple large urban cities, such as Cleveland, Dallas, Detroit, Houston, Memphis, Los Angeles, New Orleans, and New York City (6).

The discovery of these so-called "rape kit backlogs" often sparks community outrage and calls from prosecutors and victim advocates to submit these kits for forensic DNA testing (12,13). Studies of previously unsubmitted SAKs suggest that the biological evidence in these kits is still viable, and testing yields a sizable number of DNA profiles that meet FBI eligibility for upload to the national criminal DNA database CODIS (Combined DNA Index System) (14–17). Uploaded profiles have returned a substantial number of matches to existing criminal profiles in CODIS (often termed a "CODIS hit"), thereby providing police with promising investigational leads (14,15). Such data suggest good return-on-investment for testing previously unsubmitted SAKs (18–21), but it is important to note that these studies have been conducted exclusively in large urban cities

[1] Department of Psychology, Michigan State University, 316 Physics Road, East Lansing, MI, 48824.
[2] Center for Statistical Training & Consulting, Michigan State University, 293 Farm Lane Room, East Lansing, MI, 48824.
[3] Michigan State Police, 7150 Harris Drive, Dimondale, MI, 48821.
Corresponding author: Rebecca Campbell, Ph.D. E-mail: rmc@msu.edu

*The research in this paper was supported by a grant from Sexual Assault Kit Initiative (#2015-AK-BX-K0155), awarded by the Bureau of Justice Assistance, Office of Justice Programs, U.S. Department of Justice, and administered by the Michigan State Police.

Received 10 June 2020; and in revised form 29 July 2020; accepted 4 Aug. 2020.

with high crime rates (22). To inform local, state, and national policy on SAK testing practices, SAK forensic testing outcomes must be evaluated in a variety of jurisdictions to understand how rates vary by geographic and population density characteristics.

To address this gap in the literature, the purpose of the current study was to document forensic testing outcomes from a state census of previously unsubmitted SAKs, which included large urban–suburban centers, as well as smaller cities and rural counties. Documenting SAK forensic testing outcomes in communities of different sizes can help criminal justice system practitioners estimate how many active cases may emerge once they start processing previously unsubmitted SAKs. Such information is critical for developing staffing plans for law enforcement personnel, forensic scientists, victim advocates, and prosecutors. This may be particularly helpful for smaller communities and rural jurisdictions that may have less experience working these so-called "cold cases" and may need targeted outreach, professional development, and mentoring to increase preparedness. To set the stage for this study, we begin with a brief review of how SAKs are collected and tested, and what prior research suggests forensic DNA testing might yield in CODIS profiles and CODIS hits.

Since 1980s, it has been standard practice to direct sexual assault victims to hospital emergency departments or specialized healthcare programs (e.g., sexual assault nurse examiner [SANE] programs) for a postassault medical forensic examination (23). The purpose of this examination is to provide comprehensive healthcare to victims (11,24), which includes diagnosing and treating injuries sustained in the assault, offering emergency contraception to prevent pregnancy (if applicable), and administering prophylaxis for sexually transmitted infections that might have been contracted during the assault. In addition to these healthcare components, the medical forensic examination can include the collection of a SAK to preserve forensic evidence (11). The process of collecting a SAK is time-consuming, highly invasive, and often upsetting for victims, as it includes plucking head and pubic hairs; obtaining fingernail scrapings in the event the assailant was scratched during the attack; swabbing the genitals, anus, mouth, breasts, and/or other body areas to collect semen, blood, or saliva; and photographing injuries.

After a SAK has been collected by a healthcare professional, it is taken into custody by law enforcement personnel (11). Police are then responsible for submitting the rape kit to a forensic laboratory for testing, which includes screening the samples in the kit for biological evidence and analyzing them for DNA (25). The resulting DNA profile can be uploaded to CODIS if: it meets biological quality standards for the number of core loci; there is accompanying documentation to verify that the sample was collected from a crime scene (i.e., a police report); and the sample is from the probable perpetrator of the crime (26). Uploaded samples are compared to reference samples in CODIS's two indexing systems. First, the offender index contains known DNA profiles from arrestees/convicted offenders, obtained at their qualifying offense (i.e., a prior criminal offense that met federal requirements for CODIS entry). When a new DNA profile is uploaded to CODIS, it may match a known offender profile already in the system, which is referred to as an "offender hit." Second, the forensic index contains unknown DNA profiles obtained at crime scenes; matches to these samples are typically termed "forensic hits." Both types of hits can help police and prosecutors identify or confirm offender identity, link cases to establish patterns of repeat offending, and potentially exonerate those wrongfully accused.

Forensic DNA testing can provide tremendously useful information to police and prosecutors, but for decades, law enforcement personnel have not been routinely submitting SAKs for testing (6,9). The reasons why this has become standard practice in so many LEAs are complex, but resource limitations are clearly a factor. Strom and Hickman (5) highlighted that police do not submit crime scene evidence for testing when they know their state, county, or local forensic crime laboratories do not have capacity to test all submitted evidence. Over a decade ago, the National Academy of Science (4) sounded the alarm that publicly funded forensic laboratories were insufficiently resourced to serve the needs of the criminal justice system. Recently, the Government Accountability Office (3) underscored that this problem is not yet fixed, as many state and local government crime labs have too many testing requests to resolve all in a timely manner. However, laboratory capacity is not the sole reason why SAKs have not been submitted for testing, as current research also indicates that negative stereotypes about victims and their credibility affect detectives' decisions on how scarce testing resources should be used (27,28).

As a growing number of communities throughout the United States have discovered stockpiles of unsubmitted SAKs, community pressure has prompted LEAs to submit these kits for testing to determine whether the evidence is still viable and potentially actionable. To gauge the potential value of testing previously unsubmitted kits, researchers have tracked the number of kits submitted for forensic DNA testing, the number of SAKs that yield a CODIS eligible DNA profile, and whether the uploaded profile matched to an existing CODIS sample (i.e., a CODIS hit). Some studies have also reported whether the CODIS hit is to another sexual assault case, indicating a pattern of suspected serial sexual offending.

In the first study to track testing outcomes of previously unsubmitted SAKs, Peterson, Johnson, Herz, Graziano, and Oehler (16) reviewed 1320 SAKs randomly sampled from 10,895 "backlogged/untested" rape kits from Los Angeles. In this sample, 699 DNA profiles were entered into CODIS (53% of the total sample tested), resulting in 347 CODIS hits (50% of profiles uploaded to CODIS; 26% of the total sample tested). Given these promising results, in 2010 the National Institute of Justice funded two action research projects in cities with large numbers of unsubmitted SAKs, one in Detroit, one in Houston (14,17). From the Detroit site, Campbell et al. (14) tested a stratified random sample of 1595 previously unsubmitted SAKs from a stockpile of approximately 11,000 kits. Testing yielded 785 DNA profiles for upload to CODIS (49% of the total sample tested), resulting in 455 CODIS hits (58% of profiles uploaded to CODIS; 28.5% of the total sample tested); 127 hits (28% of the hits) were to a reference sample in CODIS from another sexual assault crime and/or another SAK, revealing a pattern of suspected serial sexual offending. In follow-up study, Campbell, Feeney, Goodman-Williams, Sharma, and Pierce (29) tested 7287 previously unsubmitted Detroit SAKs: 2938 SAKs had a DNA profile that met eligibility for upload into CODIS (40% of the total sample tested), and 1675 SAKs yielded a CODIS Hit (57% of profiles uploaded to CODIS; 23% of the total sample tested); 775 SAKs (46% of the SAKs with hits) revealed suspected serial offending. From the Houston site, Wells, Campbell, and Franklin (17) tested a sample of 493 SAKs from their inventory of 6663 previously unsubmitted kits, which yielded 203 DNA profiles for CODIS upload (43% of the total sample tested) and 104 CODIS hits (51% of profiles uploaded; 21% of the total sample tested). As these action research projects were in progress, Cleveland also began testing their unsubmitted SAKs (15). From a sample of 4966 SAKs, there were 2943 DNA profiles for CODIS upload (59% of the total sample tested) and 1935 CODIS hits (66% of profiles uploaded; 39% of the total sample tested). Lovell, Luminais, Flannery, Overman, Huang, Walker, and Clark (30) studied a subsample of 433 SAKs, of which 245 (56%) were connected to a suspected serial sexual offender, defined as a DNA match to another SAK or to a prior arrest for a sexual assault documented in an offender's criminal history record.

Taken together, these results from rape kit testing initiatives in urban communities indicate that previously unsubmitted SAKs do indeed contain valuable evidence: approximately half (40–59%) of SAKs tested have DNA profiles that meet eligibility for upload to CODIS, and at least half of those uploaded profiles (50–66%) yield a match to a criminal DNA profile (14–17,31). As these studies were unfolding, and as more cities in the United States began reporting that they too had a large number of unsubmitted SAKs, the Department of Justice's Bureau of Justice Assistance (BJA) created a national-scale project, the Sexual Assault Kit Initiative (31), to support inventorying SAKs, testing, prosecution, and victim advocacy. SAKI represented an intentional effort to address the problem of unsubmitted SAKs not only in urban cities, but also in smaller towns and rural communities. In its inaugural funding in 2015, SAKI funded 20 sites, nine of which were entire states (31). Currently, SAKI funds 63 sites, 22 of which are state-wide initiatives.

In Michigan, high-profile media coverage of the 11,000 unsubmitted SAKs in Detroit prompted state-level leaders to assess the extent of this problem in other jurisdictions throughout the state. The Michigan State Police (MSP) applied for and received BJA SAKI 2015 funding to conduct a state census of all previously unsubmitted SAKs in Michigan LEAs and to submit all inventoried kits to an outside vendor laboratory for forensic DNA testing. This grant also included forming a partnership with a research team (the authors of this paper) to document the forensic testing results from this state census. Michigan has 83 counties, three of which comprise the Detroit urban-suburban community, but most metropolitan counties are substantially smaller, and a sizable proportion of the state's counties are rural, which creates an opportunity to evaluate how SAK forensic testing results vary by geographic and population density measures.

Specifically, we explored three research questions in this study. First, from the state census, how many SAKs were tested for DNA, how many yielded a CODIS eligible DNA profile, how many yielded a CODIS hit, and how many of those hits matched to another sexual assault case, indicating suspected serial sexual offending? We examined whether forensic testing rates patterns identified in prior studies of urban communities were consistent in our state census, with its inclusion of smaller cities and rural areas. In other words, would approximately half of SAKs tested yield a CODIS eligible profile, and approximately half of those uploaded profiles yield a CODIS hit?

Second, how do DNA testing rates, CODIS eligible rates, CODIS hit rates, and CODIS serial sexual assault hit rates vary by geographic and population density characteristics? We created county-level grouping variables and then compared forensic testing results *among* those groups to evaluate whether rates differ among urban counties of varying population densities, as well as among urban, suburban, and rural counties.

Third, focusing *within* each geographic and population density group, how do our obtained rates for DNA testing, CODIS eligibility, CODIS hits, and CODIS serial sexual assault hits compare to hypothetical threshold values for each of these phases of testing? We evaluated whether the rates documented in each of our county-level groupings exceeded commonly used heuristic criteria that practitioners and policymakers may consider when developing testing plans for previously unsubmitted SAKs. The process of inventoring unsubmitted kits is a considerable financial undertaking, and criminal justice system personnel may wonder whether testing these kits is truly necessary and whether it will yield results at some level and quantity that would justify the time, effort, and expense. For example, at a minimum threshold, are testing outcome rates significantly greater than zero? Are they greater than 33%? Greater than 50%? In low-density rural counties, for instance, if it is unlikely that CODIS hits would exceed the lower thresholds, stakeholders may need to consider how best to use limited laboratory resources. In high-density urban counties, if CODIS hit rates may exceed the higher thresholds, police and prosecutors will need careful planning to determine how they will take on a large number of new, active cases. Comparing obtained rates to hypothetical thresholds can inform scenario planning for communities of varying geographic and population density characteristics.

Methods

Sample

The Michigan State Police received 2015 BJA SAKI funding to complete a state census of all previously unsubmitted SAKs, *excluding* SAKs in custody of the City of Flint Police Department and the City of Detroit Police Department because separate 2015 BJA SAKI grants funded efforts to inventory and test those kits. The state census began in February 2016 when the state SAKI coordinator sent letters to all LEAs in Michigan's 83 counties (except the two noted above; n = 592 LEAs) instructing them to count all previously unsubmitted SAKs collected before March 1, 2015. This date was selected to align with recent state legislation, the Michigan Sexual Assault Evidence Kit Submission Act, MCL 752.931, which required all kits collected on or after March 1, 2015, to be submitted for testing and analyzed for DNA within statutorily defined time periods; thus, an inventory of all SAKs collected *prior to* March 1, 2015, would reflect a state census of all previously unsubmitted SAKs. LEAs were provided a standardized definition of "previously unsubmitted SAKs" to ensure uniform counting: "Under the terms of the grant, all unsubmitted kits must be accounted for and audited regardless of the reason why the kits were not previously submitted. For example, the following kits must be included in your inventory: kits where the complainant has refused to prosecute; kits believed to be beyond the statute of limitations; kits where a determination has been made that the charges are unfounded; and kits where the underlying case was adjudicated by trial or plea." SAKs released by victims for testing that were still in the physical custody of a medical facility were also to be included in the inventory, so the LEAs were instructed to reach out to medical facilities in their jurisdiction to collect and count any stored kits. LEAs were provided a standardized spreadsheet to record the requested information. This spreadsheet also tracked the specific law enforcement agency that had custody of each kit and the county in which that agency was located. Though it is possible that the reported crime occurred in a different county, stakeholders indicated that it was common practice to redirect victims to the appropriate law enforcement agency that had jurisdiction and the SAK would be taken into evidence by the LEA that had jurisdiction. The SAKI coordinator made repeated emails and phone calls to LEAs throughout spring and summer 2016 to monitor the inventory process. The inventory closed September 2016.

Of the 83 counties in this state, 25 counties had no LEAs with previously unsubmitted SAKs in custody; all of these counties are rural and are among the least densely populated rural counties in this state. The remaining 58 counties had LEAs with previously unsubmitted SAKs in custody. The total number of previously unsubmitted SAKs inventoried was $N = 3422$. We were able to determine the year the SAK was originally collected for nearly all SAKs ($n = 72$ kits had missing date information), and approximately 1% had been collected between 1980 and 1989, 9% from 1990 to 1999, 40% from 2000 to 2009; and 50% from 2010 to the close date of the census.

Procedures

The state SAKI coordinator conducted in-person site visits at urban and suburban LEAs to inspect the inventoried kits and work with local personnel to prepare the kits for shipment to an outside vendor laboratory for testing. Each kit was physically examined in the presence of an LEA property officer, and the information on the kit was compared to the data submitted on the inventory spreadsheet to ensure accuracy and that the kit met the parameters of the project. Kits were then bar coded and shipped to the vendor. For LEAs in rural jurisdictions, police personnel were instructed to ship their inventoried SAKs to the Michigan State Police Forensic Science Division (MSP-FSD), which completed the same accuracy and eligibility checks, and then bar coded and shipped the kits to the vendor lab for testing.

Sexual assault kits began shipping to the vendor in March 2016 and testing was largely completed by July 2018. Testing results were posted in a secure web portal. CODIS ineligible and negative kit reports (i.e., no DNA present) were recorded by MSP-FSD, and then, those SAKs were returned directly to each law enforcement agency by mail. CODIS eligible results were downloaded by MSP-FSD for independent technical review. The technical reviews were completed by September 2019; all CODIS uploads and local, state, and national DNA databases searches were completed in October 2019; and the data were released to the research team for analysis in November 2019.

Measures

Forensic DNA testing is a multi-stage process and the outcomes at each stage can be quantified for statistical analysis. We defined Stage 0 as the process of screening the samples in the SAKs to determine which can progress for DNA testing. In this study, the vendor laboratory used the y-screen method and recorded how many SAKs progressed to DNA testing; thus, the probability that a kit will pass from Stage 0 to Stage 1 is the *DNA testing rate*. In Stage 1, forensic scientists attempt to extract the DNA from the sample cells. If the resulting DNA profile has the requisite number of core loci for that specimen type, and there is reasonable belief that the sample is from the person who committed the reported crime, then in Stage 2, the profile can be uploaded into CODIS. The probability that a kit will pass from Stage 1 to Stage 2 is the *CODIS eligible rate*. Once a DNA profile is uploaded to CODIS, it is compared to other DNA samples that have been previously entered into the indexing systems (i.e., the offender index and the forensic index). If the DNA matches to an existing profile, it is termed a CODIS hit (Stage 3) and the probability that a DNA profile will pass from Stage 2 to Stage 3 is the *CODIS hit rate*. Finally, for each SAK that had a CODIS Hit, we checked what the hit was "hitting to," specifically, whether the hit was to a DNA profile from another sexual assault case. Consistent with operational definitions used in past research (32), a hit was coded as a CODIS serial sexual assault hit if: (i) the identified offenders had a qualifying offense in CODIS from another (different) sexual assault crime; or (ii) the forensic association in CODIS was to another sexual assault crime (e.g., another previously unsubmitted SAK or to a DNA profile in the forensic index from an unsolved sexual assault). The probability that a SAK will pass from Stage 3 to Stage 4 is the *CODIS serial sexual assault hit rate*.

To compare how DNA testing rates, CODIS eligible rates, CODIS hit rates, and CODIS serial sexual assault rates varied by geographic and population density characteristics, we created a county-level grouping variable (based on the county in which each SAK was collected, per the census spreadsheet) informed by the U.S. Office of Management and Budget's definition of metropolitan and micropolitan statistical areas (33). Per these definitions, metropolitan counties include at least one urban area with a population >50,000, and micropolitan counties include one or more urban clusters between 10,000 and 50,000 (33). We first grouped the counties in our sample according to whether they were metropolitan, micropolitan, or neither. Based on this initial categorization, we had a relatively high number of counties that met the definition of a metropolitan statistical area ($n = 26$) but varied substantially in population density, so we further divided the metropolitan counties based on population density per the 2010 Census (34). Thus, our final combined statistical area/population density grouping variable was comprised of five categories: category 1 = metropolitan counties with a population density of over 1000 people per square mile ($n = 3$ counties from the $N = 58$ counties that had previously unsubmitted SAKs in the state census; $n = 1329$ SAKs in these three counties); category 2 = metropolitan counties with a population density of 400–1000 people per square mile ($n = 6$ counties; $n = 1031$ SAKs in these six counties); category 3 = metropolitan counties with a population density of less than 400 people per square mile ($n = 17$ counties; $n = 858$ SAKs in these 17 counties); category 4 = micropolitan counties ($n = 19$ counties; $n = 157$ SAKs in these 19 counties); and category 5 = counties with no metropolitan or micropolitan areas ($n = 13$ counties; $n = 47$ SAKs in these 13 counties).

Results

Research Question 1: State-Level Forensic Testing Rates

Using the state census of $N = 3422$ SAKs, we counted how many kits progressed through each stage of forensic DNA testing and conducted continuation ratio modeling to obtain unconditional and conditional rate estimates, as well as their 95% confidence intervals (35,36) using R software, version 3.6.2 (37). The continuation ratio models require a set of stage-specific binary variables showing whether a SAK progressed past each stage to the next one (0 = no, 1 = yes) with any SAK reaching one of the later stages, by definition having progressed past every previous stage. All $N = 3422$ SAKs were submitted for testing and screened, and $n = 2193$ SAKs progressed to DNA testing (64.1% unconditional DNA testing rate, 95% CI 0.624–0.657). A total of $n = 1239$ SAKs had a DNA profile that met eligibility for upload to CODIS (36.2% unconditional and 56.5% conditional CODIS eligible rate, 95% CI 0.544–0.586), n = 585 SAKs yielded a CODIS hit (17.1% unconditional and 47.2% conditional CODIS hit rate, 95% CI 0.444–0.500), and n = 152 SAKs produced a CODIS serial sexual assault hit (4.4% unconditional and 26.0% conditional CODIS serial sexual assault hit rate, 95% CI 0.225–0.297).

Research Question 2: Forensic Testing Rates by Geographic and Population Density Characteristics

We examined how DNA testing rates, CODIS eligible rates, CODIS hit rates, and CODIS serial sexual assault hit rates varied by geographic and population density characteristics. Table 1 summarizes the proportions for each stage of forensic DNA testing by each of the five county-level groupings of metro- and micropolitan population density. The DNA testing rate in the overall sample was 64% (see above), and as seen in Table 1, this rate was quite similar across all groups, including metropolitan areas (1 = 0.64, 2 = 0.65, 3 = 0.64) and micropolitan areas (4 = 0.64). The DNA testing rate was lower in category 5, the rural counties that contained neither metro- nor micropolitan areas (0.47). We computed pairwise two-tailed chi-square tests of proportions of success probabilities in pairs of categories (38) using R software, version 3.6.2. These results indicated the DNA testing rate in category 5 was significantly lower than in categories 1, 2, 3, and 4. No other pairwise comparisons between groups were significant.

With respect to CODIS eligibility rates, the conditional rate in the overall sample was 57% (see above), and the rates in category 1 (the largest metro counties) and category 5 (the least densely populated rural counties) were even higher (0.60 and 0.64, respectively). The other county groupings had conditional CODIS eligible rates ranging from 0.49 to 0.54. Pairwise tests of proportions indicated the rates were significantly higher in group 1 compared to groups 2, 3, and 4 (i.e., the largest metro group had higher CODIS eligible rates than the other metro and micro groups). Due to the small number of SAKs in category 5 that progressed to this stage, the tests of proportions do not reflect significant differences.

Once those DNA profiles were uploaded to CODIS, 47% yielded a CODIS hit in the state census (see above). The CODIS hit rate was higher in category 2 (the second-largest metro group) at 0.52, which was significantly higher than category 5 (the least densely populated rural group) (0.31). All other groups had CODIS hit rates ranging from 0.42 to 0.46, and no other pairwise comparisons were statistically significant.

When we checked whether these CODIS hits were hitting to other sexual assault cases, we found that 26% of the hits overall were to another SAK or another sexual assault case in CODIS (see above). This rate was markedly higher in the category 1 counties (0.40) and lower in categories 2–4 (0.15–0.21). There were no CODIS serial sexual assault hits in the category 5 counties. Pairwise tests comparing rates across groups indicated significant differences between the largest metro counties (category 1) and the smaller metro county groupings (categories 2 and 3).

TABLE 1—*SAK forensic testing outcomes by metropolitan/micropolitan/population density category.*

Unconditional DNA Testing Rates

	DNA Test = 0 (no)		DNA Test = 1 (yes)		
Category	Frequency	Proportion	Frequency	Proportion	Pairwise Differences*
1	483	0.36	846	0.64	1 vs. 5 (p = 0.028)
2	358	0.35	673	0.65	2 vs. 5 (p = 0.015)
3	307	0.36	551	0.64	3 vs. 5 (p = 0.024)
4	56	0.36	101	0.64	4 vs. 5 (p = 0.047)
5	25	0.53	22	0.47	

Conditional CODIS Eligibility Rates

	CODIS Eligible = 0 (no)		CODIS Eligible = 1 (yes)		
Category	Frequency	Proportion	Frequency	Proportion	Pairwise Differences
1	335	0.40	511	0.60	1 vs. 2 (p = 0.018) 1 vs. 3 (p = 0.032) 1 vs. 4 (p = 0.029)
2	308	0.46	365	0.54	
3	251	0.46	300	0.54	
4	52	0.52	49	0.49	
5	8	0.36	14	0.64	

Conditional CODIS Hit Rates

	CODIS Hit = 0 (no)		CODIS Hit = 1 (yes)		
Category	Frequency	Proportion	Frequency	Proportion	Pairwise Differences
1	275	0.54	236	0.46	
2	174	0.48	191	0.52	2 vs. 3 (p = 0.048)
3	167	0.56	133	0.44	
4	29	0.59	20	0.41	
5	9	0.64	5	0.36	

Conditional CODIS Serial Sexual Assault (SA) Hit Rates

	Serial SA Hit = 0 (no)		Serial SA Hit = 1 (yes)		
Category	Frequency	Proportion	Frequency	Proportion	Pairwise Differences
1	142	0.60	94	0.40	1 vs. 2 (p < 0.001) 1 vs. 3 (p < 0.001)
2	162	0.85	29	0.15	
3	107	0.81	26	0.20	
4	17	0.85	3	0.15	
5	5	1.00	0	0.00	

Category 1 = metropolitan counties with a population density of over 1000 people per square mile; category 2 = metropolitan counties with a population density of 400–1000 people per square mile; category 3 = metropolitan counties with a population density of less than 400 people per square mile; category 4 = micropolitan counties; category 5 = counties with no metropolitan or micropolitan areas.

*Differences between categories are reported if p < 0.05.

Research Question 3: Obtained Forensic Testing Rates Vs. Heuristic Threshold Rates

For each of the five geographic and population density groups, we compared the obtained rates for DNA testing, CODIS eligibility, CODIS hits, and CODIS serial sexual assault hits to a series of hypothetical threshold values (zero, operationalized as 0.01; 0.33, and 0.50; see Table 2). We computed binomial exact tests of success probability for each rate being greater than the hypothetical threshold values (39), using

TABLE 2—*Comparing obtained SAK forensic testing rates to hypothetical thresholds.*

		p-Values		
Category	Proportion	Proportion > 0	Proportion > 0.33	Proportion > 0.5
Unconditional DNA Testing Rates				
1	0.64	<0.001	<0.001	<0.001
2	0.65	<0.001	<0.001	<0.001
3	0.64	<0.001	<0.001	<0.001
4	0.64	<0.001	<0.001	<0.001
5	0.47	<0.001	0.034	0.720
Conditional CODIS Eligibility Rates				
1	0.60	<0.001	<0.001	<0.001
2	0.55	<0.001	<0.001	0.003
3	0.52	<0.001	<0.001	0.173
4	0.53	<0.001	<0.001	0.309
5	0.56	<0.001	0.001	0.244
Conditional CODIS Hit Rates				
1	0.46	<0.001	<0.001	0.962
2	0.52	<0.001	<0.001	0.201
3	0.44	<0.001	<0.001	0.978
4	0.41	<0.001	0.156	0.924
5	0.36	<0.001	0.514	0.910
Conditional CODIS Serial Sexual Assault Hit Rates				
1	0.40	<0.001	0.016	0.999
2	0.15	<0.001	1.000	1.000
3	0.20	<0.001	1.000	1.000
4	0.15	0.001	0.981	1.000
5	0.00	1.000	1.000	1.000

Category 1 = metropolitan counties with a population density of over 1000 people per square mile; category 2 = metropolitan counties with a population density of 400–1000 people per square mile; category 3 = metropolitan counties with a population density of less than 400 people per square mile; category 4 = micropolitan counties; category 5 = counties with no metropolitan or micropolitan areas.

R software, version 3.6.2. For DNA testing rates, county groups 1, 2, 3, and 4 (i.e., all metropolitan and micropolitan areas) had obtained rates significantly higher than 0%, 33%, and 50%. For category 5 (the least densely populated group), DNA testing rates significantly exceeded the 0% and 33% threshold. The DNA testing rate in category 5 was 51%, which was not significantly greater than a rate of 50%. With respect to CODIS eligibility rates, all category groups exceeded the 0% and 33% thresholds, and categories 1–3 (the metropolitan counties) also exceeded the 50% threshold. The obtained CODIS hit rates exceeded the zero threshold in all groups and the 33% threshold for the metropolitan counties (categories 1–3). Because CODIS hit rates hovered around 50%, it not surprising that none significantly exceeded the 50% threshold. The CODIS serial sexual assault hit rates in this study exceeded the 0% threshold in all county groups, except category 5, which had no suspected serial sexual assault cases. Only category 1 counties exceeded the 33% threshold, and no groups exceeded the 50% threshold.

Discussion

Prior research on previously unsubmitted SAKs suggests that these kits contain potentially actionable information for law enforcement personnel and prosecutors. In studies of large urban cities with high crime rates (e.g., Cleveland, Detroit, Houston, Los Angeles), approximately half of SAKs tested have yielded a CODIS eligible profile and approximately half of those uploaded profiles yielded a CODIS hit. These findings suggest good immediate return on investment (i.e., first-search hits against current CODIS profiles) and longer-term returns by populating CODIS so that future uploaded samples will be searched against larger profile pools. Yet, whether these rates of return are unique to large urban areas was unknown, so in this study we tested a state census of previously unsubmitted SAKs that included kits collected in smaller cities and rural communities. Our census identified $N = 3422$ previously unsubmitted SAKs in Michigan LEAs; this sample did *not* include previously unsubmitted SAKs from Detroit, MI and from Flint, MI, which were inventoried and tested in separate grant projects. When the SAKs in our sample were outsourced for forensic DNA testing, 57% tested yielded a CODIS eligible profile and 47% of those uploaded profiles yielded a CODIS hit. Thus, even with the exclusion of two large cities with high crime rates, we still found strong yield rates for CODIS eligible profiles and CODIS hits that were consistent with prior research.

We compared forensic DNA testing outcomes by geographic and population density characteristics to explore whether the obtained rates were "pulled up" by results from larger metropolitan areas. We created five county-level groupings: three metropolitan groups varying in population density, one micropolitan group, and one group of rural areas so sparsely populated they contained neither metro- nor micropolitan centers. The DNA testing rates in the metropolitan groups and the micropolitan group were quite similar, but the rural counties had significantly lower rates. The sample size for category 5 was small ($n = 47$ SAKs), so we are hesitant to draw conclusions about why DNA testing rates were lower in these communities. One possibility is that healthcare providers in rural areas may have had less training on evidence collection techniques, which could affect DNA recovery. SANE program staff have extensive training in forensic evidence collection procedures, but if patients in rural areas have difficulty accessing these programs, that could affect DNA testing rates.

The CODIS eligibility rates across the five groups ranged from 0.49 to 0.64; rates were significantly higher in the largest metro groups relative to smaller metro- and micropolitan areas, but all groups were near the 50% mark. For practitioners and policymakers, these results suggest that when communities test previously unsubmitted SAKs, they can certainly expect CODIS eligibility rates above 0.33, and most likely rates near 0.50, with rates above 0.50 in larger metro areas. Once these profiles were uploaded, the CODIS hit rates across county groupings ranged from 0.42 to 0.52, with significantly lower rates in the least densely populated rural areas (0.31). Yet, even in these most rural regions of the state, this is a notable rate of immediate return, and more hits may come later as CODIS is continuously populated over time.

In this study, we delved deeper into each of these CODIS hits to ascertain whether the DNA match was to another sexual assault case, indicating a pattern of suspected serial sexual offending. Previous studies in Cleveland and Detroit have found that 29–56% of CODIS hits match to other sexual assault cases, highlighting the utility of forensic testing to police and prosecutors for identifying possible serial offenders. In this state census, we found a lower overall CODIS serial sexual assault hit rate (26%), with statistically significant differences between the largest metro group (40%) and the smaller metro and micropolitan areas (15–20%). These numbers can inform scenario planning, both at the local level and at the state level, as police and prosecutors should anticipate at least 15–20% of their hits will be linked to other sexual assault cases. These findings may be surprising to practitioners in smaller cities and rural communities, as serial sexual offending is often thought of as an "urban problem" (40), but our data suggest that even in communities as small as 10,000–50,000 people, one-in-five CODIS hits may identify a suspected serial sexual offender. These findings suggest that mid-sized and rural jurisdictions may need additional staffing, particularly well-trained investigators and advocates, to work these cases. Connecting practitioners in these smaller communities to each other—and to "mentor" practitioners in larger communities—may be helpful. The SART (sexual assault response team)/MDT (multidisciplinary team) model provides a useful structure for this kind of collaboration (41). The BJA SAKI program recommends the SART/MDT model for communities addressing previously unsubmitted SAKs to increase collaboration and competencies of practitioners (31). Our findings emphasize that investing in SART/MDT development in smaller communities is particularly important, as they may have less initial preparedness to respond to cold case sexual assaults, including cold case serial sexual assaults.

To the best of our knowledge, this is the first study in the literature to evaluate SAK forensic DNA testing outcomes across markedly diverse community types, but there are several key limitations of this study that must be noted. First, the state census does not include Detroit, MI or Flint, MI, as those communities were inventoried in other grant projects. Thus, we do not have a true state-wide count of all unsubmitted SAKs. Other published studies have reported that Detroit had 11,217 unsubmitted SAKs (32), which is clearly an extreme outlier from any other LEA in the category 1 county group. If Detroit's census had been included in this project, it would have been necessary to model those data separately anyway because it is a statistical outlier (and that work has already been done in other published studies [32]). In public reports, Flint officials have indicated they have approximately 250 unsubmitted SAKs (42), which is a comparable volume of kits to other LEAs in our category 2 county group. There are no indications that Flint's inventory or testing results are anomalous, so even if we were able to include their data with other category 2 SAKs in this study, it is unlikely that their data would change our findings, given their reported sample size. As such, there is no reason to believe that these data would have skewed our findings regarding forensic testing rates or heuristic threshold rates. As such, we acknowledge the absence of Detroit's and Flint's data as a limitation, but not one that meaningfully calls into question the validity of the analyses presented in this paper.

Second, although the state census of 3422 previously unsubmitted SAKs is a large number, once SAKs progress through the forensic testing stages, the numbers reduce substantially in some of the county-level groupings (particularly category 5). Small sizes limited what tests of significance we could use to evaluate group differences, and we urge caution in interpreting chi-square tests of proportions and binomial exact tests with small sample sizes. Future studies should evaluate the feasibility of pooling data across multiple sites (e.g., multiple SAKI-funded sites) to increase sample sizes for rural jurisdiction groupings, which would expand options for statistical analysis and increase the validity of the findings.

Third, the research team did not have access to some types of data that would have been useful to including in our statistical analyses. County-level factors that may have affected forensic testing outcomes include the level of resources in the different LEAs to respond to reported sexual assaults (e.g., staffing levels) and the availability of specialized forensic nursing programs/sexual assault nurse examiner (SANE) programs for the collection of the evidence itself. Future research should employ multi-level research designs to assess how contextual variables affect individual-level testing outcomes. Large sample sizes will be needed for such work, so again, pooling data across sites would allow for more sophisticated research designs and analyses. We also did not have full information about each offender's CODIS-qualifying offenses. We know whether there was at least one other separate sexual assault case that necessitated DNA collection and upload to CODIS, but these offenders may have been linked to additional sexual assault cases. We did not have criminal history records for the identified offenders, so we do not know their arrest, charge, and conviction histories for other types of crime. Research that has delved into sex offenders' full criminal records indicates they often have diverse offending patterns across multiple types of crime (43), but such work often focuses on offenders in large urban areas. Future research needs to examine how levels and types of crossover offending vary by geographic and population density characteristics as well.

Finally, we emphasize that this study did not assess how CODIS hits and CODIS serial sexual assault hits were utilized by police and prosecutors in these jurisdictions and whether the forensic DNA testing results contributed to new arrests and convictions. Emerging data suggest that DNA evidence is instrumental in prosecutorial decision making for charging decisions in sexual assault case (44,45) so this is an important area for continued study. We have follow-up studies planned in both metro- and micropolitan jurisdictions in this state to explore these questions.

With these limitations noted, the results of this study can still inform local, state, and national policy regarding the problem of unsubmitted SAKs. When BJA established the SAKI project in 2015, it was clear the issue of unsubmitted SAKs was a national-scale issue that required strong, sustained support for inventorying and testing SAKs, and investigating and prosecuting reported sexual assaults. The intentional focus to serve not only large urban areas, but also smaller communities and rural counties have provided the opportunity to study how forensic testing rates vary by community characteristics. Michigan's mix of large urban–suburban and multiple small rural communities provides empirical guidance for scenario planning in other jurisdictions that have large numbers of unsubmitted SAKs. Our results are consistent with prior studies indicating strong return on investment for testing rape kits and for CODIS as an investigatory tool—in urban and non-urban community contexts. Our findings also underscore the importance of populating CODIS to maintain high return on testing investments, which, in turn, requires increasing support for publicly funded crime laboratories so they can keep pace with growing demand for their services from the criminal justice system.

Acknowledgments

The authors thank the members of the Michigan State Police SAKI team for their support of this project, and Lucas Walters for his contributions to data coding.

References

1. Bureau of Justice Statistics. Census of publicly funded forensic crime laboratories. Washington, DC: US Department of Justice, 2009.
2. Bureau of Justice Statistics. Publicly funded forensic crime laboratories: resources and services. Washington DC: US Department of Justice, 2014.
3. Government Accountability Office. DNA: DOJ should improve performance measurement and proper design control for national wide grant program. Washington, DC: Government Accountability Office, 2019.

4. Committee on Identifying the needs of Forensic Sciences Community, National Research Council. Strengthening forensic science in the United States: a path forward. Washington, DC: National Academies Press, 2009.
5. Strom KJ, Hickman MJ. Unanalyzed evidence in law-enforcement agencies: a national examination of forensic processing in police departments. Criminol Public Policy 2010;9(2):381–404. https://doi.org/10.1111/j.1745-9133.2010.00635.x
6. Campbell R, Feeney H, Fehler-Cabral G, Shaw J, Horsford S. The national problem of untested sexual assault kits (SAKs): scope, causes, and future directions for research, policy, and practice. Trauma Violence Abuse 2017;18(4):363–76. https://doi.org/10.1177/1524838015622436
7. Pinchevsky GM. Criminal justice considerations for unsubmitted and untested sexual assault kits: a review of the literature and suggestions for moving forward. Crim Justice Policy Rev 2018;29(9):925–45. https://doi.org/10.1177/0887403416662899
8. Strom KJ, Hendrix JA, Parish WJ, Melton P.Efficiency in processing sexual assault kits in crime laboratories and law enforcement agencies. Washington, DC: National Institute of Justice, 2017 Aug. Report No.: 2013-NE-BX-0006.
9. Human Rights Watch. Testing justice: the rape kit backlog in Los Angeles City and County. New York, NY: Human Rights Watch, 2009.
10. Human Rights Watch. “I used to think the law would protect me”: Illinois’s failure to test rape kits. New York, NY: Human Rights Watch, 2010.
11. Department of Justice. A national protocol for sexual assault medical forensic examinations: adults and adolescents, 2nd rev edn. Washington, DC: Department of Justice, 2013.
12. Campbell R, Shaw J, Fehler-Cabral G. Shelving justice: the discovery of thousands of untested rape kits in Detroit. City Community 2015;14 (2):151–66. https://doi.org/10.1111/cico.12108
13. Tofte S.Guest blog Sarah Tofte: police often abandoned rape kits and investigations. http://www.cleveland.com/rape-kits/index.ssf/2013/08/guest_blog_sarah_tofte.html (accessed May 12, 2020).
14. Campbell R, Fehler-Cabral G, Pierce S, Sharma D, Bybee D, Shaw J, et al.The Detroit sexual assault kit action research project final report. Washington, DC: National Institute of Justice, 2015 Dec. Report No.: NIJ-2011-DN-BX-0001.
15. Lovell R, Luminais M, Flannery DJ, Bell R, Kyker B. Describing the process and quantifying the outcomes of the Cuyahoga County sexual assault kit initiatives. J Crim Justice 2018;57:106–15. https://doi.org/10.1016/j.jcrimjus.2018.05.012
16. Peterson JL, Johnson D, Herz D, Graziano L, Oehler T.Sexual assault kit backlog study. Washington, DC: National Institute of Justice, 2012 June Report No.: 2006-DN-BX-0094.
17. Wells W, Campbell BC, Franklin C.Unsubmitted sexual assault kits in Houston, TX: case characteristics, forensic testing results, and the investigation of CODIS hits. Washington, DC: National Institute of Justice, 2016 Apr. Report No.: 2011-DN-BX-0002.
18. Davis RC, Wells W. DNA testing in sexual assault cases: when do the benefits outweigh the costs? Forensic Sci Int 2019;299:44–8. https://doi.org/10.1016/j.forsciint.2019.03.031
19. Lovell R, Singer M, Flannery D. Cost savings and cost effectiveness of the Cuyahoga County sexual assault task force. Begun Center for Violence Prevention Research and Education, 2016https://digital.case.edu/islandora/object/ksl:2006061446
20. Speaker PJ. The jurisdictional return on investment from processing the backlog of untested sexual assault kits. Forensic Sci Int Synergy 2019;1:18–23. https://doi.org/10.1016/j.fsisyn.2019.02.055
21. Wang C, Wein LM. Analyzing approaches to the backlog of untested sexual assault kits in the USA. J Forensic Sci 2018;63(4):1110–21. https://doi.org/10.1111/1556-4029.13739
22. Federal Bureau of Investigations (FBI). Crime in the United States: 2010 uniform crime reports. https://ucr.fbi.gov/crime-in-the-u.s/2010/crime-in-the-u.s.-2010 (accessed May 12, 2020).
23. Martin PY. Rape work: victims, gender, and emotions in organization and community context. New York, NY: Routledge, 2005;73–93.
24. Lynch VA. Forensic nurse examiners. Am J Nurs 1989;89(2):176.
25. Butler JM. Fundamentals of forensic DNA typing. Waltham, MA: Elsevier, 2010;3–16.
26. Department of Justice. The FBI’s combined DNA index system program: CODIS. Washington, DC: Department of Justice, 2000.
27. Campbell R, Fehler-Cabral G. Why police “couldn’t or wouldn’t” submit sexual assault kits (SAKs) for forensic DNA testing: a focal concerns theory of analysis of untested rape kits. Law Soc 2018;52(1):73–105. https://doi.org/10.1111/lasr.12310.
28. Lovell R, Luminais M, Flannery D. Perceptions of why the sexual assault kit backlog exists in Cuyahoga County, Ohio and recommendations for improving practice. Begun Center for Violence Prevention Research and Education, 2017. https://digital.case.edu/islandora/object/ksl:2006061457 (accessed July 24, 2020).
29. Campbell R, Feeney H, Goodman-Williams R, Sharma DB, Pierce SJ. Connecting the dots: identifying suspected serial sexual offenders through forensic DNA evidence. Psychol Violence 2019;10(3):255–67. https://doi.org/10.1037/vio0000243
30. Lovell R, Luminais M, Flannery DJ, Overman L, Huang D, Walker T, et al. Offending patterns for serial sex offenders identified via the DNA testing of previously unsubmitted sexual assault kits. J Crim Justice 2017;52:68–78. https://doi.org/10.1016/j.jcrimjus.2017.08.002.
31. Bureau of Justice Assistance. The sexual assault kit initiative (overview). https://bja.ojp.gov/program/sexual-assault-kit-initiative-saki/overview (accessed May 12, 2020).
32. Campbell R, Fehler-Cabral G, Pierce SJ, Sharma D, Bybee D, Shaw J, et al.The Detroit sexual assault kit (SAK) action research project (ARP). Washington, DC: National Institute of Justice, 2015 Nov. Report No: 2011-DN-BX-0001.
33. United States Census Bureau. Metropolitan and micropolitan: about. https://www.census.gov/programs-surveys/metro-micro/about.html (accessed May 12, 2020).
34. United States Census Bureau. Michigan: 2010 population and housing unit counts, CPH-2-24, 2012. https://www.census.gov/prod/cen2010/cph-2-24.pdf (accessed May 12, 2020).
35. Agresti A. Categorical data analysis, 2nd rev edn. New York, NY: John Wiley & Sons Inc, 2002.
36. Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression, 3rd rev edn. Hoboken, NJ: John Wiley & Sons Inc, 2013;291–3.
37. R Development Core Team. R: a language and environment for statistical computing (Version 3.6.2) [Computer program]. Vienna, Austria: R Foundation for Statistical Computing, 2019.
38. Newcombe RG. Interval estimation for the difference between independent proportions: comparisons of eleven methods. Stat Med 1998;17 (8):873–90. https://doi.org/10.1038/s41598-020-59747-0
39. Clopper CJ, Pearson ES. The use of confidence of fiducial limits illustrated in the case of the binomial. Biometrika 1934;26(4):404–13. https://doi.org/10.2307/2331986
40. Edelstein A. Rethinking conceptual definitions of the criminal career and serial criminality. Trauma Violence Abuse 2015;17(1):62–71. https://doi.org/10.1177/1524838014566694
41. Greeson MR, Campbell R, Bybee D, Kennedy AC. Improving the community response to sexual assault: an empirical examination of the effectiveness of sexual assault response teams (SARTs). Psychol Violence 2016;6(2):280–91. https://doi.org/10.1037/a0039617
42. Keefer W.Flint gets over $1.6 million to help local law enforcement process sexual assault kits. Mlive 2019 Sept 20; https://www.mlive.com/news/flint/2019/09/flint-gets-over-16-million-to-help-local-law-enforcement-process-sexual-assault-kits.html (accessed July 27, 2020).
43. Lussier P, Tzoumakis S, Cale J, Amirault J. Criminal trajectories of adult sex offenders and the age effect: examining the dynamic aspect of offending in adulthood. Int Crim Justice Rev 2010;20(2):147–68. https://doi.org/10.1177/1057567710368360
44. Alderden M, Cross TP, Vlajnic M, Siller L. Prosecutors’ perspectives on biological evidence and injury evidence in sexual assault cases. J Interpers Violence 2018. https://doi.org/10.1177/0886260518778259
45. Henry TK, Jurek AL. Identification, corroboration, and charging: examining the use of DNA evidence by prosecutors in sexual assault cases. Fem Criminol 2020. https://doi.org/10.1177/1557085120940795

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14508
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Jessica M. McLamb,[1,2] *M.S.; Lara D. Adams*,[3] *M.S.; and Mark F. Kavlick*,[2] *Ph.D.*

Comparison of the M-Vac® Wet-Vacuum-Based Collection Method to a Wet-Swabbing Method for DNA Recovery on Diluted Bloodstained Substrates*'†'‡

ABSTRACT: A wet-vacuum-based collection method with the M-Vac® was compared to a wet-swabbing collection method by examining the recovery of diluted blood on 22 substrates of varying porosity. The wet-vacuum method yielded more total nuclear DNA than wet-swabbing on 18 porous substrates, recovering on average 12 times more DNA. However, both methods yielded comparable amounts of total DNA on two porous and two nonporous substrates. In no instance did wet-swabbing significantly recover more DNA. The wet-vacuum method also successfully collected additional DNA on previously swabbed substrates. Mitochondrial DNA yields were assessed, and outcomes were generally similar to the nuclear DNA outcomes described above. Results demonstrate that wet-vacuuming may serve as an alternative collection method to swabbing on difficult porous substrates and could potentially recover additional DNA on previously swabbed substrates. However, swabbing remains the preferred collection method on substrates with visible stains and/or nonporous surfaces for reasons of convenience, simplicity, and lower cost relative to the wet-vacuum method.

KEYWORDS: wet-vacuum, M-Vac®, wet-swab, blood, DNA collection, DNA extraction, DNA quantification, forensic analysis

The sensitivity of forensic DNA testing has steadily increased and improved over the last 20 years through advances in DNA extraction, detection, and analysis. Yet the routine use of conventional collection methods, for example, swabbing (1–4), cutting (5,6), taping (7–9), means that improvements in sensitivity have been limited to post-collection processing. While these conventional collection techniques are effective for some substrates, they have limited efficacy for large, porous, absorbent, rough, and/or creviced substrates where the DNA may be too diffuse or unavailable for surface sampling. An alternative collection method which utilizes wet-vacuum technology has been developed to optimize DNA recovery from challenging items of interest where DNA may be absorbed within the substrate matrix.

The wet-vacuum-based collection system is designed for recovering DNA from porous substrates (10). The system consists of a vacuum, a hand-held collection device, a sample collection bottle, and sterile solution. It functions by dispensing the sterile solution onto a substrate while simultaneously vacuuming cellular material into the sample collection bottle. The liquid contents of the bottle are then filtered through a 0.45 μM polyethersulfone (PES) membrane in a two-stage filter unit, which traps and concentrates cellular material on the filter. Lastly, the filter is cut from the unit and processed for DNA extraction using common procedures.

There are published (11–14) and other academic research (15–17) studies on the use of a wet-vacuum-based collection system for forensic purposes. In one study, the wet-vacuum method was shown to perform better than double swabbing and taping methods for bloodstain collection on denim and carpet (11). In others, the wet-vacuum approach was more successful at collecting dried saliva from bricks (12) and laminated wood (13) compared to swabbing. It was also demonstrated that DNA quantities recovered with a wet-vacuum were comparable to those recovered with swabbing on nonporous materials, that is, tiles and glass (11,13), as well as human skin before and after showering (16,14). For touch DNA samples on cotton t-shirts, the wet-

[1] Visiting Scientist Program, Laboratory Division, Federal Bureau of Investigation, 2501 Investigation Parkway, Quantico, VA, 22135.

[2] Research & Support Unit, Laboratory Division, Federal Bureau of Investigation, 2501 Investigation Parkway, Quantico, VA, 22135.

[3] DNA Casework Unit, Laboratory Division, Federal Bureau of Investigation, 2501 Investigation Parkway, Quantico, VA, 22135.

Corresponding author: Mark F. Kavlick, Ph.D. E-mail: mfkavlick@fbi.gov

*Presented in part at the Southern Association of Forensic Scientists, May 3, 2019, in Asheville, NC; the International Symposium on Human Identification, September 25, 2019, in Palm Springs, CA; and the Northeastern Association of Forensic Scientists, November 14, 2019, in Lancaster, PA.

†Supported in part by an appointment to the Visiting Scientist Program at the FBI Laboratory Division, administered by the Oak Ridge Institute for Science and Education, through an interagency agreement between the U.S. Department of Energy and the FBI.

‡This is publication number 20-07 of the Laboratory Division of the Federal Bureau of Investigation (FBI). Names of commercial manufacturers are provided for information only and inclusion does not imply endorsement by the FBI or the U.S. Government. The views expressed are those of the authors and do not necessarily reflect the official policy or position of the FBI or the U.S. Government.

Received 24 April 2020; and in revised form 18 June 2020; accepted 22 June 2020.

[Correction made 21 October. The Highlights section was removed.]

vacuum recovered more DNA than direct fabric cuttings (17). Although the higher yielding wet-vacuum samples provided DNA profiles more consistently than the fabric cuttings, some alleles belonging to individuals outside of the study were observed which increased the degree of mixed profiles (17). Touch DNA samples on bricks were also examined; however, the variability inherent to touch DNA studies limited the author's ability to draw conclusions (12). These previous studies provided some insights as to the performance of wet-vacuum-based collections; however, the variety of substrates tested was limited. Thus, sampling efficiency remains somewhat unclear on many difficult, forensically relevant substrates.

This study endeavored to expand evaluation of the wet-vacuum system as a possible DNA recovery method for use on multiple challenging substrates. Blood was deposited on 22 substrates in a diluted concentration designed to allow the differences in DNA recovery efficiency to be evaluated. First, DNA recovery from items collected with the wet-vacuum and the wet-vacuum manufacturer's extraction protocol was compared to DNA recovery using a conventional wet-swabbing and an automated magnetic bead-based extraction technique. Second, the wet-vacuum was also used on 10 previously swabbed substrates to recover potentially uncollected DNA. Lastly, efficiency of the collection techniques was assessed by using the same downstream extraction method for both wet-vacuum and wet-swab collections. Total DNA yields obtained from wet-vacuuming and wet-swabbing were quantitatively compared to assess each method's capability to recover DNA on challenging substrates. By overcoming some of the limitations associated with traditional collection techniques for specific substrate types, the wet-vacuum approach may be an effective alternative for forensic examiners when conventional methods yield poor DNA results.

Materials and Methods

Substrate Preparation

The 22 substrates of varying porosity examined in this study included household items (glass, wood countertop, drywall painted with flat, satin, semi-gloss, and gloss paints, carpet padding, and outdoor carpet), construction materials (pressure-treated wood, oak, pine, plywood, brick, hemp rope, nylon rope, cinderblock, and unpainted drywall), and automotive items (carpet, seat cushion insert, seat cushion collar, trunk liner, and trunk mat). Glass, which served as a control substrate, and wood countertop were the only nonporous substrates examined.

Blood was voluntarily obtained from a single human subject with informed consent, and the resultant DNA extracts were quantified but not sequenced or typed in this study. Blood was diluted 1/100 in sterile Butterfield's buffer (0.3 mM monobasic potassium phosphate, pH 7.2, M-Vac® Systems, Inc., Sandy, UT). Substrates were obtained in new condition, except for automotive items, and were wiped with disposable low-lint laboratory wipes to remove dust and/or loose debris then UV irradiated for 15 min before sample application. The automotive carpet and seating were laundered and UV irradiated before sample application.

A 12-multichannel pipet was used to evenly distribute the diluted blood into 12 rows of 10 μL drops in an approximate 100 cm^2 area on most substrates for a total volume of 1.44 mL (14.4 μL of whole blood). The nylon and hemp rope substrates were cut into one-foot segments and spotted with 1.44 mL diluted blood along the horizontal length of the rope. Spotting was performed in triplicate for both the wet-swabbing and wet-vacuum methods on each substrate. The bloodstains were allowed to air dry overnight prior to collection. Reagent blank controls consisting of Butterfield's buffer were prepared for the wet-vacuum and wet-swab methods on all substrates.

Wet-Swab Method

Dried bloodstains were collected using a single wet-swab method with a sterile wooden-stemmed cotton swab (Puritan, Guilford, ME) moistened with 50 μL molecular biology grade (MBG) water. Then, the swab was rubbed and rolled over the spotted area in back-and-forth motions with pressure until the visible stain was transferred. Some substrates, for example, carpet, cinderblock, and drywall, required more wetting for stain retrieval; therefore, an additional 50 μL of MBG water was pipetted onto the swab head. Swab heads were then cut off from the wooden stem using sterile scissors and transferred into Investigator™ Lyse&Spin baskets (Qiagen, Hilden, Germany). DNA extraction was performed according to an automated magnetic bead-based method wherein swab heads were digested in 423 μL of Buffer G2 (Qiagen), 13.5 μL of 20 mg/mL proteinase K, and 13.5 μL of 1 M dithiothreitol (DTT) at 56°C for 1 h with 200 rpm shaking (18). After incubation, the samples were centrifuged at 16K × *g* for 5 min. The baskets containing the swab heads were discarded and the lysates were processed on the EZ1 Advanced XL (Qiagen) using the large volume protocol and eluted in 50 μL of TE buffer (10 mM Tris-HCl, 0.1 mM EDTA, pH 8.0).

Wet-Vacuum Method

The M-Vac® (M-Vac® Systems, Inc.), consisting of a Support Equipment Case (SEC) 100, a sampling device with collection bottle, Butterfield's buffer, and plastic tubing, was used for wet-vacuum collections per the manufacturer's instructions (19). Collection consisted of applying the vacuum force and spray dispersal while repeatedly moving the hand-held sampling device in forward and backward motions over the entire spotted area (12 rows × 12 columns of 10 μL spots – 1440 μL of 1/100 diluted blood in total – over an approximate 100 cm^2 area). The sampling device was also dragged over the spotted area with only the vacuum force applied, that is, solution spray off, to collect the residual liquid remaining on the substrate. A new sampling device was used for each collection. The collection volume was approximately 150 mL of Butterfield's buffer per sample, which permitted three passes of the ~100 cm^2 substrate area with the wet-vacuum collecting about 50 mL for each pass. The dried bloodstains required repeated collection cycles to release the stain from the substrates. Additionally, the technique and volume used here may not be optimal in all scenarios; collection can vary depending on the type of sample and surface it is deposited on.

The head of the sampling device is flat, and therefore, the spraying action and suction of the wet-vacuum system worked optimally when it was in complete contact with the substrate; indeed, many of the substrates in this study were relatively flat. Because ropes have irregular and rounded surfaces, they were placed inside a sterile plastic tray for wet-vacuum collection. The wet-vacuum was applied to the surface of the ropes and the residual, run-off liquid which was not captured during collection on the surface was retrieved in the tray via suction as well.

The collected buffer solution was concentrated by pouring the liquid contents of the collection bottle through a Nalgene™ Rapid-Flow™ two-stage filter unit with a 0.45 μM PES membrane filter (Thermo Fisher Scientific, Waltham, MA). To maximize the recovery of cellular material, the filtrate was poured back into the collection bottle, swirled to dislodge cells from the bottle walls and poured over the same PES membrane filter once again. The membrane filter, while still damp, was cut into eight strip pieces using a sterile scalpel and transferred into two Investigator® Lyse&Spin baskets (Qiagen) with sterile forceps, four filter strips in each, for efficient digestion. To maximize surface area exposure, the filter strips were stacked and loosely rolled into coils when they were placed inside the baskets.

The filter samples were then extracted following the manufacturer's recommended automated magnetic bead-based method (M-Vac® Systems, Inc., personal communication, Jan. 24, 2018). Briefly, samples were digested in 490 μL of 1:1 diluted Buffer G2 (Qiagen) in MBG water and 10 μL of 20 mg/mL proteinase K at 56°C for 15 min with shaking at 850 rpm. After incubation, samples were centrifuged at 16K × *g* for 5 min to collect the filtered lysate. The spin basket containing the filter strips was removed, and Buffer MTL plus 1 μg/μL carrier RNA (Qiagen) was added. Finally, the samples were purified on the EZ1 Advanced XL (Qiagen) using the large volume protocol and eluted in 50 μL of TE buffer. The eluates from the same filter were then combined into a single tube for a 100 μL total elution volume.

DNA Quantification

Samples were quantified in duplicate using the Quantifiler™ Human Plus (HP) DNA Quantification Kit (Applied Biosystems, Foster City, CA) in accordance with the manufacturer's protocol and analyzed using the small autosomal target concentration. Samples were also quantified using a published and validated real-time quantitative PCR assay for human mitochondrial DNA (mtDNA) (20). Both assays were performed on an ABI 7500 Real-Time PCR System with the HID Real-Time PCR Software (Applied Biosystems). Average total DNA yields from the wet-vacuum method and the wet-swabbing method were compared to approximate relative collection efficiencies. Although mtDNA analysis is not typically performed on surface-deposited DNA, mtDNA was used as another quantitative measure to determine collection efficiency. MtDNA outcomes were generally similar to nuclear DNA (nDNA) outcomes and are therefore presented as Figures S1–S4. Additionally, all reported total DNA yields are detailed in Tables S1–S4.

Wet-Vacuum Recovery Following Wet-Swabbing

Ten of the previously swabbed substrates were subsequently subjected to wet-vacuum collection in an attempt to recover additional DNA. The following substrates were chosen based on relatively low DNA yields obtained via swabbing: pressure-treated wood, pine, brick, automotive seating (cushion collar), automotive carpet, trunk liner, trunk mat, carpet padding, and painted drywall (flat and satin paint). The length of time which passed between the initial swabbing and subsequent wet-vacuuming of these substrates ranged from 2 to 79 days. Wet-vacuum collection, concentration via filtration, extraction, and quantification were performed in the same manner as previously described.

Collection Efficiency

To isolate cellular collection efficiency (while holding DNA isolation efficiency constant), diluted 1/100 bloodstains on glass were collected with the wet-vacuum and wet-swabbing techniques in triplicate and were extracted using the same automated magnetic bead-based extraction protocol previously described under the wet-swab method (18). Quantification was performed as previously described for nDNA recovery.

Statistical Analysis

Unpaired, two-tailed *t* tests, assuming equal or unequal variance as determined by *F* tests, were carried out to determine significant differences between wet-vacuum or wet-swab data sets at a 0.05 significance level.

Results

Comparing Total DNA Yields from Wet-Vacuum and Wet-Swab Methods

The average total DNA yields obtained with a wet-vacuum on 22 substrates were compared to those recovered with a wet-swab method. Figures 1–3 show the average total nDNA yields for the two methods from household items, construction materials, and automotive items, respectively. There was no indication of PCR inhibition or degradation in any of the samples according to the qPCR results and no reagent blank yielded DNA as expected (data not shown).

Of the 20 porous substrates, the wet-vacuum method resulted in consistently greater nDNA yields than the wet-swab method on all but two surfaces, that is, cinderblock and unpainted drywall. In total, wet-vacuuming recovered more nDNA on 18 porous substrates compared to wet-swabbing, eight of which were significantly greater. Additionally, the amount of DNA recovered with the wet-vacuum was generally several-fold greater, ranging from 3× to 66× on the household items, 2×–28× on the construction materials, and 10×–47× on the automotive items. Overall, the wet-vacuum yielded an average of 12 times more nDNA compared to the wet-swab. Average mtDNA yields were 17 times greater overall for wet-vacuuming than for wet-swabbing (Figures S1–S3).

Wet-swabbing did not recover significantly more DNA on any substrate. However, both methods yielded comparable DNA amounts on two nonporous (glass and wood countertop) and two porous substrates (unpainted drywall and cinderblock). Yields from both methods on the unpainted drywall and cinderblock were generally lower than the yields obtained from other porous substrates.

Wet-Vacuum Recovery Following Wet-Swabbing

Average nDNA yields from the wet-vacuum method only, the wet-swab method only, as well as the wet-vacuum following the wet-swab method were compared for 10 of the originally tested substrates (Fig. 4). For nine substrates, the wet-vacuum recovered additional DNA that was, at minimum, equivalent to the initial swabbing, and maximally 46-fold more. For three substrates, the wet-vacuum recovered significantly greater yields than the initial swabbing. The satin-painted drywall was the only substrate where the wet-vacuum after wet-swab DNA yields were lower than the initial swabbing. Altogether, wet-vacuuming

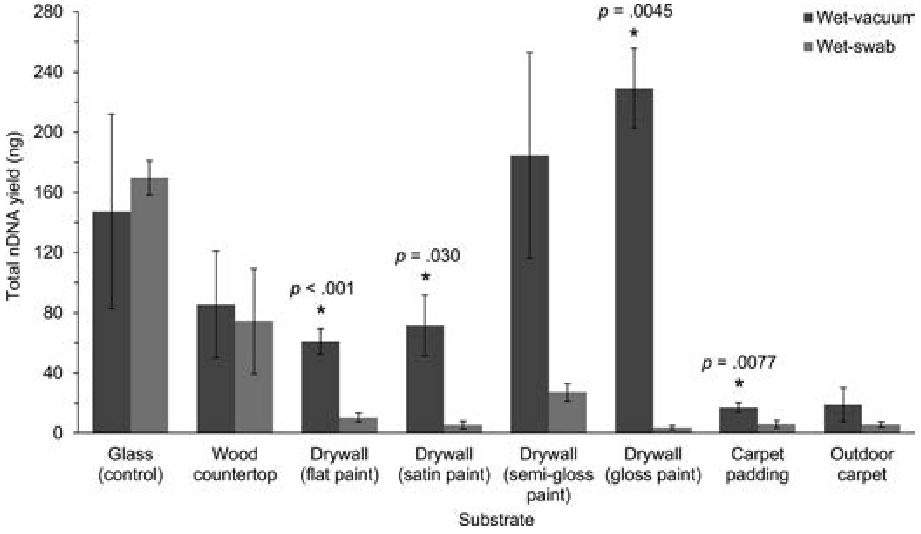


FIG. 1—*Average total nDNA yields recovered with the wet-vacuum or wet-swab methods for 1/100 bloodstains applied onto household items. Error bars represent the standard deviation of three replicates. Asterisks indicate significantly greater mean yields; the* p-*values are reported above. [Color figure can be viewed at wileyonlinelibrary.com]*

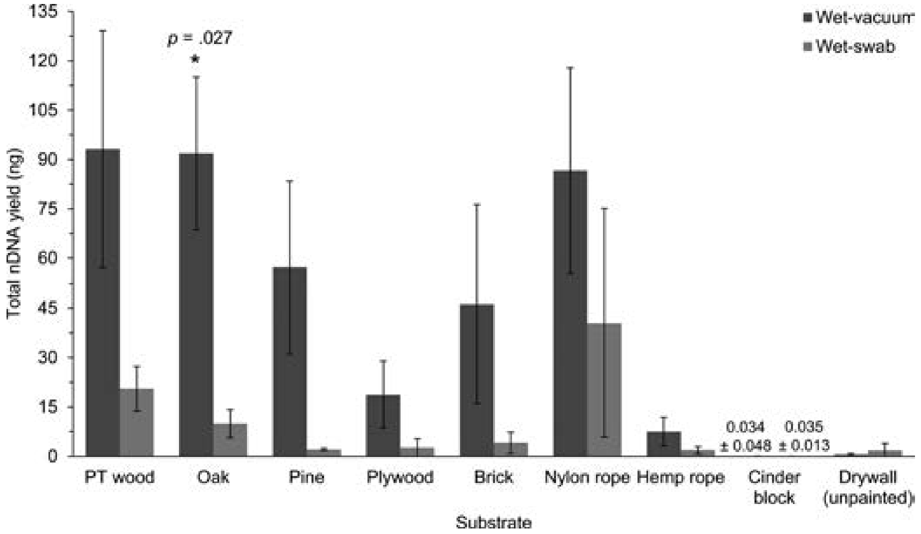


FIG. 2—*Average total nDNA yields recovered with the wet-vacuum or wet-swab methods for 1/100 bloodstains applied onto construction materials. The average yields from both methods were very low for cinderblock; therefore, those quantities (mean ± SD, ng) are reported within the figure. Error bars represent the standard deviation of three replicates. The asterisk indicates a significantly greater mean yield; the* p-*value is reported above. PT wood, pressure-treated wood. [Color figure can be viewed at wileyonlinelibrary.com]*

after swabbing yielded an average of 10 times more nDNA and nine times more mtDNA (Figure S4) as compared to the initial wet-swabbing. These results demonstrated that considerable DNA remained in or on these substrates after wet-swab collection.

Collection Efficiency

Results from evaluating the different collection techniques on glass with the same DNA isolation protocol showed a modest increase in DNA yields for the wet-vacuum samples, 189 ± 33 ng (mean ± SD), compared to the wet-swab samples, 149 ± 33 ng; however, the difference was not significant (Figure S5). Thus, the overall increase in DNA yields from wet-vacuuming compared to wet-swabbing on the other substrates tested may be attributed to the collection technique - to include recovery from the membrane or swab - and that the differences between the two extraction methods were negligible.

Discussion

The efficiency of DNA collection methods is largely dependent on the physical characteristics of the substrate being sampled. The size, absorbency, irregular shape, and coarse nature of a particular substrate can present challenges to traditional DNA collection methods, such as swabbing of the surface or direct cuttings of small areas. The data presented in this study have shown that a wet-vacuum-based collection method not only has the ability to successfully recover biological material from a variety of challenging porous substrate types, but also often exhibits improved performance over a conventional swabbing method. Furthermore, the wet-vacuum technique not only

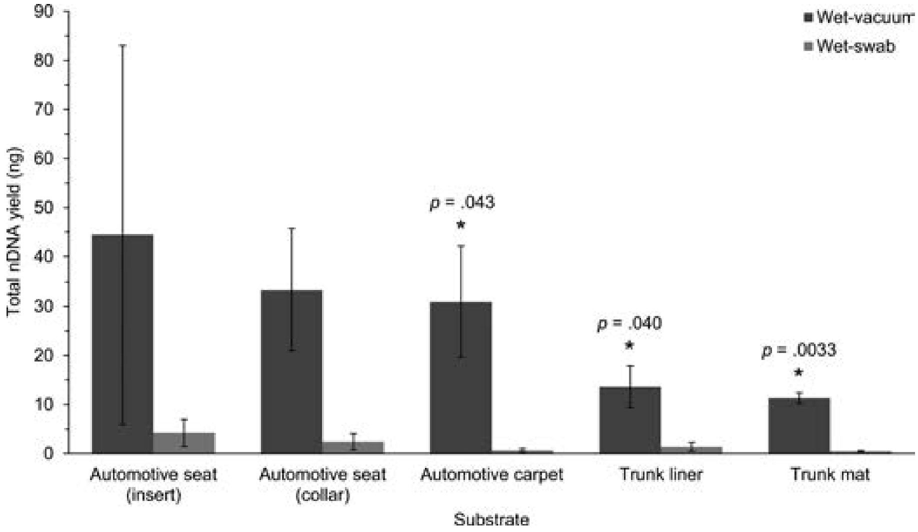


FIG. 3—*Average total nDNA yields recovered with the wet-vacuum or wet-swab methods for 1/100 bloodstains applied onto automotive items. Error bars represent the standard deviation of three replicates. Asterisks indicate significantly greater mean yields; the* p-*values are reported above. [Color figure can be viewed at wileyonlinelibrary.com]*

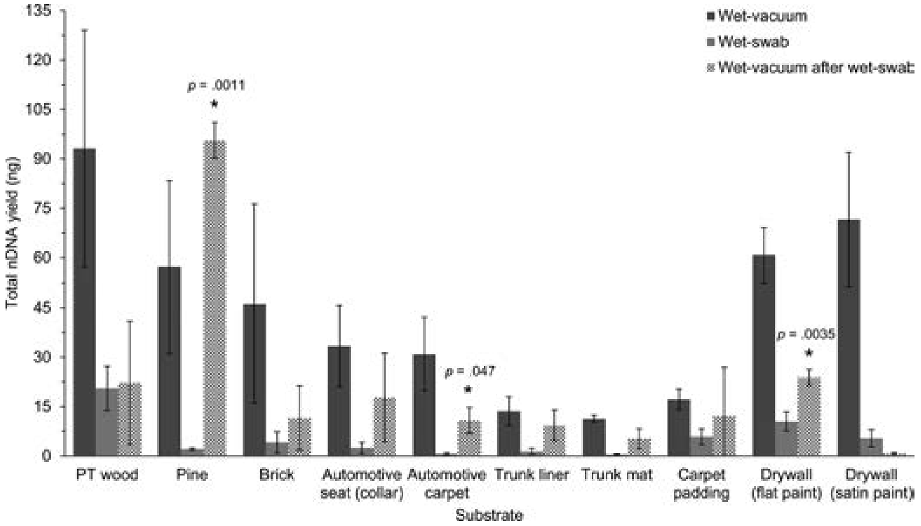


FIG. 4—*Comparison of total nDNA yields recovered from the wet-vacuum, wet-swab, or wet-vacuum after wet-swab collection methods. Error bars represent the standard deviation of three replicates. Asterisks indicate significantly greater mean yields from the wet-swab versus wet-vacuum after wet-swab methods; the* p-*values are reported above. [Color figure can be viewed at wileyonlinelibrary.com]*

recovered additional DNA from previously swabbed substrates, but it also frequently recovered more DNA than the initial swabbing.

There were two instances during this study where efficacies of both collection methods were greatly influenced by the physical characteristics of substrates. Unpainted drywall, an absorbent material, negatively affected DNA yields of the wet-vacuum more so than the wet-swab, possibly because it required substantially more liquid for collection. Also, the abrasive surface of cinderblock—the most rough and porous material tested—caused the cotton head of the swab to fray or break off. When wet-vacuumed, the solution passed through the pores/holes of the cinderblock quicker than the vacuum's capability to suction it back up. Therefore, and not surprisingly, substrate characteristics should be considered when deciding upon an appropriate collection technique.

For other difficult or porous substrates, the use of a wet-vacuum technique may be advantageous compared to a traditional wet-swab collection method. First, the amount of biological material that is collected is potentially increased due to the force and agitation of collection buffer which is applied to the substrate when using a wet-vacuum system. The wet-vacuum technique could allow for better retrieval of DNA hidden within fibers or crevasses of absorbent and/or porous items that a swab would otherwise leave behind. In addition, while cuttings can address cells trapped within porous materials, they are limited to smaller areas with visible staining, whereas the wet-vacuum method is not constrained as such. The latter may enable large substrates to be more efficiently processed compared to swabbing or cutting, particularly when biological material is diffuse or not isolated to a specific area of the substrate.

From a practical standpoint, swabbing is more convenient, simple, time-efficient, and inexpensive relative to the wet-vacuum method, and can be very effective in collection from visible stains or areas of repeated handling/contact. Swabbing and other traditional collection techniques should remain the preferred collection method on surfaces with visible stains since DNA quantities may be relatively abundant, whereas the wet-vacuum collection and processing method would require greater effort and cost. The estimated start-up cost for the wet-vacuum system ranges from US$43,000–45,000, while the cost per sample is about US $90 (M-Vac® Systems, Inc., personal communication, Apr. 25, 2019) compared to less than US$15 for the wet-swab method. Additionally, even when a stain is not visible, these results suggest that swabbing may yet be most appropriate and efficient for flat, smooth, or nonporous surfaces. For example, both wet-swab and wet-vacuum techniques had similar collection efficiencies when collecting blood on glass and wood countertop substrates.

The increased level of collection efficiency over larger areas offered by the wet-vacuum collection approach may allow diffuse or diluted staining to be more efficiently collected, which also makes case scenario consideration of paramount importance. While wet-vacuuming methods have the advantage of expanding the area of sampling, it can also potentially recover more DNA, which may unnecessarily increase the complexity of the DNA mixtures obtained and negate the probative nature of a more targeted sampling. As always, care must be taken when selecting items and areas for sampling when used in a forensic context.

This study demonstrated that a wet-vacuum-based collection system is capable of collecting diluted blood on multiple types of challenging substrates, often with increased collection efficiency over traditional swabbing techniques. While blood was chosen for this project specifically for convenience, homogeneity, and reproducibility, the results described here likely serve as a proxy for other cell types or body fluids. Although this study focused on substrates that may absorb biological material, the wet-vacuum method, owing to its high efficiency of collection, might also be suitable for collection of touch DNA.

While this study focused on efficacy of the wet-vacuum collection method on multiple substrates, there is room for further studies to improve this method and expand its forensic applications. As observed by Vickar *et al*., cell-free fragments of DNA can be lost during the wet-vacuum filtration step (12). Indeed, the 0.45 μM PES membrane has very low DNA-binding properties and filters solely by particle size. Therefore, while whole cells are captured on the membrane, smaller cell fragments and/or cell-free DNA may easily pass through, suggesting the need to optimize their retention. The wet-vacuum method might also be successful for recovering water-insoluble explosive residue as well as trace evidence, for example, hair, fibers, soil, polymer particles, etc., with appropriate extraction procedure modifications. Additional research and modifications in these areas are warranted in order to optimize the wet-vacuum method for forensic use.

Acknowledgments

The authors would like to acknowledge Thomas R. Myers, Eugene M. Peters, and Libby A. Stern for their helpful discussions and support of this research as well as Susannah C. Kehl and Leslie D. McCurdy for their critical review of this manuscript.

References

1. Comte J, Baechler S, Gervaix J, Lock E, Milon MP, Delemont O, et al. Touch DNA collection – performance of four different swabs. Forensic Sci Int Genet 2019;43:102113. https://doi.org/10.1016/j.fsigen.2019.06.014.
2. Verdon TJ, Mitchell RJ, van Oorschot RA. Swabs as DNA collection devices for sampling different biological materials from different substrates. J Forensic Sci 2014;59(4):1080–9. https://doi.org/10.1111/1556-4029.12427.
3. Adamowicz MS, Stasulli DM, Sobestanovich EM, Bille TW. Evaluation of methods to improve the extraction and recovery of DNA from cotton swabs for forensic analysis. PLoS One 2014;9(12):e116351. https://doi.org/10.1371/journal.pone.0116351.
4. Mulligan CM, Kaufman SR, Quarino L. The utility of polyester and cotton as swabbing substrates for the removal of cellular material from surfaces. J Forensic Sci 2011;56(2):485–90. https://doi.org/10.1111/j.1556-4029.2010.01659.x.
5. Petricevic SF, Bright JA, Cockerton SL. DNA profiling of trace DNA recovered from bedding. Forensic Sci Int 2006;159(1):21–6. https://doi.org/10.1016/j.forsciint.2005.06.004.
6. Dong H, Wang J, Zhang T, Ge JY, Dong YQ, Sun QF, et al. Comparison of preprocessing methods and storage times for touch DNA samples. Croat Med J 2017;58(1):4–13. https://doi.org/10.3325/cmj.2017.58.4.
7. Barash M, Reshef A, Brauner P. The use of adhesive tape for recovery of DNA from crime scene items. J Forensic Sci 2010;55(4):1058–64. https://doi.org/10.1111/j.1556-4029.2010.01416.x.
8. Verdon TJ, Mitchell RJ, van Oorschot RA. Evaluation of tapelifting as a collection method for touch DNA. Forensic Sci Int Genet 2014;8(1):179–86. https://doi.org/10.1016/j.fsigen.2013.09.005.
9. Hess S, Haas C. Recovery of trace DNA on clothing: a comparison of mini-tape lifting and three other forensic evidence collection techniques. J Forensic Sci 2017;62(1):187–91. https://doi.org/10.1111/1556-4029.13246.
10. M-Vac® Systems, Inc.How does the M-Vac work? https://www.m-vac.com/why-mvac/how-it-works (accessed November 7, 2019).
11. Garrett AD, Patlak DJ, Gunn LE, Brodeur AN, Grgicak CM. Exploring the potential of a wet-vacuum collection system for DNA recovery. J Forensic Identif 2014;64(5):429–48.
12. Vickar T, Bache K, Daniel B, Frascione N. The use of the M-Vac® wet-vacuum system as a method for DNA recovery. Sci Justice 2018;58(4):282–6. https://doi.org/10.1016/j.scijus.2018.01.003.
13. Hedman J, Agren J, Ansell R. Crime scene DNA sampling by wet-vacuum applying M-Vac. Forensic Sci Int Genet Suppl Ser 2015;5:e89–e90. https://doi.org/10.1016/j.fsigss.2015.09.036.
14. Williams S, Panacek E, Green W, Kanthaswamy S, Hopkins C, Calloway C. Recovery of salivary DNA from the skin after showering. Forensic Sci Med Pathol 2015;11(1):29–34. https://doi.org/10.1007/s12024-014-9635-7.
15. Gunn LE. Validation of the M-Vac® cell collection system for forensic purposes [master's thesis]. Boston, MA: Boston University School of Medicine, 2013.
16. Caswell KL. A comparison of M-Vac and surface swabs for collecting saliva and touch DNA from skin [master's thesis]. Davis, CA: University of California Davis, 2014.
17. Wander MJ. An investigation of touch DNA collection methods from clothing: traditional cutting techniques versus a wet vacuum system [master's thesis]. Davis, CA: University of California Davis, 2014.
18. Federal Bureau of Investigation Laboratory Quality System Documents. DNA standard operating procedures – procedures for the semi-automated extraction of DNA. https://fbilabqsd.com (accessed February 28, 2018).
19. M-Vac® Systems, Inc. SEC Series 100 and 150 user guide. Sandy, UT: M-Vac® Systems Inc, 2018.
20. Kavlick MF. Development of a triplex mtDNA qPCR assay to assess quantification, degradation, inhibition, and amplification target copy numbers. Mitochondrion 2019;46:41–50. https://doi.org/10.1016/j.mito.2018.09.007.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Average total mtDNA yields recovered with the wet-vacuum or wet-swab methods for 1/100 bloodstains applied onto household items.

Figure S2. Average total mtDNA yields recovered with the wet-vacuum or wet-swab methods for 1/100 bloodstains applied onto construction materials.

Figure S3. Average total mtDNA yields recovered with the wet-vacuum or wet-swab methods for 1/100 bloodstains applied onto automotive items.

Figure S4. Comparison of total mtDNA yields recovered from the wet-vacuum, wet-swab, or wet-vacuum after wet-swab collection methods.

Figure S5. Comparison of wet-vacuum and wet-swab collection efficiencies.

Table S1. Average total nDNA and mtDNA yields recovered with the wet-vacuum or wet-swab methods for 1/100 bloodstains applied onto household items.

Table S2. Average total nDNA and mtDNA yields recovered with the wet-vacuum or wet-swab methods for 1/100 bloodstains applied onto construction materials.

Table S3. Average total nDNA and mtDNA yields recovered with the wet-vacuum or wet-swab methods for 1/100 bloodstains applied onto automotive items.

Table S4. Average total nDNA and mtDNA yields recovered with the wet-vacuum after wet-swab method.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14534
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Yacine Boumrah,[1] *Ph.D.; Salem Baroudi,*[1] *M.Sc.; Mohamed Kecir,*[1] *M.Sc.; and Sabrina Bouanani,*[1] *M.Sc.*

Characterization of Algerian-Seized Hashish Over Eight Years (2011–2018). Part I: Physical Categorization

ABSTRACT: Data on the physical characteristics of North African hashish are scarce. This article exploits hashish seizure data collected over an 8-year period (2011–2018) in Algeria in order to establish a physical profile of North African hashish. The collected data were subjected to statistical analysis in order to characterize the three main forms in which North African hashish is packaged, namely hashish bags, hashish packets, and hashish units (slab bars). The study revealed that 82% of hashish bags weigh 25 kg that hashish bags are made either as a handbag or as a back bag that they are most often wrapped with woven plastic. Two hashish bag configurations were identified—Bag-Packet-Unit (79%) and Bag-Unit (21%)—and 81% of the total studied bags featured logos. Hashish packets contain the units, which are wrapped with three to five different types of packaging to constitute packets of 0.5 kg (65%) or 1 kg (32%), with two different configurations including 100- or 250-g units. Hashish packets are mainly covered with adhesive tape, and only 18% of them feature inscriptions. Hashish units are found in three different shapes: slab bar (most common), soap bar, and egg-shaped bar. Sixty-five percent of the North African hashish slab bars have a weight of 100 g; other weights are 250 g (20%), 90 g (10%), and 200 g (2%). Most of the 90–100 g units have a light brown color, and 200–250 g units are dark brown in color. Sixty-four percent of hashish units contain logos. Five logo classes were identified: letters (37%), numbers (27%), symbols (23%), animals (11%), and, more recently, "paper logos" at just 2%, exclusively reserved for high-quality hashish and entirely intended for the European market. The findings of this work allow for the establishment of a profiling platform of hashish seizures in this region and can be generalized to all countries that report this region as the primary source of seized hashish within their territory.

KEYWORDS: North African hashish, cannabis resin, physical characteristics, seizure, criminalistics, forensic science

Cannabis continues to be the most widely illicitly produced drug worldwide, in terms of both size and geographical spread of the areas of cultivation and the volume actually produced (1,2). While the cannabis herb is grown in almost every country in the world, the production of cannabis resin, also known as hashish, is confined to a small number of countries in North Africa, the Middle East, and South-West Asia (3).

Drug flow is characterized by three main points along a continuum: countries of production, countries of trans-shipment, and countries of targeted consumption (4). Europe and a large share of countries in Africa report Morocco as the main source of hashish (1,5,6).

North African hashish is prepared by collecting and isolating the resin glands of the female cannabis plant, producing a fine sticky powder. This powder is pressed either by hand for small amounts or by mechanical pressing devices for commercial quantities to facilitate transporting and smoking (7,8).

The specific geographical situation of Algeria results in tremendous quantities of hashish seizures, sometimes reaching more than ten tons in just one find. Algeria does not represent an exception to the worldwide situation. Cannabis is by far the most seized drug, hashish being exclusively the marketed type, with strong and growing evidence of an increase in domestic consumption within the country (9).

Drug use became recognized as a widespread scourge in the 1998 United Nations General Assembly Special Session (UNGASS) on the world drug problem (10). Drug use reduction is therefore a common responsibility and requires credible data on the patterns and trends of its consumption and trafficking.

The United Nations Office on Drugs and Crime (UNODC) revealed a lack of information and reporting on issues concerning cannabis cultivation, production, use, and seizures in many countries, especially in the regions of Africa and Asia (1). This fact is deemed problematic as obtaining a comprehensive picture of a global phenomenon requires the commitment of all the affected entities.

As hashish is an internationally controlled illicit drug scheduled by the 1961 Single Convention on Narcotics Drugs (11), data about its trafficking and consumption originate either from law enforcement or from health authorities (10).

In the present research, the data were extracted from a study of hashish case seizures conducted by law enforcement across the Algerian soil during an 8-year period (2011–2018). Both physical and chemical profiles were assessed as every hashish seizure is submitted to the National Institute of Criminalistics and Criminology of the Algerian Gendarmerie (NICC/NG). A technical and a police report summarizing the circumstantial information and the physical characteristics along with a representative sample of the whole seizure for chemical analysis are sent to the forensic laboratory. The goal of this research work is

[1] Département de Toxicologie, Institut National de Criminalistique et de Criminologie (INCC-GN), Boite Postale 194, Bouchaoui, Algiers, 16000, Algeria.

Corresponding author: Yacine Boumrah, Ph.D. E-mail: y.boumrah1978@hotmail.com

Received 22 April 2020; and in revised form 7 June 2020, 10 July 2020; accepted 13 July 2020.

to obtain the physical and chemical identity of the Algerian-seized hashish.

Due to the large amount of data, and to organize this research in a structured way, a series of two articles was needed to cover all the found results: Part I of the research will be devoted to the physical characterization of the Algerian-seized hashish, and Part II will concern the chemical constitution of the seized hashish.

In the present article, the most pertinent physical characteristics of the Algerian hashish seizures, including shapes, weights, packaging, colors, dimensions, configurations and logos, will be extracted. The data will then be analyzed to establish the physical profile of the Algerian-seized hashish.

As we cannot fight against a threat that remains unknown and ambiguous, the present research is intended to be a contribution to the global efforts of reducing drug trafficking and consumption by providing credible data about hashish seizures in Algeria. It is the first report that summarizes almost all the physical characteristics of North African hashish, allowing for the establishment of a profiling platform of hashish seizures in this region.

Methods

Methodology

Hashish seizures constitute four different types: bags seizures, packet seizures, unit seizures, and hashish piece seizures (Fig. 1). It should be noted that hashish pieces are made from hashish units; a group of units constitutes a packet, and a number of packets are assembled to compose a hashish bag.

At each hashish seizure, specialized agents examine the seizure, take pictures, and conduct a proper sampling of the bulk (seized material), taking into account its heterogeneity to get a representative analytical sample and ensure its preservation during and after sampling but also to elaborate a technical report containing all the physical information about the seizure. Samples, the technical report, and the police report that contains circumstantial information are then sent to a police forensic laboratory for exploitation by drug analysis experts.

For this research, the physical data were subjected to statistical analysis to characterize the three main entities in which the North African hashish is found, namely hashish bags, hashish packets, and hashish units (slab bars). Shapes, weights, packaging, colors, logos, and configurations of hashish bags were studied and characterized. Hashish packets were characterized from the standpoint of packaging, logos, and configurations, and finally, hashish units, which are the main entity, were fully characterized in terms of shapes, weights, dimensions, colors, and logos.

The data sets have an unequal number of observations; notched box plots have been used to visualize the distribution of units' dimensions and densities by unit weights, as explained by Chambers et al. (12) and Benjamini (13) and inspired by the research of Dujourdy and Besacier (3).

In order to apprehend the different molds used in the production of hashish slab bars, hierarchal ascending classification (HAC) was applied to the hashish units' dimensions, considering three variables (length, width, and thickness of the units by weight). HAC is an agglomerative cluster analysis method. The analysis begins with as many groups as there are individuals. The groups are formed from these initial units in ascending order until the end of the process, when all focused cases are grouped into the same conglomerate (14).

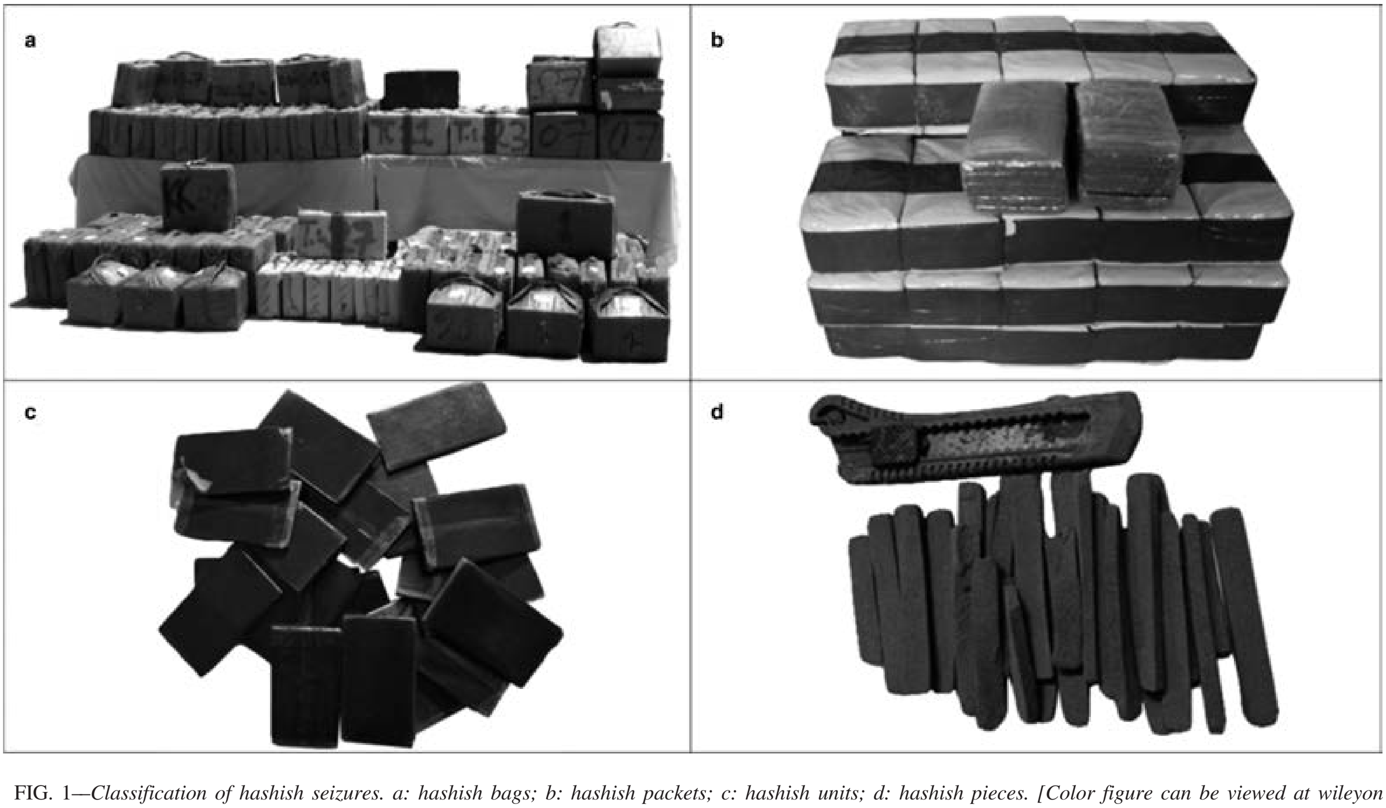


FIG. 1—*Classification of hashish seizures. a: hashish bags; b: hashish packets; c: hashish units; d: hashish pieces. [Color figure can be viewed at wileyonlinelibrary.com]*

Presentation of the Study Population

The hashish seizures studied for this research were conducted by law enforcement units across Algeria at the urban and suburban levels over 8 years (2011–2018). A large amount of data has been acquired from the seizures and is characterized by a large variability, both in terms of the quantities and types of seizures. The study concerned 2232 hashish seizures comprising 3265 batches, with weights that vary from a few grams to tens of tons, and contains different types of seizures (Table 1).

Results and Discussion

Characterization of the Hashish Bag

The hashish bag is the largest entity that can be found in a hashish seizure and is linked to commercial importation and distribution on a large scale of significant quantities, since a large portion of the bag seizures (45%) range from 100 to 1000 kg.

The external morphological characteristics of a hashish bag are shape, type, weight, color, packaging, and inscription (Fig. 2*a*).

Hashish Bag Shapes and Weights

Several bag shapes are encountered with slight, subtle differences in geometrical patterns; some are more common than others. Nevertheless, three main geometrical shapes are distinguishable, as indicated in Fig. 3: cubic bags (Fig. 3*c*), cuboid bags (Fig. 3*a*, *b*,*d*,*e*), and pyramidal bags (Fig. 3*f*). Shapes a, b, and c are the most popular ones, while shape f is very rare.

For ergonomic purposes, and to make the transportation of hashish bags easy and convenient, the bags are made in the shape of a handbag or, much more frequently, in the shape of a back bag, as described in Fig. 3. Hashish bags appear in three main weights: The most common is 25 kg, the weight of more than 82% of the total studied hashish bags; other weights include 30-kg bags (11%) and 20-kg bags (7%).

Hashish Bag Packaging and Colors

To wrap hashish bags, four different types of packaging were used: woven plastic, burlap, adhesive tape, and fabric. Over the 2011–2018 period, woven plastic was the most used packaging material for hashish bags (66%). This fact is probably due to the compromise it offers between an acceptable protection against environmental hazards (moisture and UV lights), a good resistance against tearing and low cost. Fabric was the least used (1%), mainly because of its uneconomical character, while burlap and adhesive tape were almost equally used, at 16% and 17% of the total hashish bag packaging, respectively (Fig. 4).

Both woven plastic and burlap packaging materials have four color classes, while adhesive tape counts only two-color classes. Regarding fabric packaging, no color classification was considered due to the limited number of hashish seizures recorded with this type of packaging. Details on the hashish bag color classes

TABLE 1—*Studied hashish seizures.*

Year	2011	2012	2013	2014	2015	2016	2017	2018	Total
Number of seizures	77	86	98	152	142	530	621	526	2232
Number of batches	128	132	165	259	189	783	806	806	3265
% of seizures' types									
Bags	27	31	36	47	44	33	24	22	33%
Packets	44	36	36	30	38	34	32	30	35%
Units	28	34	26	20	18	25	31	34	27%
Pieces	3	2	2	3	1	7	10	12	5%
% of seizures' weight classes									
<1 kg	25	25	20	29	29	33	34	37	29%
1–10 kg	10	11	11	12	14	18	17	19	14%
10–100 kg	30	31	32	35	34	39	37	42	35%
100–1000 kg	21	21	25	19	19	16	16	15	19%
>1000 kg	3	3	5	4	4	2	2	1	3%

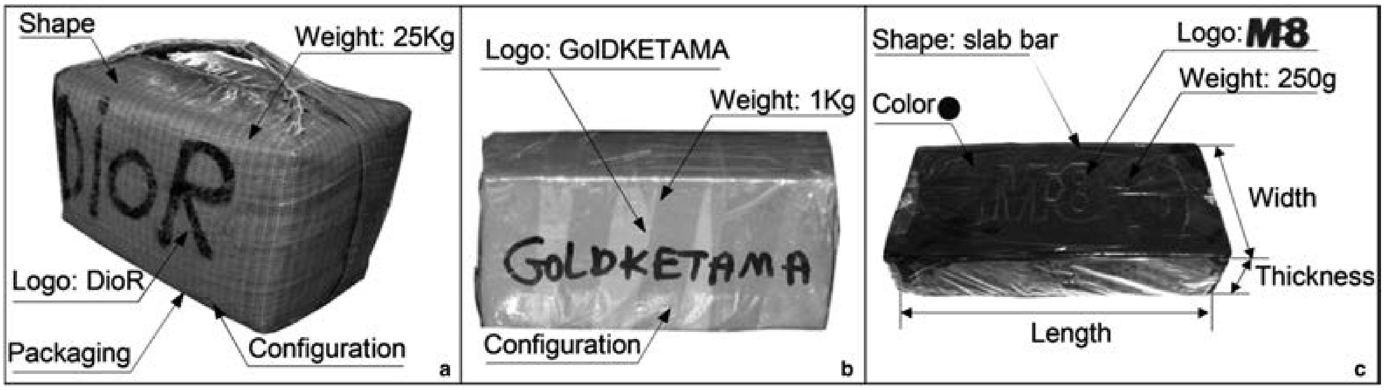


FIG. 2—*Studied physical characteristics of the different hashish entities. a: hashish bag; b: hashish packet; c: hashish unit. [Color figure can be viewed at wileyonlinelibrary.com]*

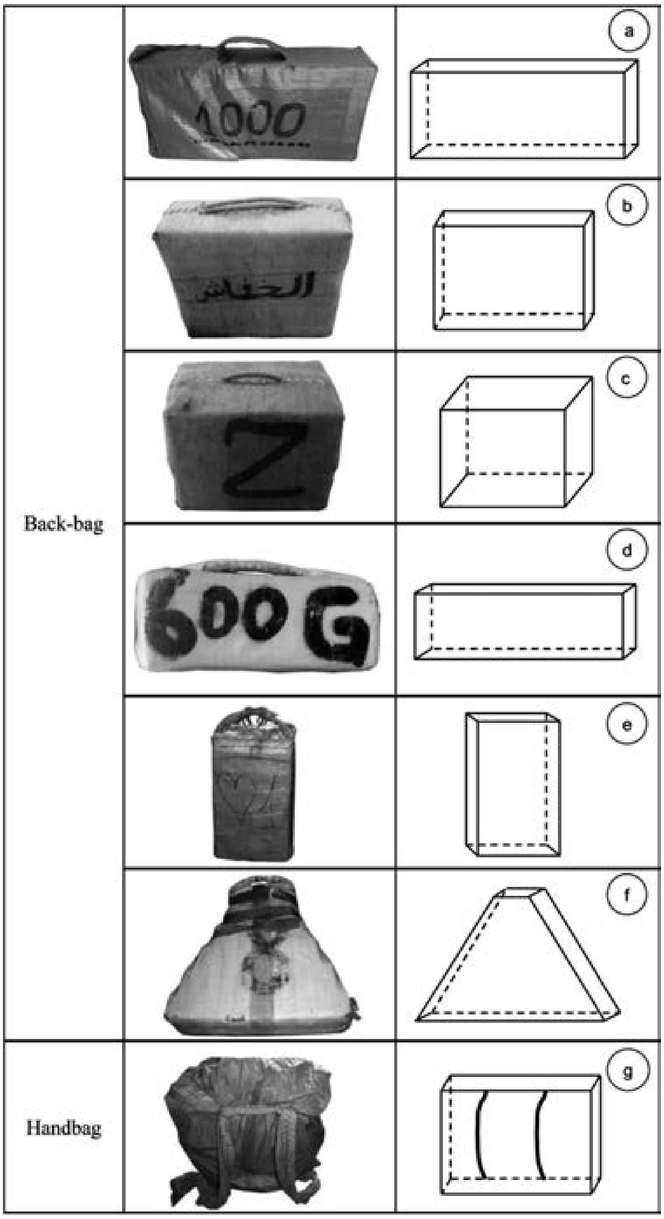


FIG. 3—*Shapes and types of hashish bags. [Color figure can be viewed at wileyonlinelibrary.com]*

and their distribution are provided as Supplemental Information in Figure S1.

Hashish Bag Inscriptions (Logos)

Inscriptions are one of the key elements of hashish bags' external characteristics. Hashish bags contain an inscription 81% of the time, which cannot be fortuitous. From our point of view, these inscriptions are likely a kind of language or messages between hashish producers and traffickers.

Three main classes of hashish bag logos were identified: the most important are those relating to letters—"letters class"—and numbers—"numbers class." The last identified class was designated as "others" and includes all the inscriptions that do not fall within either of the previously cited classes. "Number logos" are by far the most encountered hashish bag inscriptions (67%), followed by "letter logos," which are found on 32% of the inscribed bags; Arabic letters were the most prevalent in this class, with 35% of "letter logos" consisting of Arabic writing. Some examples of bag inscriptions and the detailed distribution of the different identified inscriptions are provided as Figure S2.

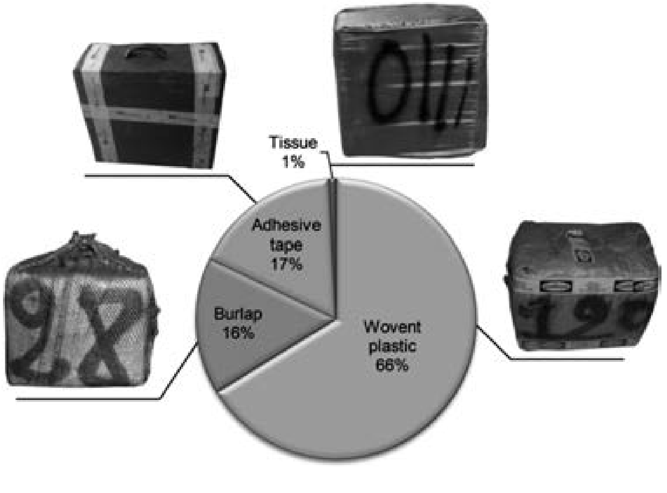


FIG. 4—*Distribution of hashish bags' packaging. [Color figure can be viewed at wileyonlinelibrary.com]*

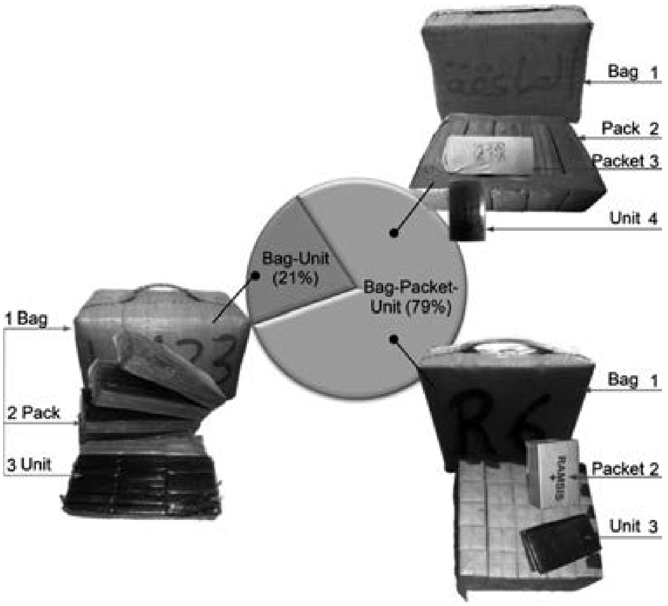


FIG. 5—*Main configurations of hashish bags. [Color figure can be viewed at wileyonlinelibrary.com]*

Hashish Bag Configurations

A bag of hashish has a structured configuration; its components are ordered from the largest entity, the bag, to the smallest, the unit, with intermediate elements, the packs, and the packets, in between. A precise number of units are grouped in the packets; the packets are then gathered in a definite number to form the packs. Each pack contains the same number of packets, and each packet contains the same number of units. Finally, the packs are arranged to form a compact block, which is exclusively wrapped with adhesive tape at first and then wrapped with one of the previously cited types of packaging. This classical

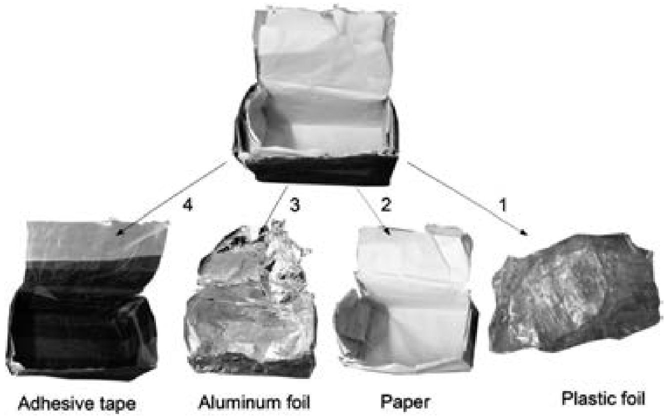


FIG. 6—*Stratified packaging of hashish packets. [Color figure can be viewed at wileyonlinelibrary.com]*

hierarchy (Bag-Packs-Packets-Units) is not always respected; in some cases, the packs are missing, while in other cases, the packets are absent. Because the packs have no particular characteristics, two hashish bag configurations have been retained: (i) Bag-Packet-Unit and (ii) Bag-Unit. The configuration that includes packets is the most wide spread, constituting almost 79% of the total studied bag seizures (Fig. 5).

Hashish bags' composition falls within one of the previously cited configurations, and yet there are a multitude of arrangements for the components of the bag. The configurations were identified according to the weight of the hashish bag, the weight of the hashish packet or hashish unit (in the case of a Bag-Unit configuration) found in the bag, and the distinctive arrangement of the packets or units of the same weight. Table S1, given as Supplemental Information, provides the detailed arrangements of the different hashish bags.

Characterization of the Hashish Packet

Hashish packets are mostly seized in large (10–100 kg) to middle amounts (0.5–10 kg). These hashish entities are mainly impounded from upper-level dealers and, to a lesser extent, from traffickers.

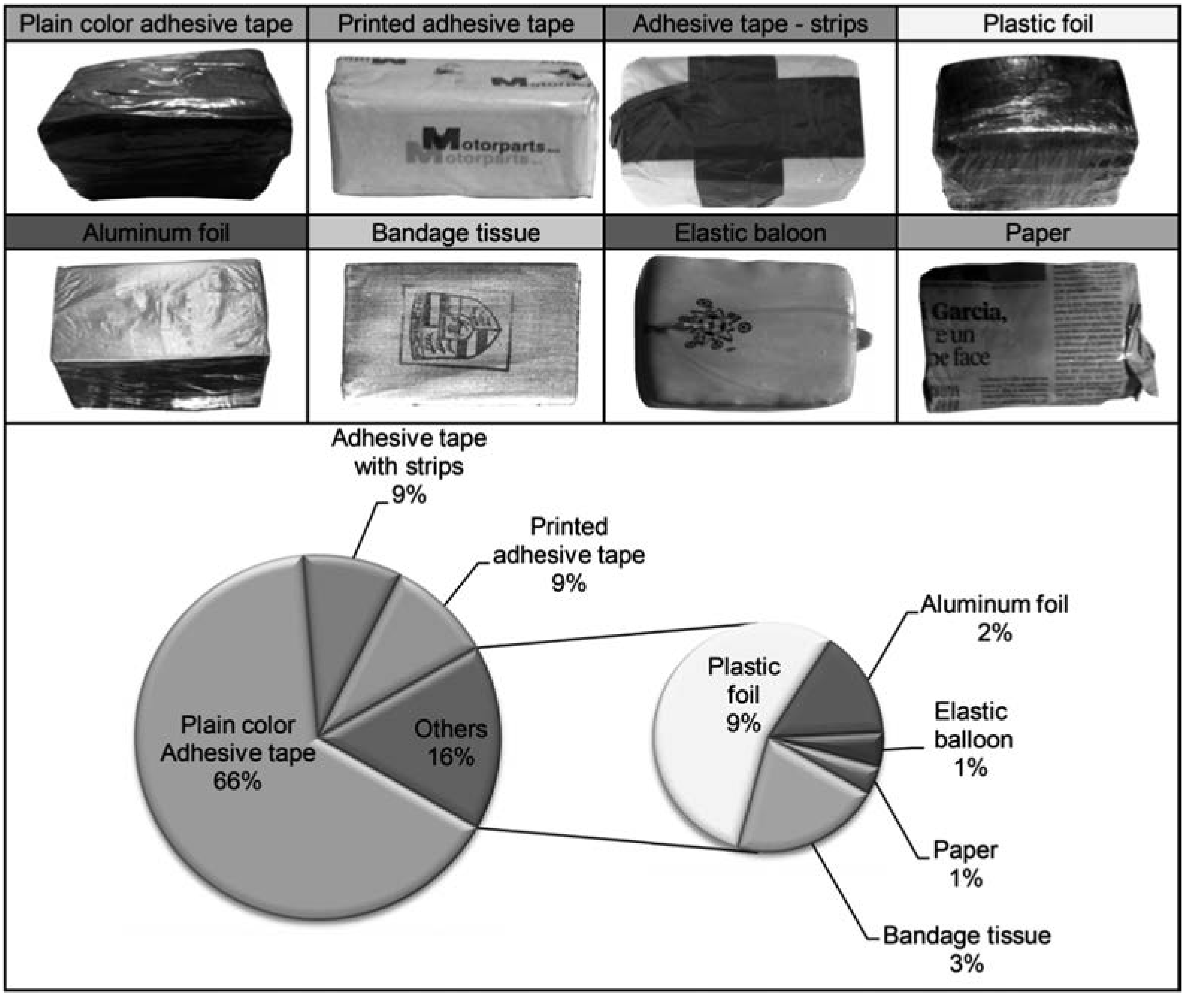


FIG. 7—*Distribution of hashish packets' packaging. [Color figure can be viewed at wileyonlinelibrary.com]*

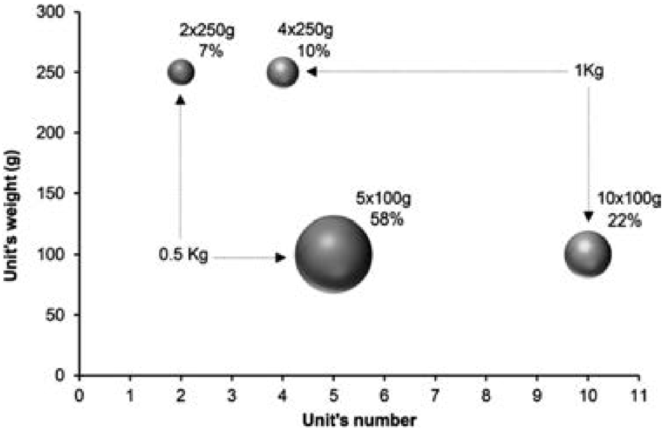


FIG. 8—*Main configurations of the majority (97%) of hashish packets. [Color figure can be viewed at wileyonlinelibrary.com]*

Hashish packets are the component that contains the units; the elements of their physical characterization are weight, inscription, configuration, and packaging (Fig. 2*b*).

Hashish Packet Packaging

A multilayer packaging provides protection against moisture and limits contact with the external environment, thus protecting the chemical makeup and maintaining the quality and potency of the hashish for the longest time possible. The most frequent wrappers of hashish packets are shown in Fig. 6. There is no particular rule in their sequencing except that some of the wrappers, like fabric and adhesive tape, generally appear first when unpacking the packet of hashish while others, like plastic or aluminum foil, are more likely to constitute the internal layers of the packaging. In the sections that follow, "packaging" will exclusively refer to the external wrap.

Hashish packets are mainly covered with adhesive; it was found in 84% of cases. For the rest, plastic foil, bandage tissue, paper, or elastic balloons were used as hashish packet packaging. The detailed distribution of hashish packet packaging is shown in Fig. 7.

Hashish Packet Inscriptions (Logos)

While the presence of an inscription on hashish packets is not usual, only 18% of the packets were inscribed, perhaps because, the packet being an intermediate component, the inscription is affixed either upstream or downstream of it. However, some practices in the marking of the packets were identified repeatedly, like the use of printed adhesive tape with specific mentions; thus, as minor as they are, not considering the packets' inscriptions would lead to a qualitative loss of information.

The studied seizures reveal that inscribed hashish packets are classified fall into two main classes, the "Affixed Inscriptions" class, which is characterized by a logo affixed by the hashish makers, and the "Printed in Adhesive tape" class, which encompasses those with particular mentions found on the printed adhesive tape used to wrap packets. Some examples of typical hashish packet inscriptions and the detailed distribution of each inscription class are provided in Figure S3.

Hashish Packet Configurations

The studied seizures allowed the identification of four main configurations of hashish packets, as summarized in Fig. 8, which shows 97% of the possible arrangements of the hashish packets. These packets are principally found in two weight classes, 0.5 and 1 kg, with shares of 65% and 32%, respectively. Each type has two main configurations that consist of units of 100 or 250 g.

The grouping of five 100-g units represents the majority of packet configurations (58%); the second most encountered configuration is the grouping of ten 100-g units to form 1-kg packets (22%). Ten percent of the hashish packets consist of four 250-g units, and 7% are two 250-g units.

Some other configurations have been observed, such as 0.4, 0.8, and 1.25-kg packets, but on very rare occasions. These configurations have just one composition that consists of units of no less than 200 g.

Hashish packets of 0.5 and 1 kg have other configurations encountered at a much lower frequency, such as the assembling of 50 10-g units in the case of 0.5-kg packets and the assembling of 5 200-g units or 100 10-g units in the case of 1-kg

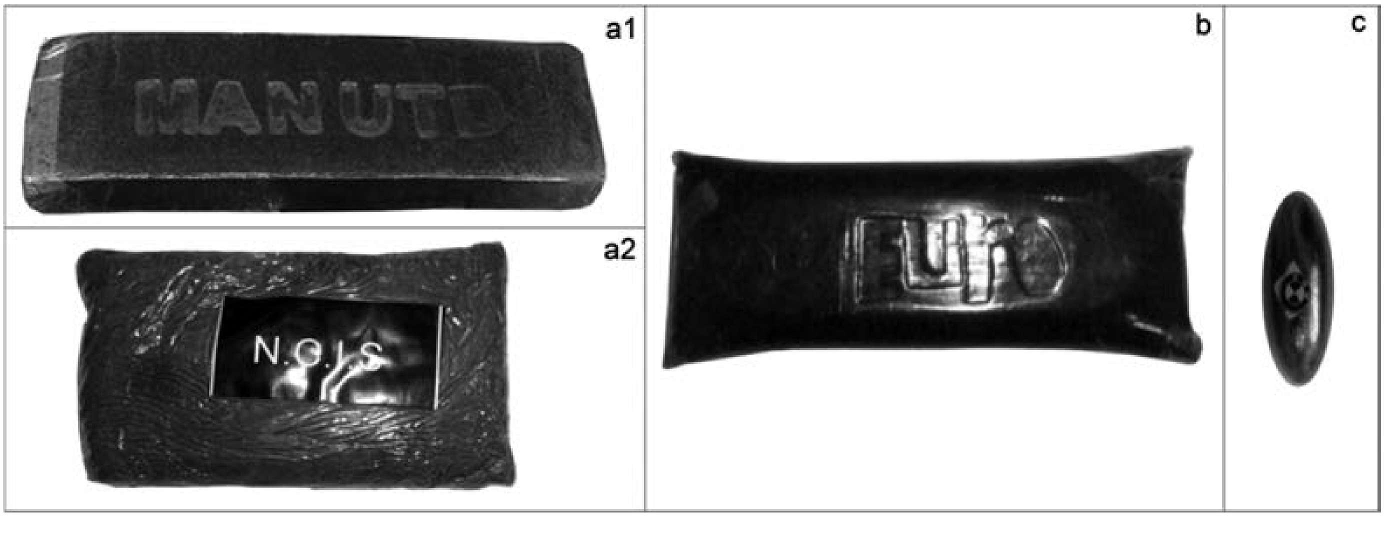


FIG. 9—*Shapes of hashish units. a1 and a2: Slab bars (Flat bars); b: Soap bar; c: Egg-shaped bar. [Color figure can be viewed at wileyonlinelibrary.com]*

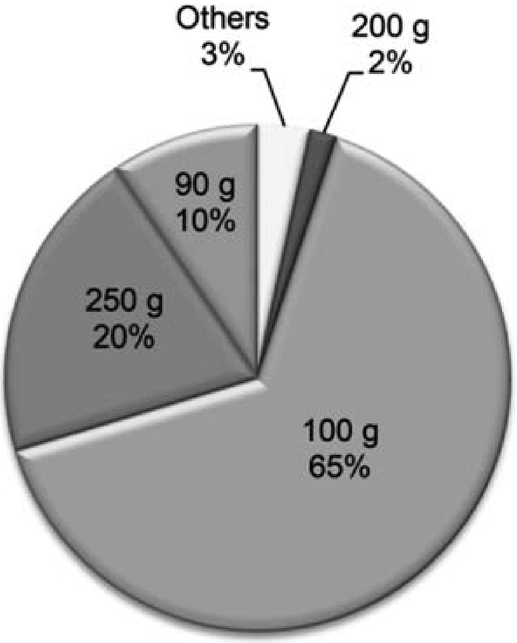


FIG. 10—*Distribution of hashish slab bars' weight. [Color figure can be viewed at wileyonlinelibrary.com]*

packets. Table S2, given as Supplemental Information, provides the detailed arrangements of the different hashish packets.

Characterization of the Hashish Unit

Hashish units' seizures are conducted at the urban and suburban levels and rarely exceed 10 kg. These seizures involve low-level dealers and thus correspond to hashish distribution on a small scale.

The unit is a structured agglomeration of cannabis resin powder; it is the basic product resulting from the process of hashish making. The elements included in a complete physical characterization of the hashish unit are shape, weight, dimension, density, color, and logo (Fig. 2*c*).

Hashish Unit Shapes

Three main hashish unit shapes were identified: slab bar, soap bar, and egg-shaped bar (Fig. 9). Hashish slab bars are the most common (98% of the total studied hashish units); they have regular straight edges that suggest the use of molds and mechanical devices in their manufacture and have an exclusively cuboid shape. Hashish soap bars and egg-shaped bars are much less common.

In some cases, slab bar units have an irregular shape with no defined edges, and this shape could approximately be related to a cuboid (Fig. 9*a2*). Recently, it was observed that an increasing number of unmolded hashish slab bars were seized, and the investigations revealed that these kinds of units are of high-potency hashish.

Hashish Unit Weights

Hashish units are made in specifically defined weights. The obtained data revealed that the seized hashish soap bars had the same 250-g weight and the egg-shaped bars the same 10-g weight. Concerning hashish slab bars, different weights were encountered: The most abundant was a 100-g weight, found in 65% of the total studied hashish slab bars; 250-g hashish slab bars are also common, at 20%, while 10% of slab bars were 90 g, and 200-g slab bars are also confectioned but in a lower percentage (2%). The least frequent weights, like 10 g, or the intermediate ones, like 80, 150, or 220 g, are grouped in one class designated as "Others" and representing a share of 3%, as shown in Fig. 10.

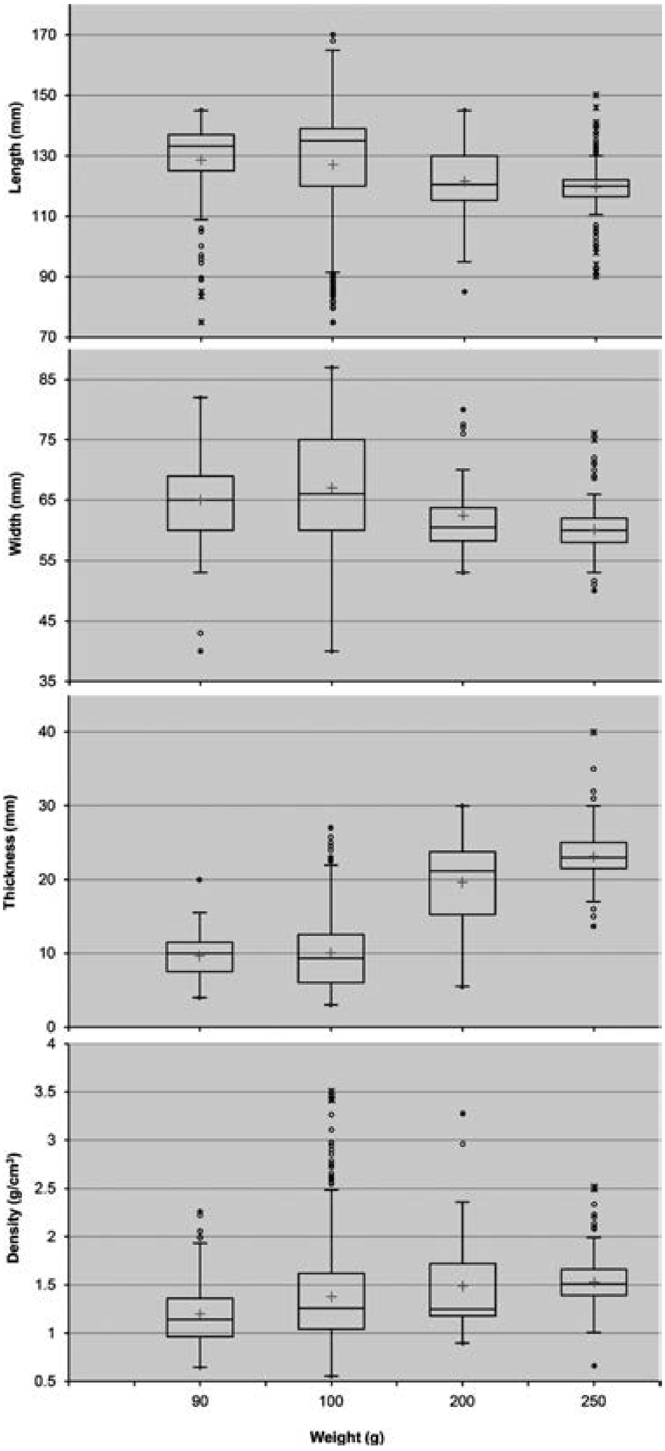


FIG. 11—*Notched box plots of the hashish units' dimensions and density by units' weight. [Color figure can be viewed at wileyonlinelibrary.com]*

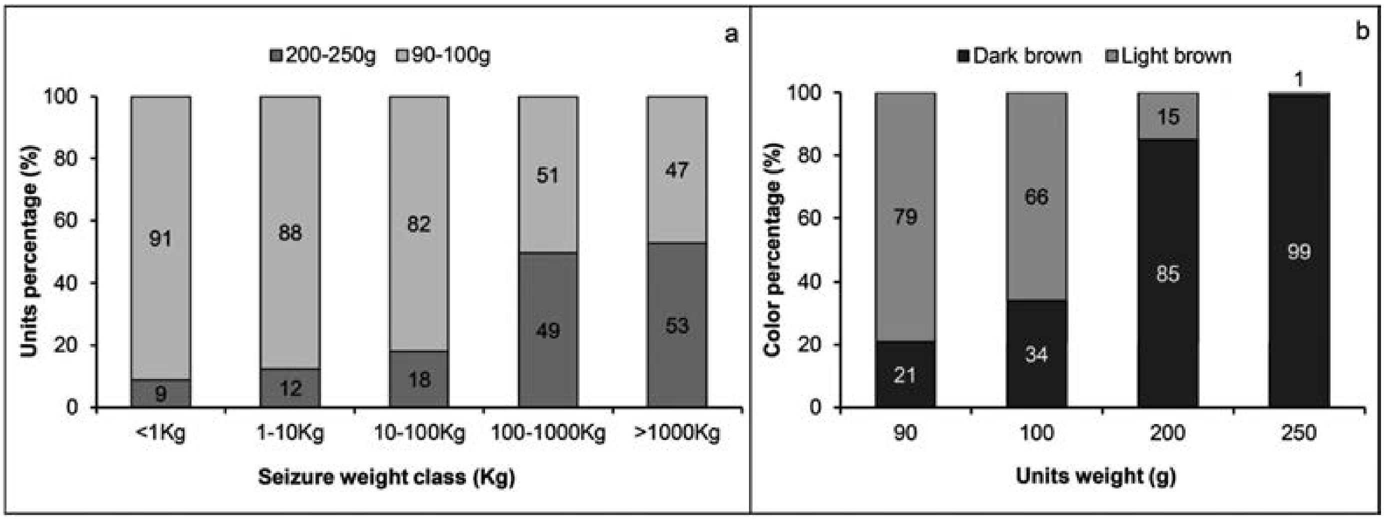


FIG. 12—*Hashish units' weight ratios regarding units' color & seizures' weight. a: units' weight versus seizures' weight classes; b: units' colors versus units' weight. [Color figure can be viewed at wileyonlinelibrary.com]*

TABLE 2—*Estimated molds of hashish units.*

	Mold (Length × Width × Thickness)			Variance Decomposition of the Optimal Classification*	
Weight (g)	Obtained	Percentage (%)	Retained	Intra-Class (%)	Inter-Class (%)
90	137 × 69 × 09	54	140 × 70 × 10	21.4	78.6
	129 × 62 × 11	22	130 × 60 × 10		
	116 × 60 × 12	16	115 × 60 × 10		
	093 × 57 × 13	8	95 × 55 × 15		
100	139 × 77 × 06	34	140 × 75 × 5	17.4	82.6
	126 × 62 × 11	26	125 × 60 × 10		
	138 × 67 × 09	21	140 × 65 × 10		
	096 × 57 × 16	19	95 × 55 × 15		
200	117 × 60 × 22	57	115 × 60 × 20	39.7	69.3
	134 × 67 × 16	37	135 × 65 × 15		
	090 × 57 × 18	07	90 × 55 × 20		
250	121 × 60 × 24	42	120 × 60 × 25	34.6	65.4
	116 × 58 × 22	33	115 × 60 × 25		
	135 × 61 × 25	11	135 × 60 × 25		
	118 × 64 × 24	11	120 × 65 × 25		
	100 × 65 × 25	05	100 × 65 × 25		

*Variance calculation is based on dissimilarity by measuring the Euclidian distance and using the Ward method for aggregation.

Dujourdy and Besacier (3) stated that most of the cannabis resin seized in France comes from Morocco, which is the same situation in Algeria, and reported the existence of the same 10, 100, 200, and 250-g hashish units.

Hashish Unit Dimensions

Hashish units are produced in specific dimensions (length, width, and thickness). Though they do not have the accuracy of an industrial process, units' dimensions have typical values.

Figure 11 shows the distribution of hashish unit dimensions using notched box plots. The dimensions of 90–100-g units and those of 200–250-g units are spread out in two distinct regions. Additionally, there is an overlap between the notches of 90 and 100-g units on one hand and those of 200 and 250-g units on the other hand. This clearly indicates dimensional similarities between units of a close weight, which implies an existing "weight-dimension" correlation.

The 90–100-g units are more likely to be found at a length between 130 and 140 mm, yet the length of those units is characterized by a large variability; some of them have unusually small length, and hence, the median, which is about 135 mm, is a good indicator of their length. On the other hand, the length of 200–250-g units varies less and tends toward 110 to 130 mm, the mean value being around 120 mm.

However, hashish units tend to have the same width, the most common values ranging between 60 and 70 mm, except 250-g units, whose most frequent widths range between 50 and 60 mm.

Thickness is the dimension that makes the difference between the two pairs of units' weight (90–100 g and 200–250 g). While 10 mm is the distinctive thickness of 90–100-g units, the vast majority of the values being in the interval of 5–15 mm, the 200–250-g units are more likely to have a thickness of around 20 mm, in the 20–25-mm interval. This constitutes a dimensional difference of about 50%.

Thus, hashish slab bars have the same width, slightly different lengths (90–100-g units being longer) and doubtlessly different thicknesses (200–250-g units being twice as thick as 90–100-g ones).

To evaluate the weight-to-volume relation, the distribution of hashish units' density by weight is plotted. Hashish units' density is directly proportional to the units' weights, since it steadily increases, as is described in Fig. 11. This translates to the intention of smuggling maximum hashish quantity in minimum volume when hashish units of superior weight are produced. This trend is highlighted by the fact that the share of 200–250-g units sharply increases in hashish seizures involving large quantities, as illustrated in Fig. 12*a*.

The Euclidian distance was used as a metric and the Ward method as an aggregation criterion; the obtained molds used for the making of hashish slab bars are listed in Table 2. The intra- and inter-class variances give an appreciation of the classification quality by respectively assessing the dissimilarity between the individuals within the same class and the dissimilarity

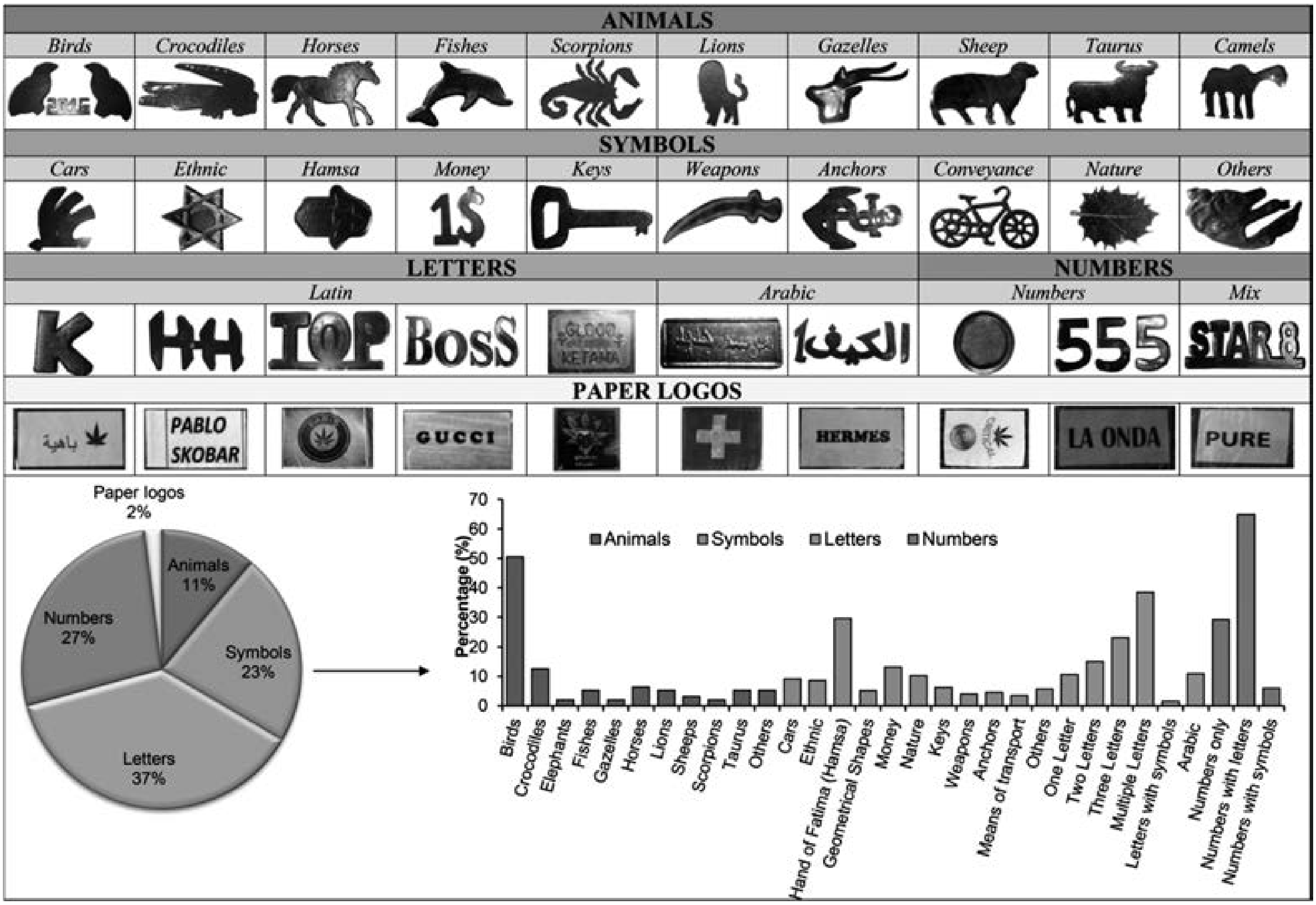


FIG. 13—*Classification and distribution of hashish units' logos. [Color figure can be viewed at wileyonlinelibrary.com]*

between the obtained classes. A relatively good classification would have an intra-class variance that tends toward 0% and an inter-class variance that tends toward 100%.

The HAC (with the chosen parameters) displayed at least three molds for each unit's weight, some being more frequent than others. This mold variability could be explained either by the existence of multiple hashish producers who use different molds or, more probably, as the dimensional differences are about a few millimeters, by manipulation inaccuracies. As hashish production does not unfold like an industrial process, the dimensions of the obtained molds have been rounded to the nearest possible mold used in the hashish manufacture process.

Hashish Unit Colors

Hashish is encountered as light to dark brown-colored units, depending mainly on the production circumstances. Dark hashish has been heated and pressed more than light-colored hashish. Heat and pressure lead to the rapid oxidation and degradation of the resinous contents and reduction of the THC levels, while in light-colored hashish, resin glands remain intact (7).

The tendency is clear: The 90–100-g units have a light brown color, and 200–250-g units are dark brown-colored, as shown in Fig. 12*b*. Exceptions to this trend are possibly due to old, improperly stored and transported 90–100-g units whose color turned from light to dark brown because of oxidation. These findings allow the deduction that the 90–100-g hashish units are processed differently than 200–250-g units. Indeed, the molding of higher amounts of hashish requires dense, stronger pressing and sometimes the use of heat and moisture to overcome the stubbornness of a voluminous hashish bulk.

Hashish Unit Logos

Logos are writings or drawings directly affixed on the center of the unit, serving as a seal; they were found on 64% of the total studied hashish units. Five main logo classes were identified after the visualization of all the studied units' logos, namely "letters," "numbers," "symbols," "animals," and "paper logos."

Figure 13 shows examples of the different main classes of hashish slab bars' logos and provides the percentage distribution of the different logos. The most encountered logos fall within the "letters" or "numbers" classes, comprising 37% and 27% of inscriptions, respectively, this is probably because of the ease of their seal conception. "Symbols" are the third most encountered logos (23%), dominated by symbols of the Hand of Fatima (Hamsa), followed by the "animals" class, which gathers all drawings of animals (11%), dominated by bird logos.

The least encountered class was that of "paper logos," with a share of only 2%. Such logos are increasingly used for high-potency hashish and entirely intended for the European market. The "paper logos" are only found on the unmolded hashish units, since these units are so sticky and resinous (high-quality hashish) that it has become impossible to affix them with a logo by means of a seal.

Logos are a physical characteristic of great importance; they are a common feature of the hashish entities. Furthermore, hashish producers sometimes affix the same inscription, or they affix different inscriptions that convey the same meaning in bags, packets and units. This inscription similarity can be found among the three main constituents of hashish (bag-packet-unit). In most cases, the inscription similarity is between the bags and units. All other inscription similarities involve hashish packets and are encountered to a much lower degree.

Inscription analogy allows the linkage of hashish seizures that have been conducted in different locations but have the same inscriptions. Figure S4 provides some examples of such similarities.

There is no ambiguity in spotting and linking hashish entities that have identical inscriptions. However, some inscriptions that at first look different in fact have the same meaning. For example, the inscription "555" refers to the "Hand of Fatima (Hamsa)," called "Khamsah, خمسة" in Arabic, a word that means "five" (15),"1D" refers to "$1," "17" to "2017" and "TT" to the car model "Audi TT."

Conclusion

The physical characteristics and features of North African hashish were investigated based on law enforcement data collected over 8 years (2011–2018) concerning Algerian-seized hashish.

North African hashish has particular psychical characteristics, which indicates that its production is not a random process but rather subjected to precise requirements intended to make it a more easily transported and marketed commodity.

This fact drives the diversity in physical features of the seized hashish in Algeria but, at the same time, allows the establishment of definite classes of common physical aspects, primarily related to weight and packaging.

Units, packets, bags, and pieces are hashish entities that may be encountered in a North African hashish seizure. Units of 90, 100, 200, and 250 g are molded in specific dimensions, affixed with a logo, gathered, and coated in a protective packaging, generally consisting of three to five protective layers, to form in most cases 0.5 or 1-kg hashish packets. The packets are arranged into a compact mass with a specific configuration to form a bag, which is then wrapped first and exclusively with adhesive tape and secondly with an additional packaging of a distinct color and nature, most commonly made of woven plastic.

Logos and inscriptions are pertinent physical characteristics as they allow a relevant linkage between geographically scattered seizures.

The present paper is the first report to have examined the physical attributes of North African hashish. The findings of this work can be generalized to all countries that report this region as the primary source of the seized hashish within their territory.

References

1. United Nations Office on Drugs and Crime. World drug report 2018. Vienna, Austria: United Nations Publications, 2018. Sales No. E.18.XI.9.
2. United Nations Office on Drugs and Crime. World drug report 2017. Vienna, Austria: United Nations Publications, 2018. Sales No. E.17.XI.8.
3. Dujourdy L, Besacier F. A study of cannabis potency in France over 25 years (1992–2016). Forensic Sci Int 2017;272:72–80. https://doi.org/10.1016/j.forsciint.2017.01.007
4. Singer M. Drugs and development: the global impact of drug use and trafficking on social and economic development. Int J Drug Policy 2008;19(6):467–78. https://doi.org/10.1016/j.drugpo.2006.12.007
5. European Monitoring Centre for Drugs and Drug Addiction. European drug report 2018: trends and developments. Luxembourg: Publications Office of the European Union, 2018. http://www.emcdda.europa.eu/publications/edr/trends-developments/2018 (accessed January 15, 2020).
6. Afsahi K, Darwich S. Hashish in Morocco and Lebanon: a comparative study. Int J Drug Policy 2016;31:190–8. https://doi.org/10.1016/j.drugpo.2016.02.024
7. Clarke RC. Hashish! Los Angeles, CA: Red Eye Press, 1998;61–87.
8. United Nations Office on Drugs and Crime. Recommended methods for the identification and analysis of cannabis and cannabis products. Vienna, Austria: United Nations Publications, 2009. Sales No. E.09.XI.15.
9. Office National de Lutte Contre la Drogue et la Toxicomanie. Activité de lutte contre la drogue et la toxicomanie (Bilan annuel 2017) [Drug and drug addiction activity (annual review 2017)]. 2018. https://onlcdt.mjustice.dz/onlcdt_fr/?p=donnees (accessed January 25, 2020).
10. Vandam L, Matias J, McKetin R, Meacham M, Griffiths P. Illicit drug trends globally. In: Quah SR, Cockerham WC, editors. International encyclopedia of public health, 2nd edn, vol. 4. Cambridge, MA: Academic Press, 2017;146–56. https://doi.org/10.1016/B978-0-12-803678-5.00223-X
11. United Nations Office on Drugs and Crime. The international drug control conventions. United Nations. 2013. https://www.unodc.org/unodc/en/commissions/CND/conventions.html (accessed January 20, 2020).
12. Chambers JM, Cleveland WS, Kleiner B, Tukey PA. Graphical methods for data analysis. Pacific Grove, CA, USA: Duxbury Press, 1983;60–3.
13. Benjamini Y. Opening the box of boxplot. Am Stat 1988;42:257–62.
14. Lis-Gutiérrez JP, Reyna-Niño HE, Gaitán-Angulo M, Viloria A, Santander Abril JE. Hierarchical ascending classification: an application to contraband apprehensions in Colombia (2015-2016). In: Tan Y, Shi Ying Y, Tang Q, editors. Data Mining and Big Data: Proceedings of the Third International Conference (DMBD 2018); 2018 June 17–22; Shanghai, China. Cham, Switzerland: Springer International Publishing, 2018;168–78.
15. Nozedar A. The element encyclopedia of secret signs and symbols: the ultimate A-Z guide from alchemy to the zodiac. London, U.K.: Harper Collins Publishers, 2010;132–3.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Distribution of hashish bags' packaging and colors.

Figure S2. Classification and distribution of hashish bags' inscriptions.

Figure S3. Classification and distribution of hashish packets' inscriptions.

Figure S4. Examples of similar inscriptions among the different hashish entities.

Table S1. Detailed configurations of the hashish bags (Pa: pack; P: packet; U: unit).

Table S2. Detailed configurations of the hashish packets (U: unit).

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14533
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Yacine Boumrah,[1] Ph.D.; Salem Baroudi,[1] M.Sc.; Mohamed Kecir,[1] M.Sc.; and Sabrina Bouanani,[1] M.Sc.

Characterization of Algerian-Seized Hashish Over Eight Years (2011–2018). Part II: Chemical Categorization

ABSTRACT: In Algeria, large quantities of hashish are seized every year. This study aimed to investigate the total content of major cannabinoids in the illicit seized hashish in Algeria over an 8-year period (2011–2018) in order to establish the chemical profile of North African hashish. A total of 3265 hashish samples were analyzed using a validated high-performance liquid chromatography–diode array detection (HPLC-DAD) method, allowing the simultaneous quantification of both the acidic and the neutral forms of Δ9-tetrahydrocannabinol (THC), cannabidiol (CBD), and cannabinol (CBN). The results revealed a slight upward trend in the mean THC content, from 7.0% in 2011 to 9.4% in 2018, with an overall mean value of 8.4%. The overall means of CBD and CBN content were 3.5% and 0.8%, respectively. The number of high-potency hashish samples gradually increased to reach 6% in 2018. Two distinct hashish chemotypes were identified: the highly populated chemotype II, corresponding to the traditional medium-potency hashish ([THC + CBN]/CBD ~ 2.16), and chemotype I, containing hashish samples of relatively high THC levels and low levels of CBD (ratio ~ 4.90). Both chemotypes I and II were characterized in the ternary plot, and the proportions (THC:CBD:CBN) were about 85%:13%:2% and 60%:35%:5%, respectively.

KEYWORDS: North African hashish, Δ9-tetrahydrocannabinol, cannabidiol, cannabinol, potency, high-performance liquid chromatography–diode array detection, criminalistics, forensic science, cannabis

The United Nations Office on Drugs and Crimes (UNODC) has reported a lack of systematic reporting and information about cannabis cultivation, production, use, and seizures in the regions of Africa and Asia. While the cannabis herb can be found all over the world, cannabis resin, commonly called hashish, is produced in only a few regions of the world, among them North Africa, which is considered to be one of the main sources of this drug. The North African country of Algeria is one that faces this hashish phenomenon, with several tons of the drug seized every year in the country.

The overall aim of this study was to supply reliable data about the hashish seized in Algeria by scientifically investigating its physical and chemical features along with the potential changes in the trafficking patterns and potency. The study was conducted with the aim of participating in the global efforts of developing a broader understanding of the most trafficked drug across the world (1).

Part I of this study focused on the physical features of the hashish seized in Algeria. Prominent traits were addressed, and shifts in production practices were assessed, as were the patterns that exist among the physical characteristics.

For the present article (Part II), the chemical characterization of the hashish seized in Algeria over the 8-year period of 2011–2018 was performed through a study of its chemical compositions.

Hashish contains more than 400 components (2,3), including more than 60 cannabinoids as well as terpenoids, phenols, and other additives (adulterants and contaminants). Cannabinoids are defined as a group of terpenophenolic compounds uniquely produced by cannabis (4).The major cannabinoids are considered to be Δ9-tetrahydrocannabinol (THC), cannabinol (CBN), and cannabidiol (CBD) (4–8), and what's left are minor cannabinoids.

Cannabis potency is defined as the percentage by weight of the primary active ingredient, THC. High-potency cannabis increases risks to health (9), making the scientific testing of the available cannabis to monitor current and ongoing trends in cannabis potency of paramount importance. However, a multivariate approach that takes into account the existing chemical pathways between the acidic cannabinoids—the biosynthesis products of the cannabis plant, their decarboxylation counterparts, and their degradation products—provide an extensive understanding of the psychotropic power of the studied hashish. Thereby, cannabidiol is considered as a major nonpsychoactive ingredient in cannabis that thwarts the psychoactive effect of Δ9-tetrahydrocannabinol (10), whereas degradation of Δ9-tetrahydrocannabinol results in the formation of cannabinol (11).

As explained in Part I of this study, a trained member of staff is designated to conduct a sampling every time hashish is seized in the Algerian territory. Laboratory samples are then analyzed according to the UNODC-recommended methods for the identification of cannabis and cannabis products using high-performance liquid chromatography (12).

The total content of individual major cannabinoids THC, CBD, and CBN, taking into account their acidic and neutral form, is monitored, in addition to the ratio (THC + CBN)/CBD

[1] Département de Toxicologie, Institut National de Criminalistique et de Criminologie (INCC-GN), Boite Postale 194, Bouchaoui, Algiers, 16000, Algeria.

Corresponding author: Yacine Boumrah, Ph.D. E-mail: y.boumrah1978@hotmail.com

Received 22 April 2020; and in revised form 10 July 2020; accepted 13 July 2020.

to evaluate possible existing chemotypes and their potential shifts and, as such, to assess global tendencies in hashish production. The relative concentration of CBN to THC is calculated to evaluate the freshness of the samples.

Chemical composition of seized hashish has never been assessed before in Algeria, so the obtained results will be compared with studies conducted in other countries such as Morocco, France, Italy, the U.K., the Netherlands, the U.S.A., and Japan (13–20). Although the temporal patterns of major cannabinoids' levels are the same over different regions of the world, potency discrepancies have been observed in the mean values.

Methods

Sample Acquisition

This study exploits hashish seizure data collected over an 8-year period (2011–2018) in Algeria. A total of 3265 hashish samples obtained from 2232 seizures made across the Algerian territory were analyzed.

Cannabis Surveyed: Sampling Procedure

Hashish seizures can be composed of multiple batches whose weight can reach into the tons. Hence, a proper sampling that ensures the representativeness of all the seizure batches must be conducted to obtain depictive test samples. Figure 1 provides a full example of a proper sampling procedure conducted when dealing with hashish seizures; all sampling steps are described for a real hashish seizure weighing 3320 kg. The seizure was visually inspected to identify the batches that had the same physical characteristics. Once the batches were separated, a sample (n) equal to the square root of the initial population (N), rounded to the next integer, was taken from the seizure elements in a stratified way (bags, packets, and units), according to the standard methods (21).

Regarding the sampling of the hashish slab bars, many studies have been conducted to establish the best sampling procedure (2,3,8,20). Most of the studies stated that cannabis resin on the surface differs from the interior because of the possible oxidation of the surface, which is in direct contact with air, and the heating effect on the surface during the pressing of the hashish slab bars. However, the surface layers of the North African hashish slab bars are thin in comparison with the hashish of different origins (20), and the studies conducted on the subject have indicated that two plugs of 0.5 g each, taken along the diagonal by taking a section through the entire slab bar (without removing the surfaces), are sufficient to ensure the hashish slab bar's representativeness.

Chemicals and Reagents

Cannabinoid reference materials—methanol solutions (1 mg/mL) of CBD and CBN, ethanol solution (1 mg/mL) of THC and isopropanol solution (1 mg/mL) of THCA—were purchased from Lipomed® (Arlesheim, Switzerland). Ibuprofen, used as internal standard (IS), was purchased from Saidal® (Algiers, Algeria).

High-performance liquid chromatography-grade methanol and chloroform purchased from VWR (Fontenay-sous-Bois, France) were used as extraction solvents. For the mobile phase, HPLC-grade methanol, acetonitrile, and orthophosphoric acid were purchased from VWR. Pure water was prepared in the laboratory using ELGA station, model: MEDICA RFII 7.

Sample Preparation

Hashish samples were finely ground and mixed using a porcelain mortar and pestle. Fifty milligram of the resulting powder was weighed in a centrifuge tube, and then, 10 mL of a methanol/chloroform 80:20 (v/v) mixture used as an extraction solvent was added. The tube was then vortexed for 1 min and subjected to an ultrasonic bath for 15 min. Subsequently, samples were mechanically agitated for 30 min and then centrifuged at **1370 × *g*** for 1 min. A 100-μL aliquot of the clear supernatant was transferred in a vial and diluted with 800 μL of the extraction solvent and 100 μL of IS.

Chromatographic Analysis

All the samples were analyzed by a fully validated in-house method using a high-performance liquid chromatography–diode array detector (HPLC-DAD). The method used a specific THERMO HPLC system, FINNIGAN SURVEYOR model, with a photodiode array detector (DAD). Chromatographic separation was achieved using a Phenomenex® (Macclesfield, Cheshire, UK) C18 analytical column (5 μm, 150 mm × 4.6 mm). Equipment control, data acquisition, and integration were performed with ChromQuest 4.2 software (Thermo Fisher Scientific, San Jose, CA, USA).

The mobile phase consisted of acetonitrile (A), methanol (B), and 0.1% orthophosphoric acid (C). The initial setting was 38% A, 33% B, and 29% C for 1 min, which was linearly increased to 53% A, 47% B, and 0% C over 20 min. Then, the column was set to initial conditions in 0.5 min and re-equilibrated under these conditions for 5 min. The following parameters were employed: flow rate, 1.2 mL/min; injection volume, 10 μL; detection wavelength, 220 nm; column temperature, 30°C. Figure 2 displays a typical chromatogram of an authentic hashish sample.

Calculation of Concentrations

A six-point calibration was used for the quantification of the neutral form of THC and its acidic form (THCA) with a linear range from 0.4% to 10%, and a six-point calibration was used for the quantification of the neutral forms of CBD and CBN with a linear range from 0.06% to 1%. Acidic forms of CBD (CBDA) and CBN (CBNA) were quantified using the THCA calibration curve. The accuracy of results was continuously monitored by the analysis of quality control (QC) samples. These QC samples consisted of positive samples taken from previously analyzed seizures, used later as inter-laboratory tests before their conversions to QC samples, stored at −20°C and changed every month.

For a convenient chemical characterization of the hashish samples studied, cannabinoids are classified into three groups: acidic cannabinoids produced by plant metabolism, neutral cannabinoids resulting from the natural decarboxylation of acidic cannabinoids, and degradation products resulting from various influences, such as UV light, heat, or prolonged storage (11).

To overcome the influence of variable conditions during processing, storage, and transportation of hashish, the quantification of the cannabinoids was conducted by adding the concentrations

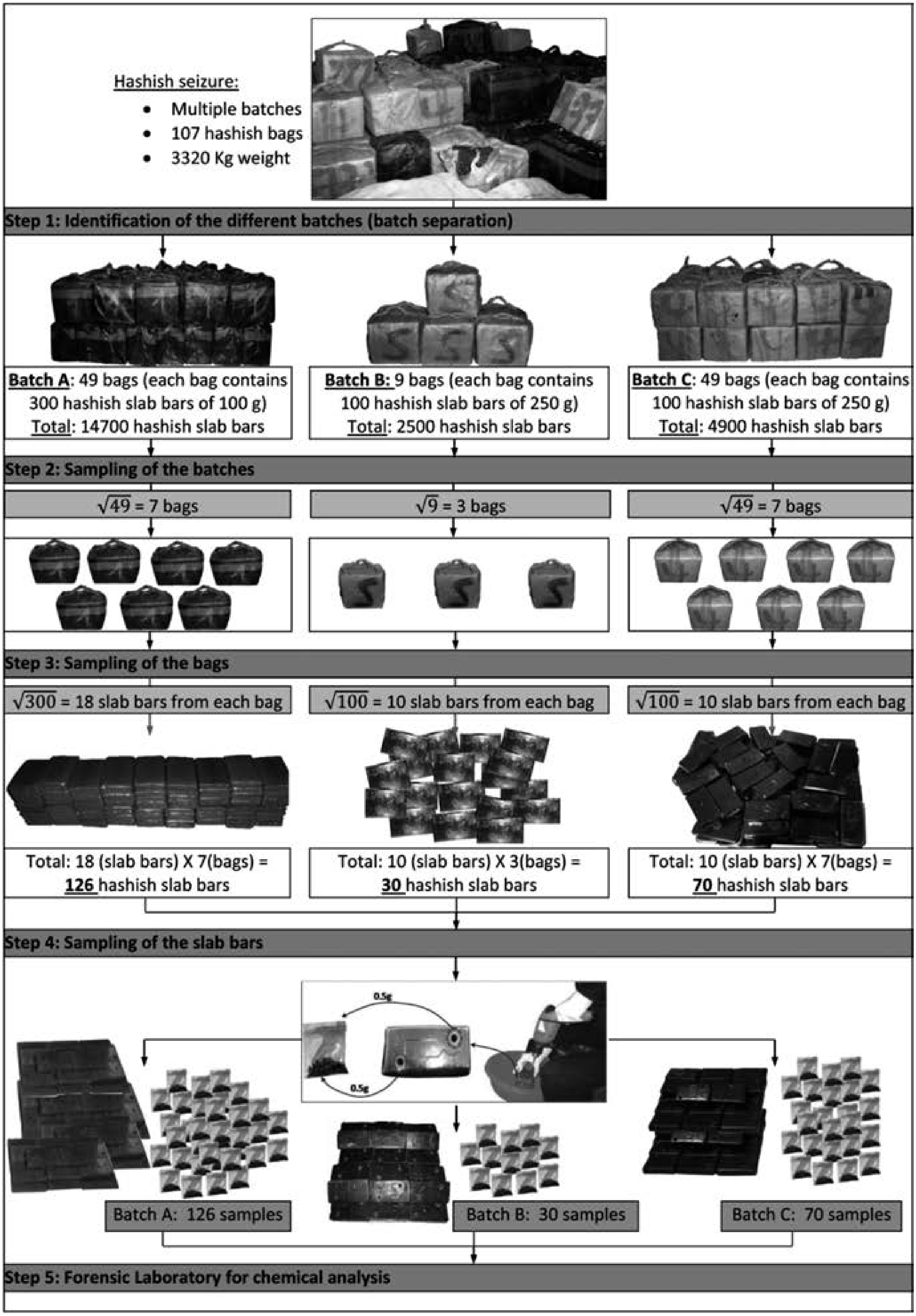

Hashish seizure:
• Multiple batches
• 107 hashish bags
• 3320 Kg weight
Step 1: Identification of the different batches (batch separation)
Batch A: 49 bags (each bag contains 300 hashish slab bars of 100 g)
Total: 14700 hashish slab bars
Batch B: 9 bags (each bag contains 100 hashish slab bars of 250 g)
Total: 2500 hashish slab bars
Batch C: 49 bags (each bag contains 100 hashish slab bars of 250 g)
Total: 4900 hashish slab bars
Step 2: Sampling of the batches
√49 = 7 bags
√9 = 3 bags
√49 = 7 bags
Step 3: Sampling of the bags
√300 = 18 slab bars from each bag
√100 = 10 slab bars from each bag
√100 = 10 slab bars from each bag
Total: 18 (slab bars) X 7(bags) = 126 hashish slab bars
Total: 10 (slab bars) X 3(bags) = 30 hashish slab bars
Total: 10 (slab bars) X 7(bags) = 70 hashish slab bars
Step 4: Sampling of the slab bars
0.5g
0.5g
Batch A: 126 samples
Batch B: 30 samples
Batch C: 70 samples
Step 5: Forensic Laboratory for chemical analysis

FIG. 1—*Example of sampling procedure conducted on a real hashish seizure. [Color figure can be viewed at wileyonlinelibrary.com]*

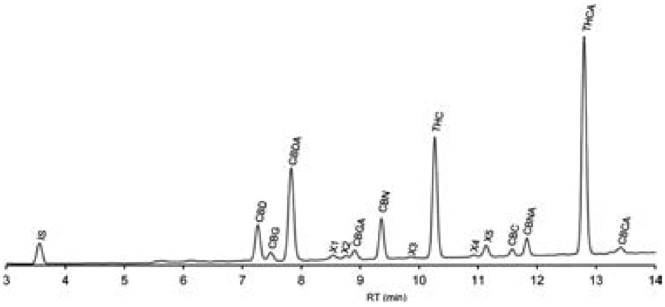


FIG. 2—*Example of HPLC/DAD chromatogram of authentic hashish sample. CBC, cannabichromene; CBCA, cannabichromenic acid; CBD, cannabidiol; CBDA, cannabidiolic acid; CBG, cannabigerol; CBGA, cannabigerolic acid; CBN, cannabinol; CBNA, cannabinolic acid; IS, internal standard; THC, tetrahydrocannabinol; THCA, tetrahydrocannabinolic acid; X1 to X5, unidentified cannabinoids. [Color figure can be viewed at wileyonlinelibrary.com]*

of their acidic and neutral forms:

$$\mathrm{THC}\,(\%) = \mathrm{THC}^{\mathrm{neutralform}}\,(\%) + \mathrm{THC}^{\mathrm{acidicform}}\,(\%)$$

$$\mathrm{CBD}\,(\%) = \mathrm{CBD}^{\mathrm{neutralform}}\,(\%) + \mathrm{CBD}^{\mathrm{acidicform}}\,(\%)$$

$$\mathrm{CBN}\,(\%) = \mathrm{CBN}^{\mathrm{neutralform}}\,(\%) + \mathrm{CBN}^{\mathrm{acidicform}}\,(\%)$$

Statistical Analysis

The mean and standard deviation (SD) of the sample concentrations were calculated for the combined data set by year. Normal and outlier hashish samples were determined based on the mean and SD of the THC concentration for each year (22) and also by the degrees of freshness of the samples. Normal samples are defined as fresh samples (CBN/THC ≤ 0.5), with potencies in the range: mean ± 2.5 × SD, and outliers are all samples considered not fresh as well as those with potencies that fall outside the range (mean ± 2.5 × SD). The precision of the mean was determined through 95% confidence intervals (CIs). The CI was calculated using the method described by Mehmedic et al. (23)

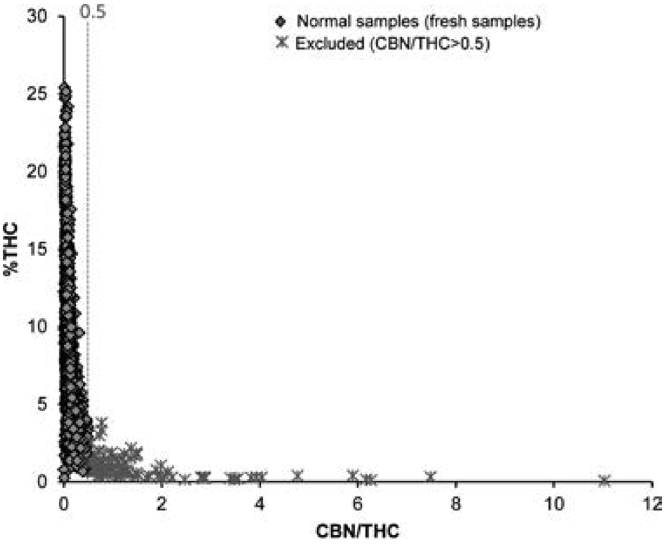


FIG. 3—*Degree of freshness of the studied hashish samples. [Color figure can be viewed at wileyonlinelibrary.com]*

Results and Discussion

Normal and Outlier Hashish Samples

The degree of freshness of the studied hashish samples is displayed in Fig. 3. Under the influence of heat and light, CBN is produced by oxidative degradation of THC (20). CBN is the most commonly found degradation product in aged cannabis; the relationship between THC (%) and the CBN/THC ratio, then, demonstrates the degradation process, which is directly proportional to the age of the samples. The CBN/THC ratio increases as THC levels decrease, which means that degraded hashish samples are characterized by a high CBN/THC ratio.

From the total 3265 studied hashish samples, 138 were eliminated because they were classified not fresh, and 94 other samples were eliminated because their potencies fell outside the range (mean ± 2.5 × SD).

TABLE 1—*Mean and SD Δ9-THC% with outliers* excluded, prevalence of low- (<10%), medium- (10 < THC < 20), and high- (>20%) potency hashish samples and means CBD% & CBN% by year.*

		Total THC%					Prevalence Samples Potency			CBD %	CBN %
			All Samples		Outliers Excluded						
Year	*n*	Outliers (%)	Mean	SD	Mean	SD	THC < 10% (%)	10 < THC < 20 (%)	THC > 20 (%)	All Samples Mean	All Samples Mean
2011	128	3.9	7.0	3.8	7.0	3.5	78.1	21.9	0.0	3.7	0.6
2012	132	3.0	8.9	3.9	8.8	3.6	64.1	35.1	0.8	4.5	0.7
2013	165	6.0	7.3	3.8	7.2	3.2	75.6	23.8	0.6	3.6	0.6
2014	259	7.7	7.8	3.9	7.5	2.8	78.4	19.7	1.9	3.5	0.6
2015	189	8.9	7.9	4.0	7.6	3.0	75.0	23.9	1.1	3.6	0.7
2016	783	8.6	9.4	5.0	8.9	3.9	62.5	32.7	4.9	3.7	0.8
2017	806	8.9	8.1	3.8	7.9	3.0	81.3	18.2	0.5	3.2	0.8
2018	806	9.4	9.4	5.5	9.1	4.4	60.3	33.7	6.0	3.3	0.9
Total	3265	7.1	8.4	4.7	8.3	3.8	71.9	26.1	2.0	3.5	0.8
95% CI range			8.3–8.6		7.4–8.3					3.5–3.6	0.7–0.8

95% CI range, range of values that contains the true mean with 95% certainty; SD, Standard deviation.

*Outliers are samples that (Mean − 2.5 × SD > %THC > Mean + 2.5 × SD) + samples that (CBN/THC) > 0.5.

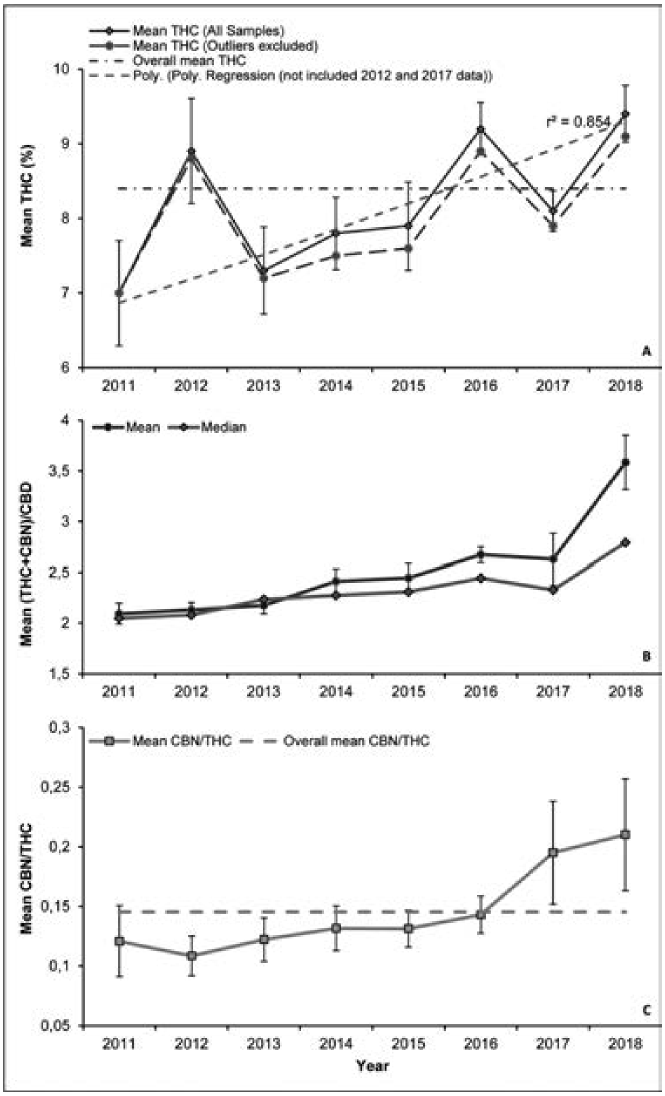


FIG. 4—*Temporal evolution of the potency of the studied hashish samples over the period 2011–2018 with 95% confidence intervals. A: Mean THC concentration for all the studied hashish samples and all the samples with outliers excluded; B: Mean (THC + CBN)/CBD ratio (mean and median) for all the studied hashish samples; C: Mean CBN/THC ratio for all the studied hashish samples. [Color figure can be viewed at wileyonlinelibrary.com]*

Potency of the Studied Hashish

In general, the yearly arithmetic mean THC concentration for the different hashish samples displays a relative stability over the 8-year period (2011–2018) with an overall mean THC content of 8.3% and SD of 4.7% (Table 1). However, a slight increase in mean THC content has been observed ranging from (7.0% ± 3.8%) in 2011 to (9.4% ± 5.5%) in 2018. In fact, despite hashish samples seized in 2012 and late 2017 that deviate from the general tendency, the hashish THC concentration appeared to gradually increase from 2011 to 2018, with a second-order polynomial regression correlation coefficient of 0.854.

This tendency toward increased potency also applies to other countries and regions across the world. However, discrepancies in the expressed distribution of mean THC have been identified. The existence of extremely potent forms of hashish was recorded a decade ago in Europe and the U.S.A., where mean THC values were already above 10% (5,18).

The potency of confiscated hashish in France, which mostly comes from Morocco, as it does in Algeria, had a mean THC concentration of 18.2% over the period of 2011–2016 (15). This suggests that hashish samples of the same geographical origin do not necessarily have the same potency, leading to the fact that hashish producers process the cannabis according to the targeted consuming countries.

The influence of the outlier samples on the overall mean concentration of THC was investigated. The outlier hashish samples represent 7.1% of the total samples studied, including degraded samples (4.22%) and those with THC concentrations that fall outside the range of mean ± 2.5 × SD (2.88%). A comparison of the mean potency of hashish samples calculated for all samples versus for samples with outliers excluded indicates that the mean THC concentration decreases weakly for each year when the outliers are excluded (Table 1, Fig. 4). However, the general pattern of the slightly increasing potency of hashish samples since 2011 appears to exist even when outliers are excluded.

Δ9-tetrahydrocannabinol content was assessed by categorizing hashish seizures according to three types: low-potency hashish (THC < 10%), medium-potency hashish (10% < THC < 20%), and high-potency hashish (THC > 20%). The results were compared to those obtained by Stambouli et al. (13) and Dujourdy and Besacier (14), who published statistics on hashish in Morocco and France, respectively.

Irrespective of the period over which statistics were calculated, the results obtained were in agreement with those observed by Stambouli et al.: Low-potency hashish is predominant, here representing more than 70% of the hashish seizures, and high-potency hashish represents the lowest percentage at just 2% of the total hashish samples examined for this study (Table 1).

Disregarding the fluctuations observed in 2012 and 2017, the results shown in Fig. 5 indicate that high- and medium-potency hashish samples increased from 0% and 21% in 2011 to 6% and 33% in 2018, respectively. Conversely, low-potency hashish samples decreased from 78% (2011) to 60% (2018). Contrary to what was reported in Europe and the U.S.A., concentrated hashish in Algeria appears sporadically, suggesting that the local market is scarcely supplied with this kind of hashish.

The mean concentrations of the two other major cannabinoids in cannabis (CBD and CBN) were monitored (Table 1). The yearly arithmetic mean CBD and CBN concentrations for the hashish samples remained relatively stable, with an overall mean content of 3.5% and 0.8%, respectively. However, in recent years, more hashish samples have been found with lower CBD content (<3%) in parallel with the rise of high-potency hashish samples. Figure 4B shows the evolution of the (THC + CBN)/CBD ratio by year; there is a clear tendency toward higher THC and lower CBD content over the years, which highlights increasing health risks and harms from the use of this type of hashish. From 2011 to 2013, the mean and median are almost the same, averaging around 2, but since 2014 the mean steadily deviates from the median toward higher values, indicating a constant increase in the appearance of high-potency hashish samples.

By plotting the relative concentration of CBN to THC (Fig. 4), the freshness of the seized hashish was also investigated. The CBN/THC ratio was used to estimate the age of plant materials (marijuana) with indicative values; fresh marijuana samples, considered to be <6 months old, have a CBN/THC ratio below 0.013, and those with a ratio between 0.04 and 0.08 are from 1 to 2 years old (14,15). The age of hashish can be

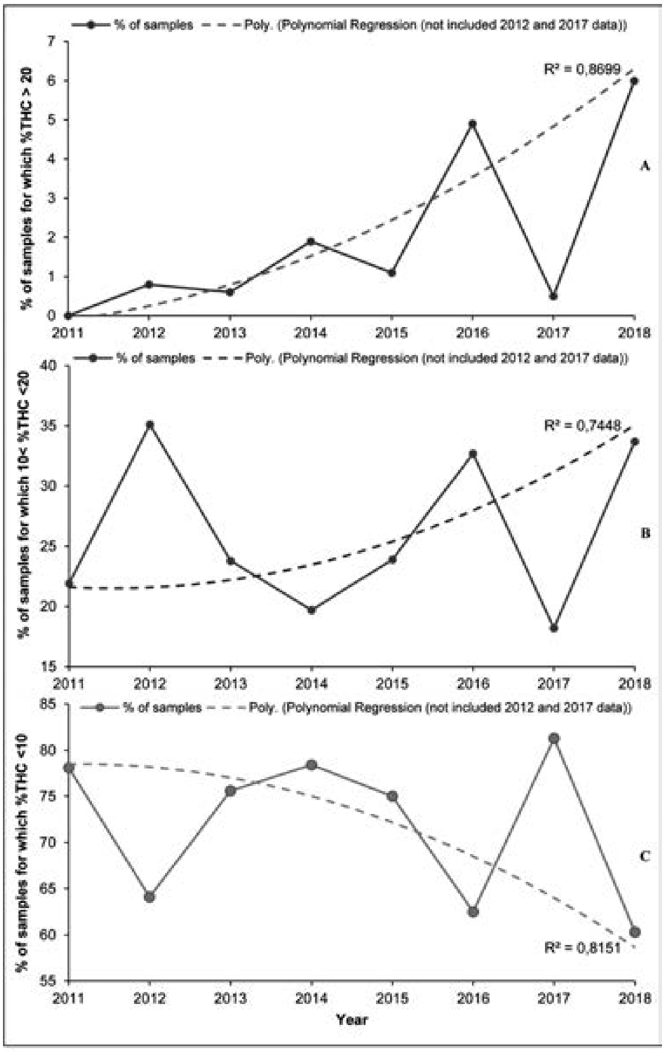


FIG. 5—*Distribution frequency of hashish samples over the period 2011–2018 according to their potency classes. A: Prevalence of high-potency hashish samples (THC > 20%); B: Prevalence of medium-potency hashish samples (10 < THC% < 20); C: Prevalence of low-potency hashish samples (THC < 10%). [Color figure can be viewed at wileyonlinelibrary.com]*

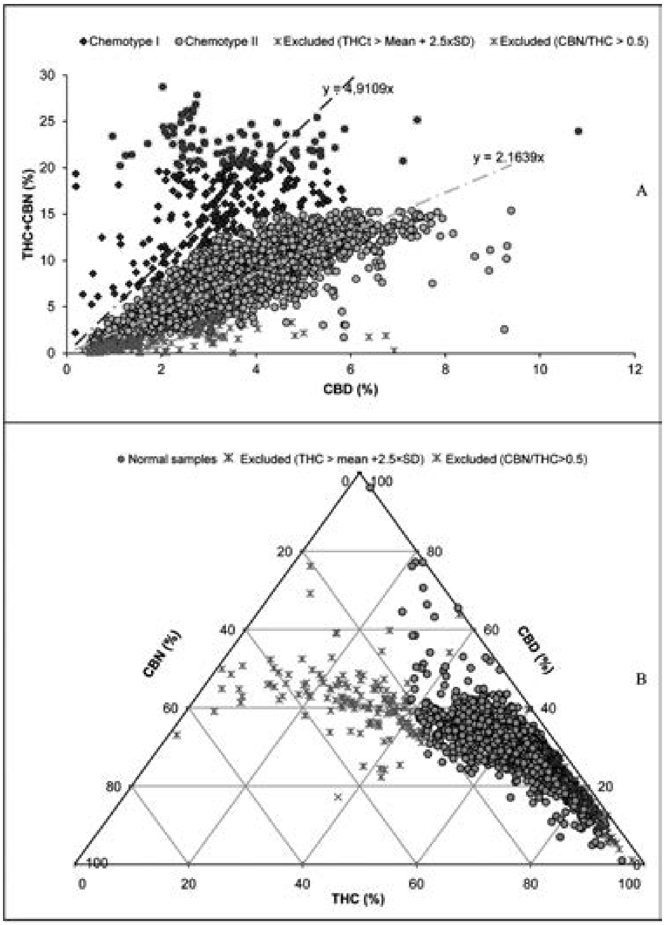


FIG. 6—*Chemotypes and ternary plot of the studied hashish. A: (THC + CBN)% vs. CBD% of all the studied hashish sample; B: Triangular co-ordinate graph (THC:CBD:CBN) of all the studied hashish samples. [Color figure can be viewed at wileyonlinelibrary.com]*

estimated in the same way as marijuana samples, as mentioned by Zamengo et al. (7). However, hashish commonly contains greater amounts of THC than marijuana (8) and is more subject to degradation because of the manufacturing method, including the pressing step, which makes THC more prone to degradation. Hence, CBN/THC ratios of hashish samples are of greater magnitude.

The Algerian-seized hashish evolves toward a decreasing freshness, with an overall mean CBN/THC ratio of about 0.15. This decrease is probably due to longer periods between the harvesting and smuggling of hashish, meaning longer transport periods toward the Algerian territory, which indicates a shift in the smuggling dynamic but also a constant increase in the appearance of high-potency hashish samples.

Chemotypes of the Studied Hashish

In order to comprehend the possible changes in the chemical profile of North African hashish, more investigations were carried out by introducing ratios that take into account the different chemical pathways that exist between major cannabinoids, such as the evolution of the factor THC + CBN with respect to CBD content. The results described in Fig. 6 indicate the existence of two separate chemotypes: Most of hashish samples fall into chemotype II, characterized by a (THC + CBN)/CBD ratio of about 2.16, corresponding to the traditional North African hashish. Chemotype I includes fewer samples and is characterized by high THC concentration and relatively low levels of CBD (ratio ~ 4.90). All outlier samples with potencies higher than the mean potency (THC concentrations fall outside the range of mean ± 2.5 × SD) are included in chemotype I, as they follow the same trend of chemotype I samples.

As a result of this new trend, a higher risk of psychotic effect is induced, as studies conducted on cannabinoids indicate that CBD may counteract some of the adverse psychological effects induced by THC (6). This rather novel fact is an indicator of the introduction of new cultivars and new methods of resin production, which is consistent with findings in Europe and the U.S.A. (7,14,16,18).

Jenkins and Patterson (24) developed the concept of relating the results of the chemical analysis of a cannabis sample to its geographical origin. The authors concluded that there were differences in the relative amounts of THC, CBD, and CBN between samples of different geographical origins, which can be useful from an intelligence perspective. Evident differences in THC content and THC/CBD ratios have been observed between the hashish seized in Algeria and in France even if they were from the same geographical origin. In this scope, the ternary plot of THC, CBD, and CBN is used to characterize the hashish seized in Algeria rather than to investigate its geographical origin (Fig. 6).

The results showed that Algerian-seized hashish has a mean THC:CBD:CBN factor of 60%:35%:5%. In almost 80% of the total samples, the relative proportion of THC ranges between 55% and 75%, CBD between 25% and 45%, and CBN between 2% and 10%. All outliers considered as degraded samples (CBN/THC > 0.5) were excluded, as the trend of these samples did not follow the pattern of the large majority of the studied samples, this being the reason why they are considered as statistical artifacts. Meanwhile, outliers for which THC concentrations fall outside the range of mean $\pm$ 2.5 $\times$ SD were included into the hashish chemotype I and characterized by a THC:CBD:CBN factor of an average value of 85%:13%:2%.

Conclusion

The content of major cannabinoids in Algerian-seized hashish was investigated. Mean THC content showed a slight increasing trend from 7.0% in 2011 to 9.4% in 2018 with an overall mean value of 8.4%. High-potency hashish is not abundant in Algeria as it is in European countries and the U.S.A. However, the proportion of high-potency hashish samples gradually increased to reach 6% in 2018.

The monitoring of the (THC + CBN)/CBD ratio revealed the existence of two hashish chemotypes: The traditional one (chemotype II) includes the majority of the studied samples and is characterized in the ternary plot by THC:CBD:CBN proportions of about 60%:35%:5%. Chemotype I includes all hashish samples that tend to have high THC and low CBD content and is characterized by THC:CBD:CBN proportions of about 85%:13%:2%.

References

1. United Nations Office on Drugs and Crime. World drug report 2018. Vienna, Austria: United Nations Publications, 2018. Sales No. E.18.XI.9.
2. Lewis R, Ward S, Johnson R, Burns DT. Distribution of the principal cannabinoids within bars of compressed cannabis resin. Anal Chim Acta 2016;538:399–405. https://doi.org/10.1016/j.aca.2005.02.014
3. McDonald PA, Gough TA. Determination of the distribution of cannabinoids in cannabis resin from the Lebanon using HPLC. Part III. J Chromatogr Sci 1984;22:282–4. https://doi.org/10.1093/chromsci/22.7.282
4. De Backer B, Debrus B, Lebrun P, Theunis L, Dubois N, Decock L, et al. Innovative development and validation of an HPLC/DAD method for the qualitative and quantitative determination of major cannabinoids in cannabis plant material. J Chromatogr B Analyt Technol Biomed Life Sci 2009;877:4115–24. https://doi.org/10.1016/j.jchromb.2009.11.004
5. Freeman TP, Groshkova T, Cunningham A, Sedefov R, Griffiths P, Lynskey MT. Increasing potency and price of cannabis in Europe, 2006–16. Addiction 2019;114:1015–23. https://doi.org/10.1111/add.14525
6. Marcu JP. An overview of major and minor phytocannabinoids. In: Victor R, editor. Neuropathology of drug addictions and substance misuse: foundations of understanding, tobacco, alcohol, cannabinoids and opioids, vol. 1. London, U.K.: Elsevier Inc/Academic Press, 2016;672–8. https://doi.org/10.1016/B978-0-12-800213-1.00062-6
7. Zamengo L, Frison G, Bettin C, Sciarrone R. Cannabis potency in the Venice area (Italy): update 2013. Drug Test Anal 2015;3:255–8. https://doi.org/10.1002/dta.1690
8. Cascini F, Aiello C, Di Tanna G. Increasing delta-9-tetrahydrocannabinol (Δ9-THC) content in herbal cannabis over time: systematic review and meta-analysis. Curr Drug Abuse Rev 2012;5:32–40. https://doi.org/10.2174/1874473711205010032
9. Di Forti M, Sallis H, Allegri F, Trotta A, Ferraro L, Stilo SA, et al. Daily use, especially of high-potency cannabis, drives the earlier onset of psychosis in cannabis users. Schizophr Bull 2014;40:1509–17. https://doi.org/10.1093/schbul/sbt181
10. Devinsky O, Cilio MR, Cross H, Fernandez-Ruiz J, French J, Hill C, et al. Cannabidiol: pharmacology and potential therapeutic role in epilepsy and other neuropsychiatric disorders. Epilepsia 2014;55:791–802. https://doi.org/10.1111/epi.12631
11. Hazekamp A, Fischedick JT, Llano DM, Lubbe A, Ruhaak RL. Chemistry of cannabis. In: Mander L, Liu H-W, editors. Comprehensive natural products II:chemistry and biology, vol. 3. Kidlington, U.K.: Elsevier Ltd, 2010;1033–84.
12. United Nations Office on Drugs and Crime. Recommended methods for the identification and analysis of cannabis and cannabis products: manual for use by national drug analysis laboratories. Rev. and updated. Vienna, Austria: United Nations Publications, 2009;41–2.
13. Stambouli H, El Bouri A, Bouayoun T. Évolution de la teneur en Δ9-THC dans les saisies de résines de cannabis au Maroc de 2005 à 2014 [Evolution of the Δ9-THC content in cannabis resin seizures in Morocco from 2005 to 2014]. Toxicol Anal Clin 2016;28:146–52. https://doi.org/10.1016/j.toxac.2015.11.001
14. Dujourdy L, Besacier F. A study of cannabis potency in France over a 25 years period (1992–2016). Forensic Sci Int 2017;272:72–80. https://doi.org/10.1016/j.forsciint.2017.01.007
15. Zamengo L, Frison G, Bettin C, Sciarrone R. Variability of cannabis potency in the Venice area (Italy): a survey over the period 2010–2012. Drug Test Anal 2014;6:46–51. https://doi.org/10.1002/dta.1515
16. Potter DJ, Hammond K, Tuffnell S, Walker C, Di Forti M. Potency of Δ9–tetrahydrocannabinol and other cannabinoids in cannabis in England in 2016: implications for public health and pharmacology. Drug Test Anal 2018;10:628–35. https://doi.org/10.1002/dta.2368
17. Niesink RJM, Rigter S, Koeter MW, Brunt TM. Potency trends of Δ9-tetrahydrocannabinol, cannabidiol and cannabinol in cannabis in the Netherlands: 2005–15. Addiction 2015;110:1941–50. https://doi.org/10.1111/add.13082
18. Chandra S, Radwan MM, Majumdar CG, Church JC, Freeman TP, ElSohly MA. New trends in cannabis potency in USA and Europe during the last decade (2008–2017). Eur Arch Psychiatry Clin Neurosci 2019;269:5–15. https://doi.org/10.1007/s00406-019-00983-5
19. Tsumura Y, Aoki R, Tokieda Y, Akutsu M, Kawase Y, Kataoka T, et al. A survey of the potency of Japanese illicit cannabis in fiscal year 2010. Forensic Sci Int 2012;221:77–83. https://doi.org/10.1016/j.forsciint.2012.04.005
20. Baker PB, Gough TA, Wagstaffe PJ. Determination of cannabinoids in cannabis resin from Morocco using high-performance liquid chromatography. Part II. J Anal Toxicol 1983;7(1):7–10. https://doi.org/10.1093/jat/7.1.7
21. Williams DG. Drugs and poisons. In: Topics in renal disease: an illustrated guide. Dordrecht, the Netherlands: MTP Press Limited, Springer, 1981;73–6. ISBN-13: 978-0852004210.
22. Barnett V, Lewis T. Outliers in statistical data, 3rd edn. New York, NY: John Wiley & Sons Ltd., 1994.
23. Mehmedic Z, Chandra S, Slade D, Denham H, Foster S, Patel AS, et al. Potency trends of Δ9-THC and other cannabinoids in confiscated cannabis preparations from 1993 to 2008. J Forensic Sci 2010;55(5):1209–17. https://doi.org/10.1111/j.1556-4029.2010.01441.x
24. Jenkins RW, Patterson DA. The relationship between chemical composition and geographical origin of cannabis. Forensic Sci 1973;2:59–66. https://doi.org/10.1016/0300-9432(73)90014-9

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14553
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Jacqueline A. Speir,[1] *Ph.D.; Nicole Richetelli,*[1] *M.S.; and Lesley Hammer,*[2] *M.S.*

Forensic Footwear Reliability: Part I—Participant Demographics and Examiner Agreement*

ABSTRACT: In order to assess the extent of agreement between forensic footwear examiners in the United States, a reliability study was performed by West Virginia University between February 2017 and August 2018. Over the span of 19 months, 70 examiners each performed 12 comparisons and reported a total of 840 conclusions. For each comparison, participants were queried on a number of factors in order to determine the degree to which different types of features were identified, evaluated, and weighted, before arriving at a final decision regarding the strength of the association or disassociation between questioned and test impressions. Preliminary results from this study are divided into a series of three summaries. This manuscript (Part I) describes participant demographics as well as community agreement in both feature identification/annotation, and final reporting. Results indicate considerable variation in feature identification/annotation (as low as 66.5% agreement), but higher consistency in the reporting of overall conclusions. For mated pairs, this agreement was 79.7% ± 14.1% (median of 85.7% and a 90% confidence interval between 75.9% and 83.2%). For nonmated pairs, the equivalent overall agreement was 89.8% ± 6.69% (median of 91.4% and a 90% confidence interval between 87.4% and 92.1%). These estimates of agreement are further compared with previous published findings, and collectively, the work extends the body of knowledge concerning reliability in forensic footwear comparisons and conclusions.

KEYWORDS: forensic footwear evidence, reliability, gray box study, footwear examiners, participant demographics, feature identification agreement, expert agreement

Research supporting the foundational validity of feature-comparison methods is of tremendous importance to the forensic community (1–3). In the field of footwear, much of the research effort has focused on the discrimination potential of the physical evidence itself (4–18), with less emphasis on the evaluation of examiner consensus and variation in feature identification, feature evaluation, and source attribution conclusions. In fact, research aimed at understanding the expert decision-making process in the field of forensic footwear analysis is somewhat limited to 3 published studies, performed in the past 20 years, each with different intended audiences and project design (although the authors acknowledge that additional work on this topic has been presented at conferences, but with results that are much less universally available for review and interpretation). Of these 3 studies, the first was by Majamaa and Ytti (19) in 1996, the second by Shor and Weisner (20) in 1999, and the most recent by Hammer et al. (21) in 2013.

The earliest of these studies (Majamaa and Ytti [19]) was conducted internationally, using 6 simulated crime scene impressions, sent to 34 laboratories, with a 97% completion rate (responses received from 33 analysts). Each of the 6 questioned impressions was collected using an electrostatic dust lift of an impression on paper, to be compared to a single known. In 2 simulated cases, the known was a worn shoe, and in the remaining 4 cases, the knowns were new/unused. In total, 2 "types" of shoes were evaluated (i.e., the worn shoes had the same class characteristics, and the new/unused had the same class characteristics). All examiners were given a 5-level conclusion scale for reporting (19). For some laboratories, this required translation. Each analyst had to individually define the criteria (i.e., the type of features that must be present, and the degree of similarity/dissimilarity expressed by each) in order to reach a specific conclusion (e.g., "*probable*," "*very probable*," or "*identification*"). In addition, examiners were informed that class characteristics and identified randomly acquired characteristics were in agreement, eliminating the tasks of feature identification and comparison, and instead, asked to limit their comparison to the weight of evidence associated with a given conclusion. Despite the alleviation of several sources of variation (feature identification and comparison), the outcome of this review indicated a relatively high level of variability in conclusions for examiners evaluating identical cases (although it is important to note that the degree of variation was case specific). To illustrate the extreme or most variable case (denoted as case #3), 5 examiners reported "*possible*," 9 reported "*probable*," 11 reported "*very probable*," and 8 reported "*identification*" (19).

[1] West Virginia University, 208 Oglebay Hall, PO Box 6121, Morgantown, WV, 26506.

[2] Hammer Forensics, LLC, 10601 Prospect Drive, Anchorage, AK, 99507.

Corresponding author: Jacqueline A. Speir, Ph.D. E-mail: Jacqueline.Speir@mail.wvu.edu

*This project was supported by Award No. 2016-DN-BX-0152, awarded by the National Institute of Justice (NIJ), Office of Justice Program, U.S. Department of Justice. However, the opinions, findings, conclusions, and recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the Department of Justice.

Received 18 Mar. 2020; and in revised form 15 June 2020; accepted 31 July 2020.

In contrast, the Shor and Weisner (20) study involved 2 actual case impressions (ground truth not available), evaluated by 20 experts from 7 different laboratories, across 6 different countries (plus 3 examiners from the respective authors' own laboratory). The cases were selected due to their difficulty (the questioned impressions were deemed ambiguous and controversial by experts), and each analyst was permitted to use his or her own conclusion scale when providing results. Again, the major observation from this report was the degree of variability in conclusions provided by the 23 participants, including intra- and inter-laboratory variability. For both impressions, the responses spanned 5 categories (of a possible 6) and the largest agreement for 1 category was 7 responses of "*possible*" and 7 responses of "*highly probable*" out of 23 for the first shoe (30% each), and 9 responses of "*possible*" out of 23 for the second shoe (39%).

The most recent study by Hammer et al. (21) was conducted in 2013, and although modeled after Majamaa and Ytti (19)'s work from 1996, the project design included 3 notable differences. First, the targeted participants were restricted to examiners located in North America. Second, all participants were certified as footwear examiners by the International Association for Identification (IAI) as of July 2008. Finally, all participants were expected to use a standardized 7-category conclusion scale published by the Scientific Working Group for Shoeprint and Tire Tread Evidence (SWGTREAD) (at the time, the recommended scale allowed for one of the following conclusions: "*unsuitable*," "*elimination*," "*probably did not make*," "*inconclusive*," "*could have made*," "*probably made*," or "*identification*" [22] but has since been updated and the current permissible categories are now: "*lacks sufficient detail*," "*exclusion*," "*indications of non-association*," "*limited association of class characteristics*," "*association of class characteristics*," "*high degree of association*," and "*identification*" [23]). Despite these 3 differences, 1 major similarity persisted: the examiners were told to assume that the design and physical size of the questioned and test impressions corresponded, and finally, the research team highlighted the features of interest in the questioned impressions to be used during examination, comparison and reporting. The entire study was comprised of 6 simulated cases, each consisting of a gelatin lift questioned impression. Four different pairs of shoes were used to create the 6 simulated cases, with variation in class and individualizing characteristics, as well as quality/totality. Compared to former studies, Hammer et al. (21) found that experienced examiners using the standardized scale and predetermined features, expressed less variability in their conclusions, with the most variation occurring in a single case regarding the degree of confidence an examiner was willing to place on his or her final conclusion. For this case, 82% of the examiners found the suspected shoe "*probably made*" the impression, 15% concluded with "*could have made*," and another 3% chose "*identification*."

Using the Majamaa and Ytti (19) and Hammer et al. (21) studies as motivation, an additional reliability study on footwear examiner conclusions was conducted between February 2017 and August 2018 by West Virginia University (WVU). The intended audience were footwear examiners in the United States, who were asked to review 7 simulated cases, requiring a total of 12 questioned-test impression comparisons. The goals of this project were several-fold, including an evaluation of (i) reproducibility in feature identification, (ii) reproducibility in feature evaluation, (iii) consensus and inter-rater reliability, and (vi) overall accuracy in conclusions regarding source attribution. In terms of organization, initial results are dispersed in a series of three manuscripts. Part I (this summary) describes reproducibility in feature identification and generalized community agreement.

Materials and Methods

Participant Demographics

One hundred and fifteen (115) forensic footwear examiners were recruited through a variety of media, including electronic solicitation, word-of-mouth, and in-person announcements during regional and international conferences. Enrolled participants completed a background survey providing information regarding their education, experiences, job capacity, certification status, as well as details concerning the nature and frequency of training, research, teaching, and professional development activities. Of the 115 enrolled participants, conclusions were submitted by 77 individuals, resulting in a 67% response rate.

Case Variety

Each participant performed 12 pairwise comparisons, spanning 7 simulated cases (of which 5 required the analysis of 2 exemplars, and 2 required the analysis of a single exemplar). Each case (Table 1) was comprised of 1200 PPI digital and print imagery, collected using a flatbed Epson Expression 11000XL Graphic Arts Scanner, and printed using a Canon Pixma Pro-1. Case materials consisted of a single questioned impression, 1–2 outsole exemplars, and 2 Handiprint exemplar replicates per known shoe (Table 2). The questioned impressions were created under natural conditions (walking at a regular pace/stride length) using a range of media (blood, dust, wax), substrates (linoleum/ceramic/vinyl tiles, paper), and processing techniques (lifting, chemical/digital enhancement). Effort was expended to create "crime scene-like" impressions of the type, variety, and quality encountered by analysts during routine casework. However, it is acknowledged that the wearer creating the questioned impressions had a smaller foot size than the actual outsoles used in this study, which may have created experimental limitations.

Case Analyses

Each participant received a package via USPS of all relevant case materials, including high-resolution color prints, a set of blank acetates for overlay annotation, a CD containing the electronic reporting software, a copy of the Scientific Working Group for Shoeprint and Tire Tread Evidence (SWTREAD) 2013 Conclusion Standard (23) and an instruction document (with additional weblinks to access electronic copies of all case materials, including digital files of 1200 PPI imagery still accessible as of 04-Sep-2019 as BlackBox.zip from http://4n6chemometrics.com/Downloads/). Participants were asked to process the simulated cases as if each were routine casework, and analyze the case materials according to their training and expertise, assuming that no time had passed between collection of the questioned and test impressions (i.e., the absence of any change due to continued usage/wear). After performing a routine analysis, participants were asked to respond to a series of questions using a customized software reporting interface that solicited responses regarding the similarity, dissimilarity, clarity, and value of manufacturing and wear-acquired features used when reaching conclusions. Based on these instructions, the only anticipated deviation from typical casework was the absence of

TABLE 1—*Shoes, substrates, media, and processing techniques used to create simulated case materials.*

Case	Manufacturer of Known(s)	Size & Style of Known(s)	Substrate of Unknown	Medium of Unknown	Processing of Unknown	# of Known(s)
001	Converse	All Star (9)	Ceramic tile	Blood	Leucocrystal violet	2
002	Nike	Lebron James (10)	Vinyl tile	Dust	Digitally enhanced gel lift	1
003	Nike	Rosherun (9)	Ceramic tile	Blood	Leucocrystal violet	2
004	Nike	Air Max (10.5)	Linoleum tile	Wax	Magnetic powder & gel lift	2
005	Nike	Air Max (11)	Vinyl Tile	Dust	Digitally enhanced gel lift	1
006	Nike	Air Max Cage (10)	Paper	Dust	Digitally enhanced	2
007	Under Armour	Unknown (10 & 11)	Ceramic Tile	Blood	Leucocrystal violet	2

TABLE 2—*Manufacturing details for each known shoe.*

Known(s)	Manufacturer	Style	Size	Additional Details
001K1, 001K2	Converse	All Star	9	–
002K1	Nike	Lebron James	10	–
003K1, 003K2	Nike	Rosherun	9	Microcellular material, Manufacturer reports same size, but measureable size difference for 003K2
004K1, 004K1	Nike	Air Max	10.5	–
005K1	Nike	Air Max	11	Manufacturer reported (and measurable) size difference
006K1, 006K2	Nike	Air Max Cage	10	–
007K1	Under Armour	–	11	Manufacturer reported (and measurable) size difference
007K2A, 007K2B	Under Armour	–	10	007K2B has an apparent manufactured defect/heel anomaly

consultation and/or any type of independent verification of examiner conclusions prior to reporting.

Results and Discussion

Participant Demographics

The cumulative demographics for all 77 individuals that participated in this study are summarized in Table S1 (top row). Self-reports revealed that the majority of participants (83%) were actively working in a crime lab at the time of participation. Self-reports also revealed that 7 of the 77 participants had either never performed a comparison and/or were still in training. Since this evaluation was meant to determine the accuracy and conformity in reporting for footwear examiners actively performing casework, the results from these 7 participants were excluded when creating summary statistics.

Table S2 reveals that 36% of all participants had completed 11–50 comparisons when they agreed to participate in this study, while another 20% had completed 51–100 comparisons, and 23% had completed more than 100 comparisons. Participants were also asked to report the frequency at which different types of activities were performed using a Likert scale (Table S3). It appears that few participants collect or develop impressions (presumably at scenes), but more frequently enhance, photograph and compare impressions (presumably in laboratories), which fits with anecdotal reports within the field.

Since footwear examiners may be asked to perform comparisons on multiple types of evidence, and since experts can cross-train and/or move from one discipline to another throughout their careers, Table S4 reports additional current and past activities that participants have performed. Results indicate that many footwear examiners have been or still are involved with crime scene related work (processing, investigation, reconstruction, bloodstain pattern analysis, etc.). Additionally, many have worked in trace evidence or firearms and toolmark fields, but a majority co-listed fingerprints as an area of current occupation.

Table S5 reports the number of years of footwear experience and the number of years of total forensic experience for examiners that participated in this study; 53% had 8 or more years of experience in footwear, and the majority (86%) have been in the forensic field for more than 8 years.

Table S6 reports the frequency of training in the last 5 years, as well as the types of training providers. Note that the majority of participants (97%) have attended 1 or more training sessions beyond laboratory-specific activities. In addition to employment and professional experiences, each examiner's traditional academic history was queried; Table S6 also reports the highest level of education earned for each participant, with 54% possessing a Bachelor's degree, and 40% having earned a Master's degree.

Finally, Table S7 reports that half of the participants in this study use the SWGTREAD (2013) scale (23) (without modification) in their laboratory, and approximately half are certified.

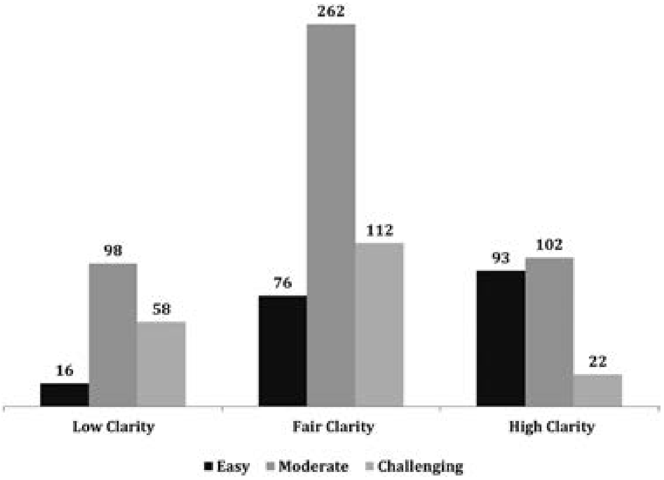


FIG. 1—*Overall reports of clarity (low, fair, and high) and difficulty (easy, moderate, and challenging) for* n = *839 comparisons (note that one examiner reported that case 005Q was unsuitable for comparison; this observation is excluded from this figure). [Color figure can be viewed at wileyonlinelibrary.com]*

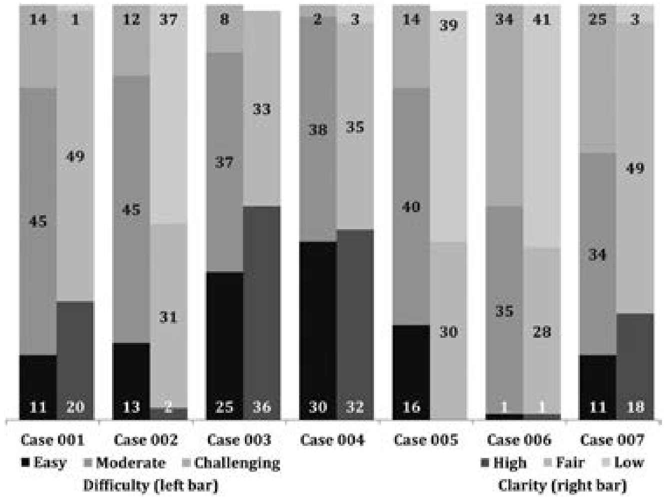


FIG. 2—*Reports of the frequency (*y*-axis) of quality (low, fair and high) and difficulty (easy, moderate and challenging) for each questioned impression (note that 1 examiner reported insufficient detail for 005Q, and an examiner switched his or her clarity report for 003Q when comparing it against 003K1 and 003K2; these 2 reports are excluded from this figure). [Color figure can be viewed at wileyonlinelibrary.com]*

Moreover, 87% participated in a proficiency test in the past year, and 60% have taught courses in the field of forensic footwear.

Case Variety

The details concerning each simulated case in this study were previously summarized in Table 1. As discussed, the goal was to create medium and high quality questioned impressions that simulate casework with regard to substrate and medium. In an effort to assess the perceived suitability of the simulated cases as a proxy for real casework, examiners were asked to assess the clarity of the questioned impressions, and the perceived difficulty in the comparison process. Figure 1 reports the frequency of difficulty ratings (easy, moderate, and challenging) as a function of the clarity of the questioned impressions (low, fair, and high). Note that no instructions or definitions were provided to guide the examiner in forming a clarity or difficulty rating.

A global chi-square test indicated that an expert's rating of impression quality impacted their reported case difficulty ($p < 0.05$). A Bonferroni adjusted post hoc analysis revealed that high clarity images were reported less challenging to analyze than expected if independent, and low clarity impressions were reported to be more challenging to analyze than would be expected if the two variables were independent. However, one might assume that high clarity impressions were always rated the least challenging, but this was not always found to be the case. To illustrate this, Fig. 2 provides frequency information (clarity and difficulty), but on a per case basis. Each bar in this histogram sums to $n = 70$ except for cases 003 and 005 (1 examiner switched his or her clarity assessment for 003Q (questioned impression) when reviewing 2 possible knowns (003K1 and 003K2), and 1 examiner reported *insufficient detail* to perform an analysis of 005Q). Note that every single questioned sample was given a range of clarity ratings, and a range of difficulty ratings, and not all clarity ratings are directly correlated with difficulty. For example, one examiner reported that questioned impression 006Q was of high clarity, and only one examiner reported that case 006 was easy. However, this was not the same examiner; the examiner that reported high clarity reported that the case analysis was challenging, and the examiner that reported the case analysis was easy, reported the image clarity as fair. Thus, an individual examiner's assessment of clarity and difficulty are not always related. Another illustration of this can be observed for 005Q, wherein 39 examiners (56%) reported that 005Q was of low clarity, but only 14 (20%) reported that the case analysis was challenging overall. Thus, clarity and difficulty are not necessarily standardized concepts across examiners. Hypothesizing that perhaps experience impacts difficulty, a chi-square test of self-assessed difficulty versus self-assessed experience (binned frequency of cases completed between 0 and 50, 51 to 100, 101 to 150, 151 to 200 and >200) was computed, but likewise failed to detect any convincing trend (global *p*-value of 0.5535), suggesting that additional hypotheses should be further explored.

Evaluation of Class Characteristics

Table 3 reports the value that examiners assigned to the class characteristics of outsole design, physical size, and the size of individual or grouped tread elements when comparing a questioned impression with an exemplar. For each row in Table 3, the same feature (e.g., design) if summed across all values (association, exclusion, not evaluated or insufficient) will equal

TABLE 3—*Evaluation of class features when comparing questioned and test impressions for* n = *70 examiners (note that "design" refers to the geometric pattern, "size" refers to the physical size of the outsole, and "tread size" refers to the size of individual tread elements or groups of tread elements). The same sub-column header across all columns (an example is highlighted in orange) will sum to* n = *70.*

	Association			Exclusion			Not Evaluated			Insufficient		
Case	Design	Size	Tread Size	Design	Size	Tread Size	Design	Size	Tread Size	Design	Size	Tread Size
001K1	65	55	58	5	9	10	0	2	2	0	4	0
001K2	68	63	65	2	2	2	0	3	3	0	2	0
002K1	66	54	61	4	9	5	0	1	1	0	6	3
003K1	68	60	62	2	8	8	0	1	0	0	1	0
003K2	62	27	34	8	41	32	0	2	3	0	0	1
004K1	66	51	55	4	9	12	0	3	2	0	4	1
004K2	69	65	68	1	2	1	0	1	0	0	2	1
005K1	61	27	40	8	34	25	1	1	0	0	8	5
006K1	66	57	67	3	2	3	0	3	0	1	8	0
006K2	63	51	54	7	8	15	0	3	1	0	8	0
007K1	61	21	23	9	47	44	0	2	3	0	0	0
007K2	65	46	46	5	21	21	0	2	3	0	1	0

TABLE 4—*Count (percentage) of examiners providing SWGTREAD (2013) (23) conclusions for mated (M) pairs, nonmated (NM) pairs, and for all data combined, including community agreement (IQR) and its 90% confidence interval as a function of sample size.*

Comparison	Exclusion	Indications	Limited	Association	High Degree	Identification	IQR Count (Median %) (Mean% ± SD%)	IQR % Lower	IQR % Upper
Combined	370	76	87	135	64	100	715 (89.3) (85.6 ± 11.1)	83.5	87.6
Nonmates	350	66	31	35	2	0	436 (91.4) (89.8 ± 6.69)	87.4	92.1
Mates	20	10	56	100	62	100	279 (85.7) (79.7 ± 14.1)	75.9	83.2

TABLE 5—*Count (percentage) of examiners providing SWGTREAD (2013) (23) conclusions for each Q versus K comparison, and whether or not the known is a mated (M) or nonmated (NM) pair (the remaining columns as defined in the caption for Table 4).*

Comparison	Exclusion	Indications	Limited	Association	High Degree	Identification	IQR Count (%)	IQR % Lower	IQR % Upper
001K1 NM	49 (70)	15 (21)	2 (3)	4 (6)	0 (0)	0 (0)	64 (91.4)	83.8	96.2
003K2 NM	66 (94)	2 (3)	0 (0)	2 (3)	0 (0)	0 (0)	66 (94.3)	87.4	98.0
004K1 NM	65 (93)	4 (6)	1 (1)	0 (0)	0 (0)	0 (0)	65 (92.9)	85.6	97.1
005K1 NM	34 (49)	13 (19)	13 (19)	9 (13)	0 (0)	0 (0)	60 (85.7)	77.0	92.0
006K2 NM	33 (47)	18 (26)	12 (17)	7 (10)	0 (0)	0 (0)	63 (90.0)	82.0	95.2
007K1 NM	68 (97)	1 (1)	1 (1)	0 (0)	0 (0)	0 (0)	68 (97.1)	91.3	99.5
007K2B NM	35 (54)	13 (20)	2 (3)	13 (20)	2 (3)	0 (0)	50 (76.9)	66.7	85.2
001K2 M	2 (3)	0 (0)	5 (7)	12 (17)	28 (40)	23 (33)	63 (90.0)	82.0	95.2
002K1 M	6 (9)	8 (11)	25 (36)	30 (43)	0 (0)	0 (0)	55 (78.6)	68.9	86.3
003K1 M	11 (16)	2 (3)	4 (6)	19 (27)	20 (29)	14 (20)	39 (55.7)	45.2	65.9
004K2 M	0 (0)	0 (0)	1 (1)	0 (0)	7 (10)	62 (89)	62 (88.6)	80.3	94.2
006K1 M	1 (1)	0 (0)	21 (30)	39 (56)	7 (10)	1 (1)	60 (85.7)	77.0	92.0

The gold-shaded cells represent the conclusions that span the interquartile range (IQR), and the final three columns of the table report the number (percentage) of participants that reported conclusions within the interquartile range, and the 90% confidence interval for this estimate. Note that comparisons 002Q versus 002K1, 005Q versus 005K1, and 006Q versus 006K1 all had 1 response of "insufficient detail," which is not shown in this table; all other rows will sum to 70 (and result in percentages that sum to 100% barring rounding) except 007K2B which sums to 65 since 5 examiners reviewed a different impression (denoted as 007K2A) that had only limited circulation before being replaced with 007K2B (researchers felt that 007K2A was too easy owning to a patent/prominent RAC that spanned almost a full lug and decommissioned this impression within a month of starting the study).

n = 70 (illustrated as a series of gold-shaded cells for row 001K1). From the summary data, it is clear that the majority of examiners routinely evaluate these class characteristics in terms of association or exclusion. In fact, 99.9% ± 0.41% (mean percentage ± 1 standard deviation) compared overall design (median of 100%), 96.8% ± 2.0% compared the overall physical size (median of 97.1%), and 97.9% ± 1.9% compared individual/grouped tread size between questioned and test impressions (median of 97.9%). However, the only shoes with observable class differences (and therefore value for exclusion) exist for comparisons of appropriate questioned impressions with 003K2, 005K1, 007K1, and 007K2B. Thus, all other selections of "*value for exclusion*" are not fully understood.

In contrast, a limited number of examiners chose not to evaluate physical size of the outsole and/or physical size of tread features. This observation was not anticipated, since by definition, physical size refers to the "dimensions, shapes, spacing and relative positions of the footwear outsole design components" (24). In hindsight, the structure of the reporting interface may have created confusion, but moving forward, additional study may be warranted to determine how examiners define physical size, if the unevaluated observations are a product of varying interpretations in the definition, and under what circumstances these features would not be evaluated during a comparison.

Features Marked

In total, 3524 features of interest were annotated by examiners when comparing 840 questioned-test impressions. Not surprisingly, wear patterns and RACs accounted for the majority of annotations (46% and 36%, respectively and resulting in 82% combined). Table S8 reports the feature type and frequency of marking per case. The first nine items in the table could be selected by the user from a pull-down menu and included features such as stippling, mold defect, die-cut variation, air bubble, foxing strip, etc. The tenth option was "other," which required the examiner to provide input (a label) for the selected feature. After reviewing these inputs, some of the items marked as "other" could be remapped to existing features for the purpose of summarization. For example, an examiner selected "other" and typed "specific wear," but for the purpose of an overall summary, this was remapped to "wear" in Table S8. After this remapping, 210 annotated features marked as "other" persisted. Figure S1 illustrates that nearly a quarter of the 210 that remained were listed as "cannot determine" (suggesting that an examiner noted a difference or similarity between the questioned and known impression, but was unable to label the feature's identity, possibly because they did not have access to the physical outsole for the known). A smaller percentage were grouped as miscellaneous (e.g., "wear turning into RAC," "movement/slippage," "void," and "possible incomplete mixing of outsole material"). Finally, just over half (56%) of the features marked "other" could be categorized as class characteristics (e.g., "spacing of elements," "same physical shape and size," "design element difference," and "smaller element size").

Annotation Maps

Tables 4 and 5 report examiner conclusions for all mated pairs, nonmated pairs, the combined dataset, and each individual

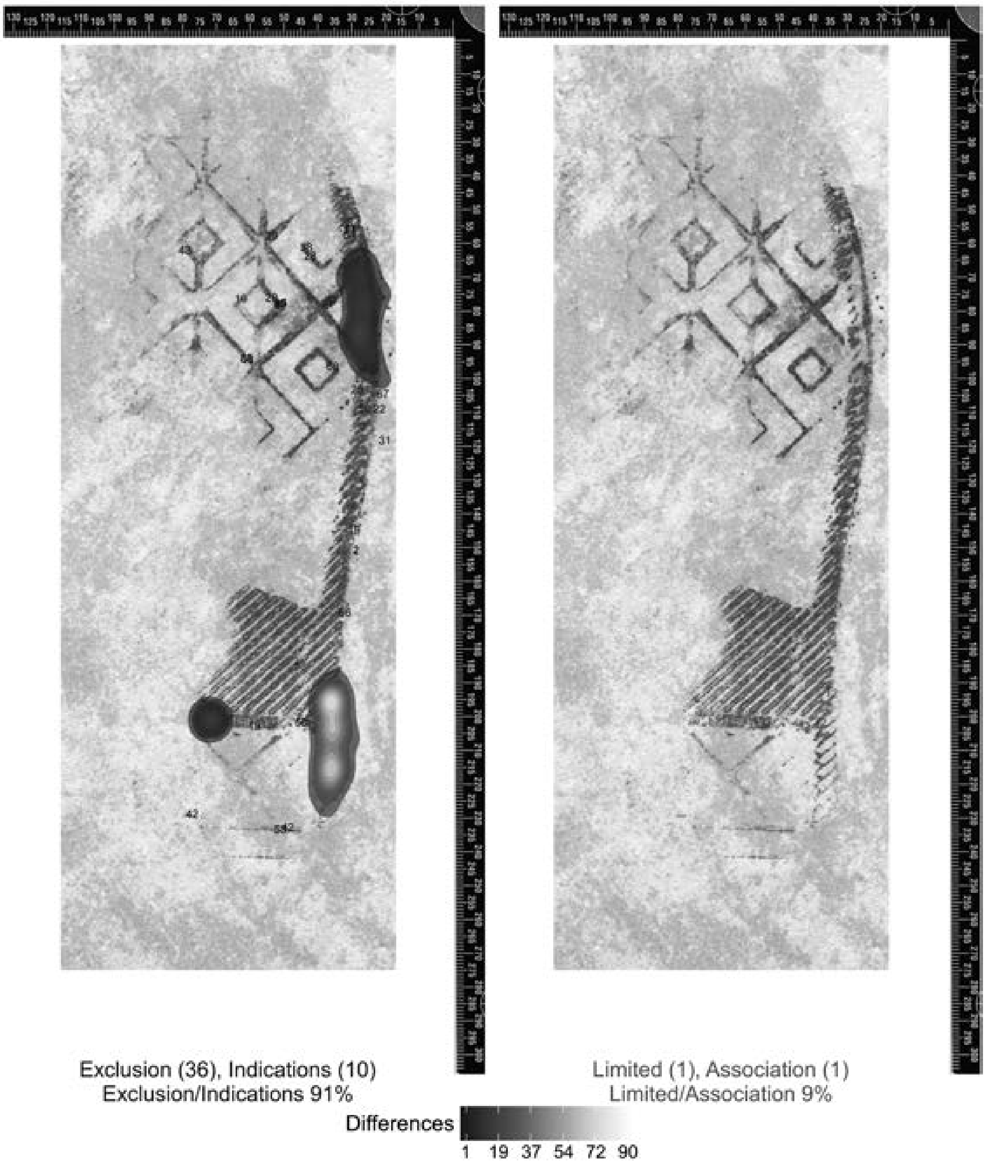

Exclusion (36), Indications (10)
Exclusion/Indications 91%
Limited (1), Association (1)
Limited/Association 9%
Differences
1 19 37 54 72 90

FIG. 3—*Difference maps for the comparison of 001Q versus 001K1 (nonmated pair) illustrating analyst annotations of wear-type features made on the questioned impression only. A total of 91% of examiners reached a conclusion within the IQR, but only 81% justified their conclusions through comments and/or annotations (36 reaching exclusion and 10 reaching indications of non-association as illustrated in the figure). Additionally, 9% provided a decision outside of the IQR (a weak dissociation or weak association). Of these, 2 of the 6 analyst marked 2.0 features each (right), while the remaining 4 outside of the IQR did not mark any wear features. [Color figure can be viewed at wileyonlinelibrary.com]*

comparison (Q versus K). Results are presented as both frequency/count and percentage. The gold-shaded cells in Table 5 correspond to community agreement (which is roughly defined as the interquartile range (IQR). By definition, the interquartile range is intended to capture the middle 50% of the data. However, as applied to the categorical conclusions here, the IQR is intended to approximate the community agreement in conclusions, which means although it cannot capture less than the middle 50% of all conclusions, for comparisons with high agreement, it can extend and capture a higher degree of consensus among responses. This is illustrated in the final three columns of Tables 4 and 5, which report the number (percentage) of participants whose conclusions were within the minimum of the interquartile range. Results indicate that with 90% confidence (based on the Clopper–Pearson Exact method) (25), the community agreed upon IQR includes 75.9–83.2% of all responses for mated pairs, 87.4–92.1% for all nonmated pairs, and 83.5–87.6% for all combined data, with a low of 56% for 003Q versus 003K1, and a high of 97% for 007Q versus 007K1. In an attempt to relate observed results with three previously published studies on footwear reliability, Table S9 was constructed to provide an overview of IQR or community agreement across all comparisons. The mean agreement for the Majamaa and Ytti (19) study was 83.8% ± 12.4% (1 standard deviation). The equivalent value for the Shor and Weisner (20) and Hammer et al. (21) studies was 78.3% and 94.3% ± 7.36%, respectively. The mean agreement for the WVU study is 85.6% ± 11.1%. Based on study design, these values match intuition; the Shor and Weisner (20) study was limited to questioned impressions deemed very challenging and resulting in the lowest IQR. The Majamaa and Ytti (19) study was performed internationally, with limited instructions on how to interpret a prescribed scale, and

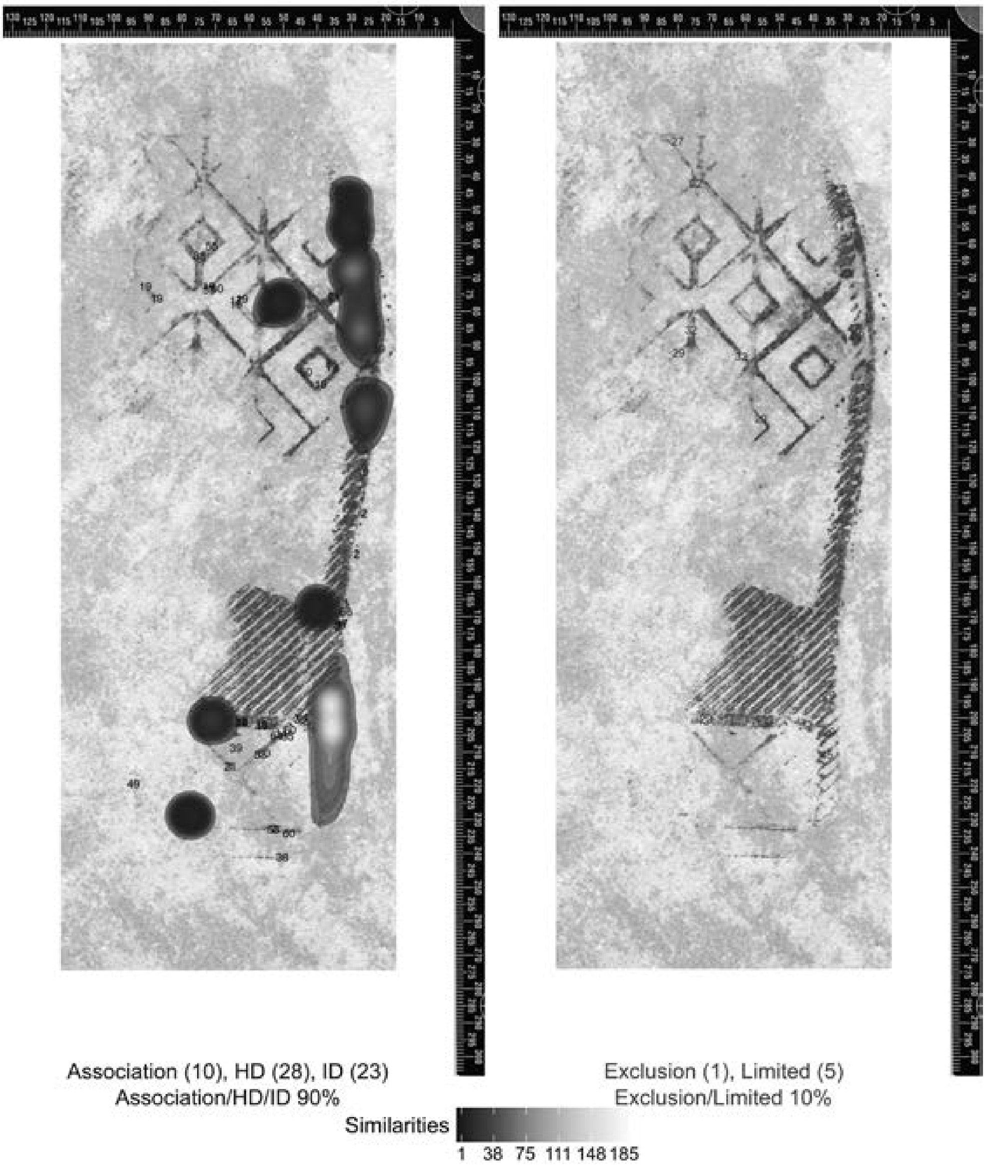


FIG. 4—*Similarity maps for the comparison of 001Q versus 001K2 (mated pair) illustrating analyst annotations of wear-type features made on both the questioned and known impressions. A total 90% of all responses comprised the IQR and 61 of 63 reporting a strong or weak association marked 6.6 ± 3.8 wear features (left); the remaining 2 did not mark any features (and reported an association). Conversely, 10% of examiners reported a disassociation for this known mated pair; 6 out of these 7 analysts marked 3.0 ± 1.5 wear features (right). [Color figure can be viewed at wileyonlinelibrary.com]*

with a fixed impression type (electrostatic dust lifts from paper) leading to a moderate IQR. Conversely, the Hammer et al. (21) study was limited to certified examiners in North America, and excluded class comparisons and the need for feature identification, resulting in the highest IQR.

In contrast to the former studies, the current work includes a larger number of comparisons, using a variety of media, substrates, qualities and totalities, no requirement for certification, but limited to examiners in the United States. On one hand, these variations in study design are unfortunate, making direct comparison and/or a meta-analysis impossible. However, the differences are also fortuitous since it sheds light on the natural range in variation that exists among evidence types (media, substrates, difficulty, clarity, quality, etc.), as well as the differences in examiner demographics, including training, certification status, and variations in workload that can impact confidence and the amount of variety encountered during the duration of an examiner's career.

It is also important to note that although 100% community agreement would be the ideal, the footwear community typically uses 6- and 7-point conclusion scales, and an anticipated consequence of an increasing number of ordinal conclusion categories is increased variability. In addition, many of the results are variations in degrees of confidence (e.g., "*probably made*," "*high degree of association*," and "*identification*" are variations of strong associations, which is very different from one examiner reporting an "*exclusion*" and another reporting "*identification*," where the latter reflects clear disagreement, and the former represents subtle variations in a sliding scale without salient boundaries).

In addition to the summary data provided in Tables 4 and 5, features denoted as wear, RAC, and/or Schallamach patterns (comprising more than 86% of the items marked by participants) were

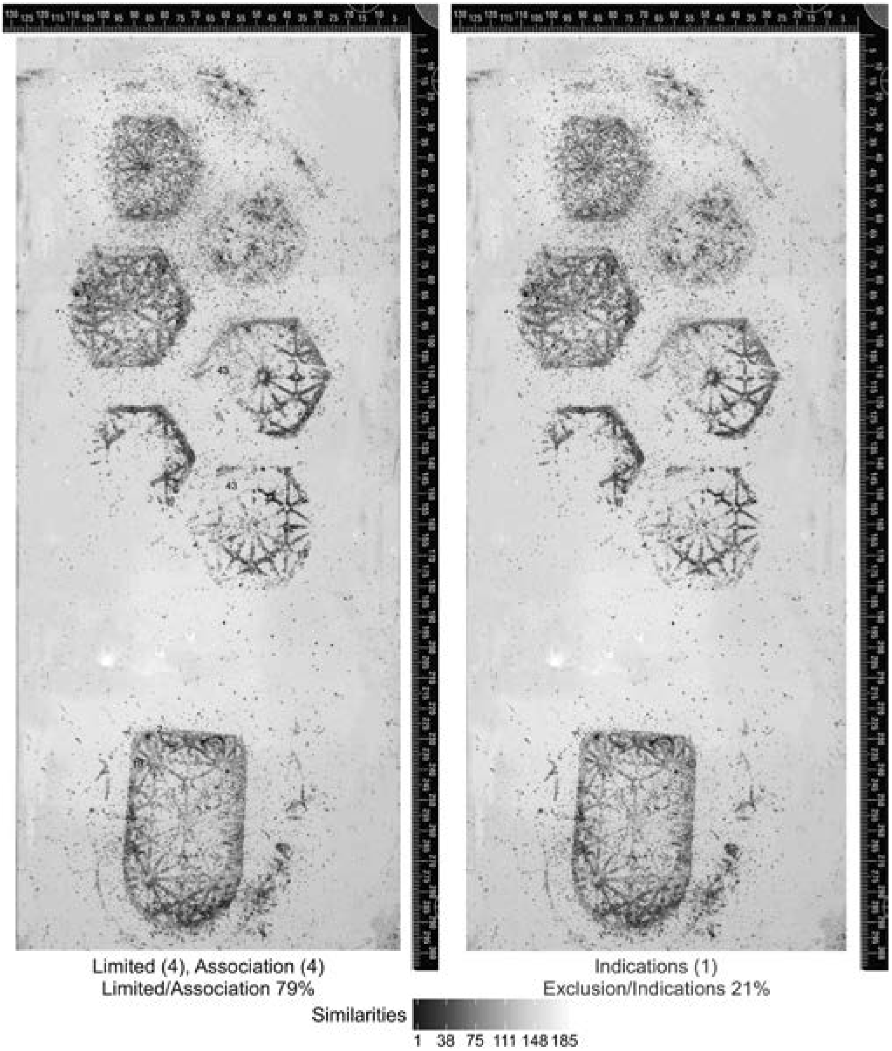


FIG. 5—*Similarity maps for the comparison of 002Q (contrast-reversed) versus 002K1 (mated pair) illustrating analyst annotations of wear-type features made on both the questioned and known impressions. A total of 79% of examiners reported a decision within the IQR, but only 8 of these 55 unique examiners marked 1.6 ± 0.5 features each (left); the remaining 47 examiners did not mark any wear features. Additionally, 1 out of the 15 analysts who reported a decision outside of the IQR categories marked 3.0 features (right), while the remaining 14 did not mark any wear features. [Color figure can be viewed at wileyonlinelibrary.com]*

grouped and converted into 2 annotation maps per case (Fig. 3–14). Except for overlays (Fig. 7, 10 and 13), which were created for comparisons exhibiting an observable size difference, the annotation map on the left in Fig. 3–14 reveals the frequency of features marked by examiners that reported a final SWGTREAD (2013) (23) conclusion that fell within the interquartile range. Conversely, the map on the right reveals the frequency of features marked by examiners that reported a final SWGTREAD (2013) (23) conclusion outside the community agreed upon range. Thus, inspection of each annotation map describes the frequency of features marked by different examiners, wherein each examiner was randomly assigned a number between 1 and 70. For any location on the outsole with at least 7 marked features (10% of the 70 respondents), instead of reproducing all examiner numbers associated with the feature, an actual frequency map was plotted. Thus, the maps reveal features marked by several examiners with the total number of marks revealed by the frequency color code, while individual numbers reveal features marked by a limited or fewer number of examiners.

The purpose of each map is to allow for inspection of the features deemed relevant to comparison among both groups of examiners (those reporting a conclusion in agreement with the community, and those reporting a conclusion that is inconsistent with the community IQR). In some cases, examiners reaching non-IQR conclusions marked the same features as those reaching IQR conclusions, but since these examiners arrived at non-IQR conclusions, the implication is that the total number and/or weight attributed to marked features differs.

Figure 3 is a difference map (red-yellow) for 001Q versus 001K1 (nonmated pair). In the left-most image, 46 of 64 unique examiners who reached a conclusion within IQR) marked 3.5 ± 2.1 features each. Wear features on the lateral edge of the shoe, in both the toe and heel, showed significant flocking differences, which are clearly highlighted by the red/yellow frequency maps. Interestingly, the remaining 18 examiners within IQR did not mark any wear features, but 11 out of 18 noted an overall difference in wear between the questioned and test impression in their open-ended responses (while 6 of 18 offered

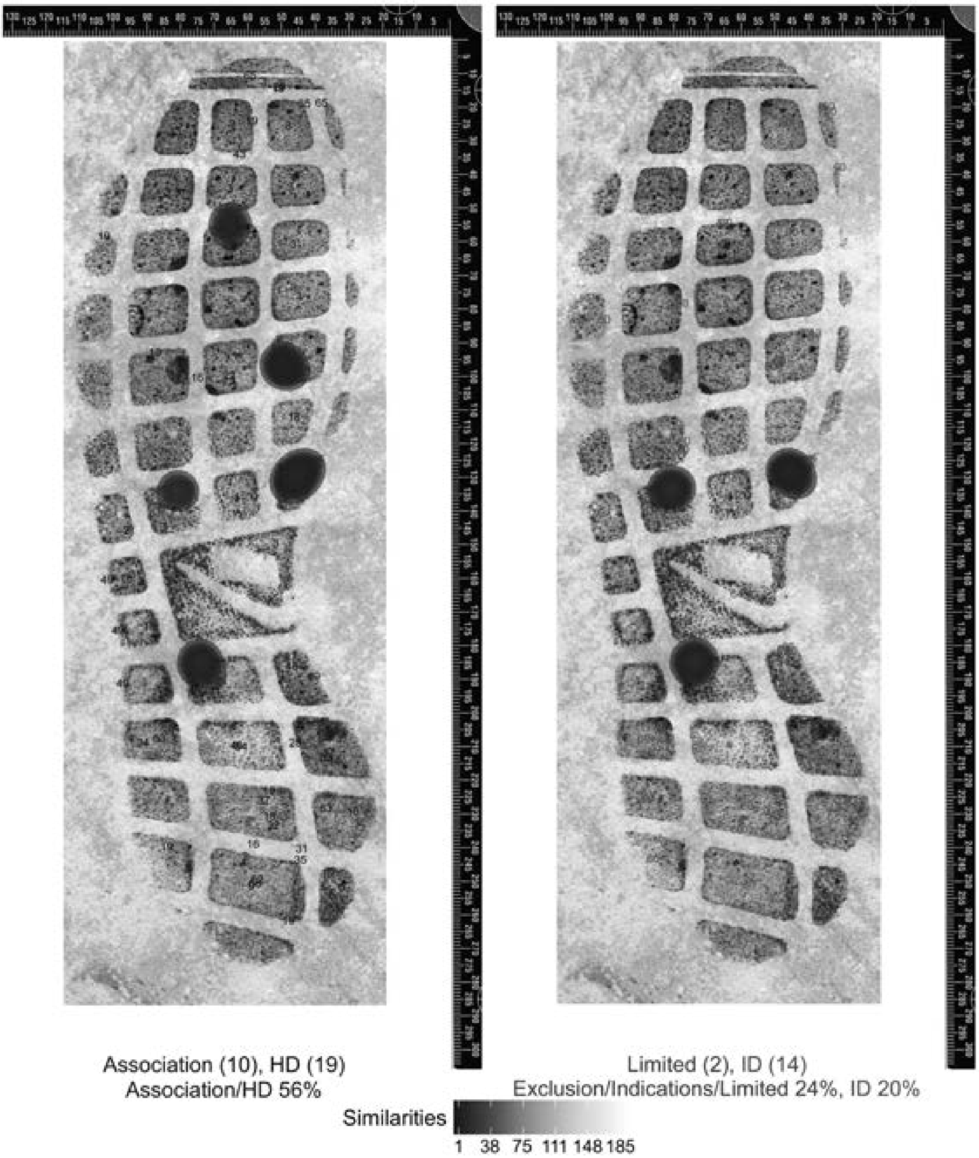
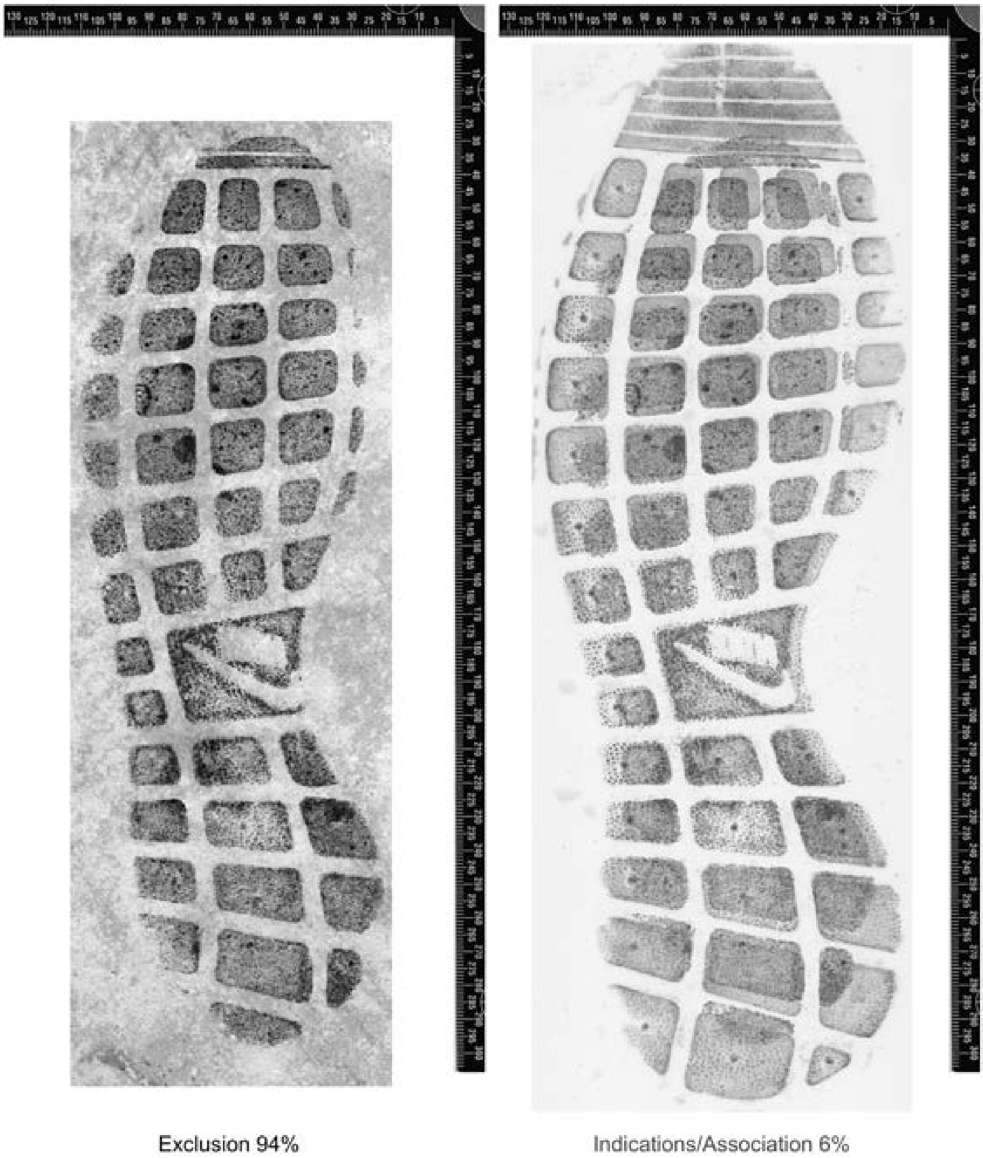


FIG. 6—*Similarity maps for the comparison of 003Q versus 003K1 (mated pair) illustrating analyst annotations of wear-type features made on both the questioned and known impressions. A total of 56% of all responses comprise the IQR, and 29 of the 39 unique examiners who reached a conclusion within IQR marked 4.2 ± 3.8 features (left), while the remaining 10 examiners did not mark any wear features (9 of these 10 reported association, but 1 reached high degree of association and reported that wear patterns agreed in various locations throughout the outsole). Additionally, 16 out of the 31 total analysts who reported a decision outside of the IQR categories marked 4.9 ± 4.8 features (right), while the remaining 15 did not mark any wear features. [Color figure can be viewed at wileyonlinelibrary.com]*

no commentary on what informed their decisions and 1 of 18 excluded based on a size difference [which is not apparent, but he or she reported that the edges of the known extended past the edges of the questioned impression, perhaps giving undue merit to the partiality of the impression]). The right-most image in Fig. 3 depicts the wear features marked by the 6 examiners that reported conclusions outside the IQR. Of these 6, 2 examiners marked 2 features each, and the remaining 4 did not mark any wear features when reporting their results. One concern here was inexperience in evaluating flocking material, but of the 6 examiners that reported outside of the IQR, none indicated an issue or unfamiliarity with flocking in their open-ended responses regarding limitations for this comparison.

The overall summary is that although a total of 91% of the respondents reached a conclusion within the IQR and matching ground truth, only 81% (57/70) justified their conclusions using comments and annotation, and 1 examiner came to a valid IQR conclusion, but for the wrong reason (erroneously reported a size difference). Conversely, 4 examiners outside IQR did not note or elected not to mark any differences, and 2 examiners did not place sufficient weight on the observed differences that exist between 001Q and 001K1.

Figure 4 is a similarity (blue-green) map for 001Q versus 001K2 (mated pair). On average, 61 of 63 unique examiners who reached a conclusion within IQR marked 6.6 ± 3.8 features each (left). More specifically, examiners that concluded *association of class* or *high degree of association* marked 5.5 ± 3.4 features, whereas analysts who reported *identification* marked 8.0 ± 4.2 features. Additionally, 6 out of the 7 analysts who reported a decision outside of IQR marked 3.0 ± 1.5 features each (1 examiner did not mark any wear features). The overall summary is that 90% of respondents reached a conclusion within


Exclusion 94%

Indications/Association 6%

FIG. 7—*Case 003 questioned impression (left) overlaid on the test impression from nonmated pair 003K2 (right) illustrating the measurable difference in size. [Color figure can be viewed at wileyonlinelibrary.com]*

the IQR and matching ground truth, and the vast majority justified their conclusions with annotations (wherein stronger conclusions were anecdotally associated with more marked features). Seventy percent of the remaining 10% of respondents seemed to mark similar areas in agreement between the questioned and known test impression, but did not attribute sufficient weight in order to reach a strong enough association. Conversely, 30% of the remaining 10% of examiners reported an *exclusion*—which is clearly erroneous (either typographical or actual false negatives—note that the research team reviewed all responses in an attempt to identify apparent typographical errors, but acknowledge that such errors could exist and yet be undetectable by the research team).

Figure 5 is a similarity (blue-green) map for a contrast-reversed image of 002Q versus 002K1 (mated pair). On average, 8 of 55 unique examiners who reached a conclusion within IQR marked 1.6 ± 0.5 features each (left). Additionally, 1 out of the 15 analysts who reported a decision outside of the IQR categories marked 3.0 features (right) (the remaining 14 examiners outside the IQR did not mark any wear features). It is important to note that 9 of the 15 outside of IQR noted disagreement in wear and/or RACs between the questioned and known impressions, which may indicate that some examiners interpreted voids as features of interest, rather than appropriately regarding them as a matrix interference (although 4 out of 15 noted or selected substrate texture and/or matrix effects as a limitation). Of the 15 examiners reporting a false disassociation for this known mate, 6 reported the strongest disassociation possible (*exclusion*); upon inspection of their open-ended responses, 5 of these 6 examiners reported a size difference, and 3 of 6 reported a wear difference (2 of these 3 also reported a size difference). Thus, it is hypothesized that these examiners may have used unreliable measurements or tread elements when attempting to align overlays and/or compare measurements (note that this is a dust impression with incomplete tread reproduction, which should not be misidentified as wear or size differences).

Figure 6 is a similarity (blue-green) map for 003Q versus 003K1 (mated pair). On average, 29 of 39 unique examiners

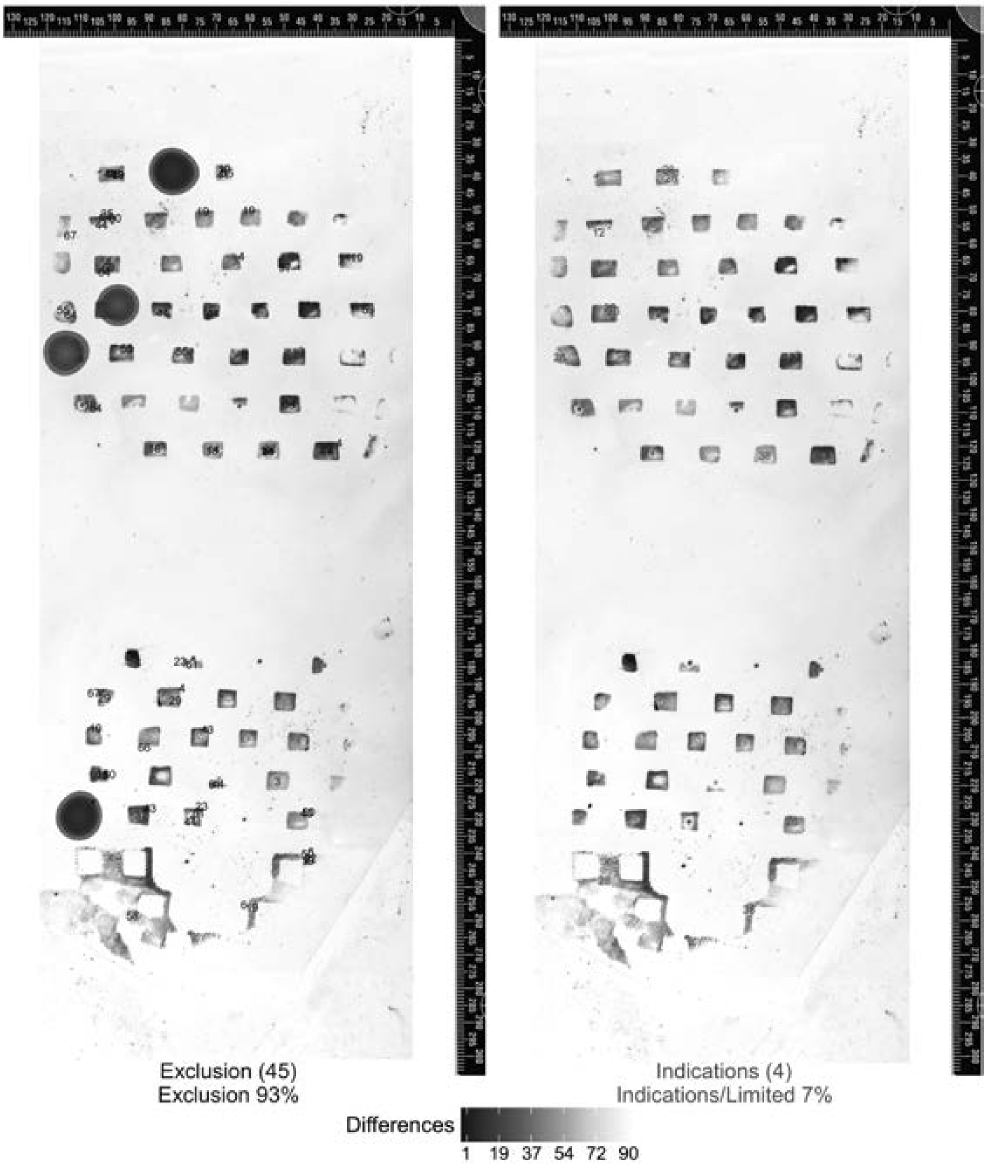


FIG. 8—*Difference maps for the comparison of 004Q versus 004K1 (nonmated pair) illustrating analyst annotations of wear-type features made on the questioned impression only. Forty-five of 65 unique examiners who reached a conclusion within IQR (exclusion) marked 3.7 ± 2.0 features each (left). Additionally, 4 out of the 5 analysts who reported a decision outside of the IQR category marked 2.8 ± 1.3 features (right). [Color figure can be viewed at wileyonlinelibrary.com]*

who reached a conclusion within IQR marked 4.2 ± 3.8 features. However, 10 examiners did not mark any wear features. A total of 31 examiners (44%) reported conclusions outside the IQR and the results were extremely variable (24% leaning toward the exclusionary side of the conclusion standard, and 20% reporting *identification*). Interestingly, those that reported a value outside the IQR and marked features (4.9 ± 4.8 features illustrated in the right-most map of Fig. 6) were largely the 14 examiners (20%) that came to a conclusion of *identification* and the remaining 15 examiners that did not mark any features and leaned toward weak and/or strong exclusions seemed to perceive a size and/or wear difference (11 of 15), with the former perhaps being a function of the partiality of the questioned impression. Thus, examiners in this latter group may have been using unreliable perimeter details or substrate interference when reaching an erroneous conclusion (4 of the 15 did not offer any additional comments regarding their reasoning).

In summary, this comparison had the greatest variability in responses. Since this is a mated pair, the respondents that reported *identification* (20%) are indeed correct as a function of ground truth, although outside of community agreement. Inspection of the annotation maps reveals that most respondents identified and compared many of the same features within the questioned and known test impression (especially in the upper and lower instep), suggesting that differences in conclusions are more likely a function of the weight applied to features. Thus, it is the 16% of all participants that reported a false negative (*exclusion*) that are of greater interest, and results suggest an unreliable size and/or wear difference may have led to erroneous conclusion. In addition, it may be relevant to highlight that this was the only case involving microcellular material,

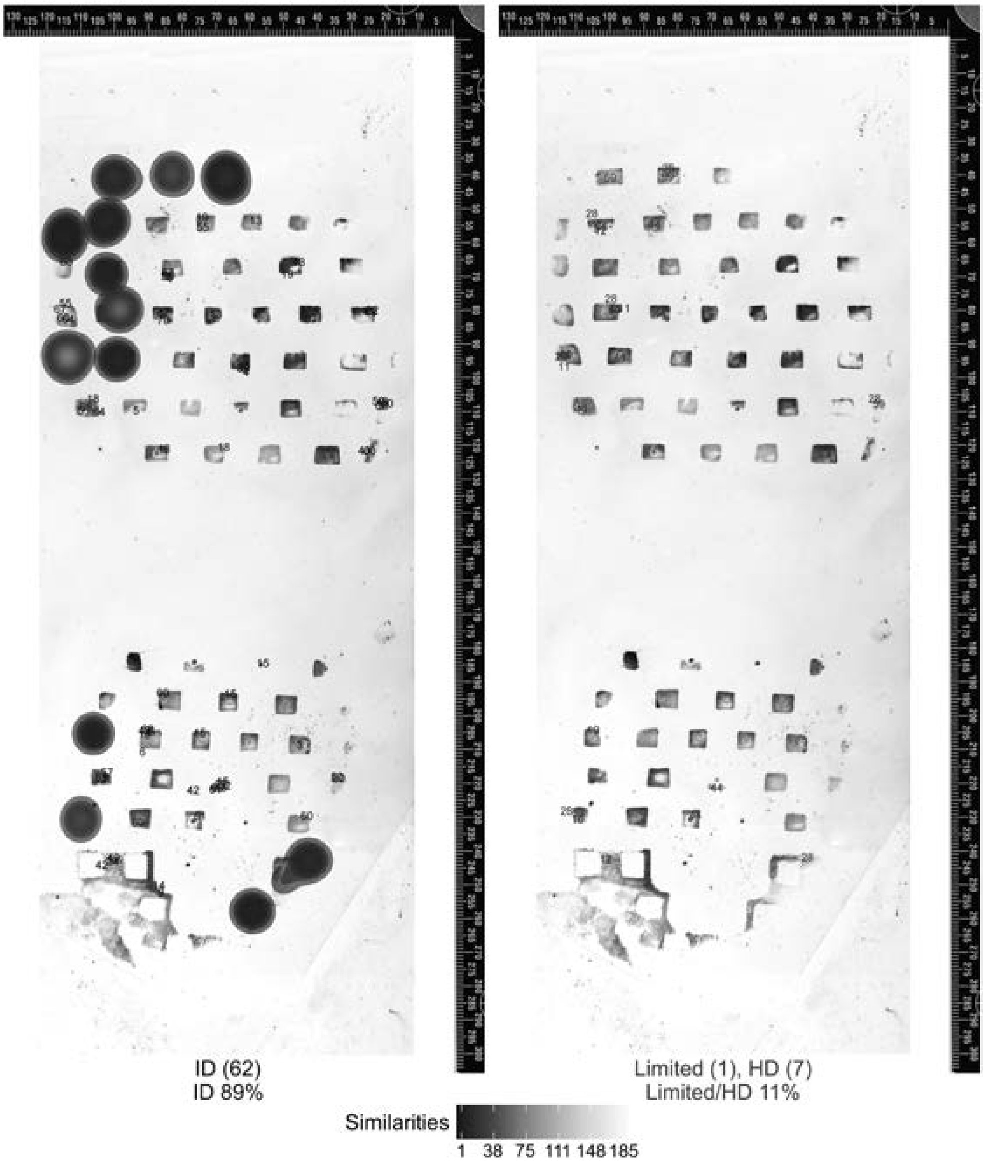


FIG. 9—*Similarity maps for the comparison of 004Q versus 004K2 (mated pair) illustrating analyst annotations of wear-type features made on both the questioned and known impressions. All 62 unique examiners who reached a conclusion within IQR (identification) marked 6.4 ± 3.6 features each (left). Additionally, all 8 analysts who reported a decision outside of the IQR category (7 reporting high degree of association and 1 reporting limited association) marked 4.3 ± 2.4 features (right). [Color figure can be viewed at wileyonlinelibrary.com]*

which may suggest that additional studies are needed in order to fully characterize if this material presents a unique challenge for examiners, or if it is simply coincidence that this outsole created more variability in responses than some of the other case comparisons.

Figure 7 is an overlay of questioned impression 003Q and known test impression 003K2, which is a nonmated pair with a noticeable size difference. Given the nature of the impression (blood on a hard tile substrate), the media and substrate cannot be used as a means to explain the observed size difference, meaning a small percentage of examiners erroneously failed to observe the size difference, or if observed, erroneously believed it was an explainable variation in reproduction, and/or did not complete their analysis in order to reach a valid conclusion (note that the manufacturer reports that both 003K1 and 003K2 are the same size, but 003K2 has a measurable size difference despite the manufacturer's report).

Figure 8 is a difference (red-yellow) map for 004Q versus 004K1 (nonmated pair) and Fig. 9 is a similarity (blue-green) map for 004Q versus 004K2 (mated pair). For each pair, one examiner (but not the same examiner) reported an erroneous *limited association*. In both instances, the examiner reported agreement in size, design, and limited agreement of wear, but did not mark any RACs or Schallamach patterns. Finally, of the 7 examiners that reported *high degree of association* (rather than *identification*) for the mated pair, 3 indicated that they could not confirm their observations without having the actual shoe that created the test impression, and 1 indicated that the number of features observed was insufficient to reach *identification*. Thus, conclusions for both comparisons in this case exhibited a high level of agreement, which is matched by the annotation maps, and any variation in the actual ordinal category selected by different examiners was clearly articulated.

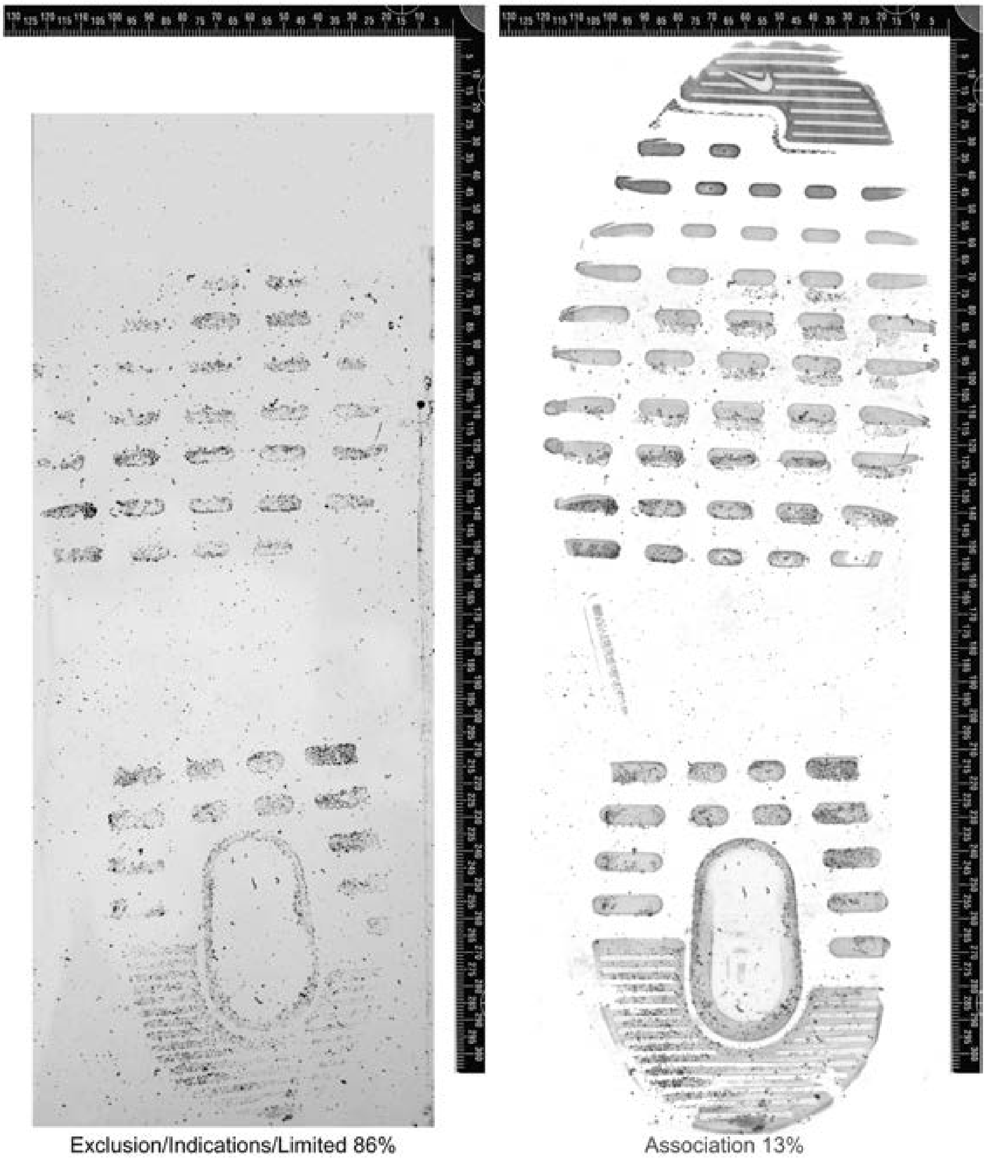


FIG. 10—*Case 005 questioned impression (left) overlaid on the test impression from mated pair 005K1 (right) illustrating the measurable difference in size. [Color figure can be viewed at wileyonlinelibrary.com]*

Figure 10 is an overlay of questioned impression 005Q and known test impression 005K1, which is a nonmated pair with a noticeable size difference. Ten of the 13 examiners reaching *indications of non-association* reported a potential size difference, but in general, also reported that they were unable to confirm this observation. Additionally, 4 of the 13 examiners that reported *limited association* also noted a size discrepancy, but were unable to reach the strongest exclusion. Since this case includes a gel lift, analysts' ability to verify and weigh sizing information with gelatin lifters was further investigated as a possible confounding factor. Cases 002, 004, and 005 each included gelatin lifters, resulting in a total of 280 gel lift evaluations (4 knowns × 70 examiners). Of the 280 reports, only 6 examiners noted that the gel lift may lead to potential sizing or distortion issues, and only 3 of the 280 directly noted a perceived distortion. Therefore, the general conclusion is that examiners were not overly concerned with the possibility of any size or distortion issues when presented with gel lifts in this study, and thus examiners do not seem to attribute less weight to perceived size differences when examining gel lifts.

Figure 11 is a similarity (blue-green) map for 006Q versus 006K1 (mated pair) and Fig. 12 is a difference (red-yellow) map for 006Q versus 006K2 (nonmated pair). For the mated pair, 8 comparisons resulted in stronger associations than expected (7 as *high degree of association* and 1 *identification*). However, none of these 8 examiners marked RACs, which is typically expected in order to justify such a strong association. One of the 8 reported "specific wear," which does adhere with the SWGTREAD (2013) standard for conclusions (23), and 7 marked at least one wear feature with value for *high degree of association*, suggesting that these examiners afforded substantial weight to wear similarities, resulting in very strong associations.

When comparing 006Q against a nonmated pair, the IQR is comprised of 90% of all conclusions. Since differences in wear are evident, examiners were expected to report weak to strong disassociation. Twelve examiners reached a conclusion of *limited*

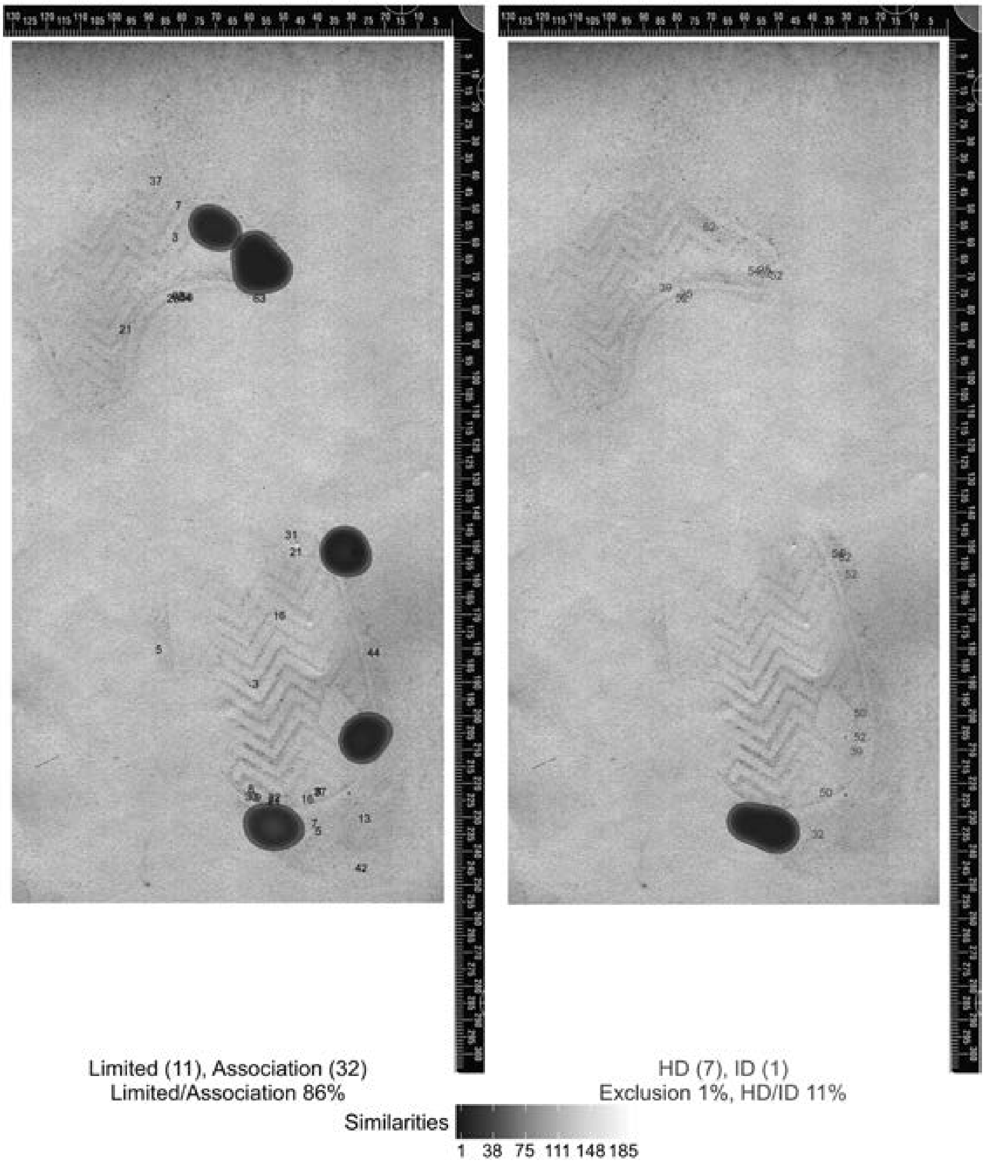


FIG. 11—*Similarity maps for the comparison of 006Q versus 006K1 (mated pair) illustrating analyst annotations of wear-type features made on both the questioned and known impressions. Forty-three of 60 unique examiners who reached a conclusion within IQR (limited association or association) marked 2.6 ± 1.4 features (left). Additionally, 8 out of the 10 total analysts who reported a decision outside of the IQR categories marked 3.1 ± 0.8 features (right). [Color figure can be viewed at wileyonlinelibrary.com]*

association of class, and 9 of these 12 reported that their evaluation was limited to class due to detail/quality limitations. However, the comparison does show differences in wear, which should push examiners toward *indications of non-association*, perhaps suggesting an incomplete comparison was undertaken (although the IQR includes decisions of *exclusions*, *indications of non-association* and *limited association*). Interestingly, 6 of the 12 examiners that marked features, marked them in disagreement, but again, did not reach a stronger exclusionary decision.

Figure 13 is an overlay of questioned impression 007Q and test impression 007K1, which is a nonmated pair with a noticeable size difference. Ninety-seven percent of examiners reported *exclusion*, presumably because the size difference cannot be explained by substrate or matrix effects given that this is a blood impression on tile. Conversely, the remaining examiners reported either *indications of non-association* or *limited association* (these examiners reported that size did not agree in their open-ended responses, but did not use this information to reach a stronger disassociation).

Figure 14 is a difference (red-yellow) map for 007Q versus 007K2B (also a nonmated pair). A total of 13 examiners (20%) reported an *association of class* when their peers reported stronger disassociations. All of these examiners noted agreement in design, physical size, and the size of tread elements, despite the manufacturing anomaly in the heel (although 1 examiner did note an unexplained variance or shift between impressions). Moreover, 2 of the 13 marked a wear pattern in disagreement, and 5 marked at least 2 RACs in disagreement, but none attributed sufficient weight to these differences in order to reach a disassociation.

Table 6 is a summary of the annotation map data, comparing the percentage of conclusions within the IQR, versus the percentage of examiners with responses within the IQR that marked features to justify their conclusions (the annotation map

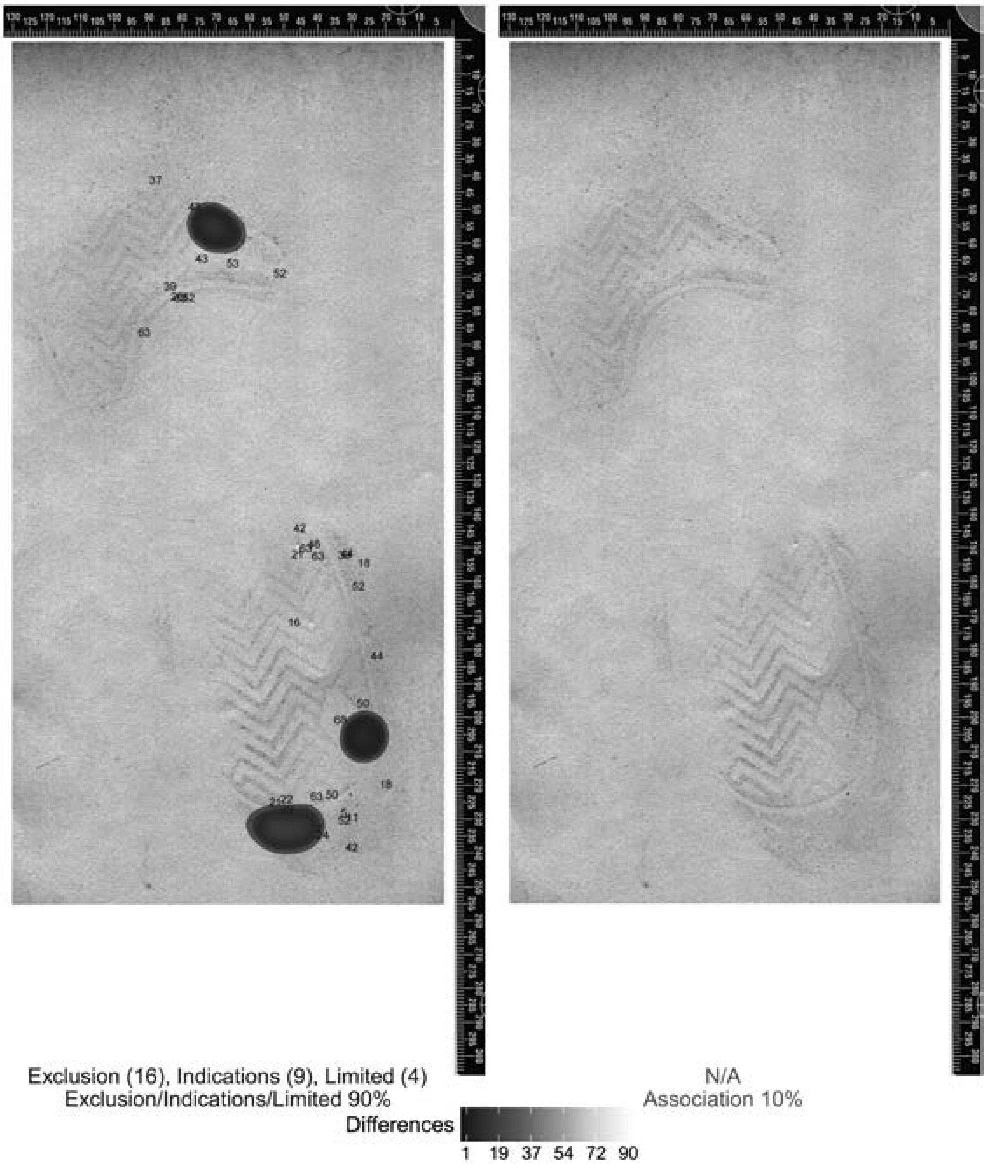

Exclusion (16), Indications (9), Limited (4)
Exclusion/Indications/Limited 90%
N/A
Association 10%
Differences
1 19 37 54 72 90

FIG. 12—*Difference maps for the comparison of 006Q versus 006K2 (nonmated pair) illustrating analyst annotations of wear-type features made on the questioned impression only. Twenty-nine of 63 unique examiners who reached a conclusion within IQR (exclusion, limited association, or association) marked 2.5 ± 1.5 features (left). Conversely, none of the 7 analysts who reported a decision outside of the IQR categories marked any features (right). [Color figure can be viewed at wileyonlinelibrary.com]*

itself illustrates which features were actually marked). If all examiners within the IQR (similarity in conclusions) also marked features, then the column labeled “% in IQR & Marked” would always be 100%. However, inspection of the summary data indicates considerable variability with an average of 66.5%. This roughly indicates that agreement in conclusion is only matched with agreement in electing to mark one or more features to explain/justify conclusions at a rate of about 67%. However, of those that do agree in conclusion and elect to mark features, a large percentage mark the same or similar features (as evidenced by inspection of the annotation maps in Fig. 3–14).

Notation of Limitations

Examiners were prompted to report if they experienced any of the following limitations during their comparisons: substrate texture, photographic distortion, improper lighting, and/or improper scale position. Table S10 details the frequency of responses to this question. In total, just over one-fifth of comparisons noted a limiting factor (187/840 = 22%) with the majority of these reporting a substrate texture limitation (approximately 80%). Note that “improper scale” was not selected as an issue affecting any of the 840 analyses, and is therefore not included in Table S10.

Participants were also permitted to provide comments in open-ended responses about their comparisons and any issues experienced during analyses. In these open-ended responses, several additional limitations were discussed by examiners, including preferences for overlays or transparencies (3 examiners), access to digital copies (7 examiners), additional digital processing of images, and/or additional photographs at different lighting angles (8 examiners). With regard to these requests, (i.) digital copies of all images were provided to examiners (at 1200PPI), and (ii.)

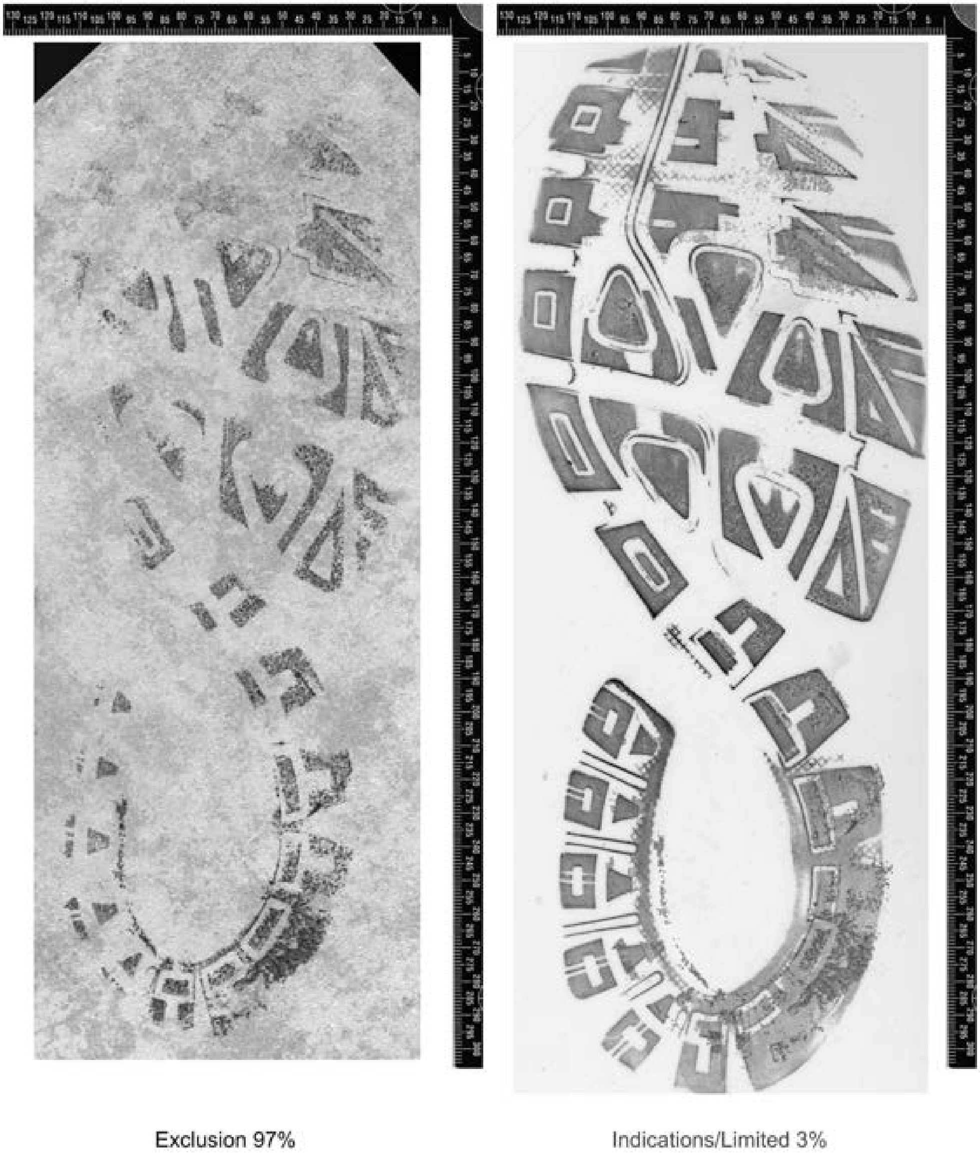


FIG. 13—*Case 007 questioned impression (left) overlaid on the test impression from nonmated pair 007K1 (right) illustrating the measurable difference in size. [Color figure can be viewed at wileyonlinelibrary.com]*

in hindsight, the research team wishes printed transparencies were included upfront and would recommend this to anyone preparing studies of this nature in the future.

Finally, the open-ended examiner responses also informed us that 3 examiners (a total of 7 times) reached an *exclusion* for a specific known because they concluded *identification* for the other known shoe provided in the case. Clearly, this was not an anticipated response; it purports that two shoes with the same class characteristics could never have similar types of wear and/or RACs, which has not been proven. Moreover, it does not fit with adherence to the SWGTREAD (2013) standard (23) (an *association of class* may very well be a valid conclusion for the second shoe).

Examiner-Specific Impact on Results

In order to determine the degree to which a single or specific examiners impacted reliability results, Fig. 15 reports the frequency (percent) of examiners with 0, 1, 2, etc. responses outside of the IQR range (out of a total of 12 responses across 7 cases). Inspection indicates that 19% of all respondents were always within the IQR, while 33% were outside for a single conclusion. In contrast, and of more concern, are the examiners that are consistently outside of the IQR (e.g., the 6 examiners with 4 or 5 conclusions outside of the IQR). Possible explanations for this discrepancy are numerous, but may include disparities in training, examiner inexperience, and/or a persistent variation in interpretation of the SWGREAD 2013 conclusion standard (23). Regardless of the origin, these variations should be addressed in order to allow these analysts to self-calibrate against community norms. In addition, if the 12 comparisons in this study are considered representative of typical casework, then the 18 analysts with 3 conclusions outside of the IQR are also expected to form conclusions and opinions that are consistently different from the majority of their peers approximately 25% of the time.

Conclusions

The purpose of this summary was to evaluate and report community demographics and consistency in both feature

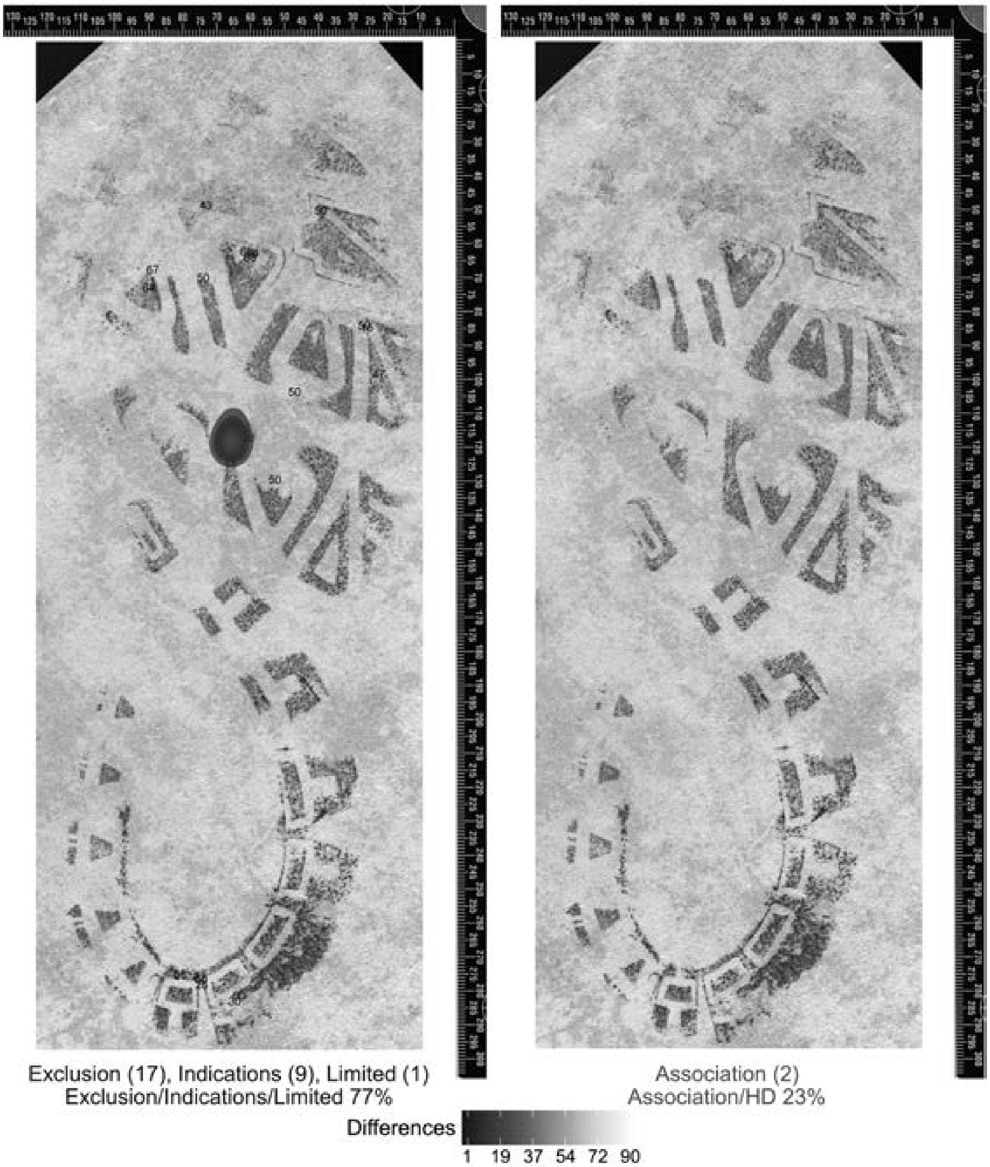

Exclusion (17), Indications (9), Limited (1)
Exclusion/Indications/Limited 77%
Association (2)
Association/HD 23%
Differences
1 19 37 54 72 90

FIG. 14—*Difference maps for the comparison of 007Q versus 007K2B (nonmated pair) illustrating analyst annotations of wear-type features made on the questioned impression only. Twenty-seven of 50 unique examiners who reached a conclusion within IQR (exclusion, limited association, or association) marked 1.6 ± 1.3 features (left). Additionally, 2 out of the 15 total analysts who reported a decision outside of the IQR categories marked 1.0 feature (right). [Color figure can be viewed at wileyonlinelibrary.com]*

identification and conclusion for a specific reliability study conducted in the United States between 2017 and 2018, based on 835 individual reports, collected from 70 examiners, performing 12 comparisons, across 7 simulated cases (840 less 5 examiners that compared 007K2A before it was retired). In response, four major conclusions are supported. First, the issue of case-assessed clarity versus difficulty should be further investigated. Second, the majority of examiners evaluate and report on consistencies and inconsistencies in class characteristics during their analyses (on average, between 99.9% and 96.8%). However, this does not mean they afforded the same weight to observed consistencies and inconsistencies in final weight of evidence conclusions.

Third, for those examiners that agree in conclusion, only moderate agreement (66.5%) was found in whether or not this cohort elected to mark wear-acquired features to justify their conclusions. At this point, no consideration of feature weight has been evaluated, and only summary statistics (66.5%) are provided for feature agreement in marking, but this variation warrants additional study and review in order to more fully understand how examiners reach conclusions. Moreover, this statistic must be evaluated within the confines of the instructions given to examiners that participated in this study. More specifically, examiners were asked to perform the comparison as if it were casework, without any other directives. Thus, the results should reflect typical variation in feature annotation between analysts, which is likely to be higher than if the examiners were asked to exhaustively identify and mark all relevant features for comparison.

Finally, this study found reasonable consistency in reporting (an average of 85.6%) which is in alignment with former studies of this nature (78.3% (20), 83.8% (19), and 94.3% (21)) when considering differences in project/study design.

With regard to limitations, several additional observations are also offered. First, examiners were not given the actual shoe and permitted to collect their own impressions, in this or any of the

TABLE 6—*Examiners that were consistent in conclusion (within the IQR) and the degree of their consistency in electing to mark features.*

Comparison	% Conclusions in IQR	Total # Features Marked	% in IQR & Marked	(D) ifferent/(S) imilar
001Q-001K1	91.4	304	71.9	D
001Q-001K2	90.0	447	96.8	S
002Q-002K1	78.6	83	14.5	S
003Q-003K1	55.7	253	74.4	S
003Q-003K2	94.3	136	–	–
004Q-004K1	93.0	358	69.2	D
004Q-004K2	89.0	440	100	S
005Q-005K1	85.7	72	–	–
006Q-006K1	85.7	176	71.7	S
006Q-006K2	90.0	188	46.0	D
007Q-007K1	97.1	254	–	–
007Q-007K2B	76.9	310	54.0	D

Note that 003Q versus 003K2, 005Q versus 005K1, and 007Q versus 007K1 each have a physical size difference, and the column labeled "Total # of Features Marked" refers to wear features, RACs and/or Schallamach patterns from Table S8.

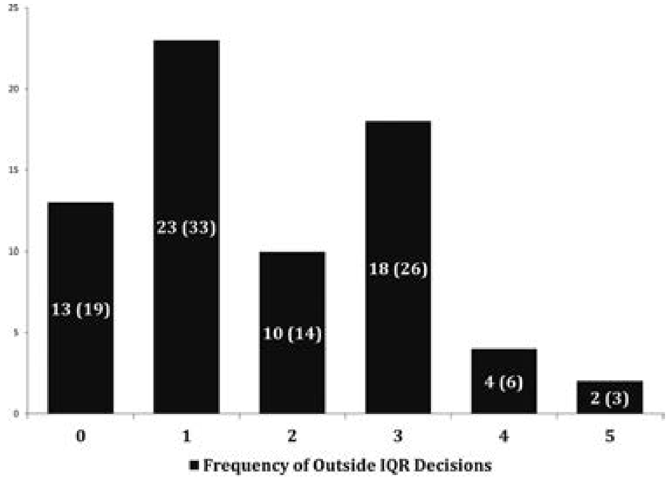


FIG. 15—*Frequency (percent) of examiners with 0, 1, 2, etc. responses outside of the IQR range. [Color figure can be viewed at wileyonlinelibrary.com]*

previously discussed reliability studies, which is typically a requirement of operational casework. Although two replicate test impressions were provided per known shoe, the ability to collect additional impressions, using additional methods, may have impacted results (all exemplars were created statically rather than dynamically). Second, examiners were not given transparencies for overlays. Although there was nothing stopping an examiner from creating his or her own transparencies, in hindsight, supplying these upfront would have been a worthwhile expense because it may have made it less time-intensive to participate and helped to increase overall quality control. Third, only 12 comparisons were requested, across 7 simulated cases. Moreover, all questioned and test impressions were made using men's athletic shoes, of sizes 9 through 11, which may indicate that the results found here are not transferable to other shoe types. Fourth, of the simulated cases, the study included 3 blood impressions, 3 gelatin lifts, and 1 paper impression. This means that the summary statistics span a wide number of media and substrates, but since each media and substrate is of limited sample size individually, the results cannot be subdivided to look for consistencies and inconsistencies that may be media/substrate specific. In addition, 5 of the 7 cases included 2 knowns, and 2 included a single known, and the cases with a single known for comparison included a questioned impression collected in dust. In addition, participants were instructed to treat the cases as if "casework." This would imply no guarantee that a true mate was present, but in instances with 2 knowns, if 1 test impression was deemed an *identification*, results indicated that *at least* 3 examiners (a total of 7 times) "gamed" the system by concluding the second test impression was an *exclusion*. Fifth, the overall quality of the simulated comparisons is at least moderate or high, suggesting that the results may not reflect consistency when examiners are confronted with lower quality impressions. Sixth, a variable not accounted for in this study was any mismatch between the size of the footwear and the actual foot size of the person preparing the questioned impressions. Unfortunately, research records do not allow for a detailed analysis of the degree and type of variability this factor may have introduced into the presented results, but it is recommended that this potential shortcoming be either negated or tracked in any future studies. Seventh, analysts were asked to perform the analysis of the simulated cases as they would for typical casework, but a customized reporting interface was provided in order to collect examiner opinions and conclusions. Although the research group performed consistency checks to search for typographical errors or inconsistencies in using the software as intended, no guarantee can be made that some variation is not a function of the customized reporting interface. Finally, no verification step was permitted. In other words, each examiner was asked to submit his or her own conclusions, without consultation or independent verification. This latter restriction suggests that the results presented here are extreme for the types of impressions provided since verification of casework, as required by many laboratories, is likely to increase examiner consensus in reporting.

Acknowledgments

The authors would like to express our sincerest gratitude to the forensic footwear examiners who generously shared their time and expertise in order to participate in this West Virginia University (WVU) Institutional Review Board (IRB) approved (1602021821) study. Thank you for making this work possible.

References

1. National Research Council. Strengthening forensic science in the United States: a path forward. Washington, DC: National Academies Press, 2009.
2. President's Council of Advisors on Science and Technology (PCAST). Report to the President – Forensic science in criminal courts: ensuring scientific validity of feature-comparison methods. Forensics Addendum. Washington, DC: Executive Office of the President of the United States, 2017.
3. President's Council of Advisors on Science and Technology (PCAST). Report to the President – Forensic science in criminal courts: ensuring scientific validity of feature-comparison methods. Washington, DC: Executive Office of the President of the United States, 2016.
4. Cassidy MJ. Footwear identification. Salem, OR: Lightning Powder Company Inc, 1995;98–103.
5. Sheets HD, Gross S, Langenburg G, Bush PJ, Bush MA. Shape measurement tools in footwear analysis: a statistical investigation of accidental characteristics over time. Forensic Sci Int 2013;232:84–91. https://doi.org/10.1016/j.forsciint.2013.07.010
6. Wilson HD. Comparison of the individual characteristics in the outsoles of thirty-nine pairs of Adidas Supernova Classic shoes. J Forensic Identif 2012;62(3):194–203.

7. Petraco NDK, Gambino C, Kubic TA, Olivio D, Petraco N. Statistical discrimination of footwear: a method for the comparison of accidentals on shoe outsoles inspired by facial recognition techniques. J Forensic Sci 2010;55(1):34–41. https://doi.org/10.1111/j.1556-4029.2009.01209.x
8. Hannigan TJ, Fleury LM, Reilly RB, O'Mullane BA, DeChazal P. Survey of 1276 shoeprint impressions and development of an automatic shoeprint pattern matching facility. Sci Justice 2006;46(2):79–89. https://doi.org/10.1016/s1355-0306(06)71578-7
9. Stone RS. Footwear examination: mathematical probabilities of theoretical individual characteristics. J Forensic Identif 2006;56(4):577–99.
10. Richetelli N, Bodziak WJ, Speir JA. Empirically observed and predicted estimates of chance association: estimating the chance association of randomly acquired characteristics in footwear comparisons. Forensic Sci Int 2019;302:109833. https://doi.org/10.1016/j.forsciint.2019.05.049
11. Champod C, Voisard R, Girod A. A statistical study of air bubbles on athletic shoe soles. Forensic Sci Int 2000;109:105–23. https://doi.org/10.1016/S0379-0738(99)002233-6
12. Davis RJ, Keeley A. Feathering of footwear. Sci Justice 2000;40 (4):273–6. https://doi.org/10.1016/S1355-0306(00)71997-6
13. Davis R, DeHaan J. A survey of men's footwear. J Forensic Sci Soc 1977;17:271–85. https://doi.org/10.1016/S0015-7368(77)71161-2
14. Fawcett AS. The role of the footmark examiner. J Forensic Sci Soc 1970;10:227–44. S0015-7368(70)70613-0
15. Adair TW, LeMay J, McDonald A, Shaw R, Tewes R. The Mount Bierstadt study: an experiment in unique damage formation in footwear. J Forensic Identif 2007;57(2):199–205.
16. Hamburg C, Banks R.Evaluation of the random nature of acquired marks on footwear outsoles. 2010 Impression and Pattern Evidence Symposium; 2010 Aug 2–5; Clearwater, FL. Washington, DC: Office of Justice Programs' National Institute of Justice/Bureau of Justice Assistance/Federal Bureau of Investigation Laboratory Division, 2010. https://projects.nfstc.org/ipes/presentations/Hamburg_random-acquired-marks.pdf (accessed July 31, 2020).
17. Mancini JM, Wilson E. The evolution of individuality: a 3-year comprehensive footwear study. In: Ropero-Miller JD, Daye CM, Eldridge H, editors. Conference Proceedings: 2015 Impression, Pattern, and Trace Evidence Symposium; 2014 Aug 25–27; San Antonio, TX. Research Triangle Park, NC: RTI Press Publication, 2015;42.
18. Yekutieli Y, Shor Y, Wiesner S, Tsach T.Expert assisting computerized system for evaluating the degree of certainty in 2D shoeprints. Technical Report. Washington, DC: National Institute of Justice, Office of Justice Programs, U.S. Department of Justice, 2016.
19. Majamaa H, Ytti A. Survey of the conclusions drawn of similar footwear cases in various crime laboratories. Forensic Sci Int 1996;82:109–20. https://doi.org/10.1016/0379-0738(96)01972-X
20. Shor Y, Weisner S. A survey on the conclusions drawn on the same footwear marks obtained in actual cases by several experts throughout the world. J Forensic Sci 1999;44:380–4. https://doi.org/10.1520/JFS14468J
21. Hammer L, Duffy K, Fraser J, Daeid NN. A study of the variability in footwear impression comparison conclusions. J Forensic Identif 2013;63:205–18.
22. SWGTREAD. Standard terminology for expressing conclusions of forensic footwear and tire impression examinations. Scientific Working Group for Shoeprint and Tire Tread Evidence. 2006. https://treadforensics.com/images/swgtread/standards/archived/swgtread_10_terminology_conclusions_200603_201302.pdf (accessed July 31, 2020).
23. SWGTREAD. Range of conclusions standard for footwear and tire impression examinations. Scientific Working Group for Shoeprint and Tire Tread Evidence, 2013. https://www.nist.gov/system/files/documents/2016/10/26/swgtread_10_range_of_conclusions_standard_for_footwear_and_tire_impression_examinations_201303.pdf (accessed July 31, 2020).
24. SWGTREAD. Standard for terminology used for forensic footwear and tire impression evidence. Scientific Working Group for Shoeprint and Tire Tread Evidence, 2013. https://www.nist.gov/system/files/documents/2016/10/26/swgtread_15_standard_for_terminology_used_for_forensic_footwear_and_tire_impression_evidence_201303.pdf (accessed July 31, 2020).
25. Tobi H, van den Berg PB, de Jong-van den Berg LTW. Small proportions: what to report for confidence intervals? Pharmacoepidemiol Drug Saf 2005;14(4):239–47. https://doi.org/10.1002/pds.1081

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Frequency of qualitative description for 210 features marked as "other."

Table S1. Examiner self-reported casework experience.

Table S2. Type of work performed by $n = 70$ participants, including database searches and comparison of impressions (note that 10 examiners reported that they had never performed a database search).

Table S3. Frequency of activities performed as assessed using a Likert scale for $n = 70$ participants.

Table S4. Examiner ($n = 70$) reports of current and past forensic activities (examiners were asked to report all that applied, so row totals can eclipse 70).

Table S5. Years of experience for $n = 70$ participants (note that 2 examiners did not give a response for "Total Years of Forensic Experience").

Table S6. Participant training and education for $n = 70$ examiners. Note that 1 examiner selected "other" for their education level, and 11 examiners selected "other" for their training provider.

Table S7. Summary of $n = 70$ gray box participants' backgrounds, including use of the SWGTREAD (2013) (23) conclusion standard (without modification), certification, proficiency testing, and further activities related to teaching and research (note that one examiner did not give a response regarding past research).

Table S8. Summary of all features marked per comparison, totaling 3524 annotations (note that examiners were permitted to mark on the questioned impression only, the known impression only, or both simultaneously; regardless of which option they selected, each marking was counted as a single feature).

Table S9. Community agreement/IQRs for current and former reliability studies.

Table S10. Frequency of comparison limitations reported (note that "improper scale" was not selected as an issue affecting any of the 840 analyses, and is therefore not included in this table).

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14551
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Nicole Richetelli,[1] M.S.; Lesley Hammer,[2] M.S.; and Jacqueline A. Speir,[1] Ph.D.

Forensic Footwear Reliability: Part II—Range of Conclusions, Accuracy, and Consensus*

ABSTRACT: Between February 2017 and August 2018, West Virginia University conducted a reliability study to determine expert performance among forensic footwear examiners in the United States. Throughout the study's duration, 70 examiners each performed 12 comparisons and reported a total of 840 conclusions. In order to assess the accuracy of conclusions, the similarities and differences between mated and nonmated pairs were evaluated according to three criteria: (i) inherent agreement/disagreement in class, wear, and randomly acquired features, (ii) limitations as a function of questioned impression quality, clarity, and totality, and (iii) adherence to the Scientific Working Group for Shoeprint and Tire Tread Evidence (SWGTREAD) 2013 conclusion standard. Using these criteria, acceptable/expected categorical conclusions were defined. Preliminary results from this study are divided into a series of three summaries. This manuscript (Part II) reports accuracy and reproducibility. For mated pairs, accuracy equals 76.3% ± 13.0% (median of 78.6% and a 90% confidence interval between 72.2% and 80.0%). For nonmated pairs, accuracy equals 87.4% ± 9.24% (median of 91.4% and a 90% confidence interval between 84.7% and 89.8%). In addition, the community assessed agreement (denoted by IQR) of reported results equals the research team's accepted/expected conclusions for 10 out of 12 comparisons. In terms of reproducibility, the 90% confidence interval for consensus was computed and found to equal 0.71–0.86 (median of 0.77) for the combined dataset. Although based on a limited sample size, these results provide a baseline estimate of accuracy and consensus/reproducibility as a function of the existing seven-point SWGTREAD 2013 conclusion standard.

KEYWORDS: forensic footwear evidence, reliability, gray box study, footwear examiners, accuracy, consensus, reproducibility

Between February 2017 and August 2018 West Virginia University (WVU) conducted a reliability study to investigate performance among forensic footwear examiners in the United States. The goals of this project were several-fold, including an evaluation of accuracy, consensus, and reproducibility in feature identification, feature evaluation, and overall conclusions regarding source attribution. In terms of organization, initial results are dispersed in three manuscripts; Part I (1) described the reproducibility in feature identification, feature evaluation, and generalized community agreement, while Part II (this summary) reports on examiner accuracy and reproducibility (or consensus) of conclusions. Within the context of this study, accuracy is defined according to the President's Council of Advisors on Science and Technology (PCAST) 2016 report, or the known probability (or frequency) at which "an examiner obtains correct results both (1) for samples from the same source (true positives) and (2) samples from different sources (true negatives)" (2). Unfortunately, this is not trivial to calculate when the conclusion standard for the community is a seven-point scale (such as the Scientific Working Group for Shoeprint and Tire Tread Evidence [SWGTREAD] 2013 conclusion standard [3]), rather than binary conclusive determinations (such as *identification* and e*xclusion*, as exists within some other forensic pattern sciences). Moreover, there is little guidance on how to handle this nuance when attempting to determine the accuracy for assignment to decisions within a categorical scale that mimics a Likert scale varying from strong to weak dissociations (i.e., *exclusion*, *indications of non-association*) and weak to strong associations (i.e., *association of class*, *high degree of association*, and *identification*). In addition, this complication is not alleviated by a research study with known ground truth. For example, even though ground truth was known for every simulated comparison pairing a crime scene-like questioned impression with a known test impression within this reliability study, the questioned prints were deposited and collected under natural conditions, and thus vary in both quality and clarity, as well as inherent discrimination potential (degree and type of wear, presence/absence of randomly acquired characteristics (RACs), etc.), possibly resulting in outcomes that span a range of SWGTREAD (2013) conclusion categories (3). Thus, a contrived research paradigm with ground truth does not solve the issue of defining a reasonable or accurate accepted conclusion. Accordingly, although the research team knew which shoe created which impression, binary conclusions such as *identification* and *exclusion* were not anticipated for each and every mated pair and nonmated pair, respectively. Instead, to define an expected/accepted conclusion, each questioned/known impression combination was independently evaluated with respect to ground truth, observable features (and their

[1] West Virginia University, 208 Oglebay Hall, PO Box 6121, Morgantown, WV, 26506.

[2] Hammer Forensics, LLC, 10601 Prospect Drive, Anchorage, AK, 99507.

Corresponding author: Jacqueline A. Speir, Ph.D. E-mail: Jacqueline.Speir@mail.wvu.edu

*This project was supported by Award No. 2016-DN-BX-0152, awarded by the National Institute of Justice (NIJ), Office of Justice Program, U.S. Department of Justice. The opinions, findings, conclusions, and recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the Department of Justice.

Received 18 Mar. 2020; and in revised form 15 June 2020; accepted 31 July 2020.

associated reliability), and the SWGTREAD (2013) conclusion criteria (3). For example, consider a known nonmated shoe without any significant characteristics of use, and that agrees in both outsole design and physical size with a questioned impression. Under this scenario (assuming no differences between impression features and outsole characteristics) a conclusion of *association of class* would be defined as reasonable (or acceptable) according to the SWGTREAD (2013) guidelines (3). In other words, "…the known footwear is a possible source of the questioned impression and therefore could have produced the impression" (3), noting (importantly) that other outsoles with the same characteristics observed in the impression are also included in the population of possible sources.

As a result, the research team was presented with a difficulty not believed to be present in many other forensic reliability studies. In order to address this challenge, solutions for similar problems in other fields were considered. This revealed that consensus is typically the major study goal in subjective judgment analysis, while accuracy is relegated for idealized scenarios (e.g., the accuracy of a weather forecast or a financial prediction that can be assessed by gathering additional information after a time delay). Thus, the research team approached the accuracy assessment problem using an accepted technique employed in other fields, such as the evaluation of surgical procedures or images, wherein a small number of individuals "establish a gold standard" (4,5). In other words, the research team was afforded an "oracle" status, and permitted to define what would be considered accurate and inaccurate, while still allowing for some degree of opinion evolution (see methods for a full description). Conversely, consensus became the focus of reproducibility. Fortunately, measuring consensus with ordinal scales is somewhat easier than assessing accuracy, but its quantification differs from both consensus estimation/group decision making (6–11) and crowd ranking (4,12–14), where the goal of the latter is to reach consensus or agreement through discussion and opinion evolution, while the goal of the former is to quantify the degree of agreement reached by independent observers during a single round of decision making. Thus, for the purpose of this study, consensus and dissension are considered a proxy for reproducibility, where reproducibility is defined according to the PCAST (2016) report, or the known probability (or frequency) at which "different examiners obtain the same result, when analyzing the same samples" (2). As with accuracy, this metric is likewise complicated by the use of a seven-point conclusion standard. For example, if a participant can select between two binary categories (i.e., *agree* or *disagree*), then if an actual ranking or agreement model exists for the decision (which is assumed to be true for expert opinions within scientific disciplines, vs., say, users' preferences in movies) then agreement should be higher for these types of binary decisions, than for experts presented with Likert scales with increasing numbers of categories (i.e., *strongly disagree*, *disagree*, *neutral*, *agree*, and *strongly agree*). With regard to the SWGTREAD (2013) conclusion standard (3), after removing *insufficient detail*, the remaining conclusions represent an ordinal scale, ranging from strong to weak exclusionary statements, followed by weak to strong associative statements. When presented with similar scales, Tastle and Wierman (15) illustrate that measures of agreement are poorly described by typical metrics, such as the mean, standard deviation, and entropy. As a specific example, consider a five-point Likert scale; if the mean response is near the end points of the scale (one or five) the variance must be smaller than if the mean is at the midpoint (three) (16). Thus, a more appropriate measure of consensus and dispersion was sought in order to address the question of reproducibility in footwear expert opinions, as is further described in the following section.

Materials and Methods

Participant Demographics

Participant demographics are fully characterized in Part I (1). In total, 115 forensic footwear examiners were recruited, and 77 submitted results (resulting in a 67% response rate). However, the remainder of all statistical analyses is based on the responses from 70 participants who reported that they had previously completed training and performed one or more footwear comparisons (the results from seven participants were excluded moving forward based on self-reports of never having completed a comparison, or still in-training).

Case Variety

The details of each simulated case are also fully described in Part I (1), but for convenience, Table 1 repeats limited information concerning the shoes, substrates, media, and processing techniques used to create case materials. Note that across all seven cases, participants were required to perform 12 pairwise comparisons and that each case consisted of 1200 PPI digital and print imagery, comprised of a single questioned impression, one-two exemplars (outsoles), and two Handiprint replicate exemplars per known shoe. Of the 12 pairwise comparisons provided to participants, five were true mates, and seven were nonmates. Of the two cases with a single known, one included a true mate and the other, a nonmate. Of the remaining five cases with two knowns, four had a true mate and a nonmate, and one had two nonmates.

Based on participant ratings of case difficulty, examiners reported that the 22% of the cases were easy, 56% were moderate, and 22% were difficult. In terms of media, three cases involved blood, three involved dust, and one was in wax. Unfortunately, the authors are unaware of the exact ratio of case types submitted to laboratories and therefore cannot say whether the ratio of 3:3:1 for substrate/media is representative of casework.

Case Analyses

Each participant received a package via USPS of all relevant case materials, including high-resolution color prints, a set of blank acetates for overlay annotation, a CD containing the digital reporting software, a copy of the SWTREAD (2013) conclusion scale (3) and an instruction document (with additional weblinks to access digital copies of all case materials). Participants were asked to treat the simulated cases as if each were routine casework and given that no time had passed between collection of the questioned and test impressions. In terms of analysis, each examiner was asked to review the material and perform their analyses according to their training and expertise, but without the option for any external or independent verification. After performing routine analyses, participants were asked to respond to a series of questions using a customized software reporting interface that solicited responses regarding the identity, similarity, dissimilarity, clarity and value of class, manufacturing, and characteristics of use (wear and RACs) significant to the comparison, before reporting a final conclusion.

TABLE 1—*Shoes, substrates, media, and processing techniques used to create simulated case materials.*

Case	Manufacturer of Known(s)	Size & Style of Known(s)	Substrate of Unknown	Medium of Unknown	Processing of Unknown	# of Known(s)
001	Converse	All Star (9)	Ceramic tile	Blood	Leucocrystal violet	2
002	Nike	Lebron James (10)	Vinyl tile	Dust	Digitally enhanced gel lift	1
003	Nike	Rosherun (9)	Ceramic tile	Blood	Leucocrystal violet	2
004	Nike	Air Max (10.5)	Linoleum tile	Wax	Magnetic powder & gel lift	2
005	Nike	Air Max (11)	Vinyl Tile	Dust	Digitally enhanced gel lift	1
006	Nike	Air Max Cage (10)	Paper	Dust	Digitally enhanced	2
007	Under Armour	Unknown (10 & 11)	Ceramic Tile	Blood	Leucocrystal violet	2

Acceptable Range of Conclusions

As previously noted, although ground truth was known for every simulated comparison, binary conclusions such as *identification* and *exclusion* were not anticipated for each and every mated pair and nonmated pair, respectively. Instead, each questioned/known impression combination was independently evaluated with respect to ground truth, observable features (and their associated reliability), and the SWGTREAD (2013) conclusion criteria (3), allowing the research team to draft an acceptable set of conclusions for each comparison. This process was repeated independently by four members of the research team (including one practitioner partner). All draft results were tabulated and through conference, discrepancies were discussed and evaluated until agreement was obtained within the team.

The process of defining an acceptable range of conclusions was repeated a second time *after* data collection and during analysis of results, during which time the research team examined the range of responses provided by the 70 members of the forensic footwear community, and predominant categories on either side of any previously accepted range were re-evaluated after consideration of participant responses. This review resulted in two changes; first, the acceptable conclusions permitted for the comparison of 003Q with 003K2 was *reduced* from *exclusion* and *indications of non-association* to *exclusion* only. Consequently, any selection of *indications of non-association* for this pairwise comparison was deemed a "failure to exclude" wherein *exclusion* is considered the *correct* answer based on the observable and reliable size differences that could be measured (varying between 3 and 8 mm) between the questioned and test impression. Second, the conclusions permitted for the comparison of 005Q with 005K1 were *expanded* from *exclusion* and *indications of non-association* to allow for *exclusion*, *indications of non-association*, and *limited association*. The extension of the permitted range for this pairwise comparison was based on participants' detection of a size difference, but many comments (made by nearly 30% of respondents) indicating an inability to confirm that the differences being observed were "reliable." Note that these two changes do not reflect any *fundamental persuasion of opinion* of the research team by the community group decisions. Rather, both are a reflection of the artificial research paradigm used in this study. More specifically, examiners were not able to prepare their own exemplars or inspect outsoles, which is typically afforded in actual casework, ergo their comments regarding reliability (i.e., an examiner is detecting a size difference, but commenting on its reliability without being able to perform additional comparisons). For comparison 003K2, the size difference was large enough that the community deemed it reliable in the absence of the outsoles, while for case 005, the community noted a size difference but expressed uncertainty in its reliability (which presumably could be rectified if afforded the actual outsoles). Thus, the research team in one instance reduced, and in another instance expanded, the accepted range of conclusions in order to account for limitations in study design.

Accuracy and Dispersion

Following the assignment of the acceptable range of conclusions for all cases and the receipt of all conclusions, expert decisions within each category on the SWGTREAD (2013) scale (3) were compiled, as illustrated in Fig. 1. For each comparison, the expected decision categories are highlighted in green, and the percentage of responses deemed accurate is reported at far right (Fig. 1). Within each conclusion category, the number and percentage of examiners selecting a specific conclusion category are enumerated. As a measure of dispersion, two additional metrics are provided; a box plot detailing the median, interquartile range (IQR), whiskers, and possible outliers, and finally, a measure of consensus (C). The metric of consensus is bounded between zero and one, and is an estimate of the variability in responses computed according to Eq. 1, where $i = 1, 2,..., n$ equals the index of the category of interest (n equals six in this study for each of the SWGTREAD [2013] conclusion categories after excluding *insufficient detail* [3]), X_i equals the value assigned to the category of interest, p_i equals the proportion of conclusions in the category of interest relative to the total, μ_x equals the mean score across all conclusion categories, and d_x equals the width of the conclusion categories ($d_x = X_{\max} - X_{\min}$ or $d_x = 6 - 1 = 5$) (15). The opposite of consensus is dissension (D) (Eq. 2), which also ranges from zero to one, such that dispersion/dissension

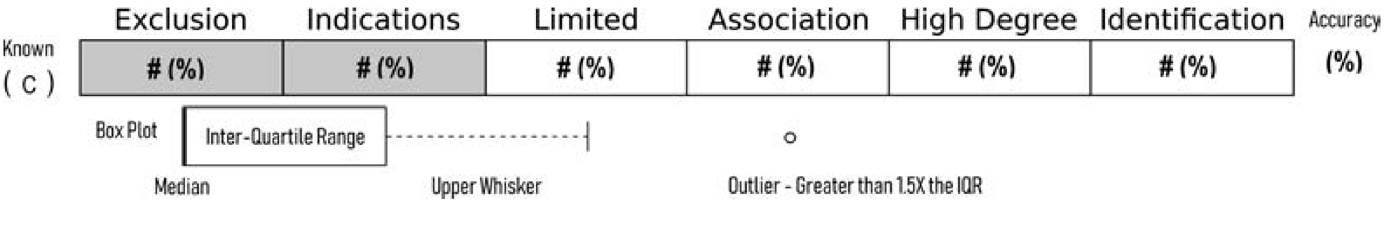


FIG. 1—*Range of expert conclusions for each questioned-known impression comparison, as a function of frequency (percentage), with the acceptable conclusions highlighted in green and reported as a total percentage at far right (accuracy). Consensus of examiner decisions (C) (15) is reported below the comparison number and visually illustrated in the form of a box plot detailing median (bold line), interquartile range (IQR), and if present, outliers (o) (18). [Color figure can be viewed at wileyonlinelibrary.com]*

approaches zero when all respondents select a single category, and one if there is a large proportion of observations at two ends of a scale (for example, if respondents selected *exclusion* and *identification* in equal proportions) (15).

$$C = 1 + \sum_{i=1}^{n} p_i \log_2 \left(1 - \frac{|X_i - \mu_x|}{d_x} \right) \quad (1)$$

$$D = - \sum_{i=1}^{n} p_i \log_2 \left(1 - \frac{|X_i - \mu_x|}{d_x} \right) \quad (2)$$

Results and Discussion

Acceptable Range of Conclusions

Table 2 reports the acceptable range of conclusions agreed on after consideration of ground truth, the opinion of three researchers, the opinion of a practitioner partner, and the reliability limitations expressed by participants through self-reporting given the contrived nature of the research experiment (examiners did not prepare the exemplars themselves, and the actual shoe was not provided for each known). Since the authors acknowledge that disagreement may be expressed concerning this process, images, overlays, and features driving each conclusion are highlighted and annotated in this publication for external review. These details are provided in an effort to be as transparent as possible, to allow readers to understand and evaluate the research team's reasoning (and/or draw his or her own conclusions), and finally, to allow downstream re-analysis of the results by external bodies should this be of interest.

Accuracy and Reproducibility

In total, 840 decisions (70 examiners × 12 comparisons) were available for interpretation. In an effort to characterize the foundational validity of forensic footwear examination as detailed by the PCAST (2016) report (2) and 2017 addendum (17), all expert conclusions were evaluated per questioned-known comparison in order to inform both accuracy ("correctness") and reproducibility ("consensus"). Figure 2 reports the frequency (percentage) of expert decisions within each SWGTREAD (2013) conclusion category (3), with the expected (accepted) decision categories highlighted in green, and consensus (C) calculated according to Eq. 1 (15). In addition, the spread of decisions per comparison is illustrated via box plots that highlight the median, interquartile range (IQR), and possible outliers (1.5 × IQR) (18).

TABLE 2—*Range of "acceptable" conclusions for each comparison.*

Comparison	Acceptable Conclusions
001Q vs. 001K1	Exclusion, Indications of Non-association
001Q vs. 001K2	High Degree of Association, Identification
002Q vs. 002K1	Limited Association, Association of Class
003Q vs. 003K1	Association of Class, High Degree of Association
003Q vs. 003K2	Exclusion
004Q vs. 004K1	Exclusion
004Q vs. 004K2	Identification
005Q vs. 005K1	Exclusion, Indications of Non-association, Limited Association
006Q vs. 006K1	Limited Association, Association of Class
006Q vs. 006K2	Exclusion, Indications of Non-association
007Q vs. 007K1	Exclusion
007Q vs. 007K2A	High Degree of Association, Identification
007Q vs. 007K2B	Exclusion, Indications of Non-association, Limited Association

In Part I (1), community agreement was defined using the IQR. In comparison to the accepted range of conclusions, Fig. 2 shows that for all but comparison 001Q versus 001K2 and 006Q versus 006K2, the acceptable range of conclusions permitted by the research team is equivalent to the community IQR. Conversely, for 001Q-001K2 and 006Q-006K2, the IQR is wider than the accepted range. The only other exception is comparison 007Q versus 007K2A; however, this comparison was only distributed to five analysts before 007K2A was retired and exchanged for 007K2B, with the latter known evaluated by the remaining 65 participants. Thus, the authors make little claim regarding its statistical utility based on the limited number of times it was reviewed. *Note that impression 007K2A was retired in order to increase the variation in cases and provide one with two nonmated pairs.*

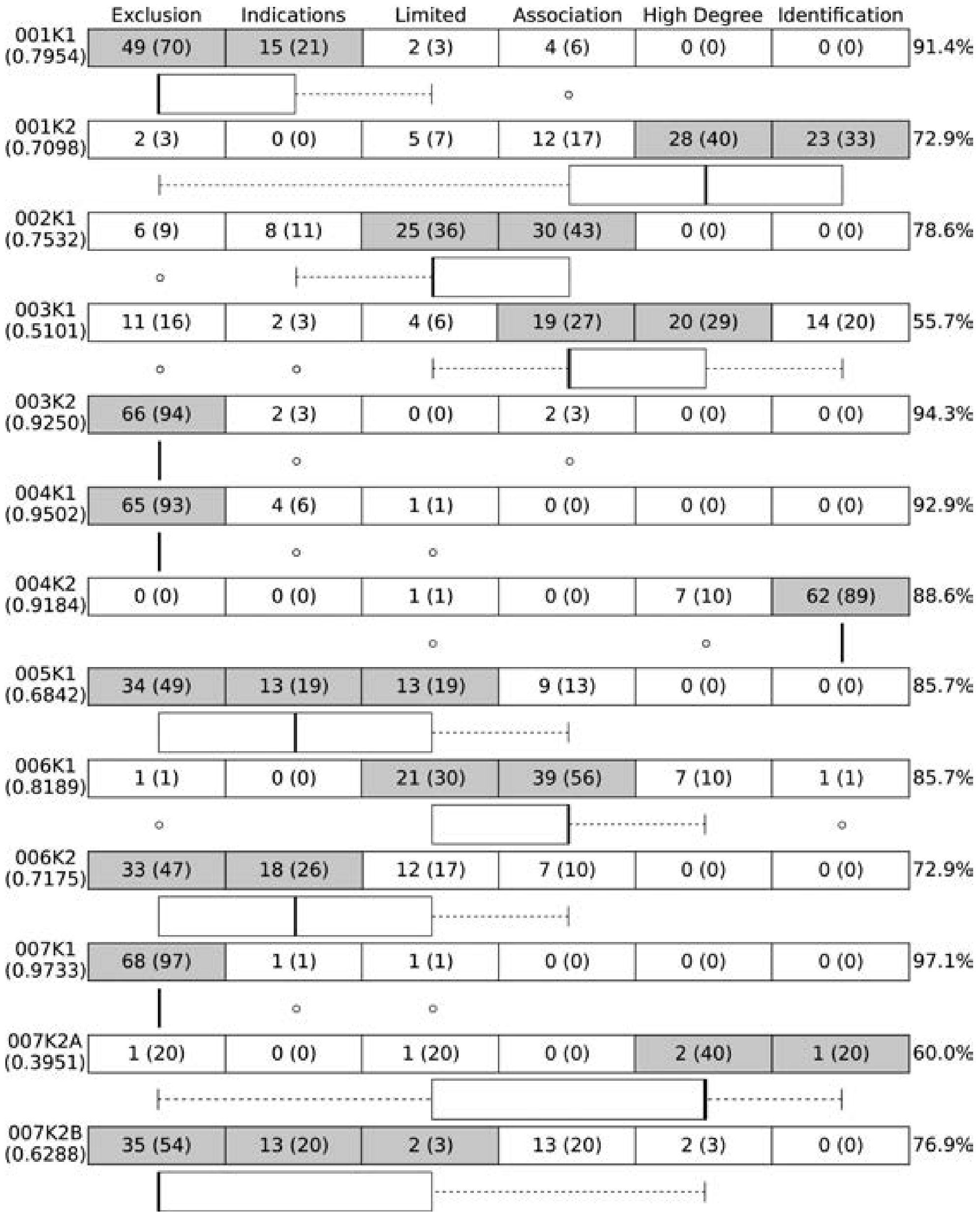
Excluding 0072KA, the overall accuracy varies from a low of 55.7% to a high of 97.1%, with the mean and one standard deviation equal to 82.8% ± 11.9% (median of 85.7% and 90% confidence interval of 80.5–84.9%). In terms of mates and nonmates, the accuracy for mated pairs equals 76.3% ± 13.0% (median of 78.6% and 90% confidence interval of 72.2–80.0%), and the accuracy for nonmated pairs equals 87.4% ± 9.24% (median of 91.4% and 90% confidence interval of 84.7–89.8%). Note that if the accepted range of conclusions for comparison 003Q versus 003K1 had not been reduced after postprocessing expert responses, the accuracy would have been 97.1% (vs. 94.3% after the range reduction). Likewise for comparison 005Q versus 005K1, if the range of conclusions had not been expanded after postprocessing expert responses, the accuracy would have been 67.1% (vs. 85.7% after the range expansion). Confidence intervals for mated, nonmated and combined data are provided in Table S1, and proportions for each case are provided in Table S2.

In terms of variation in responses, consensus ranges from a low of 0.5101 to a high of 0.9733, with the mean and one standard deviation equal to 0.7821 ± 0.1422 (median = 0.7743). In terms of mates and nonmates, the consensus among mated pairs equals 0.7421 ± 0.1516 (median of 0.7532), and the consensus among nonmated pairs equals 0.8106 ± 0.1396 (median of 0.7954). Additional standard error computations and the 90% confidence interval are provided in Table S3. One less this value is a measure of dispersion, and both collectively describe the reproducibility in responses when using an ordinal conclusion scale that varies between strong to weak disassociations and weak to strong associations.

Individual Case Summaries

Case 001

Questioned impression 001Q was a bloody shoeprint deposited on ceramic tile and enhanced with leucocrystal violet (LCV) (Fig. 3). Test impression 001K1 was a nonmated pair with 001Q, but equivalent in make, model, and size. The accepted range of conclusions (and IQR) equals *exclusion* and *indications of non-association* based on disagreement in wear as illustrated in Fig. 3. In total, 91% of analysts reported an expected conclusion, yielding an examiner consensus of 0.7954 (Fig. 2); the remaining 9% of decisions fell into the adjacent SWGTREAD



	Exclusion	Indications	Limited	Association	High Degree	Identification	
001K1 (0.7954)	49 (70)	15 (21)	2 (3)	4 (6)	0 (0)	0 (0)	91.4%
001K2 (0.7098)	2 (3)	0 (0)	5 (7)	12 (17)	28 (40)	23 (33)	72.9%
002K1 (0.7532)	6 (9)	8 (11)	25 (36)	30 (43)	0 (0)	0 (0)	78.6%
003K1 (0.5101)	11 (16)	2 (3)	4 (6)	19 (27)	20 (29)	14 (20)	55.7%
003K2 (0.9250)	66 (94)	2 (3)	0 (0)	2 (3)	0 (0)	0 (0)	94.3%
004K1 (0.9502)	65 (93)	4 (6)	1 (1)	0 (0)	0 (0)	0 (0)	92.9%
004K2 (0.9184)	0 (0)	0 (0)	1 (1)	0 (0)	7 (10)	62 (89)	88.6%
005K1 (0.6842)	34 (49)	13 (19)	13 (19)	9 (13)	0 (0)	0 (0)	85.7%
006K1 (0.8189)	1 (1)	0 (0)	21 (30)	39 (56)	7 (10)	1 (1)	85.7%
006K2 (0.7175)	33 (47)	18 (26)	12 (17)	7 (10)	0 (0)	0 (0)	72.9%
007K1 (0.9733)	68 (97)	1 (1)	1 (1)	0 (0)	0 (0)	0 (0)	97.1%
007K2A (0.3951)	1 (20)	0 (0)	1 (20)	0 (0)	2 (40)	1 (20)	60.0%
007K2B (0.6288)	35 (54)	13 (20)	2 (3)	13 (20)	2 (3)	0 (0)	76.9%

FIG. 2—*Range of expert conclusions for each questioned-known impression comparison, as a function of frequency (and percentage), with the acceptable conclusions highlighted in green and summarized as a percentage at far right (accuracy). Consensus of examiner decisions (C) (15) (or reproducibility) is reported below the comparison number and visually in the form of a box plot detailing median (bold line), interquartile range (IQR) and if present, outliers (o) (18) (note that 002Q vs. 002K1, 005Q vs. 005K1 and 006Q vs. 006K1 all had one response of insufficient detail which is not reflected in the table categories, but is reflected in the accuracy percentage). [Color figure can be viewed at wileyonlinelibrary.com]*

(2013) categories (3) (*limited association* and *association of class*, with the latter being classified statistically as an outlier based on the distribution of all conclusions). Conversely, test impression 001K2 was a mated pair with 001Q (Fig. 4). However, the expert decisions for this pairwise comparison exhibited less agreement than for 001Q versus 001K1, with 73% of analysts reporting the expected conclusion (*high degree of association* or *identification*) and a consensus of 0.7098 (Fig. 2). In alignment with the increased observed dispersion, the IQR extends outside of the range of expected conclusions and into the *association of class* category. Moreover, if the case were treated as a whole, then the overall accuracy for both comparisons would be 82%.

Case 002

Questioned impression 002Q was a dust impression on vinyl tile, collected using a black gelatin lift (Fig. 5), to be compared against a single item (002K1) which was the known source of the evidence. Despite serving as a mated pair, strong associations were not anticipated based on the nature of the impression (limited quality (clarity, contrast, reliability) and quantity of information). As illustrated in Fig. 2, a total of 55 analysts (79%) reached either *limited association* or *association of class* as anticipated (which was equivalent to the IQR). Comparatively, the consensus score equaled 0.7532. Interestingly, the remainder of the conclusions were dissociative in nature, and in

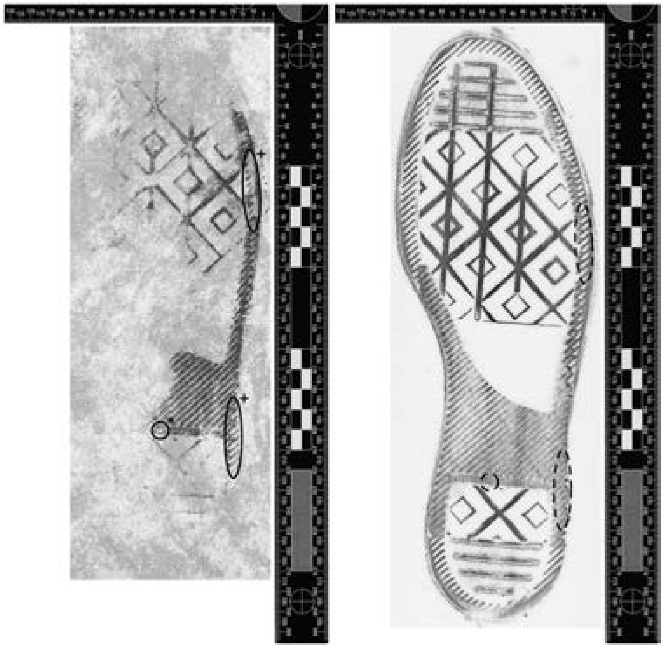


FIG. 3—*Case 001 questioned impression (left) compared to test impression from nonmated shoe 001K1 (right). Features of interest that are visible on the questioned impression are circled with a solid line (+ denotes a wear pattern and * denotes a RAC). The dashed lines on the test impression indicate that these features are not present on the known nonmated shoe. Accuracy = 91.4% and consensus = 0.7954. [Color figure can be viewed at wileyonlinelibrary.com]*

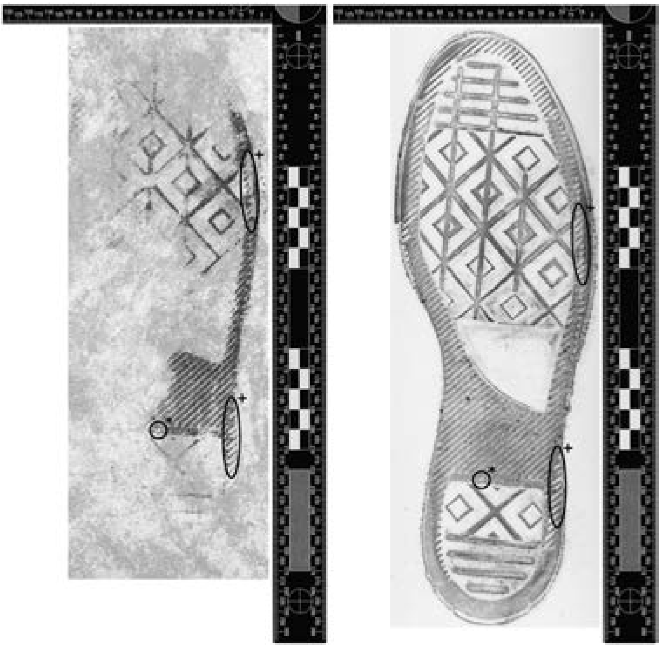


FIG. 4—*Case 001 questioned impression (left) compared to test impression from mated shoe 001K2 (right). Features of interest that are visible on the questioned impression are circled with a solid line (+ denotes a wear pattern and * denotes a RAC). The solid lines on the test impression indicate that these features are also present on the true mate. Accuracy = 72.9% and consensus = 0.7098. [Color figure can be viewed at wileyonlinelibrary.com]*

fact the *exclusion* decisions are considered statistical outliers (as was the single response of insufficient detail). Unfortunately, of the 15 examiners that reported inaccurately, only one marked features (see annotation map from the Part I summary [1]) which makes it difficult to determine the reason for the disassociations.

Case 003

Questioned impression 003Q was a bloody shoeprint deposited on ceramic tile and enhanced with LCV (Fig. 6). In contrast to case 001 (also a bloody impression on ceramic tile enhanced with LCV), comparison of the questioned impression with known source 003K1 resulted in the lowest expert accuracy for the study (with the exception of 007K1A, which is discussed in further detail below), with 56% of analysts reaching a decision of either *association of class* or *high degree of association* (which is also equal to the IQR). Not surprisingly, the consensus score is likewise low (0.5101). Statistically, this means that 19% of all conclusions were considered outliers, and this was the only comparison in this study in which every SWGTREAD (2013) (3) conclusion category was represented in the responses (the distribution of which is shown in Fig. 2). In contrast, the nonmated comparison for this case (003Q vs. 003K2) produced one of the higher accuracy and reproducibility performances in the study, likely owing to the observable disagreement in size (Fig. 7). More specifically, 66 experts (94%) reached an *exclusion*, the single acceptable decision (which is also equal to the IQR). Also, as illustrated in Fig. 2, the consensus was 0.9250, while all other reported conclusion categories (*indications of non-association* and *association of class*) were classified as statistical outliers (6%). Moreover, if the case were treated as a whole, then the overall accuracy for both comparisons would be 75%.

Case 004

Case 004 overall was arguably one of the easiest cases in this study given the relatively high quality of the questioned impression (wax impression on linoleum tile enhanced with magnetic

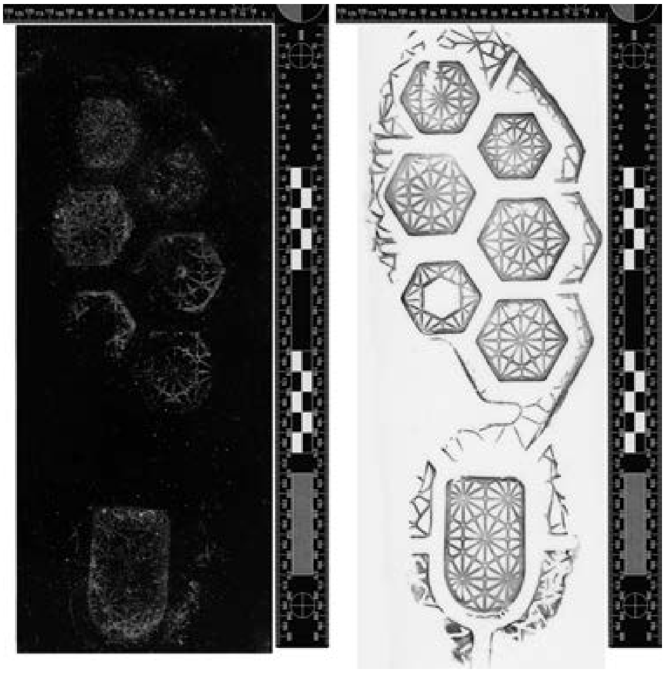


FIG. 5—*Case 002 questioned impression, reverse orientation (left) compared to test impression from mated shoe 002K1 (right). The only features of interest for this examination were general agreement of size, design (and possibly overall wear). As such, there are no specific features highlighted in these images. Accuracy = 78.6% and consensus = 0.7532.*

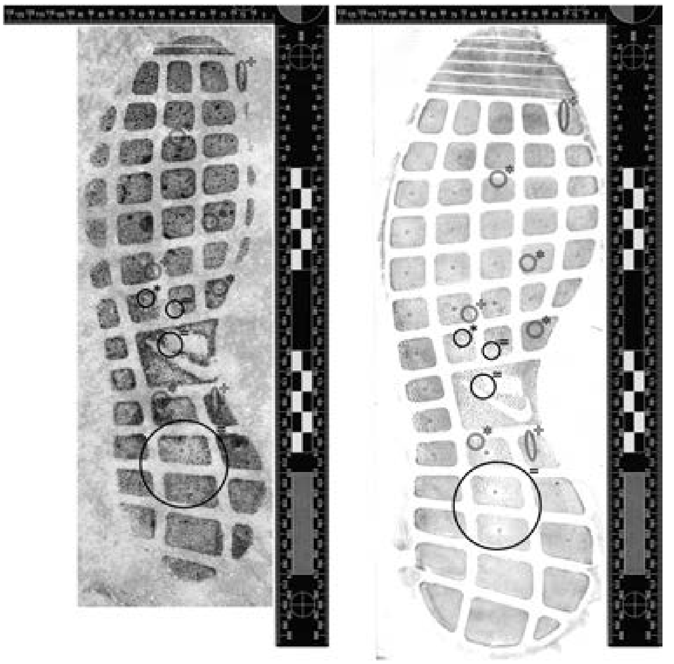


FIG. 6—*Case 003 questioned impression (left) compared to test impression from mated shoe 003K1 (right). Features of interest that are visible on the questioned impression are circled with a solid line (= denotes a texture/stippling pattern, + denotes a wear pattern, and * denotes a RAC). The solid lines on the test impression indicate that these features are also present on the true mate. Accuracy = 55.7% and consensus = 0.5101. [Color figure can be viewed at wileyonlinelibrary.com]*

powder and lifted with a white gelatin lifter) (Fig. 8), and numerous patent discriminating features (general wear, Schallamach, and RACs). Accordingly, relatively high expert performance was observed for this case as a whole. The first suspect shoe provided to experts (004K1) was not the source of the evidence (004Q). Given the high clarity, 93% of analysts reached the single acceptable outcome of *exclusion* for this comparison, as detailed in Fig. 2 (which is equivalent to the IQR). Similarly, a high level of consensus was observed for this case (0.9502). Although the remaining 7% of examiners reviewing this comparison reported a decision in the two SWGTREAD (2013) (3) categories immediately adjacent to the accepted response (*indications of non-association* and *limited association*), these outcomes were deemed outliers statistically (Fig. 2). Given the clearly observable wear, Schallamach, and RACs in impression 004Q, and the fact that the second suspect shoe (004K2) was the true source, the only accepted conclusion for the comparison of 004Q with 004K2 was *identification* (which is equivalent to the IQR) (Fig. 9). Of all 70 experts, 62 reached the expected decision (89%), with an additional 10% concluding *high degree of association*, but being statistically equivalent to outliers. Similarly, the overall consensus was high (0.9184), and if the case were treated as a whole, then the overall accuracy for both comparisons would be 91%.

Case 005

Similar to case 002, questioned impression 005Q was also a dust impression on vinyl tile collected with a black gelatin lifter (Fig. 10). A single nonmated suspect shoe (005K1) was provided to experts for comparison. As illustrated in Fig. 2, 86% of analysts reached either *exclusion*, *indications of non-association*, or *limited association*, as expected (which is equivalent to the IQR). Despite this, overall consensus was reduced to 0.6842, since the decisions were spread across a number of categories.

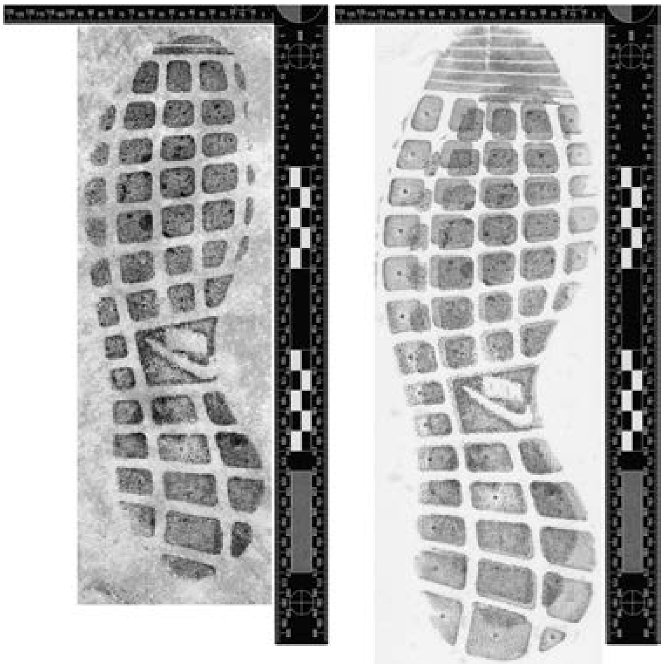


FIG. 7—*Case 003 questioned impression (left) compared to the questioned impression overlaid on a test impression from nonmated shoe 003K2 (right). Note the disagreement in size. Accuracy = 94.3% and consensus = 0.9250. [Color figure can be viewed at wileyonlinelibrary.com]*

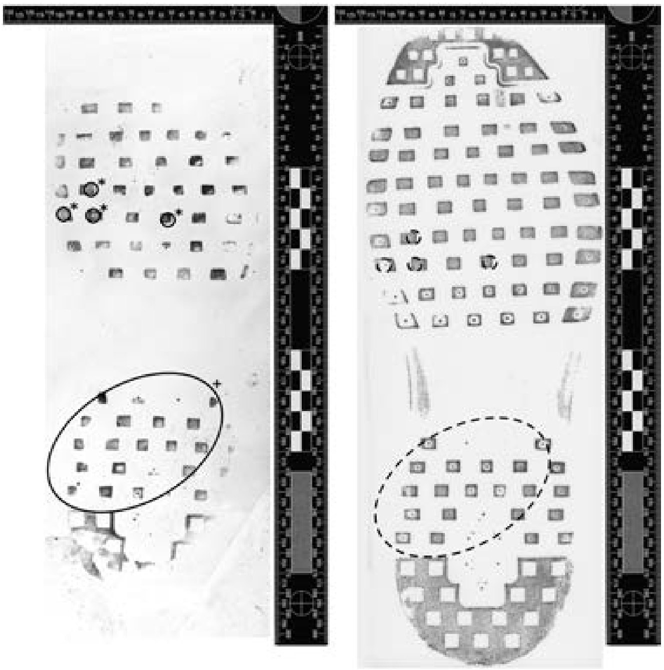


FIG. 8—*Case 004 questioned impression (left) compared to test impression from nonmated shoe 004K1 (right). Features of interest that are visible on the questioned impression are circled with a solid line (+ denotes a wear pattern and * denotes a RAC). The dashed lines on the test impression indicate that these features are not present on the true nonmate. Accuracy = 92.9% and consensus = 0.9502.*

Case 006

Questioned impression 006Q was a dust impression on paper, with 006K1 corresponding to its source (Fig. 11). However, given the print's partiality and low contrast, the acceptable range of conclusions was *limited association* and *association of class*. In total, 60 analysts (86%) reported one of these expected decisions, as detailed in Fig. 2 (which was equivalent to the IQR). This comparison yielded an examiner consensus of 0.8189, with the majority of remaining decisions equaling stronger associations and only two examiners reporting conclusions deemed outliers statistically. Conversely, suspect shoe 006K2 was a nonmated pair (Fig. 12). However, the expert decisions for the comparison of 006Q versus 006K2 were less consistent and accurate than for 006Q versus 006K1. More specifically, only 73% of analysts reported within the acceptable range of conclusions (*exclusion* and *indications of non-association*) (with the IQR being equal to *exclusion*, *indications of non-association*, and *limited association*). The consensus was similarly lower and equal to 0.7175 (Fig. 2). Consequently, if the case were treated as a whole, then the overall accuracy for both comparisons would be 79%.

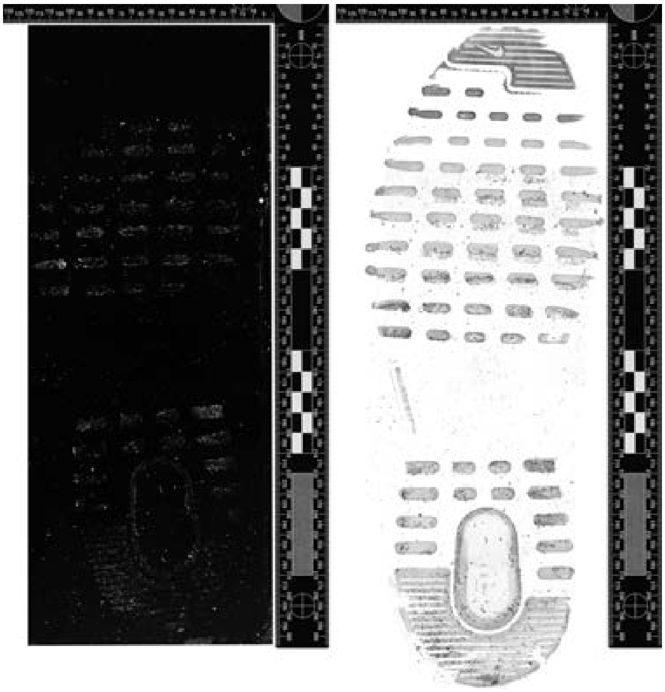


FIG. 10—*Case 005 questioned impression, reverse orientation (left) compared to questioned impression 005Q (with contrast reversal) overlaid on the test impression from nonmated shoe 005K1 (right). Note the disagreement in size and possible indications of distortion. Accuracy = 85.7% and consensus = 0.6842. [Color figure can be viewed at wileyonlinelibrary.com]*

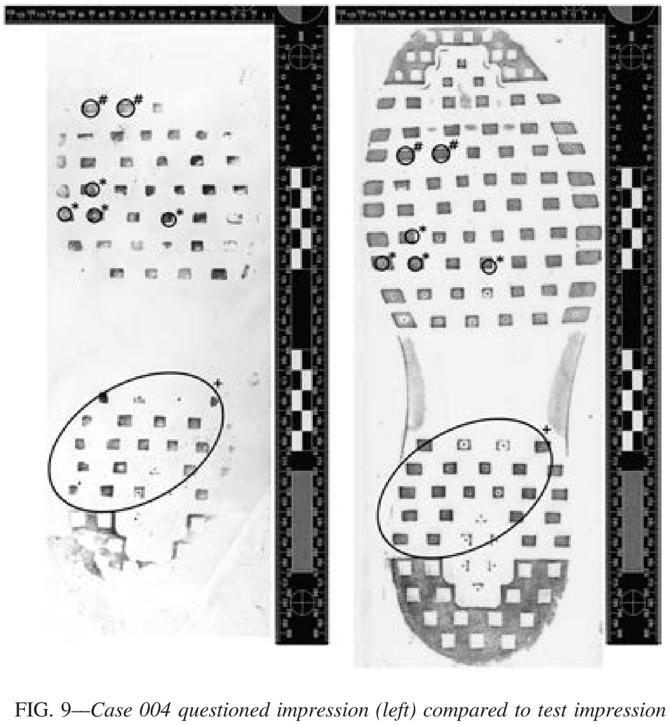


FIG. 9—*Case 004 questioned impression (left) compared to test impression from mated shoe 004K2 (right). Features of interest that are visible on the questioned impression are circled with a solid line (+ denotes a wear pattern, # denotes a Schallamach pattern, and * denotes a RAC). The solid lines on the test impression indicate that these features are also present on the true mate. Accuracy = 88.6% and consensus = 0.9184.*

Case 007

Similar to case 001 and 003, questioned impression 007Q was a bloody impression on ceramic tile enhanced using LCV. Comparison of the questioned impression versus 007K1 reveals a full size difference (Fig. 13), and resulted in the highest accuracy for this study, with 97% of examiners reporting *exclusion* (which is equivalent the IQR). Similarly, the consensus was found to be 0.9733, and only two examiners reported conclusions that were considered outliers statistically. A very small cohort of examiners (4) received 007K2A for additional comparison with 007Q. This shoe was the source of the questioned impression and also included significant characteristics of use (Fig. 14). Given that these wear-features were easily observable and comparable between the questioned and test impressions, and the desire to have an additional open-set case (wherein the true mate is not provided to examiners, as recommended by PCAST [2016] [2]), the remaining 65 examiners that participated in this study were asked to compare a third shoe, denoted as 007K2B (nonmated exemplar). Thus, the results for 007Q versus 007K2A were only from a single distribution/mailing, but are reported in Fig. 2 for completeness (60% of all decisions were considered accurate, and the consensus was computed to be 0.3951, however, the research team makes no claim regarding these statistics, which may be an artifact of sample size). The comparison of 007Q versus 007K2B is based on 65 examiners, yielding more representative statistics, including an accuracy of 77% for conclusions reported as either *exclusion*, *indications of non-association*, or *limited association* (which is equivalent to the IQR), but a lower consensus of 0.6288, owing to decisions spread across a larger number of categories. Inspection of Fig. 15 illustrates that there are several differences in characteristics of use between the questioned and known test impression, despite the fact that individual pattern and tread elements show considerable agreement in size (consistent class/manufacturing characteristics). However, it is interesting to view overlays of 007Q versus the test impression from known nonmated shoe 007K2B, which reveals a possible manufacturing inconsistency. More specifically, an overlay can align the toe of both impressions, or the heel of both

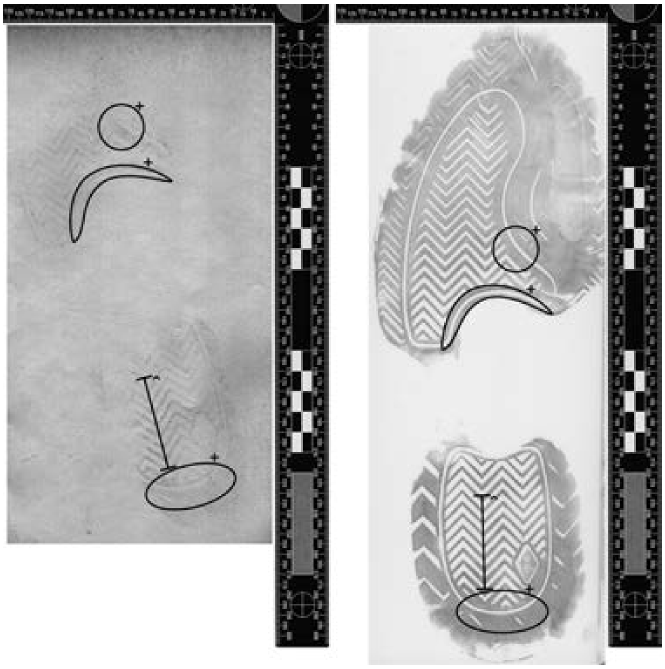


FIG. 11—*Case 006 questioned impression (left) compared to test impression from mated shoe 006K1 (right). Features of interest that are visible on the questioned impression are circled with a solid line (+ denotes a wear pattern, and ^ denotes spacing between herring bone elements). The solid lines on the test impression indicate that these features are also present on the true mate. Accuracy = 85.7% and consensus = 0.8189.*

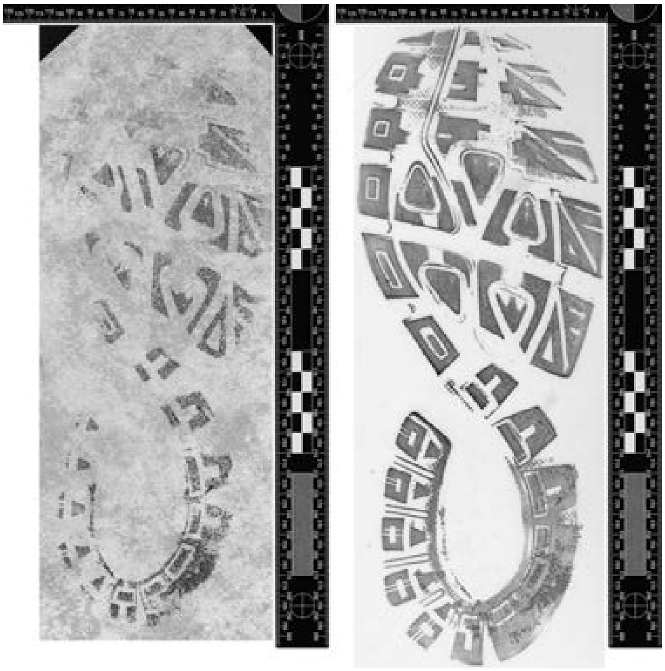


FIG. 13—*Case 007 questioned impression (left) compared to the questioned impression overlaid on a test impression from nonmated shoe 007K1 (right); note the disagreement in size. Accuracy = 97.1% and consensus = 0.9733. [Color figure can be viewed at wileyonlinelibrary.com]*

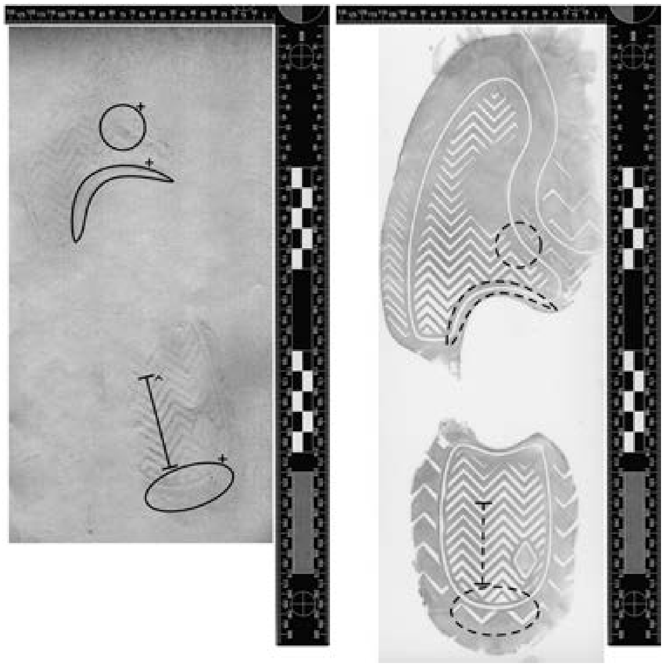


FIG. 12—*Case 006 questioned impression (left) compared to a test impression from nonmated shoe 006K2 (right). Features of interest that are visible on the questioned impression are circled with a solid line (+ denotes a wear pattern and ^ denotes spacing between herring bone elements). The dashed lines on the test impression indicate that these features are not present on the true nonmate. Accuracy = 72.9% and consensus = 0.7175.*

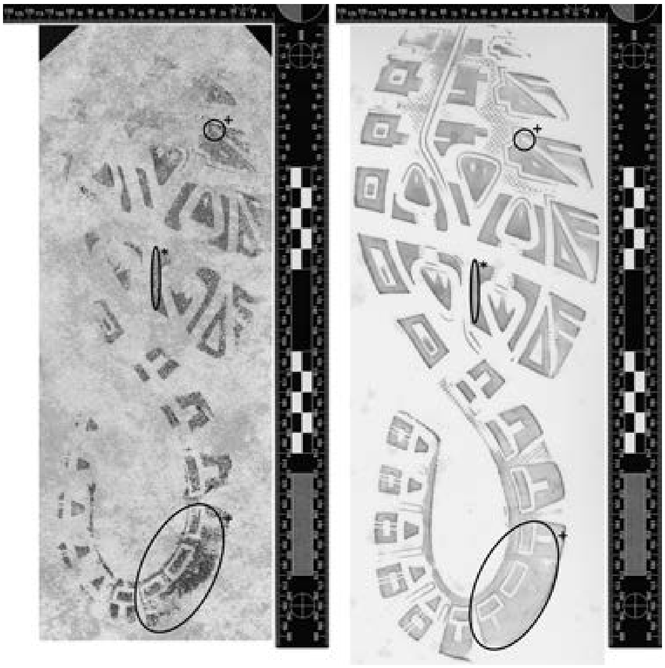


FIG. 14—*Case 007 questioned impression (left) compared to test impression from mated shoe 007K2A (right). Features of interest that are visible on the questioned impression are circled with a solid line (+ denotes a wear pattern and * denotes a RAC). The solid lines on the test impression indicate that these features are also present on the true mate. Accuracy = 60.0% and consensus = 0.3951, but results are only based on five examiners. [Color figure can be viewed at wileyonlinelibrary.com]*

impressions, but not the entire shoe (toe and heel) simultaneously (Fig. 16). This is presumed to be due to a manufacturing anomaly wherein the heel of the outsole was glued inconsistently in a rotated but rigid position to the midsole of individual shoes. Finally, if 007Q versus 007K1 and 007Q versus 007K2 are treated as a case, the overall accuracy in reporting would be 87%.

Detailed Case Analysis

In addition to the above accuracy summaries, a deep analysis of decisions falling outside the expected range was conducted for each case in an effort to ascertain the factors and observations that informed these conclusions. This examination is detailed in the Supplemental Information (Figs S1–S8 and Table S4 and S5) section that accompanies this manuscript.

Limitations

The results presented regarding accuracy and consensus should be considered within the confines of the following limitations. First, examiners were not permitted to review the shoe or make their own impressions based on study design. Second, the results are based on a limited number of using shoes of similar style (athletic) and size. Third, the impact of the number of knowns provided per case is a factor that has not been controlled in this study. For example, there were five cases with two knowns, and two cases with a single known, but the results do not allow the research team to determine if and/or how the number of knowns may have impacted accuracy and reproducibility. Moreover, the cases with a single known were all of the same substrate/matrix, which could further confound the results. Fourth, the overall quality of the questioned impressions would likely be considered moderate to high and does not include extremely low-quality impressions that may be encountered in casework. Fifth, some of the questioned impressions were created by a research analyst with a smaller foot size then most shoes, which may have inadvertently increased partiality in some questioned impressions. Sixth, all questioned impressions were created dynamically (walking), while the test impressions were created as static exemplars using a benchtop method, formerly described in (19). Lastly, the research team implemented a quality assurance mechanism to warrant that all printed images were produced using a one-to-one reproduction ratio. Although examiners should always validate the ruler/scale between impressions being compared to ensure equivalence in the print reproduction ratio before initiating their comparison, the research team had no intent to arbitrarily create and test whether examiners performed this step before beginning their analysis. Despite this goal, the team is aware of at least one packet that did not conform to the standard, which means other case packets may have likewise evaded quality control. If attempting to seek a silver lining based on this observation, one could argue that fortuitously, this may mean that the results are more robust than anticipated, testing several latent steps in the analysis and comparison process. However, it is equally reasonable to argue that this may have led to increased variation in the conclusions since the research team cannot guarantee that other packets failed to meet our quality expectations, nor can the authors guarantee that every analyst performed a full evaluation of the scale information provided in the impressions before beginning their comparisons.

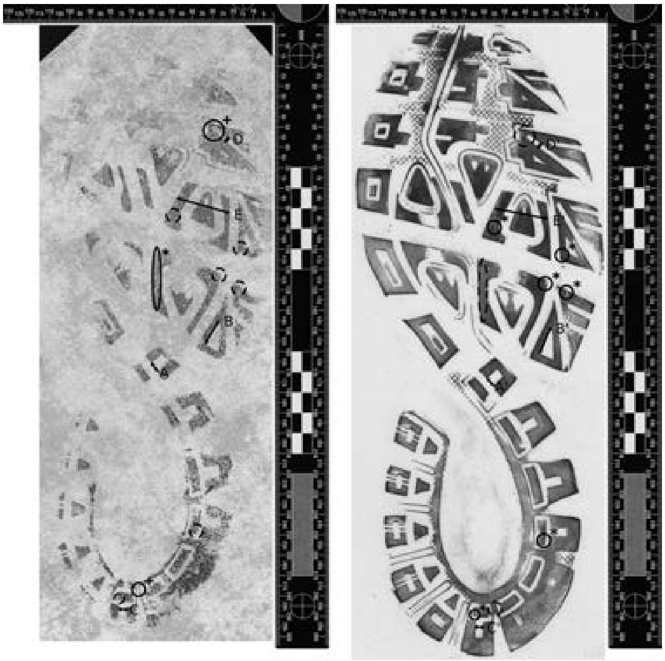


FIG. 15—*Case 007 questioned impression (left) compared to a test impression from nonmated shoe 007K2B (right). Features of interest that are visible are circled with a solid line (+ denotes a wear pattern and * denotes a RAC). The dashed lines on the test/questioned impression indicate that these features are not present in the nonmated pair. Agreement in size (mm): A = 4.80 ± 0.03 versus A′ = 4.77 ± 0.03, B = 13.22 ± 0.03 versus B′ = 13.21 ± 0.03, C = 3.56 ± 0.02 versus C′ = 3.57 ± 0.02, D = 2.32 ± 0.05 versus D′ = 2.34 ± 0.03, E = 26.84 ± 0.05 versus E′ = 26.83 ± 0.04. Accuracy = 76.9% and consensus = 0.6288. [Color figure can be viewed at wileyonlinelibrary.com]*

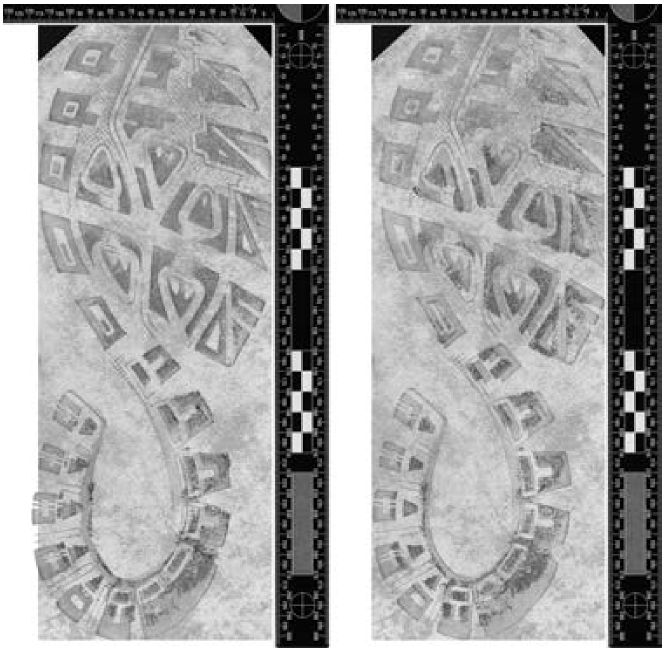


FIG. 16—*Case 007 questioned impression overlaid on a test impression from nonmated shoe 007K2B with the toe aligned (left) and additionally with the heel aligned (right). Note that due to a manufacturing anomaly wherein the heel outsole was glued in a rotated but rigid position to the midsole, the two shoes' impressions cannot be overlaid, despite otherwise having the same class characteristics. [Color figure can be viewed at wileyonlinelibrary.com]*

Conclusions

The concept of accuracy when dealing with a seven-point conclusion scale is by its very nature ambiguous. Accuracy implies right versus wrong, which can aptly handle a two-point system of *exclusion* or *identification* (after accounting for *insufficient detail*). However, it is challenging to view accuracy this way, and computed based purely on ground truth (mate or nonmate), for a seven-point categorical scale designed to express nuances in degree that can expand or limit a possible population of sources. Acknowledging this, the research team developed an accepted range of conclusions and computed accuracy accordingly, wherein expert accuracy ranged from a low of 55.7% to a high of 97.1%, with a mean of 82.8% ± 11.9% (a median of 85.7% and a 90% confidence interval between 80.5% and 84.9%). In other words, based on the impressions provided, ground truth, and adherence to the SWGTREAD (2013) conclusion standard (3), the forensic footwear community reached the expected conclusions an average of 83% of the time across a total of 835 comparisons (70 examiners × 12 comparisons = 840 less 5 reports for 007Q vs. 007K2A).

Again, acknowledging the existence of a seven-point conclusion scale, and after removal of *insufficient detail*, the remaining conclusions categories exist as an ordinal scale, ranging from strong to weak exclusionary statements, followed by weak to strong associative statements. As illustrated by Tastle and Wierman (15), agreement within an ordinal scale is poorly described by typical measures, such as the mean, standard deviation, and entropy. Thus, a more appropriate measure of dispersion was sought. First, the IQR reports the difference between the first and third quartile, essentially describing the spread or dispersion of at least the middle 50% of all conclusions per comparison. This is a useful measure of consistency, since it is insensitive to outliers. Disregarding 007Q versus 007K2A based on the limited sample size, in 10 of the remaining 12 comparisons, the IQR was consistent with the accepted range of conclusions. In other words, the bulk of the community conclusions were consistent with the research team's accepted range of conclusions. The two noted exceptions were the comparison of 001Q versus 001K2, and 006Q versus 006K2, wherein the IQR was larger than the accepted range. Shoe 001K2 was the known source of 001Q, and the research team anticipated a response of either *high degree of association* or *identification*, but the IQR included a weaker association or *association of class* which was selected by 17% of respondents. These reports are not incorrect, but based on the research team's assessments, a stronger association was warranted. Shoe 006K2 is not the source of 006Q, and the research team anticipated a response of either *exclusion* or *indications of non-association*, but the IQR included *limited association* which was also selected by 17% of respondents. Again, the research team anticipated a stronger disassociation than reported by some examiners.

If the IQR is used as the accepted range for accuracy, then the mean accuracy is 85.6% ± 11.1% (with a median of 89.3% and a 90% confidence interval between 83.5% and 87.6%). Given the research team's accepted range, the observed difference in mean accuracy is 2.8% (82.8% ± 11.9%), with a standard deviation of 16.3% (addition in quadrature). Assuming the difference is normally distributed around a mean of zero, then the observed value using IQR differs from the expected by 2.8/16.3 = 0.17 standard deviations, with a probability p(outside 0.2σ) of almost 85%, meaning a failure to detect any statistically significant difference in the accuracy estimates using either the IQR or the research team's defined range based on the assumption of normality (20). In other words, if the IQR is treated as the forensic footwear community's group decision for each comparison, then on average, the research team's expected range of conclusions (as the "oracle") is not statistically different from the community's group decision.

Using consensus and dissension to evaluate the dispersion in responses (which are metrics that are independent of the number of participants), and ignoring comparison 007Q versus 007K2A based on sample size, the remaining consensus measures range from a low of 0.5105 (for comparison 003Q vs. 003K1), a high of 0.9733 (for comparison 007Q vs. 007K1), with a mean of 0.7821 ± 0.1422 (median = 0.7743). In terms of mates and nonmates, the consensus among mated pairs equals 0.7421 ± 0.1516 (median of 0.7532), and the consensus among nonmated pairs equals 0.8106 ± 0.1396 (median of 0.7954). One less this value is a measure of dispersion, and both collectively describe the reproducibility in responses when using an ordinal conclusion scale that varies between strong to weak disassociations and weak to strong associations.

This metric considers not only the proportion of responses within a selected category, but also the distance between each category. For the purposes of computation, the distance between categories was considered constant (i.e., the data and information required to transition from *exclusion* to *indications of non-association* are equal to the data and information required to transition from *indications of non-association* to *limited association*). Stated another way, the difference in temperature between 10° and 15° is the same as the difference between 15° and 20°. Although true for a relative evaluation of changes in temperature, this is unlikely to be true for the SWGTREAD (2013) conclusion standard (3) which is ordinal, but the scaling between each category is unknown. Moreover, without further study, use of an alternative set of distance weightings is equally speculative. However, these challenges should all be considered within the larger confines of defining conclusion standards. There is a clear trade-off in terms of a community's desire to have a nonbinary scale that can express fine variations in the degree of disassociation/association, the types of conclusions that allow for inclusion or exclusion of populations (shoes of the same make, model, and size), and the use of quantitative metrics to describe accuracy and reproducibility. To compare the accuracy and reproducibility measures from this study on footwear to other forensic fields (i.e., fingerprints or firearms) with three-point conclusion standards requires a data transformation not performed in this summary. Here, the research team reports to the forensic footwear community an estimate of accuracy and reproducibility within the confines of the community's accepted seven-point system, and readers are directed to the third summary of this series in order to compare the results presented here with those collected in other forensic pattern sciences with three-point scales.

Acknowledgments

The authors would like to express our sincerest gratitude to the forensic footwear examiners who generously shared their time and expertise in order to participate in this West Virginia University (WVU) Institutional Review Board (IRB)-approved (1602021821) study. Thank you for making this work possible.

References

1. Speir JA, Richetelli N, Hammer L. Forensic footwear reliability: part I – participant demographics and examiner agreement. J Forensic Sci 2020;65:1852–70.
2. President's Council of Advisors on Science and Technology (PCAST). Report to the President – Forensic science in criminal courts: ensuring scientific validity of feature-comparison methods. Washington, DC: Executive Office of the President of the United States, 2016.

3. SWGTREAD. Range of conclusions standard for footwear and tire impression examinations. Scientific Working Group for Shoeprint and Tire Tread Evidence. 2013. https://www.nist.gov/system/files/documents/2016/10/26/swgtread_10_range_of_conclusions_standard_for_footwear_and_tire_impression_examinations_201303.pdf (accessed July 31, 2020).
4. Chen C, White L, Kowalewski T, Aggarwal R, Lintott C, Comstock B, et al. Crowd-sourced assessment of technical skills: a novel method to evaluate surgical performance. J Surg Res 2014;187(1):65–71. https://doi.org/10.1016/j.jss.2013.09.024
5. Pittayapat P, Thevissen P, Fieuws S, Jacobs R, Willems G. Forensic oral imaging quality of hand-held dental X-ray devices: comparison of two image receptors and two devices. Forensic Sci Int 2010;194(1–3):20–7. https://doi.org/10.1016/j.forsciint.2009.09.024
6. Dong Y, Zha Q, Zhang H, Kou G, Fujita H, Chiclana F, et al. Consensus reaching in social network group decision making: research paradigms and challenges. Knowl-Based Syst 2018;162:3–13. https://doi.org/10.1016/j.knosys.2018.06.036
7. Dowding D, Thompson C. Measuring the quality of judgement and decision-making in nursing. J Adv Nurs 2003;44(1):49–57. https://doi.org/10.1046/j.1365-2648.2003.02770.x
8. Keeney S, Hasson F, McKenna H. Consulting the oracle: ten lessons from using the Delphi technique in nursing research. J Adv Nurs 2006;53(2):205–12. https://doi.org/10.1111/j.1365-2648.2006.03716.x
9. Eubank BH, Mohtadi NG, Lafave MR, Wiley JP, Bois AJ, Boorman RS, et al. Using the modified Delphi method to establish clinical consensus for the diagnosis and treatment of patients with rotator cuff pathology. BMC Med Res Methodol 2016;16(1):56. https://doi.org/10.1186/s12874-016-0165-8
10. Diamond IR, Grant RC, Feldman BM, Pencharz PB, Ling SC, Moore AM, et al. Defining consensus: a systematic review recommends methodologic criteria for reporting of Delphi studies. J Clin Epidemiol 2014;67(4):401–9. https://doi.org/10.1016/j.jclinepi.2013.12.002
11. Connell LE, Carey RN, deBruin M, Rothman AJ, Johnston M, Kelly MP, et al. Links between behavior change techniques and mechanisms of action: an expert consensus study. Ann Behav Med 2018;53(8):708–20. https://doi.org/10.1093/abm/kay082
12. Saralioglu E, Gungor O. Use of crowdsourcing in evaluating post-classification accuracy. Eur J Remote Sens 2019;52(Suppl 1):137–47. https://doi.org/10.1080/22797254.2018.1564887
13. Best-Rowden L, Jain AK. Learning face image quality from human assessments. IEEE Trans Inf Forensics Secur 2018;13(12):3064–77. https://doi.org/10.1109/TIFS.2018.2799585
14. White LW, Kowalewski TM, Dockter RL, Comstock B, Hannaford B, Lendvay TS. Crowd-sourced assessment of technical skill: a valid method for discriminating basic robotic surgery skills. J Endourol 2015;29(11):1295–301. https://doi.org/10.1089/end.2015.0191
15. Tastle WJ, Wierman MJ. Consensus and dissention: a measure of ordinal dispersion. Int J Approx Reason 2007;45(3):531–45. https://doi.org/10.1016/j.ijar.2006.06.024
16. Akiyama Y, Nolan J, Darrah M, Rahem M, Wang L. A method for measuring consensus within groups: an index of disagreement via conditional probability. Inform Sciences 2016;345:116–28. https://doi.org/10.1016/j.ins.2016.01.052
17. President's Council of Advisors on Science and Technology (PCAST). Report to the President – Forensic science in criminal courts: ensuring scientific validity of feature-comparison methods. Forensics Addendum. Washington, DC: Executive Office of the President of the United States, 2017.
18. Freund RJ, Wilson WJ, Mohr DL. Statistical methods, 3rd edn. Burlington, MA: Academic Press, 2010;30–6.
19. Speir JA, Richetelli N, Fagert M, Hite M, Bodziak WJ. Quantifying randomly acquired characteristics on outsoles in terms of shape and position. Forensic Sci Int 2016;266:399–411. https://doi.org/10.1016/j.forsciint.2016.06.012
20. Taylor JR. An introduction to error analysis: the study of uncertainties in physical measurements, 2nd edn. Sausalito, CA: University Science Books, 1997;19–24, 149–51.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Summary information for case 001 (00K1 KNM, 001K2 KM).

Figure S2. Summary information for case 002 (002K1 KM).

Figure S3. Summary information for case 003 (003K1 KM, 003K2 KNM).

Figure S4. Enumeration of characteristics of use, corresponding to Tables S4 and S5 for 003Q versus 003K1.

Figure S5. Summary information for case 004 (004K1 KNM, 004K2 KM).

Figure S6. Summary information for case 005 (005K1 KNM).

Figure S7. Summary information for case 006 (006K1 KM, 006K2 KNM).

Figure S8. Summary information for case 007 (007K1 KNM, 007K2A KM, 007K2B KNM).

Table S1. Summary of accurate conclusions and the 90% confidence interval on the estimated proportion of accurate conclusions using the exact Clopper-Pearson confidence interval (1) ($np > 10$ but $n(1\text{-}p) > 10$ was only true in 8 out of 15 instances).

Table S2. Summary of accurate conclusions and the 90% confidence interval on the estimated proportion of accurate conclusions using the exact Clopper-Pearson confidence interval (1) ($np > 10$ but $n(1\text{-}p) > 10$ was only true in 8 out of 15 instances).

Table S3. Summary of consensus (mean, standard deviation (SD), median, standard error (SE) and range) where range equals $C + t \times \text{SE}$ where t represents the appropriate t-value for 90% confidence from the t-distribution for degrees of freedom = n-1.

Table S4. Summary of the number of times eight features of interest (illustrated in Fig. S4) were marked by 53 total examiners, reaching conclusions of association, high degree of association and identification, when comparing 003Q versus 003K1.

Table S5. Summary of eight features of interest (wear and RACs) present in comparison 003Q versus 003K1 marked by 53 total experts that reached a conclusion of at least association of class.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14552
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Nicole Richetelli,[1] *M.S.; Lesley Hammer,*[2] *M.S.; and Jacqueline A. Speir,*[1] *Ph.D.*

Forensic Footwear Reliability: Part III—Positive Predictive Value, Error Rates, and Inter-Rater Reliability*

ABSTRACT: Over the course of 19 months, West Virginia University collected reports from 70 footwear experts, each performing 12 questioned-test comparisons, resulting in a dataset that includes more than 1000 examiner attributes (education, training, certification status, etc.), 3500 impression features identified and evaluated (clarity, totality, and similarity), and 840 source conclusions. The results were used to estimate the performance of forensic footwear examiners in the United States, including error rates, predictive value (PV), and measures of inter-rater reliability (IRR). For the dataset and mate-prevalence (31.5%) used in this study, results indicate correct predictive value varies from 94.5% for *exclusions*, 85.0% for *identifications*, and between 70.1% and 65.2% for *limited associations* and *association of class*, respectively (with all other conclusions producing PVs between these extremes). After data transformation based on ground truth, the case study materials show a false-positive rate of 0.48%, a false-negative rate of 15.6%, a (correct) positive predictive value of 98.8%, and a (correct) negative predictive value of 93.3%. In addition to error rates and PVs, inter-rater reliability was likewise computed to describe examiner reproducibility; results indicate a Gwet AC_2 agreement coefficient of 0.751–0.692 when using a six- and four-level reporting structure, respectively, which translates into "substantial" and "moderate agreement" for a benchmarked verbal equivalent scale. The reported performance metrics are further compared against past forensic footwear reliability studies, including a discussion of how the use of a six-level reporting structure impacts results.

KEYWORDS: forensic footwear evidence, reliability, footwear examiners, error rates, predictive value, inter-rater reliability

Between February 2017 and August 2018, West Virginia University (WVU) conducted a reliability study to investigate performance among forensic footwear examiners in the United States. The goals of this project were several-fold, and preliminary results were divided into three separate publications. Part I (1) described the reproducibility in feature identification, feature evaluation, and generalized community agreement, while Part II (2) reported on examiner accuracy and reproducibility (or consensus) when using a seven-point conclusion scale (i.e., the Scientific Working Group for Shoeprint and Tire Tread Evidence [SWGTREAD] 2013 conclusion standard [3]). This summary (Part III) transforms the results in order to allow for a direct comparison between the performance of forensic footwear experts, and the performance reported for other forensic pattern disciplines that employ a three-point conclusion standard (e.g., *identification/individualization*, *exclusion*, and *inconclusive*).

According to the President's Council of Advisors on Science and Technology (PCAST) 2016 report, accuracy is defined as the known probability (or frequency) at which "an examiner obtains correct results both (1) for samples from the same source (true positives) and (2) samples from different sources (true negatives)" (4). In the presence of binary judgments (i.e., *identification/individualization* and *exclusion*) for known mates and known nonmates, the computation of accuracy is rather straightforward (at least after deciding how to handle *inconclusive* conclusions). However, at least half of the forensic footwear community participating in this reliability study reported direct use of a seven-point conclusion standard (1), while the remaining half used a scale similar to the SWGTREAD (2013) conclusion standard (3) but with some modification (four or five categories vs. seven, variants in the labels used to describe categories, etc.). Conversely, only 10% of the respondents for this study noted that their laboratories use a three-point scale for reporting footwear conclusions.

The benefit of a seven-point conclusion standard within the footwear community is the ability to succinctly describe the population of shoes that could have contributed to the questioned impression. From this "population" vantage-point, the scale is "U-shaped." At either extreme (i.e., *exclusion* and *identification*) the population is exact, while the internal categories permit associations or disassociations between a given shoe, and any other shoe of the same make, model, size, etc. Unfortunately, this increased degree of freedom in expression complicates any computation of accuracy since there are endless situations when, for example, an *association of class* is a valid conclusion for a

[1] West Virginia University, 208 Oglebay Hall, PO Box 6121, Morgantown, WV, 26506.

[2] Hammer Forensics, LLC, 10601 Prospect Drive, Anchorage, AK, 99507.

Corresponding author: Jacqueline A. Speir, Ph.D. E-mail: Jacqueline.Speir@mail.wvu.edu

*This project was supported by Award No. 2016-DN-BX-0152, awarded by the National Institute of Justice (NIJ), Office of Justice Program, U.S. Department of Justice. However, the opinions, findings, conclusions, and recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the Department of Justice.

Received 18 Mar. 2020; and in revised form 15 June 2020; accepted 31 July 2020.

known nonmated shoe (i.e., "…the known footwear is a possible source of the questioned impression… (and) other footwear with the same class characteristics observed in the impression are included in the population of possible sources") (3).

Given this complication, measurements of accuracy and reproducibility for this reliability study were tackled in two distinct manners. In a companion paper (Part II [2]), accuracy was reported as the degree to which footwear examiners reported conclusions deemed valid by the research team. Under this model, the research team was afforded an "oracle" status, defining the accepted/expected conclusions based on a combination of ground truth, observable features, feature reliability, and adherence to the SWGTREAD (2013) conclusion standard (3). Although this approach most closely upholds the community's desire to describe conclusions with a seven-point standard, it does not permit a direct comparison with other forensic pattern evidence fields that report conclusions with a three-point standard. Thus, a data transformation was desired that would allow for such a comparison, which is the topic of this summary (Part III). In addition to allowing this comparison, another goal of the following analysis is to illustrate the manner in which accuracy statistics and estimates of reliability may vary based on the number of conclusion categories permitted within a field. At present, there are jury studies and recommendation reviews designed to assess different types of reporting standards within the forensic field (5–11), and it is hoped that the data transformations provided here will give some insight into how metrics might vary depending on the acceptance of different standards, which may or may not be a factor to consider in reporting recommendations moving forward.

Materials and Methods

Participant Demographics

Participant demographics are fully characterized in Part I (1). In total, 115 forensic footwear examiners were recruited, and 77 submitted results (resulting in a 67% response rate). However, the remainder of all statistical analyses is based on the responses from 70 participants who reported that they had previously completed training and performed one or more footwear comparisons (the results from seven participants were excluded moving forward based on self-reports of never having completed a comparison, or still in-training).

Case Variety

The details of each simulated case are also fully described in Part I (1), but for convenience, Table 1 repeats limited information concerning the shoes, substrates, media and processing techniques used to create case materials. Note that across all seven cases, participants were required to perform 12 pairwise comparisons, and that each case consisted of 1200 PPI digital and print imagery, comprised of a single questioned impression, one-two exemplars (outsoles), and two Handiprint replicate exemplars per known shoe.

Case Analyses

Each participant received a package via USPS of all relevant case materials, including high resolution color prints, a set of blank acetates for overlay annotation, a CD containing the digital reporting software, a copy of the SWTREAD (2013) conclusion scale (3) and an instruction document (with additional weblinks to access digital copies of all case materials). Participants were asked to treat the simulated cases as if each were routine casework and given that no time had passed between collection of the questioned and test impressions.

Acceptable Range of Conclusions

Table 2 reports the acceptable range of conclusions agreed on after consideration of ground truth, the opinion of three researchers, the opinion of a practitioner partner, and the reliability limitations expressed by participants through self-reporting given the contrived nature of the research experiment (e.g., examiners did not prepare the exemplars themselves, and the actual shoe was not provided for each known; please see Part II [2] for additional details).

Data Preprocessing

Examiner performance statistics were evaluated using three approaches; in the first approach, performance estimates were determined based on six categories from the SWGTREAD (2013) conclusion standard (3): *exclusion*, *indications of non-association*, *limited association*, *association of class*, *high degree of association* and *identification* (disregarding the category of *insufficient detail* since all questioned samples in this dataset were of sufficient quality for comparison).

In the second approach, the SWGTREAD (2013) conclusions (3) were reformulated into a four-category conclusion hierarchy based on the size of the resulting population of outsoles that could have contributed the impression. These were labeled as *strong exclusions*, *weak exclusions*, *weak inclusions*, and *strong inclusions*. Using this transformation, *strong exclusions* were comprised of any comparisons that correlate with the traditional SWGTREAD *exclusion* category (3); accordingly, this is the most confident disassociation decision, eliminating a specific shoe (or type of shoe) from the population of possible sources

TABLE 1—*Shoes, substrates, media, and processing techniques used to create simulated case materials.*

Case	Manufacturer of Known(s)	Size & Style of Known(s)	Substrate of Unknown	Medium of Unknown	Processing of Unknown	# of Known(s)
001	Converse	All Star (9)	Ceramic tile	Blood	Leucocrystal violet	2
002	Nike	Lebron James (10)	Vinyl tile	Dust	Digitally enhanced gel lift	1
003	Nike	Rosherun (9)	Ceramic tile	Blood	Leucocrystal violet	2
004	Nike	Air Max (10.5)	Linoleum tile	Wax	Magnetic powder & gel lift	2
005	Nike	Air Max (11)	Vinyl Tile	Dust	Digitally enhanced gel lift	1
006	Nike	Air Max Cage (10)	Paper	Dust	Digitally enhanced	2
007	Under Armour	Unknown (10 & 11)	Ceramic Tile	Blood	Leucocrystal violet	2

TABLE 2—*Range of "accepted/expected" conclusions for each comparison. Note that this range was developed as a function of the features (quantity, quality, and clarity) present in the questioned impression, adherence to the SWGTREAD (2013) conclusion standard (3), and review of examiner reported limitations. Additional details can be found in Part II (2).*

Comparison	Expected/Accepted Conclusions
001Q vs. 001K1	Exclusion, Indications of Non-association
001Q vs. 001K2	High Degree of Association, Identification
002Q vs. 002K1	Limited Association, Association of Class
003Q vs. 003K1	Association of Class, High Degree of Association
003Q vs. 003K2	Exclusion
004Q vs. 004K1	Exclusion
004Q vs. 004K2	Identification
005Q vs. 005K1	Exclusion, Indications of Non-association, Limited Association
006Q vs. 006K1	Limited Association, Association of Class
006Q vs. 006K2	Exclusion, Indications of Non-association
007Q vs. 007K1	Exclusion
007Q vs. 007K2A	High Degree of Association, Identification
007Q vs. 007K2B	Exclusion, Indications of Non-association, Limited Association

that could have produced a given crime-scene impression. Conversely, any cases with expected outcomes of *indications of non-association* were considered *weak exclusions*, representing comparisons that reveal dissimilarities, but of insufficient type, quality or reliability to lead to a certain disassociation. In other words, dissimilarities were noted, but the exemplar could not be completely eliminated from the population of possible sources. Next, any cases with expected outcomes of *limited association* or *association of class* were considered *weak inclusions*. Decisions falling into this category contain the largest potential population of shoes that could have produced a questioned impression. Finally, any cases with expected outcomes of *high degree of association* and *identification* were considered *strong inclusions*, or confident associations based on the presence and reliability of associating features such as wear, Schallamach patterns, and randomly acquired characteristics (RACs). As a note for clarification, the descriptor of weak versus strong does not refer to the utility of the conclusion (a weak inclusion is a useful description of the population of shoes that could have produced a questioned impression). Instead, these are descriptions regarding the size of the resulting population—strong inclusions/exclusions have small source populations and weak inclusions/exclusions have larger source populations.

Finally, the last approach was to remap the SWGTREAD (2013) conclusion standard (3) into the Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST) 2013 three-level reporting scale used by fingerprint analysts (12) (i.e., *exclusion*, *inconclusive*, and *individualization*). The explicit purpose of this final transformation was to allow for a first order comparison against the 2011 FBI fingerprint black box study (13). To achieve this reformulation, footwear experts' decisions of *exclusion* and *indications of non-association* were re-assigned as *exclusions*, reports of *limited association* and *association of class* were reclassified as *inconclusive*, and finally, outcomes of *high degree of association* and *identification* were re-categorized as *individualization*. Moreover, ground truth (mates and nonmates) were used to create an appropriate confusion matrix. Although the authors acknowledge that this breakdown is an over-simplification of the conclusions that can be drawn in the field of forensic footwear evidence and that many may assert that these groupings are problematic (e.g., that *high degree of association* is not the same conclusion as *identification/individualization*), this segmentation nonetheless allowed for a reasonable comparison of expert error rates and predictive values among the fields of footwear and fingerprint evidence, which is the one of the primary aims of this summary.

Overall Accuracy as a Function of "Expected" Conclusions

Based on the SWGTREAD (2013) conclusion scale (3) (modified into six-, four- and three-levels), along with ground truth (mates and nonmates), and the quality, quantity, and reliability of observable features, expected conclusions were defined for each comparison (2). Using the expected outcomes (please see Part II [2] summary for additional details), a confusion matrix was prepared as a 2D frequency table of expected versus reported conclusions, and overall accuracy was computed as the number of comparisons out of 835 wherein examiners reached the expected conclusion(s) (70 examiners × 12 comparisons − 5 comparisons that were excluded owing to sample size limitations —see Part I and Part II for additional details [1,2]).

Error Rates and Predictive Value as a Function of "Expected" Conclusions and Ground Truth

Using the 2011 FBI fingerprint black box study (13) as a model, expert decision error rates and predictive values were computed for forensic footwear examiners. This was repeated using appropriately constructed four- and three-level confusion matrices, of which the three-level matrix most closely mimics the conclusion scale described by Ulery et al. (13). Table 3 outlines the adapted measures utilized in this study for evaluating expert performance via error rate and predictive value. For the data transformations employed here, converting a six-level conclusion standard into a four-level conclusion matrix presented the greatest complication, and is therefore further illustrated using Table 4.

First, the *correct within range* ($CW = \sum C|W$) describes the summation of all correct (C) decisions that fell within (W) the range of expected conclusions. For example, if interested in evaluating *weak inclusions*, then the correct within range includes all outcomes of *limited association* or *association of class* for cases in which these conclusions are expected/accepted (green cells in Table 4). In contrast, the *incorrect within range* ($IW = \sum I|W$) describes the summation of all incorrect (I) decisions that fell within (W) the range of expected conclusions. For example, if interested in evaluating *weak inclusions*, then the incorrect within range includes all outcomes of *limited association* or *association of class* for cases in which other conclusions are expected (orange cells in Table 4).

As a corollary to the *correct within range* (CW), the *correct outside range* ($CO = \sum C|O$) describes the summation of all correct (C) decisions that fell outside (O) the range of expected conclusions. Using the same example wherein *weak inclusions* are under evaluation, then the correct outside range outcomes include all decisions other than *limited association* or *association of class* for cases in which *limited association* or *association of class* were not expected outcomes (blue cells in Table 4). Finally, the *incorrect outside range* ($IO = \sum I|O$) describes the summation of all incorrect (I) decisions that fell outside (O) the range of expected conclusions. Again, when evaluating *weak inclusions*, incorrect outside range outcomes include all decisions other than *limited association* or *association of class* for cases in

TABLE 3—*Accuracy measures from the 2011 FBI black box study of fingerprint experts (13) and their adapted notation and description for this study using the SWGTREAD (2013) multi-level categorical outcomes (3).*

Fingerprint Study Binary Measures	Footwear Study Adapted Measures	Description
True-positive Rate (TPR)	Correct Within Rate (CWR)	The proportion of correct within range conclusions
False-positive Rate (FPR)	Incorrect Within Rate (IWR)	The proportion of incorrect within range conclusions
True-negative Rate (TNR)	Correct Outside Rate (COR)	The proportion of correct outside range conclusions
False-negative Rate (FNR)	Incorrect Outside Rate (IOR)	The proportion of incorrect outside range conclusions
Mate-prevalence (MP)	Within Prevalence (WP)	The proportion of all comparisons presented that have conclusions within the acceptable range
Nonmate-prevalence (NMP)	Outside Prevalence (OP)	The proportion of all comparisons presented that have conclusions outside the acceptable range
Positive Predictive Value (PPV)	Correct Predictive Value (CPV)	The proportion of correct decisions given the acceptable range of conclusions

TABLE 4—*Color-coded confusion matrix for evaluating performance on weak inclusion conclusions across 835 decisions reached by 70 examiners performing 12 comparisons (less 5 excluded based on sample size). In this example, the acceptable range of conclusions are limited association and association of class, and the color-coding represents decisions that are correct within range (CW; green), incorrect within range correct (IW; orange), correct outside of range (CO; blue), incorrect outside of range (IO; yellow), and excluded comparisons (EW) (pink).*

		Examiner Conclusion							
		Ins	Ex	Ind	Lim	Assoc	HD	ID	Total
Acceptable Range	Ex	0	199	7	2	2	0	0	210
	Ex/Ind	0	82	33	14	11	0	0	140
	Ex/Ind/Lim	1	69	26	15	22	2	0	135
	Lim/Assoc	2	7	8	46	69	7	1	140
	Assoc/HD	0	11	2	4	19	20	14	70
	HD/ID	0	2	0	5	12	28	23	70
	ID	0	0	0	1	0	7	62	70
	Total	3	370	76	87	135	64	100	835

which *limited association* or *association of class* are expected conclusions (yellow cells in Table 4).

Based on the above variable definitions, some conclusions should be excluded when attempting to determine predictive value. These *excluded comparisons* ($EW = \sum E|W$) describe all decisions that would be considered correct if they fell within *or* outside of the range of conclusions of interest and therefore should not influence the performance values associated with a particular conclusion category (for *weak inclusions* this includes the pink cells in Table 4). Finally, the *adjusted total* ($T_{\mathrm{A}} = T - EW$) is the summation of all decisions ($T = 835$), less the excluded comparisons.

Based on these definitions, the correct within rate (CWR) is similar to the true-positive rate, or the proportion of within range conclusions in which experts' decisions correctly fell within the expected range of conclusions (Eq. 1, left). Likewise, the incorrect within rate (IWR) is similar to the false-positive rate, or the proportion of out-of-range conclusions in which experts' decisions incorrectly fell within the range of expected conclusions (Eq. 1, right).

$$\mathrm{CWR} = \frac{CW}{\sum W - EW} \qquad \mathrm{IWR} = \frac{IW}{\sum O} \tag{1}$$

Since predictive value is also influenced by prevalence, within prevalence (WP) describes the proportion of all comparisons that have conclusions within the expected range (Eq. 2, left), and outside prevalence (OP) describes the proportion of all comparisons that have conclusions outside the expected range (Eq. 2, right).

$$\mathrm{WP} = \frac{\sum W - EW}{T_{\mathrm{A}}} \qquad \mathrm{OP} = \frac{\sum O}{T_{\mathrm{A}}} \tag{2}$$

Finally, the correct predictive value (CPV) is computed according to Eq. 3, which is the proportion of correct decisions given the expected range of conclusions modified by prevalence. This value was computed for the six-, four- and three-level conclusion scales previously described, wherein the three-level system most closely matches the performance statistics (true-positive rate, false-positive rate, positive predictive value, etc.) computed for the 2011 FBI fingerprint black box study (13).

$$\mathrm{CPV} = \frac{(\mathrm{WP} \times \mathrm{CWR})}{(\mathrm{WP} \times \mathrm{CWR}) + (\mathrm{OP} \times \mathrm{IWR})} \tag{3}$$

Inter-rater Reliability as a Measure of Expert Agreement

In addition to accuracy, the PCAST (2016) report highlights the importance of reproducibility (or agreement) in responses, defined as the probability (or frequency) at which "different examiners obtain the same result, when analyzing the same samples" (4). In an effort to inform this statistic, the footwear reliability results were reported using box plots, the interquartile range, and consensus (1,2). However, this summary seeks to employ a description of reproducibility based on a metric termed inter-rater reliability (IRR), which is well-suited to quantify the degree to which a series of comparisons of questioned and test impressions are categorized the same way when analyzed by different examiners. More specifically, high IRR means that examiners (raters) are interchangeable, which is desirable when the goal of the research study (and a criminal investigation/proceeding) is a measure of the similarity or dissimilarity between a questioned and test impression, regardless of the rater. Conversely, low IRR indicates that the individual rater or examiner plays a significant role in the categorization outcome (14), which in the case of forensic footwear comparisons suggests expert disagreement when presented with the same evidence, which often results in reduced clarity for the *trier-of-fact* tasked with interpreting the weight of evidence.

For nominal scales, agreement means that two raters provide identical conclusions. However, the concept of partial agreement exists when using an ordinal scale. For example, *identification* can be thought of as a *certain* conclusion, while *high degree of association* can be thought of as a *highly probable* conclusion. If some raters conclude *identification* and others conclude *high degree of association* these raters are not in total agreement, but they are also not in complete disagreement. Thus, the concept of partial agreement must be considered. In addition to partial agreement, a measure of IRR should account for chance. Agreement by chance is not considered false, but rather a "*bonus*" that inadvertently inflates an agreement metric since it is not based on an underlying process. Moreover, chance agreement is higher when fewer categories are provided (14), so the number of categories present in a scale must also be accounted for.

Given the ordinal reporting scale and the need to characterize partial and chance agreement, this summary employs the *weighted* Gwet AC_1 coefficient (or the AC_2 coefficient) as illustrated in Eq. 4 (14). The variable p_e denotes the percent chance agreement, while p_a denotes the percent realized agreement. The percent chance agreement is computed based on π_k which reports the fraction of examiners (raters) that compared questioned-test impression i and concluded k, across all comparisons n (or 835).

$$AC_2 = \frac{p_a - p_e}{1 - p_e}, \text{ where } \left(\begin{array}{c} p_a = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{q} \frac{r_{ik}(r_{ik}^* - 1)}{r_i(r_i - 1)} \\ p_e = \frac{T_w}{q(q-1)} \sum_{k=1}^{q} \pi_k (1 - \pi_k) \end{array} \right)$$
$$r_{ik}^* = \sum_{l=1}^{q} w_{kl} r_{il}$$
$$\pi_k = \frac{1}{n} \sum_{i=1}^{n} \frac{r_{ik}}{r_i} \tag{4}$$

Conversely, the percent realized agreement is computed based on r_{ik}^* which describes the number of examiners that performed comparison i and reported a conclusion k, combined with any other conclusions l that are in partial agreement with k. The degree of partial agreement is a function of a weighting factor w_{kl} wherein raters are penalized *less* for reaching decisions in categories directly adjacent to one another and *more* for decisions separated by several categorical levels (14). Although various weighting options exist (quadratic, ordinal, linear, etc.) an ordinal weighting system was employed in this analysis. The weight factor w_{kl} for two categories of interest (k and l) can be computed according to Eq. 5, where q represents the total number of categories into which conclusions can be classified ($q = 6$ for this study after removal of *insufficient detail*), and M_{kl} and $M_{\max}$ are combinations, or the number of combinations of 2 out of $(\max(k,l) - \min(k,l) + 1)$, and 2 out of q (or 15 for this study), respectively. Finally, T_w is the total of all weight factors (or 26.67 when using a six-level reporting structure and 11.00 when using a four-level reporting structure) (14).

$$w_{kl} = \begin{cases} 1 - M_{kl}/M_{\max} & \text{if } k \neq l \\ 1 & \text{if } k = l \end{cases}$$
$$M_{kl} = \binom{\max(k,l) - \min(k,l) + 1}{2}$$
$$M_{\max} = \binom{q}{2} = \frac{6!}{2!(6-2)!} = 15 \tag{5}$$
$$T_w = \sum_{l=1}^{q} \sum_{k=1}^{q} w_{kl} = 26.67 \text{ or } 11.00$$

Increasing numerical values of the AC_2 coefficient indicate increasing levels of examiner agreement. Moreover, with proper benchmarking, the magnitude of the coefficient can be related to a verbal scale, as illustrated in Table 5 (14,15). The purpose of benchmarking is to calibrate or revise Table 5 while accounting for study design (number of raters, number of comparisons, and number of response categories). To perform benchmarking, Gwet (14) proposes a four-step process. First, the agreement coefficient and its standard error are computed (see Supplemental Information Equation S1 for computation of standard error for AC_2). Second, the interval membership probability (IMP) is computed as illustrated in Eq. 6 (assuming a normal distribution) for each interval (a,b) in the verbal equivalent scale (Table 5). Third, cumulative probabilities are computed, starting from the highest benchmark level ("almost perfect"). Finally, the reported verbal equivalent for agreement for a specific study (i.e., "moderate", "substantial", etc.) is found to be equal to the agreement category that contains the smallest cumulative probability exceeding 0.95 (14).

$$\text{IMP} = P\left(\frac{AC_2 - b}{\text{SE}} \le Z \le \frac{AC_2 - a}{\text{SE}} \right) \tag{6}$$

As a matter of record, it is important to note that there are numerous agreement coefficients that can be computed. Examples include Cohen's Kappa, Fleiss' generalized Kappa, Conger's generalized Kappa, Krippendorff's alpha, Brennan—Prediger coefficient, and Scott's Pi. (14). Some are applicable to two raters, others to two or more raters, and some suffer from paradoxes; for example, Kappa can be paradoxically low when compared to the percent agreement apparent in a study. Despite this, and the fact that Gwet's AC_1 and AC_2 are paradox-resistant, Kappa is still a fairly common metric reported in the literature when examining decision tasks. Table 6 reports just a few examples (16–21) that give a rough idea of the types of tasks amenable to description using IRR, and the magnitude of the resulting agreement coefficients. Although this is helpful to

TABLE 5—*Landis and Koch's (14,15) verbal equivalent benchmark scale used to communicate the degree of agreement between raters as a function of the computed agreement coefficient.*

Agreement Coefficient	Degree of Agreement
<0.00	Poor
0.00 to 0.20	Slight
0.21 to 0.40	Fair
0.41 to 0.60	Moderate
0.61 to 0.80	Substantial
0.81 to 1.00	Almost Perfect

allow for a contextual interpretation of the agreement coefficient computed for footwear experts, it should be noted that few studies appear to benchmark their verbal equivalents. This is unfortunate since without knowledge regarding a coefficient's uncertainty, it cannot be fully interpreted (e.g., a coefficient of 0.5 ± 0.05 is not the same as 0.5 ± 0.3) (14). Nonetheless, Table 6 provides some insight into the magnitude of agreement coefficients reported in the literature for small expert studies (largest sample size of 15) and small to modest-sized novice studies (largest sample size of 120).

Results and Discussion

Overall Accuracy as a Function of "Expected" Conclusions

A summary of footwear analyst decisions on 12 questioned-known impression comparisons, as a function of the expected/acceptable range of conclusions, is provided in Table 7. Across all 835 comparisons conducted, overall accuracy was computed to be 82.8% ± 11.9% (median of 85.7% and 90% confidence interval (using the Clopper-Pearson exact method [22]) of 80.5–84.9%, excluding comparison 007Q vs. 0072KA based on sample size). This can be further divided among mates and nonmates (Table S1 and S2). For mated pairs, accuracy equals 76.3% ± 13.0% (median of 78.6% and 90% confidence interval [22] of 72.2–80.0%), and for nonmated pairs, accuracy equals 87.4% ± 9.24% (median of 91.4% and 90% confidence interval [22] of 84.7–89.8%); please see Part II (2) for additional details.

Error Rates and Predictive Value

Six-Level Data Breakdown: Predictive Value as a Function of "Expected" Conclusions

An examination of error rates and predictive value (as a function of the acceptable range of conclusions) was conducted for each of six SWGTREAD (2013) (3) decision categories (Table 8). Most notably, performance on the extremes of the SWGTREAD (2013) (3) scale (*exclusion* and *identification*) exhibit the highest correct within rates (0.85 for *exclusion* and 0.76 for *identification*) while the correct outside rates show less variation across all categories (all greater than 0.92 for the combined dataset). Conversely, as decisions become less certain (the population of shoes that could have contributed the impression increases), expert correct within rate performance declines (0.49–0.61).

With regard to correct predictive value as a function of mate-prevalence within this study, the highest correct predicted values for the combined dataset are 0.95 and 0.85, again for *exclusions* and *identifications*, respectively. Conversely, as decisions become less certain (the population of shoes that could have contributed the impression increases), expert performance declined, with *limited association* and *association of class* exhibiting the lowest correct predictive values (0.70 and 0.65, respectively). In other words, the chance that an examiner's reported outcome is correct given the acceptable range of conclusions of *limited association* and *association of class* was reduced to between 70% and 65%. With regard to *limited association* decisions, there were 560 comparisons in this study in which *limited association* was not considered a valid conclusion, and of these 560 comparisons, this category was incorrectly selected 26 times (by 20 unique/different examiners). Of the 20 different experts that committed this error, one incorrectly reached *limited association* four times and three additional experts reached this decision falsely twice (meaning almost 40% of these errors were committed by four experts [1 expert × 4 errors + 3 experts × 2 errors = 10/26]) which is less than 6% of all participants. Moreover, for approximately 60% of the 26 incorrect decisions, examiners failed to exclude the known shoe as the source of the impression (while the remainder failed to reach a stronger association). Both scenarios indicate an incomplete analysis; analysts did not reach more certain conclusions (either associative or dissociative) despite the presence of discriminating features.

Similarly, for *association of class* decisions, there were 625 comparisons in this study in which *association of class* was not considered a valid conclusion, and of these 625 comparisons, this category was incorrectly selected 47 times (by 30 unique/different examiners). Of the 30 different experts that committed this error, one incorrectly reached *association of class* five times, two experts incorrectly reached this conclusion four times, one incorrectly reached this conclusion three times, and an additional five experts incorrectly reached this conclusion twice (meaning 55% of these errors were committed by nine experts [1 expert × 5 + 2 experts × 4 + 1 expert × 3 + 5 experts × 2 = 26/47]) which is less than 13% of all participants. Moreover, 74% of the

TABLE 6—*Examples of IRR coefficients for experts and novices performing various design tasks.*

Paper	Assessment Topic	# of Raters	# Categories/# Conclusions per Case	# of Cases	Coefficient	Results	Verbal Equivalent(s)
Onate et al. (16)	LESS Landing Assessment	1 expert, 1 novice	2/15	19	Fleiss' Kappa	κ = 0.46–1.00	Moderate to Perfect
Andreasen et al. (17)	Medical Claim Compensation	15 experts	2/6	12	Fleiss' Kappa Gwet's AC_1	κ = 0.41–0.53 AC_1 = 0.43–0.54	Moderate
Gschließer et al. (18)	Diagnosis of Retinopathy	7 experts	6/1 2/2	52	Fleiss' Kappa	κ = 0.26–0.55	Fair to Moderate
Acklin & Fuger (19)	Criminal Court Decisions	3 experts	3/3	150	Fleiss' Kappa	κ = 0.24–0.81	Fair to Substantial
		3 experts	3/3	150	Krippendorff's Alpha	α = 0.18–0.51	Poor to Moderate
		2 "experts"*	2/3	150	Cohen's Kappa	κ = 0.35–1.00	Fair to Perfect
Nawrocka et al. (20)	Stage of Decomposition	120 novices	13/1 12/1 10/1	12	Krippendorff's Alpha	α = 0.81–0.85	N/A
Lee et al. (21)	Automobile Color	6 novices	18/1	1000	Fleiss' Kappa	κ = 0.22–0.98	Fair to Very Good

*Agreement was computed between two "decisions," which were a judge's verdict versus a pooled consensus decision from either two or three experts. Also note that Fleiss' Kappa is an extension of Cohen's Kappa for more than two raters and that none of these studies benchmarked their verbal scales.

TABLE 7—*Confusion Matrix of footwear expert decisions as compared to the acceptable range of conclusions for 835 total comparisons, exhibiting an overall accuracy of approximately 82.8% (shaded cells; 691/835). Please see Tables S1 and S2 for similar confusion matrices describing mates and nonmates.*

		Examiner Conclusion							
		Insufficient	Exclusion	Indications	Limited	Association	High Degree	Identification	Total
Acceptable Range	Exclusion	0	199	7	2	2	0	0	210
	Exclusion & Indications	0	82	33	14	11	0	0	140
	Exclusion & Indications & Limited	1	69	26	15	22	2	0	135
	Limited & Association	2	7	8	46	69	7	1	140
	Association & High Degree	0	11	2	4	19	20	14	70
	High Degree & Identification	0	2	0	5	12	28	23	70
	Identification	0	0	0	1	0	7	62	70
	Total	3	370	76	87	135	64	100	835

TABLE 8—*Computed error rates and predictive values per SWGTREAD (2013) (3) category, conducted by 70 forensic footwear experts as a function of the acceptable range of conclusions, as outlined in Table 7. The first six rows of the table report metrics based on 835 comparisons (a mixture of mates and nonmates denoted as "Combined") while the remaining seven rows compute the same metrics, but after dividing the dataset into nonmates (485) and mates (350). Metrics missing in the table for a specific ground truth/conclusion scenario do not exist (i.e., there were zero correct within (CW) decisions for a nonmate with an expected conclusion of identification).*

			90% Confidence			90% Confidence					90% Confidence	
Ground Truth	Conclusion	CWR	CWR		COR	COR		WP	OP	CPV	CPV	
Combined	Exclusion	0.8516	0.8196	0.8797	0.9429	0.9180	0.9618	0.5401	0.4599	0.9459	0.9244	0.9616
Combined	Indications	0.5413	0.4581	0.6228	0.9696	0.9548	0.9806	0.1629	0.8371	0.7763	0.6954	0.8406
Combined	Limited	0.5496	0.4671	0.6300	0.9536	0.9361	0.9673	0.1654	0.8346	0.7011	0.6242	0.7681
Combined	Association	0.6111	0.5395	0.6792	0.9248	0.9051	0.9414	0.1873	0.8127	0.6519	0.5920	0.7074
Combined	High Degree	0.4898	0.4028	0.5773	0.9770	0.9652	0.9855	0.1236	0.8764	0.7500	0.6589	0.8234
Combined	Identification	0.7589	0.6832	0.8241	0.9784	0.9670	0.9866	0.1388	0.8612	0.8500	0.7868	0.8970
Nonmates	Exclusion	0.8516	0.8196	0.8797	N/A	N/A	N/A	1.0000	0.0000	N/A	N/A	N/A
Nonmates	Indications	0.5413	0.4581	0.6228	0.9667	0.9383	0.9842	0.3417	0.6583	0.8939	0.8181	0.9405
Nonmates	Limited	0.3750	0.2473	0.5172	0.9543	0.9314	0.9711	0.1026	0.8974	0.4839	0.3572	0.6129
Mates	Limited	0.6479	0.5442	0.7420	0.9524	0.9206	0.9739	0.2527	0.7473	0.8214	0.7308	0.8863
Mates	Association	0.6111	0.5395	0.6792	0.9143	0.8648	0.9498	0.5070	0.4930	0.8800	0.8213	0.9212
Mates	High Degree	0.4898	0.4028	0.5773	0.9333	0.8977	0.9592	0.3182	0.6818	0.7742	0.6846	0.8442
Mates	Identification	0.7589	0.6832	0.8241	0.9286	0.8921	0.9555	0.3478	0.6522	0.8500	0.7885	0.8960

COR, correct outside rate; CPV, correct predictive value; CWR, correct within rate; OP, outside prevalence; WP, within prevalence.

47 incorrect decisions were failures to exclude a shoe as the source of the impression.

Four-Level Data Breakdown: Predictive Value as a Function of "Expected" Conclusions

Similarly, an examination of error rates and predictive value (as a function of the acceptable range of conclusions) was conducted after remapping the 835 outcomes into a four-level reporting structure as illustrated in Table 9. Using this transformation, *strong exclusions* were comprised of any comparisons that correlate with the traditional SWGTREAD *exclusion* category (3). Conversely, any cases with expected outcomes of *indications of non-association* were considered *weak exclusions*, and any expected outcomes of *limited association* or *association of class* were considered *weak inclusions*. Finally, any cases with expected outcomes of *high degree of association* or *identification* were considered *strong inclusions*.

As expected, performance based on this transformation was very similar to use of the traditional six-level SWGTREAD (2013) standard (3). Again, the extremes exhibited higher correct within rates and higher predictive values (with the latter a function of mate-prevalence provided in this study). More specifically, there was a 95% chance that an examiner's reported outcome would be correct given an expected *strong exclusion* and an 81% chance that an examiner's reported outcome would be correct given an expected *strong inclusion* (Table 9). Again, the less certain decision categories (*weak exclusions* and *weak inclusions*) exhibit lower performance with the correct predictive value ranging between approximately 67–78% for the combined dataset.

Three-Level Data Breakdown: Predictive Value as a Function of Ground Truth

Using a three-level reporting structure, decisions were remapped into *identifications*, *exclusions,* and *inconclusive*. More specifically, all conclusions of *exclusion* and *indications of non-association* were remapped to *exclusions* and all decisions of *identification* and *high degree of association* were remapped to *identification*. The remaining 225 decisions (out of 835) were considered *inconclusive*, corresponding to *insufficient detail*, *limited association*, or *association of class* reports. Since the extremes of *identification* and *exclusion* can only be reached for known mates and known nonmates, respectively, the predictive values from this transformation directly correspond to ground truth. Table 10 reports the confusion matrix based on this data transformation; in total, two false positives occurred resulting in a false-positive rate (FPR) of approximately 0.5% (FPR = 2/418). As a note for comparison, this same metric was reported as 0.1% (or six false inclusions) for past studies in fingerprint analysis (13).

In this study, both instances of false positives were committed for case 007K2B (see Part I and Part II summaries [1,2] for comparison details). For reporting, both examiners reached *high*

TABLE 9—*Computed error rates and predictive values per four-level confidence-based category, conducted by 70 forensic footwear experts as a function of the acceptable range of conclusions, as outlined in Table 7. The first four rows of the table report metrics based on 835 comparisons (a mixture of mates and nonmates denoted as "Combined") while the remaining five rows compute the same metrics, but after dividing the dataset into nonmates (485) and mates (350). Metrics missing in the table for a specific ground truth/conclusion scenario do not exist (i.e., there were zero correct within (CW) decisions for a nonmate with an expected conclusion of strong inclusion).*

Ground Truth	Conclusion	CWR	90% Confidence CWR		COR	90% Confidence COR		WP	OP	CPV	90% Confidence CPV	
Combined	Strong Exclusion	0.8516	0.8196	0.8797	0.9429	0.9180	0.9618	0.5401	0.4599	0.9459	0.9244	0.9616
Combined	Weak Exclusion	0.5413	0.4581	0.6228	0.9696	0.9548	0.9806	0.1629	0.8371	0.7763	0.6954	0.8406
Combined	Weak Inclusion	0.7304	0.6745	0.7812	0.8585	0.8308	0.8832	0.2833	0.7167	0.6711	0.6275	0.7120
Combined	Strong Inclusion	0.7824	0.7239	0.8333	0.9520	0.9358	0.9650	0.2083	0.7917	0.8110	0.7614	0.8522
Nonmates	Strong Exclusion	0.8516	0.8196	0.8797	N/A	N/A	N/A	1.0000	0.0000	N/A	N/A	N/A
Nonmates	Weak Exclusion	0.5413	0.4581	0.6228	0.9667	0.9383	0.9842	0.3417	0.6583	0.8939	0.8181	0.9405
Nonmates	Weak Inclusion	0.8333	0.6233	0.9530	0.8629	0.8301	0.8914	0.0462	0.9538	0.2272	0.1825	0.2792
Mates	Weak Inclusion	0.7204	0.6612	0.7743	0.8472	0.7891	0.8942	0.5636	0.4364	0.8590	0.8138	0.8946
Mates	Strong Inclusion	0.7824	0.7239	0.8333	0.8199	0.7626	0.8680	0.5136	0.4864	0.8210	0.7753	0.8591

COR, correct outside rate; CPV, correct predictive value; CWR, correct within rate; OP, outside prevalence; WP, within prevalence.

degree of association, reporting agreement of class characteristics and wear. However, this is invalid; wear differences are apparent, and this comparison included a possible manufacturing anomaly wherein the heel portion of the outsoles for the known nonmate and questioned impression appear to be affixed to the midsole in slightly rotated (mismatched) orientations, further precluding agreement when the questioned and known impressions are overlaid.

The false-negative rate (FNR) was higher than the false-positive rate (approximately 16%), with 30 decisions (out of 192 possible *identifications*) falsely excluding a known mated shoe; this is approximately double the error rate observed for fingerprint analysts at 7.5% (13). In total, 23 different analysts committed these 30 errors, with five examiners committing the error twice and one analyst providing three false negatives, meaning 43% of these errors were committed by six experts (5 experts × 2 errors + 1 expert × 3 errors = 13/30) which is less than 9% of all participants. As a general observation, it appears that many of the false negatives or incorrect eliminations resulted from an improper characterization of impression size wherein a size difference (physical size and/or size and spacing of outsole elements) was reported when one did not exist (2).

After considering the observed error rates for this study, computation of predictive values (posterior probabilities) was conducted. Positive predictive value reports the percentage (or probability) of strong inclusionary decisions that are true mates, while negative predictive value describes the percentage (or probability) of strong exclusionary decisions that are true nonmates (13). This evaluation is important because ground truth is not known in casework for which error rates are desired, thus an understanding of this "likelihood of correctness" in research studies represents a useful alternative. Based on 610 comparisons, the correct predictive value equals 98.8%, and the negative predictive value equals 93.3% (when 31% of comparisons are conducted on known mates and 69% on known nonmates) (Table 11).

TABLE 10—*Confusion matrix of 610 forensic footwear decisions, reclassified as binary conclusions using ground truth, for direct comparison with the fingerprint black box study (13). Note than any examiner conclusions of either insufficient detail, limited association, or association of class have been excluded from analysis as these were reclassified as inconclusive outcomes.*

Confusion Matrix of Binary Footwear Conclusions		Examiner Conclusion Identification (Positive)	 Exclusion (Negative)	 Total
True Conclusion	Identification (Positive)	162	30	192
	Exclusion (Negative)	2	416	418
	Total	164	446	610

Across all possible mate-prevalences, Fig. 1 is a plot of PPV and NPV based on ground truth for the 610 conclusive comparisons conducted in this study based on standard error computations using the Clopper-Pearson (exact method) (22) and Standard logit confidence intervals for predictive values (23). As a point of comparison, the positive and negative predictive value for fingerprint examiners at a 62% mate-prevalence was previously found to be 99.8% and 86.6%, respectively (13). At this same prevalence, the corresponding footwear performance values are 99.7% and 79.6%, with 90% confidence intervals between 98.9–99.9%, and 74.8–83.7%, respectively.

Measure of Expert Agreement: Inter-Rater Reliability

After a detailed evaluation of forensic footwear examiner accuracy as a function of the acceptable range of conclusions for each comparison in this study, an assessment of reproducibility was conducted via computation of inter-rater reliability, specifically using the Gwet AC_2 agreement coefficient (14). The results of this agreement analysis are detailed in Table 12. After benchmarking the computed coefficient, a verbal interpretation of footwear examiner performance maps between *moderate* and *substantial* agreement depending on the data transformation (six- or four-level scale, combined, mated, and nonmated, all excluding decisions of *insufficient detail* which are not part of an ordinal scale).

Interestingly, the agreement coefficient is higher in magnitude when considering the six-level reporting structure versus the four-level structure for the combined dataset (0.75 vs. 0.69 respectively, in Table 12) and corresponding to *substantial* versus *moderate* agreement. This is likely a function of the weight matrix; analysts are penalized *less* for falling into adjacent bins when there are more possible outcome categories.

The same trend for higher inter-rater reliability holds for six-level nonmates versus four-level nonmates (0.88 vs. 0.80). However, the result is reversed for mates; six-level mates have an

TABLE 11—*Computed error rates and predictive values based on ground truth across 70 forensic footwear experts performing 610 comparisons.*

			90% Confidence			90% Confidence					90% Confidence	
Ground Truth	Conclusion	CWR	CWR		COR	COR		WP	OP	CPV	CPV	
Mates	Strong Inclusion	0.8438	0.7941	0.8852	0.9952	0.9850	0.9991	0.3148	0.6852	0.9878	0.9621	0.9962
Nonmates	Strong Exclusion	0.9952	0.9850	0.9991	0.8436	0.7941	0.8852	0.6853	0.3148	0.9327	0.9132	0.9481

COR, correct outside rate; CPV, correct predictive value; CWR, correct within rate; OP, outside prevalence; WP, within prevalence.

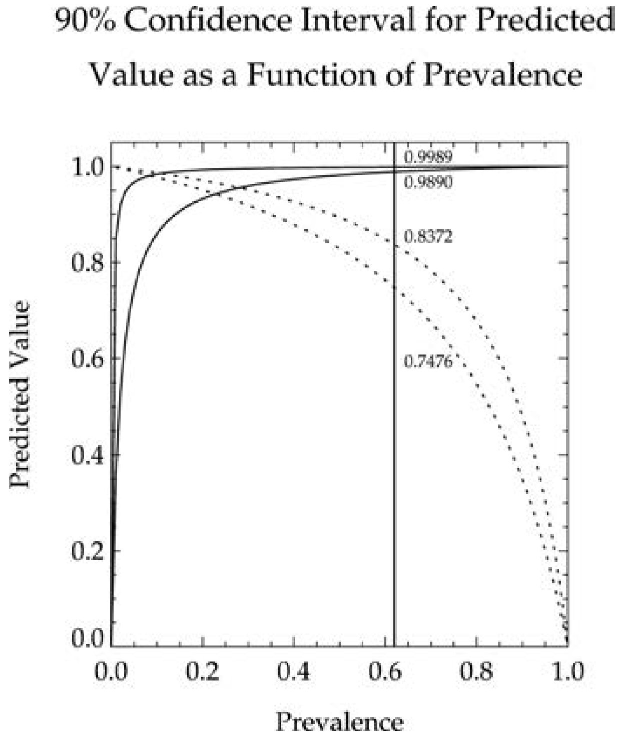


FIG. 1—*Plot of 90% confidence intervals (21,22) for positive (solid lines) and negative (dashed lines) predictive value as a function of mate-prevalence. As a point of comparison, the vertical line and reported performance metrics highlight PPV and NPV at a mate-prevalence at 62% (which corresponds to the 2011 FBI fingerprint study [13]).*

IRR of 0.66 and four-level mates 0.77). Overall, the highest inter-rater agreement is found using a six-level reporting structure for nonmates, and a four-level reporting structure for nonmates. Regardless of the subdivision, footwear examiners using the SWGTREAD (2013) conclusion standard (3) can be said to exhibit between *moderate* and *substantial* agreement (with the only other higher categorical option being *almost perfect*).

Conclusions

The primary goal of this summary was to compute error rates, predictive values, and inter-rater reliability for forensic footwear examiners participating in the WVU reliability study. Included within this goal were data transformations to allow for a comparison of reliability results across the forensic fingerprint and footwear communities. As a corollary, the secondary goal of this summary was to provide insight into how reporting structures may impact accuracy and reproducibility metrics for consideration moving forward, should revisions of conclusion standards occur downstream.

With regard to the primary goal, overall accuracy for the combined dataset was found to be 82.8% ± 11.9%, ranging from a low of 55.7% to a high of 97.1%. When using a six-level conclusion standard, correct predictive value was found to be 0.95 for *exclusions*, 0.85 for *identification*, and between 0.70 and 0.65 for *limited association* and *association of class*. After data transformation to a four-level conclusion hierarchy, similar results were found; correct predictive value equals 0.95 for *strong exclusions*, 0.81 for *strong inclusions*, and between 0.67 and 0.78 for *weak inclusions* and *weak exclusions*, respectively. Finally, using a three-level reporting structure that most closely mimics conclusion models utilized by fingerprint examiners (12), the data was transformed based on ground truth (known mates and known nonmates), allowing for computation of the study's false-positive rate, false-negative rate, and predictive value amenable to comparison with former reliability studies. These results indicate a FPR of 0.5% (vs. fingerprints at 0.1% [13]), a FNR of 16% (vs. fingerprints at 7.5% [13]), and a correct predictive value of 98.8% (mates) and 93.3% (nonmates) at a mate-prevalence of 31%. When corrected for the same mate-prevalence (62% in the fingerprint study [13]) the comparable footwear PPV is 99.7% (vs. fingerprints at 99.8% [13]) and NPV is 79.6% (vs. fingerprints at 86.6% [13]).

With regard to inter-rater reliability, the Gwet AC_2 agreement coefficient ranged from 0.88 (nonmates) to 0.65 (mates) when using a six-level ordinal reporting structure. For a four-level system, this varied between 0.80 and 0.77 for nonmates and mates, respectively. After benchmarking, all metrics translated into

TABLE 12—*Inter-rater reliability analysis results for the Gwet AC_2 agreement coefficient and the corresponding verbal equivalent for agreement after benchmarking. Computation of standard error (SE) is as described in Equation S1 of the Supplemental Information section.*

Data Transformation	# Comparison Pairs	# Possible Conclusions	# Total Decisions	Gwet AC_2	SE	90% Confidence Interval Gwet AC_2		Verbal Equivalent
Six-Level Combined	12	6	832	0.7509	0.0875	0.6070	0.8948	Substantial
Six-Level Nonmates	7	6	484	0.8818	0.0546	0.7919	0.9717	Substantial
Six-level Mates	5	6	348	0.6562	0.1369	0.4310	0.8813	Moderate
Four-level Combined	12	4	832	0.6916	0.0871	0.5483	0.8350	Moderate
Four-Level Nonmates	7	4	484	0.8002	0.0878	0.6557	0.9447	Substantial
Four-level Mates	5	4	348	0.7662	0.1120	0.5820	0.9504	Moderate

substantial and/or *moderate* agreement (note that the only category above *substantial* agreement for the Landis and Koch [15] verbal equivalency table is *almost perfect* [14,15]).

Limitations for all reported results are a function of the original study design, and have been itemized in previous summaries (Part I [1] and Part II [2]) and are not repeated here for the sake of brevity. Instead, emphasis will be placed on generalizing findings across all three summaries (Part I—III) and providing implications for future research. First, reproducibility is an elusive metric to evaluate when presented with nonbinary conclusion categories. The research team has attempted to report this using several metrics, including box plots, interquartile range, consensus (1,2), and inter-rater reliability. The mean IQR in this study was found to be 85.6% ± 11.1% (1). This is comparable to former footwear reliability studies when the results presented were reanalyzed by our team (Majamaa and Ytti [24] 83.8% ± 12.4%, Shor and Weisner [25] 78.3% ± 0.00%, and Hammer et al. [26] 94.3% ± 7.36%). Based on study design, all values match intuition and are a function of the perceived difficulty and variety of the cases presented, the extent of the geographical areas surveyed, the nature of participant training and certification, and whether or not features were pre-identified for participants.

As for measures of consensus, results indicate a low of 0.51, a high of 0.97, with a mean of 0.78 ± 0.14 (2). Although this metric does account for partial agreement when dealing with ordinal scales, there is little insight on how to interpret its magnitude. Conversely, the cumulative inter-rater reliability agreement metrics reported here adjust for study parameters (number of raters, number of comparisons, number of agreement categories, chance agreement, and partial agreement with ordinal scales). Based on the Gwet AC_2 coefficient, agreement was found to equal 0.75 with a verbal equivalence of *substantial* agreement after benchmarking and when using a six-level conclusion ordinal scale for all combined data. Unfortunately, as to which of these metrics (box plots, IQR, C, or IRR) is most appropriate to summarize the footwear reliability data, the research team is unable to provide more insight, except to purport that IRR seems to be the most comprehensive. Unfortunately, IRR measures in other expert decision studies are often not benchmarked; so Table 6 can only provide a baseline reference in the absence of uncertainty (and with relatively small sample sizes of experts as compared to this study). Despite this limitation in interpretation, the results presented here for AC_2 (*moderate* to *substantial*) align with the higher end of agreement found in other expert studies (note: $\kappa = 0.7001$ (six-level)/$\kappa = 0.6716$ (four-level), and $\alpha = 0.7031$ (six-level)/$\alpha = 0.6747$ (four-level) for this study for the combined dataset).

Second, examiner accuracy was assessed. It has been reported as both IQR and based on expected conclusions allowing the research team to serve as an "oracle," defining the accepted conclusions as a function of ground truth, observable features, feature reliability, and adherence to the SWGTREAD (2013) conclusion standard (3). Using this "oracle" status, accuracy ranges from a low of 55.7% to a high of 97.1% with a mean of 82.8% ± 11.9%, which failed to be significantly different from the same metric computed using the IQR (85.6% ± 11.1%) (2). Likewise, accuracy has been assessed using correct predicted value for a six-, four- and three-level conclusion hierarchy. For the combined dataset, CPV varies from a low of 65% for *weak inclusions* such as *association of class*, to a high of 95% for *exclusions* when using six- and four-level conclusion scales. When assessed with regard to ground truth for known mates and known nonmates and after removal of *inconclusive* decisions, a FPR of 0.5% and a FNR of 16% were found. When compared to fingerprint reliability data at various mate-prevalences, the PPV is comparable, but the NPV is lower (13). Moreover, this data transformation completely negates the forensic footwear community's desire to use a seven-point conclusion standard, and in the absence of CPV (positive or negative) for other forensic pattern evidence fields using scales that are larger than three, it is difficult to interpret the values computed here within the confines of other community- and discipline-norms. Equally important, how would such metrics be communicated to the *trier-of-fact*? Will footwear examiners report conclusions in court, along with CPVs that are function of the conclusion category provided? For example, if an examiner reports *association of class* for a questioned impression compared to exemplar "A," and *exclusion* when compared to exemplar "B," will he or she also report a CPV of 65% for the *association of class* and 95% for the *exclusion* decisions, and if so, using what prevalence estimations? Moreover, how would this be interpreted by the *trier-of-fact*, and how can these metrics be used by the footwear field to inform training and recommendations moving forward? Within the combined dataset, the lower CPVs were associated with decisions that either failed to include or failed to exclude, and of these two failures, the latter is considered more problematic. The research team hypothesizes that these errors result from *incomplete comparisons* and a lack of standardization in the qualities associated with each conclusion, and if the footwear field agrees, then this information can be used by the community moving forward to refine training. In addition, if the existing SWGTREAD (2013) standard (3) is revised, either to include more or less categories, and/or to change the description or label for each category, actual and/or latent factors associated with these changes may positively or negatively impact the CPVs reported here. Thus, although jury studies for scale interpretation are tantamount to understanding how expert opinions are digested by *triers-of-fact*, an equivalent examiner study should be performed in order to determine how the newly proposed scale is interpreted and implemented by examiners, with corresponding re-evaluations of CPVs. Finally, it is important to reiterate that except for binary conclusions combined with ground truth, all other CPVs computed here are based on the research team's decision of correct versus incorrect conclusions. The fact that the research team's assessments turned out to be consistent with the community's IQR for 10 out of 12 comparisons (2) lends support for this approach, but it is clear that this issue (an accepted means of determining correct vs. incorrect decisions for multi-level conclusion standards) should be given more thought by the forensic community as a whole, including guidance moving forward in the current and future era of forensic reliability studies.

Acknowledgments

The authors would like to express our sincerest gratitude to the forensic footwear examiners who generously shared their time and expertise in order to participate in this West Virginia University (WVU) Institutional Review Board (IRB) approved (1602021821) study. Thank you for making this work possible.

References

1. Speir JA, Richetelli N, Hammer L. Forensic footwear reliability: part I – participant demographics and examiner agreement. J Forensic Sci 2020;65:1852–70.

2. Richetelli N, Hammer L, Speir JA. Forensic footwear reliability: part II – range of conclusions, accuracy and consensus. J Forensic Sci 2020;65:1871–82.
3. SWGTREAD. Range of conclusions standard for footwear and tire impression examinations. Scientific Working Group for Shoeprint and Tire Tread Evidence. 2013. https://www.nist.gov/system/files/documents/2016/10/26/swgtread_10_range_of_conclusions_standard_for_footwear_and_tire_impression_examinations_201303.pdf (accessed July 31, 2020).
4. Holdren J, Lander S.Report to the President – Forensic science in criminal courts: ensuring scientific validity of feature comparison methods. Technical Report. Washington, DC: President's Council of Advisors on Science and Technology (PCAST), 2016.
5. McQuiston-Surrett D, Saks MJ. The testimony of forensic identification science: what expert witnesses say and what factfinders hear. Law Hum Behav 2009;33(5):436–53. https://doi.org/10.1007/s10979-008-9169-1
6. Koehler JJ. If the shoe fits they might acquit: the value of forensic science testimony. J Empir Legal Stud 2011;8:21–48. https://doi.org/10.1111/j.1740-1461.2011.01225.x
7. Thompson WC, Newman EJ. Lay understanding of forensic statistics: evaluation of random match probabilities, likelihood ratios, and verbal equivalents. Law Hum Behav 2015;39(4):332–49. https://doi.org/10.1037/lhb0000134
8. Arscott E, Morgan R, Meakin G, French J. Understanding forensic expert evaluative evidence: a study of the perception of verbal expressions of the strength of evidence. Sci Justice 2017;57(3):221–7. https://doi.org/10.1016/j.scijus.2017.02.002
9. Thompson WC. How should forensic scientists present source conclusions? Seton Hall Law Review 2018;48:773–813.
10. Thompson WC, Grady RH, Lai E, Stern HS. Perceived strength of forensic scientists' reporting statements about source conclusions. Law Probab Risk 2018;17(2):133–55. https://doi.org/10.1093/lpr/mgy012
11. Swofford HJ, Cino JG. Lay understanding of "identification": how jurors interpret forensic identification testimony. J Forensic Identif 2018;68 (1):29–41.
12. SWGFAST. Standards for examining friction ridge impressions and resulting conclusions (latent/tenprint). Scientific Working Group on Friction Ridge Analysis, Study and Technology. 2013. https://www.nist.gov/system/files/documents/2016/10/26/swgfast_examinations-conclusions_2.0_130427.pdf (accessed July 31, 2020).
13. Ulery BT, Hicklin RA, Buscaglia J, Roberts MA. Accuracy and reliability of forensic latent fingerprint decisions. Proc Natl Acad Sci USA 2011;108(19):7733–8. https://doi.org/10.1073/pnas.1018707108
14. Gwet K. Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters, 4th edn. Gaithersburg, MD: Advanced Analytics, LLC, 2014.
15. Landis J, Koch G. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–74. https://doi.org/10.2307/2529310
16. Onate J, Cortes N, Welch C, Van Lunen B. Expert versus novice interrater reliability and criterion validity of the Landing Error Scoring System. J Sport Rehabil 2010;19(1):41–56. https://doi.org/10.1123/jsr.19.1.41
17. Andreasen S, Backe B, Lydersen S, Øvrebø K, Øian P. The consistency of experts' evaluation of obstetric claims for compensation. BJOG-Int J Obstet Gy 2014;122(7):948–53. https://doi.org/10.1111/1471-0528.12979
18. Gschließer A, Stifter E, Neumayer T, Moser E, Papp A, Pircher N, et al. Inter-expert and intra-expert agreement on the diagnosis and treatment of retinopathy of prematurity. Amer J Ophthalmol 2015;160(3):553–60.e3. https://doi.org/10.1016/j.ajo.2015.05.016
19. Acklin MW, Fuger K. Assessing field reliability of forensic decision making in criminal court. J Forensic Psychol Pract 2016;16(2):74–93. https://doi.org/10.1080/15228932.2016.1148452
20. Nawrocka M, Fratczak K, Matuszewski S. Inter-rater reliability of total body score- a scale for quantification of corpse decomposition. J Forensic Sci 2016;61(3):798–802. https://doi.org/10.1111/1556-4029.13105
21. Lee K, Fatah AAA, Norizan NM, Jefrey Z, Nawi FHM, Nor WFKW, et al. Inter-rater reliability of vehicle color perception for forensic intelligence. PLoS One 2019;14(6):e0218428. https://doi.org/10.1371/journal.pone.0218428
22. Tobi H, van den Berg PB, de Jon van den Berg LTW. Small proportions: what to report for confidence intervals? Pharmacoepidem Drug Saf 2005;14(4):239–47. https://doi.org/10.1002/pds.1081
23. Mercaldo ND, Lau KF, Zhou XH. Confidence Intervals for predictive values with an emphasis to case-control studies. Statist Med 2007;26 (10):2170–83. https://doi.org/10.1002/sim.2677
24. Majamaa H, Ytti A. Survey of the conclusions drawn of similar footwear cases in various crime laboratories. Forensic Sci Int 1996;82 (1):109–20.
25. Shor Y, Weisner S. A survey on the conclusions drawn on the same footwear marks obtained in actual cases by several experts throughout the world. J Forensic Sci 1999;44(2):380–4.
26. Hammer L, Duffy K, Fraser J, Daéid NN. A study of the variability in footwear impression comparison conclusions. J Forensic Identif 2013;63: 205–18.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Confusion Matrix of footwear expert decisions as compared to the acceptable range of conclusions for 350 mated comparisons, exhibiting an overall accuracy of approximately 76.3% (green cells; 267/350).

Table S2. Confusion Matrix of footwear expert decisions as compared to the acceptable range of conclusions for 485 nonmated comparisons, exhibiting an overall accuracy of approximately 87.4% (green cells; 424/485).

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14536
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Keith L. Monson,[1] Ph.D.; Kelsey M. Kyllonen,[2] M.A.; Jeffrey L. Leggitt,[3] M.S.; Kelli E. Edmiston,[4] M.S.; Calvin R. Justus,[2] Ph.D.; Mark F. Kavlick,[1] Ph.D.; Maria Phillip,[4] B.S.Ch.E.; Maria A. Roberts,[5] M.S.; Candie W. Shegogue,[6] M.S.; and Gabriel D. Watts,[7] B.A.

Blast Suppression Foam, Aqueous Gel Blocks, and their Effect on Subsequent Analysis of Forensic Evidence*

ABSTRACT: In addition to having blast mitigation properties, aqueous foam concentrate AFC-380 blast suppression foam is designed to capture aerosolized chemical, biological, and radioactive particles during render-safe procedures of explosive devices. Exposure to aqueous environments and surfactants may negatively affect forensic evidence found at the scene, but the effects of AFC-380 foam and aqueous gel on the preservation and subsequent analysis of forensic evidence have not previously been investigated. Sebaceous finger and palm prints and DNA samples on paper, cardboard, tape, and various metal and plastic items, along with hairs, carpet and yarn fibers, and inks and documents, were exposed to AFC-380 foam. Similar mock evidence was also exposed to a superabsorbent gel of the type found in aqueous gel blocks used for shrapnel containment. Exposure to foam or aqueous gel was associated with a dilution effect for recovered DNA samples, but quality of the samples was not substantially affected. In contrast, exposure to AFC-380 foam or gel was detrimental to development of latent finger and palm prints on any substrate. Neither the hair nor the fiber samples were affected by exposure to either the foam or gel. Indented writing on the document samples was detrimentally affected by foam or gel exposure, but not inks and toners. The results from this study indicate that most types of forensic evidence recovered after being exposed to aqueous gel or blast suppression foam can be reliably analyzed, but latent finger and palm prints may be adversely affected.

KEYWORDS: forensic science, blast suppression foam AFC-380, superabsorbent aqueous gel, gel blocks, radiological dispersal device (RDD), latent fingerprints, DNA, fiber analysis, document analysis, evidence recovery

Public safety and security demand that we must defend against potential use of a chemical, biological, or radiological dispersal device (RDD) by terrorists (1,2). Part of this preparation includes maximizing the forensic information that can be gleaned from crime scenes, leading to effective prosecution of the persons responsible. The present study examines potential effects on recovery and development of conventional forensic evidence when blast suppression foam and/or an aqueous gel barrier to light shrapnel are used in conjunction with render-safe procedures. The potential effects of various radiation types on evidence have been addressed elsewhere (3).

The immediate danger posed by detonation of an RDD would be due to effects of the conventional explosive. Depending on the specifics of the explosive, the type of radioactive material, and dispersal conditions, persons nearby the event could also suffer radiation exposure. Decontamination of the site may be exceedingly costly and lengthy (1,2,4–6). Simulations comparing unmitigated and mitigated release of 10 kg of weapons-grade plutonium, requiring mitigation of 9 km^2 of land v. only the immediate area, respectively, reflected a remediation cost differential of approximately a 100-fold (7). In any case, the threat is real and dispersion of even a small amount of radioactive material by a “dirty bomb” would aptly serve the terrorist goals of inciting widespread fear and panic of the public, with long-enduring negative economic effects (2,4,8–10).

[1] FBI Laboratory, Counterterrorism and Forensic Science Research Unit, 2501 Investigation Parkway, Quantico, VA 22135.

[2] FBI Laboratory, Counterterrorism and Forensic Science Research Unit, Visiting Scientist Program, 2501 Investigation Parkway, Quantico, VA 22135.

[3] FBI Laboratory, Evidence Response Team Unit, 2501 Investigation Parkway, Quantico, VA 22135.

[4] FBI Laboratory, Latent Fingerprint Operations Unit, 2501 Investigation Parkway, Quantico, VA 22135.

[5] FBI Laboratory, Latent Fingerprint Support Unit, 2501 Investigation Parkway, Quantico, VA 22135.

[6] FBI Laboratory, Trace Evidence Unit2501 Investigation Parkway, Quantico, VA 22135.

[7] FBI Laboratory, Questioned Documents Unit, 2501 Investigation Parkway, Quantico, VA 22135.

Corresponding author: Keith L. Monson, Ph.D. E-mail: Keith.Monson@ic.fbi.gov

*This is publication 20-16 of the Federal Bureau of Investigation (FBI). This research was supported in part by an appointment to the Visiting Scientist Program at the FBI Laboratory Division, administered by the Oak Ridge Institute for Science and Education, through an interagency agreement between the U.S. Department of Energy and the FBI. Names of commercial manufacturers are provided for identification purposes only, and inclusion does not imply endorsement by the FBI. The views expressed are those of the authors and do not necessarily reflect the official policy or position of the FBI or the U.S. Government.

[Correction added 10 August 2020. The Highlights section was removed.]

Received 20 Feb. 2020; and in revised form 22 May 2020; accepted 9 July 2020.

In the mid-1980s, Sandia National Laboratories developed aqueous foams to address requirements for blast pressure reduction, fireball mitigation, particulate containment, and buoyant cloud suppression for response to an undetonated chemical, biological, or radiological dispersion device (7,11–15). Such foams have been shown to capture more than 99% of the airborne particles that may be inhaled when released by the detonation of a RDD (7).

Conventional firefighting foams are unsuitable for this application. Their low shear strength precludes piling the foam, and the foam rapidly dissipates due to draining (7,13,16,17). Sandia engineered AFC-380 to hold water in the bubble matrix so that the foam remains stable for several hours (7,11–13). The concentrate is mixed on site with water to a final concentration of 6% using a portable inline eductor that is common to nearly all fire departments for foam generation (17,18). When an undetonated RDD is encountered, trained responders place two nested enclosures over the RDD and their own render-safe equipment. The space between the two enclosures is then filled with foam to absorb the blast and radioactive particulates in the event of failure of the neutralization efforts (19).

The remarkable ability of aqueous foams to mitigate shock waves has been demonstrated empirically and modeled theoretically. A series of studies in the early 1980s showed significant decrease in shock pressure when using these foams (20). The authors proposed that the mitigation effects were likely due to internal reflections at the air–liquid interfaces, energy loss in breaking foam bubbles, and heating of the water component. These ideas were later experimentally confirmed by other researchers, further contributing insight that rupture of the foam films creates droplets, which absorb energy upon evaporation, elaborating the effect of expansion ratio and bubble size, and demonstrating the reduction in explosive noise that aqueous foams provide (16,21–25). The presence of the foam reduces overpressure by a factor of 3–30 (26,27). It also provides some reduction in velocity of light shrapnel (20). If an explosive is detonated, evidence may also be adversely affected by direct heating and/or as a result of energy confinement by the foam described above. Aqueous gel barriers or a Kevlar tent may be used to stop larger fragments (27).

Numerous studies have treated the theoretical aspects of foams and their blast mitigation (16,23,26,28). Although its primary application is preventive, AFC-380 and other aqueous foams have also been considered for roles in postdetonation remediation (29). Other foam products (e.g., CASCAD®, Allen-Vanguard, Ottawa; and MDF LSA-100/200, Modec, Inc., Denver, CO) are intended primarily for chemical, biological, and radiological decontamination, but may also offer a degree of blast mitigation. Due to the presence of an oxidizing agent in those other foams (29), DNA profiles, although quantifiable, could not be developed after their use (30) and recovery of fingerprints was significantly reduced (31,32). Evaluation of AFC-380, which lacks these oxidizers, on the preservation of evidence is a primary focus of this study.

It is well known that high-speed projectiles such as bullets are effectively stopped by water or ballistic gelatin (33,34). Thus, bomb response units sometimes use aqueous gel blocks (whether or not foam is also used) to surround an improvised explosive device to reduce the potential dispersal of light shrapnel. These blocks consist of a vinyl bladder containing highly absorbent materials such as those used in personal sanitary products (35). Filled blocks are usually placed inside a rectangular surround to facilitate stacking. The hygroscopic component of the filler is the superabsorbent polymer sodium polyacrylate with fumed silica added to reduce caking; both are considered to be nonreactive (36,37). Bomb technicians can fill the vinyl bladders with water on-site and stack the gelled blocks around an IED (Fig. 1). The gel blocks maintain some degree of integrity even when struck by multiple small fragments. Embedded fragments can be processed as evidence.

Water, the primary component of both AFC-380 foam and filled gel blocks, itself potentially threatens recovery of evidence. As the foam dissipates, physical evidence may be diluted, dissolved, or washed away. Chemical components of the foam or gel may also compromise evidentiary value. Several other factors are intrinsic to an explosive crime scene and the methods used to respond to it. The scaling law applies, that is, that damage extent is directly related to explosive charge and inversely related to distance cubed (38). When fragmentation of evidentiary material increases, the information that can be extracted decreases. Nevertheless, latent fingerprints (39–42) and DNA profiles (39,42–47) can sometimes be developed from explosive evidence.

Although DNA profiles may sometimes be developed from a deflagrated IED or surrounding material, other times there may be insufficient DNA quantity or quality, for example, due to PCR inhibitors, thus preventing analysis (42,45,46). The post-blast recovery and analysis of DNA may be further hindered by the use of blast suppression foam or gel blocks. Presumptively, DNA may be damaged, as AFC-380 contains an anionic surfactant (12). The ability of a mild detergent such as sodium dodecyl sulfate (SDS) to lyse cell membranes has been exploited for decades to isolate DNA without damaging it (48). Even commercial laundry detergents, which typically are anionic surfactants, can be used for this purpose (49). Indeed, DNA typing is often successfully carried out on laundered garments (50–54). If blood cells were lysed by foam, the DNA within them would be vulnerable to degradation by nucleases or by the foam itself. Whether lysed or not, foam may cause unexpected inhibition of the PCR, although its high concentration of water may render that effect benign. DNA profiles may (55) or may not (54,56,57) be developed after water submersion, or even after evidence in a fire has been extinguished with water (58).

Latent fingerprints may be developed after exposure to adverse environmental conditions. Fingerprints can sometimes be recovered from materials exposed to elevated temperatures, even though eccrine components are degraded by the heat of an IED deflagration (39–41,59–62). Fingerprints sometimes survive water

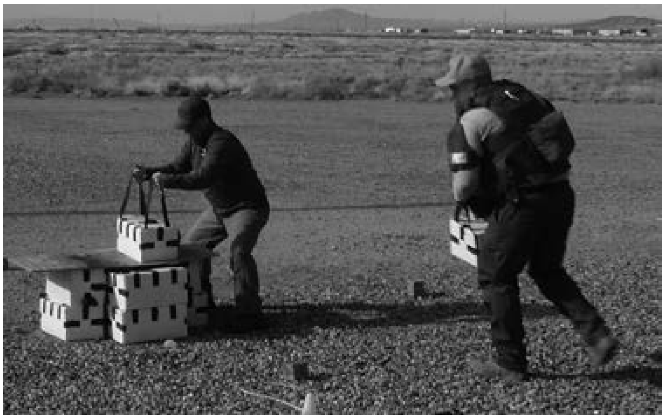


FIG. 1—*Filled gel blocks may be placed around an IED to capture small shrapnel. [Color figure can be viewed at wileyonlinelibrary.com]*

immersion or exposure, although eccrine-rich prints, consisting of water-soluble components such as amino acids, are particularly vulnerable (63,64). Since amino acid residues are likely washed away, use of ninhydrin-related developers is discouraged for wetted porous materials (65). Alternatives include physical developer, Oil Red O, and phase-transfer catalysts (65–67). Sebaceous prints on nonporous materials that have been wetted may be developed with particle suspension reagents, powder, superglue, or vacuum metal deposition, albeit with diminishing success as immersion time increases (56,57,63–65,68–75). Of particular relevance to the present study, immersion in soapy water decreases development success (31,52,56,76). In general, sebaceous residues are removed by organic solvents, particularly on nonporous surfaces (74), and AFC-380 foam contains several alcohols and an ether (12), although no detrimental effect was reported for alcohol-based hand sanitizers on subsequent development (77).

Location of trace evidence such as hairs and fibers may be compromised once the foam has dissipated into a liquid and drained from the enclosure (78,79). Wet paper becomes very fragile and may be difficult to extract from a gel block after being explosively forced into it. Though some forensic aspects of documents are recoverable after having been wetted (80), indented writing is rarely recoverable (81,82).

Materials and Methods

Mock evidence specimens were prepared to represent the types and deposition substrates that may be found at the scene of an explosive device including finger or palm prints, and human blood (deposited on paper, a cardboard box, electrical and duct tape, glass, PVC pipe and flats, and iron pipe), human hairs, natural and synthetic fibers, and inks and toners (Sharpie^{TM}, ball point, toner, and liquid ink jet). All human-derived items were collected with informed consent and approved by the FBI Institutional Review Board. After deposition of blood and/or fingerprints, all evidence samples were allowed to dry overnight and packaged according to item type. Specimens were exposed to one of four conditions: (i) AFC-380 foam used in the context of a full-scale render-safe exercise, (ii) AFC-380 foam in an isolated container, (iii) aqueous gel in laboratory conditions, and (iv) untreated controls. Evidence types and substrates subjected to each treatment are tabulated in Table 1 and detailed below. Following exposure, all specimens were recovered and processed by standard procedures as described below. All samples were handled using gloves for the entirety of this research.

DNA Samples

Venous blood from a single donor was used to represent DNA evidence left behind on various objects that could be found at the scene of an explosive device. Blood was applied with a cotton-tipped applicator to the threads of each metal pipe, to the raised logo on the side of each metal end cap, and near one end of each PVC pipe to simulate DNA recovery from nonporous objects. Blood was similarly applied to the upper corner of both paper and cardboard and to the center of sterile gauze to simulate DNA recovery from porous objects. The location of the blood on each item was marked with a permanent marker after application. Plucked hairs with root ends were collected from a second donor. For each treatment, at least five hairs were attached to an adhesive disk (to facilitate recovery) to represent both DNA and hair evidence that could be found at the scene of an explosive device, and approximately, the same number was retained as controls.

Finger and Palm Print Samples

To represent fingerprint evidence (unless stated otherwise, "fingerprint" shall henceforth be used loosely also to include palm prints) deposited on various metal and plastic objects, galvanized metal pipes with end caps, galvanized metal flats, PVC pipes, and PVC flats were obtained from various

TABLE 1—*Mock evidence and substrates exposed to AFC-380 foam; the same number was exposed to aqueous gel.*

Substrate	Foam*	Gel	Controls[†]	Finger prints[‡]	Palm prints	Whole blood	Ink
Galvanized pipe	4	2	2	+	+	+	
Galvanized pipe (cleaned)	4		2	+	+	+	
Galvanized end cap	8		4	+		+	
Galvanized end cap (cleaned)	8		4	+		+	
Galvanized flat	3	2	3		+		
Galvanized flat (cleaned)	3		3		+		
PVC pipe	4	2	2	+	+	+	
PVC pipe (cleaned)	4		2	+	+	+	
PVC flat	3 (1)	4	3	+	+		
PVC flat (cleaned)	3 (1)		3	+	+		
Glass microscope slides	6 (2)	2	6	+			
Electrical tape (black)	4 (3)	2	2	+			
Packing tape (clear)	4	2	2	+			
Duct tape (gray)	4 (3)	2	2	+			
Duct tape (black, extra strength)	4	2	2	+			
Corrugated cardboard	4 (3)	4	4		+	+	
Bond paper (toner, ink jet, 6 inks, indented writing)	4 (2)	4	4		+	+	+
Cotton gauze (soak conditions only)	5	2	5			+	
Human hair with root end	5	5	5				
Cotton fibers (red, blue)	2,2	2,2	2,2				
Nylon carpet fibers (blue, pink, red)	3,4,4	4,4,4	3,4,4				
Acrylic fibers (light blue, dark blue)	4,4	4,4	4,4				

Presence of a (+) symbol designates that a finger/hand print, blood, or ink was applied to the indicated substrate.

*Parentheses indicate the reduced number of items recovered after exposure to render-safe procedures. All items were recovered from the soak condition.

[†] Separate controls, in the number indicated, were used for the foam and gel experiments.

[‡] Fingerprints were deposited on both sides of all tapes.

hardware stores. Some items were cleaned with a mild detergent and allowed to dry prior to fingerprint deposition to remove manufacturing contaminants and any extraneous fingerprints that may have been left by suppliers, store employees, and other customers. The remaining items were not cleaned prior to fingerprint deposition. Approximately 20-cm strips each of electrical tape, packing tape, extra strength duct tape, and regular duct tape were cut from new rolls of tape to represent fingerprint evidence that could be left behind on tape, on both adhesive and nonadhesive sides. Plain copy paper and cardboard were used to represent fingerprint evidence on porous materials. The copy paper was taken from the middle of a new ream of paper to avoid extraneous prints, while the cardboard samples were cut from boxes that had already been used to ship items at least once and may have contained prints from other individuals. Microscope slides (50 × 75 mm) were used to represent fingerprint evidence on glass objects.

Because the fingerprints were to be exposed to an aqueous environment, we anticipated that eccrine deposits would be lost; therefore, groomed, primarily sebaceous, natural prints were deposited on the items by a single male donor in his mid-thirties (83). Two full-length index finger prints were deposited on each set of metal and PVC pipes. One full palm print was deposited on each of the remaining metal and PVC pipes. Two fingerprints were deposited on the top of each metal end cap. Palm prints were deposited on both the metal and PVC flats by placing two cleaned or two uncleaned flats together with the long edges touching and having the donor deposit a single palm print in the center of the two flats, creating a split palm print. One half of each split print was set aside as a control, while the other half was used in one of the two experimental conditions. Similarly, palm prints were deposited on the paper and cardboard samples by placing two pieces of paper or two pieces of cardboard together with the long edges touching and having the donor deposit a single palm print in the center, creating a split palm print. One half of each split print was set aside as a control, while the other half was used in one of the experimental conditions. Three simultaneous fingerprints were deposited on each glass slide and both the nonadhesive and adhesive sides of each piece of tape. The location of the print(s) on each item was marked with permanent marker.

Fiber Samples

Carpet fibers and both acrylic and cotton fibers were collected to represent various fiber types. Dark pink and light pink carpet fibers were cut from a sample of Masland™ (New York, NY) nylon carpet, while blue carpet fibers were cut from a sample of Gulistan™ (Aberdeen, NC) nylon carpet. Approximately 13 cm lengths of blue and red cotton yarn and dark blue and light blue acrylic yarn were cut from longer pieces of yarn and untwisted into two halves to create an experimental half and a control half of the same section of yarn. The experimental carpet and yarn fibers were attached to adhesive disks to facilitate recovery.

Document Samples

To represent various inks and indented writing that could be found on paper at the scene of an explosive device, four printer test pages were printed on copy paper, two using an inkjet printer and two using a laser printer. Additionally, the four pages each included writing samples of two different types of ballpoint pen ink, two types of Sharpie™ ink, and ink from both a pink and a yellow highlighter. Indented writing was also added to the bottom of each test page by placing a sheet of notebook paper on top of the test page and writing with a ballpoint pen at the bottom of the notebook paper. Prior to exposure to foam or gel, indented writing, as revealed by an Electrostatic Detection Apparatus (ESDA® 2, Foster + Freeman, Worcestershire, UK), was recorded for all samples (81,84).

Foam Exposure

Prepared samples and controls were shipped from Quantico, VA to Albuquerque, NM. The control items were repackaged and placed inside return shipping boxes for the duration of the training exercises, after which they were reposted. One set of experimental specimens was placed into a large plastic tub, filled with AFC-380 foam, and placed approximately 6 m outside the foam enclosure tent surrounding the mock RDD.

Another set of experimental specimens was placed adjacent to the mock RDD. One cleaned and one uncleaned metal flat, one cleaned and one uncleaned PVC flat, two glass slides, and one piece of cardboard were placed inside the small backpack containing the simulated RDD used during the training exercise. The remaining experimental specimens (except gauze specimens) were attached to the inside of a large hard-sided plastic suitcase using plastic zip ties. The large plastic suitcase was placed in close contact behind the mock RDD under an enclosure that shielded the device and items from the foam until the disruption occurred. The enclosure consisted of a rectangular prism frame built from PVC pipe with plastic sheeting stretched over five of its six sides (Fig. 2*a*). A large tent-like nylon fabric structure was placed over the cover (Fig. 2*a*). The space between the two enclosures was then filled with AFC-380 foam concentrate mixed with water at the standard 6% ratio (Fig. 2*b*,*c*). The simulated device inside the foaming tent was then rendered-safe using an explosive-based procedure.

After the area had been cleared of any hazards, the tent was dismantled, leaving behind a roughly 25-cm-thick layer of foam spread over a large area approximately 10 m in diameter (Fig. 2*a*). Although trained professionals searched the foam thoroughly for the samples as soon as it was safe to do so, both the plastic suitcase and the travel backpack were directly impacted by the disruption charge, which caused heavy damage to some of the samples and made others impossible to locate in the persistent foam layer (Fig. 2*e*,*f*). Upper portions of the foam were also carefully raked away to hasten evidence recovery from the foam and the wet ground. One of the metal pipes with end caps attached was thrown clear of the foam layer as a result of the disruption charge and was found in the dirt about 20 m away from the original device/suitcase location. In the course of recovering items from the render-safe exercise, additional duct tape fragments were located that had been used to construct the inner enclosure for the mock RDD.

After approximately 2 h, when samples from the suitcase and backpack had been recovered, the samples previously left in the plastic tub were removed. They were unaffected by the nearby disruption charge and were easily located and removed from the remaining foam. Altogether, the samples placed near the RDD and those in the plastic tub were exposed to the foam for approximately three hours. During this time, all of the experimental samples were also exposed to direct and/or indirect sunlight because there were no shaded areas near the foaming site.

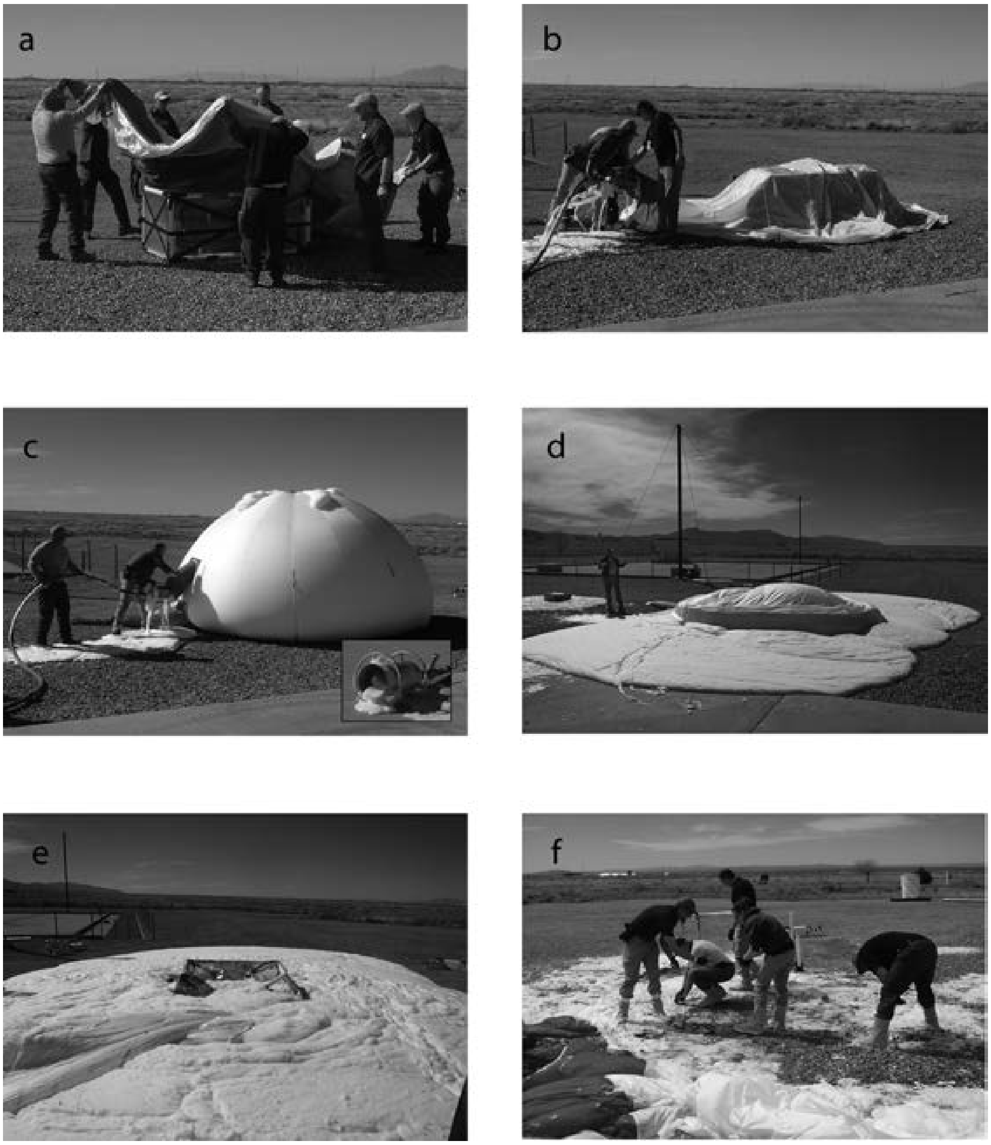


FIG. 2—*Field deployment of AFC-380 foam. (a) Outer enclosure being positioned over inner box that encloses the backpack containing the mock RDD and specimens, plus a suitcase containing more specimens; (b) filling the intervening space with foam: (c) outer tent completely filled with foam is sealed off (insert shows the foam nozzle); (d) spreading foam following explosive-based disruption (plastic tub filled with foam, but not subject to disruption shown at far left); (e) damaged inner enclosure following disruption; (f) recovery of specimens after most of the foam dissipated. [Color figure can be viewed at wileyonlinelibrary.com]*

Following standard procedures, after all possible items had been located, the samples were removed from the training site and spread out to air-dry indoors for 48 h before being packaged individually in new plastic bags and shipped back to the FBI Laboratory for processing, along with the control samples.

Gel Exposure

In an indoor laboratory setting, mock evidence items (Table 1) were placed in a new 5-gallon plastic bucket in such a way to assure full exposure to the gel (Fig. 3). Test tube racks supported the metal and PVC pipes. A gel block bladder was cut open, and the dry contents were emptied into a second new 5-gallon plastic bucket. Approximately five gallons of water was gradually added, filling the second bucket with aqueous gel. Scoops of the gel were added carefully to the first bucket, covering all items (Fig. 3). After approximately three hours had elapsed, the gel was judiciously removed to expose the items. As the items were extracted, they were placed on blotter paper without rinsing with water and allowed to air-dry overnight.

DNA Processing

Those items spotted with blood (Table 1) were swabbed with one side of a polyester swab wetted with 50 μL of a 0.01% SDS solution. The dry side of the polyester swab was then used to capture any leftover biological material, and swabs were dried overnight. Alternatively, for blood spotted on foam-exposed and foam-control cardboard, paper, and gauze only, ~5 mm × 5 mm cuttings were collected. All blood swabs and cuttings were extracted using the EZ1 DNA Investigator Kit protocol on an EZ1 Advanced XL (Qiagen, Valencia, CA). Briefly, 450 μL of a master mix containing 94% G2 buffer (Qiagen), 30 mM dithiothreitol (DTT), and 0.6 mg/mL proteinase K (Amresco,

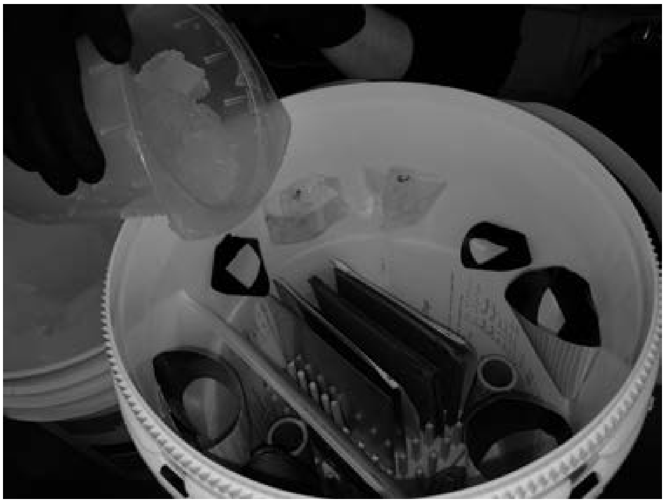


FIG. 3—*Aqueous gel being placed around mock evidence arranged in a 5-gal bucket. [Color figure can be viewed at wileyonlinelibrary.com]*

Solon, OH) were added to each swab or cutting followed by incubation at 56°C for 1 h with 200 rpm shaking. Swabs and cuttings were discarded, and lysates were processed on the EZ1 Advanced XL using the large volume protocol with 50 μL molecular biology grade (MBG) water elution, according to the manufacturer's instructions.

Hair samples were extracted using the PrepFiler Forensic DNA Extraction Kit (Applied Biosystems, Foster City, CA). Briefly, hair specimens were cleaned with a 5% Tergazyme (Alconox, Inc., White Plains, NY) solution, then rinsed with ethanol followed by MBG water. Hair specimens were digested in 320 μL of a buffer containing 93% ATL buffer (Qiagen), 66 mM DTT, and 1.25 mg/mL proteinase K (Amresco) at 56°C for 0.5–16 h with 900 rpm shaking until visibly digested. Processing continued with the addition of 300 μL AL buffer (Qiagen) followed by incubation at 70°C for 10 min with 900 rpm shaking. Samples were cooled to room temperature, then purified using the PrepFiler Forensic DNA Extraction Kit following the manufacturer's instructions with the exception of a 6–8 min air-drying step prior to elution in 65 μL PrepFiler elution buffer.

Nuclear DNA (nDNA) concentration and quality of DNA extracts were assessed using the Quantifiler Human Plus DNA Quantification Kit (Applied Biosystems). Briefly, 2 μL of extract was assayed in a 20 μL reaction on a 7500 Real-time PCR System using HID Real-Time PCR Software (Applied Biosystems) to determine concentration (ng/μL), a nDNA-specific degradation index (DI), and the presence or absence of amplification inhibitors, following the manufacturer's instructions.

Mitochondrial DNA (mtDNA) quantity and quality were assessed for all DNA extracts using a qPCR method previously described (85). Briefly, 2 μL of extract was assayed in a 20 μL reaction on a 7500 Real-time PCR System using HID Real-Time PCR Software (Applied Biosystems) as a custom assay to determine mtDNA copy number (copies/μL), a mtDNA-specific DI, and the presence or absence of amplification inhibitors.

Fingerprint Processing

Samples with fingerprints were processed using the same sequence of methods that examiners at the FBI Laboratory use to process forensic case work evidence (86). The quality of each fingerprint was assessed at each stage of development using the four-level absolute scale of Bécue et al. (87) (Table 2) (a score of 3 had been designated 4 in the original publication).

TABLE 2—*Absolute scale for fingerprint quality assessment, paraphrased from Bécue et al. (87), with the exception that a score of 3 was designated 4 in the original.*

Score	Characteristics
0	No ridge detail
1	Poor ridge detail; inadequate for comparison
2	Suboptimal, but usable, level 2 ridge detail
3	Good quality

Nonporous samples with fingerprints were first examined visually for ridge detail under white light and three forensic light sources: blue (450 nm), ultraviolet (UV, 365 nm), and laser (532 nm) light. The samples were then processed using cyanoacrylate fuming and examined for ridge detail visually and with the Reflective Ultraviolet Imaging System (RUVIS). Tape samples were placed adhesive side down onto clean acetate sheets prior to cyanoacrylate fuming, and only the nonadhesive sides were examined for this process. The adhesive sides were processed using WetWop^{TM} (Safariland LLC, Jacksonville, FL) and examined visually. Finally, the nonadhesive sides of tape, and all other nonporous samples, were treated with the cyanoacrylate fluorescent dye RAM (Rhodamine 6G, Ardrox, and MBD) and reexamined under UV, laser, and blue forensic light sources. The porous samples were examined visually under white light, then three forensic light sources before being processed with indanedione (IND) and reexamined using the laser light source. In some cases where ridge detail was observed within the processing sequence, the area of ridge detail was photographed before the sample continued on to the next processing step.

Hair and Fiber Processing for Microscopy

Hairs and a representative subset of fibers were removed from the treated samples. The hairs and fibers were mounted between a glass microscope slide and coverslip using Permount^{TM} (Thermo Fisher Scientific, Waltham, MA). The samples were then assessed using a high-magnification comparison light microscope (Leica FS CB, magnification range of 50× to 400×). Any changes in microscopic appearance relative to the controls were noted. Fibers were also examined by polarized light microscopy and fluorescence microscopy (Nikon Eclipse E600 POL, magnification range of 50× to 400×) using four excitation wavelengths: UV (330–380 nm), violet (380–420 nm), blue (450–490 nm), and green (510–560 nm). The general optical properties and the color and intensity of fluorescent emission were recorded and compared with the controls. Following polarized light microscopy and fluorescence microscopy, fiber samples were removed from slides, washed in xylene to remove residual Permount, and mounted between a quartz microscope slide and coverslip using immersion glycerin (Zeiss Immersol G). These samples were then analyzed using a Craic MSP 121 AXIO microspectrophotometer (MSP) from 240 to 800 nm. Any changes in spectra relative to controls were noted. Following MSP analysis, manufactured fibers were removed, washed in methanol to remove residual glycerin, dried, and dry-mounted between a glass microscope slide and coverslip. These samples were then analyzed using a Nicolet iN10 Fourier Transform Infrared Spectrometer (FT-IR) from

700 cm^{-1} to 4000 cm^{-1}. Any changes in spectra relative to controls were noted.

Document Processing

Following exposure and air dying (88), readability of ink and printing was assessed visually. All documents were then processed by ESDA for indented writing.

Results

Postprocessing results obtained from the same substrates were pooled and averaged, since no significant difference was detected among foamed samples exposed to render-safe (backpack and suitcase) and soak conditions nor between cleaned and as-received substrates (data not shown).

DNA Following AFC-380 Foam Exposure

DNA extracts of hair and of blood deposited on items exposed to foam were quantified, and quality was determined for both nDNA and mtDNA. Neither DNA type exhibited degradation or appeared to contain amplification inhibitors when compared to the unexposed control samples (data not shown), per the respective qPCR assays, suggesting that foam exposure did not affect DNA quality.

In contrast, foam-exposed blood extracts generally yielded less nDNA and mtDNA (Fig. 4*a*,*b*, respectively) compared with the respective unexposed controls. This seemed particularly evident for metal and PVC pipes, which have smooth intact surfaces, compared with the porous and fibrous nature of cardboard, paper, and gauze, or with the protective DNA-encapsulated nature of hair. This observation strongly suggests that foam exposure diluted or washed away some of the deposited blood and that this action was pronounced, not surprisingly, when it was deposited on hard, impervious surfaces such as pipe.

Figure 4*a* also shows that for those control items from which nDNA could be recovered in a target concentration of ≥100 ng/μL for short tandem repeat (STR) analysis, that is, cardboard, paper, gauze, and hair, nDNA could also be recovered from the respective exposed items in at least one replicate, indicating that human identification may be possible for exposed samples, provided that sufficient biological material was recovered. STR analysis may also be possible for extracts that quantify below that target, for example, from metal and PVC pipe, if first sufficiently concentrated (not performed here).

Figure 4*b* reveals that mtDNA was recovered from most foam-exposed item replicates (22 of 30) at a target concentration of ≥10 copies/μL concentration to facilitate successful mtDNA whole control region (WCR) sequence analysis via Sanger sequencing. This outcome was not unexpected since hundreds to thousands of copies of mtDNA may be present in a human cell (89). Thus, mtDNA sequence analysis may be possible when STR analysis is not for a given sample. However, while the latter may enable identification, the former permits only inclusion or exclusion.

DNA Following Aqueous Gel Exposure

DNA extracts of blood deposited on items exposed to gel were likewise quantified, and quality was determined for both nDNA and mtDNA. As with foam-exposed samples, neither DNA type exhibited degradation or appeared to contain amplification inhibitors relative to the unexposed control samples (data not shown), suggesting that, like foam, gel exposure does not negatively affect DNA quality.

As with the foam-exposed experiment (Fig. 4), gel-exposed extracts generally yielded less nDNA and mtDNA relative to the respective unexposed controls (Fig. 5*a*,*b*, respectively). Additionally, this was particularly true for smooth-surfaced substrates, but less so for the porous and fibrous cardboard, paper, and gauze (hair was not tested with gel). These results also support the assertion that exposure to gel may dilute or wash away deposited blood and that this effect is pronounced when deposited on hard, impervious surfaces.

Figure 5*a* reveals that no item replicate, gel-exposed or control, yielded extracts, which contained nDNA meeting the target concentration of ≥100 ng/μL for short tandem repeat (STR) analysis. Furthermore, most gel-exposed replicates (9 of 14) were undetectable or below the qPCR standards range for nDNA (0.005–50 ng/μL). Overall, nDNA concentrations for this experiment were well below those of the foam-exposed experiment. However, while the test items, for example, pipes, paper, were comparable between the foam and gel experiments, possible differences may have contributed to lower DNA yields in the latter: (a) the amount of blood spotted, (b) analyst swabbing and extraction technique, (c) method of collecting blood from cardboard, paper, and gauze (swabbing versus cuttings), and (d) numbers of replicates (26 versus 58) and thus variance. Nonetheless, STR analysis may be possible for any of the extracts represented in Fig. 5*a* if concentrated (not performed here).

Figure 5*b* reveals that mtDNA was recovered from most gel-exposed samples. Indeed, 11 of 14 replicates met the target concentration of ≥ 10 mtDNA copies/μL to facilitate mtDNA WCR Sanger sequencing analysis. Only 1 of 4 cardboard-, 1 of 2 metal pipe-, and 1 of 2 PVC pipe-exposed replicates did not meet the target. Thus, some forensic DNA value may be realized for DNA extracts of swabs from items exposed to gel.

Fingerprint Quality Following AFC-380 Foam Exposure

All samples with deposited latent prints were processed appropriate to their substrate after all items that also contained deposited blood samples had been processed for DNA analysis. Pieces of duct tape that had been part of the inner enclosure, and were recovered incidentally at the scene, were also processed for latent fingerprints ("undocumented tape"). The quality of fingerprints developed on various substrates following exposure to foam is shown in Fig. 6. For all substrates, at least one unexposed control reached the threshold quality score of 2 (adequate for comparison). In contrast, although some ridge detail was visible on both sides of all tape samples, only a single adhesive-side electrical tape reached quality score 2. Metal flats exposed to the foam showed some nonusable ridge detail, but the other substrates showed none.

Fingerprint Quality Following Aqueous Gel Exposure

Results from samples exposed to the aqueous gel are graphed in Fig. 7. Quality scores were averaged within substrate type, as were the controls. With the exception of the nonadhesive side of electrical tape and metal pipe, the unexposed controls attained quality score 2 for all substrates in at least one case. Among samples exposed to the gel, all tapes showed ridge detail, but not always of

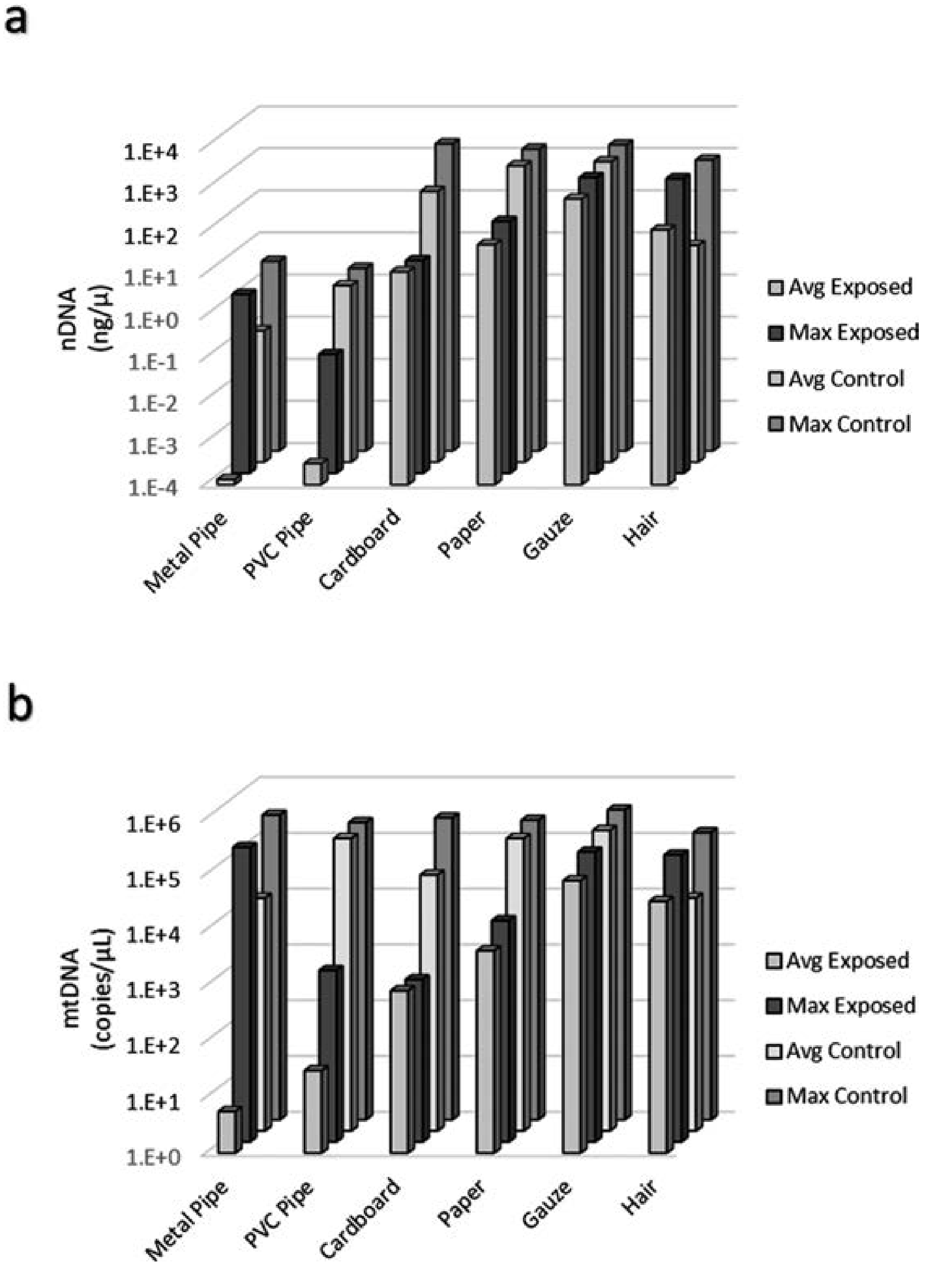


FIG. 4—*DNA recovered from blood deposited on items exposed to foam. (a) Shown are the geometric means (Avg) and maximal (Max) concentrations of nuclear DNA (nDNA) in DNA extracts of blood deposited on replicate items following exposure to foam, in comparison with the Avg and Max nDNA concentrations of extracts from blood deposited on unexposed, control replicate items. Many extracts, except those of blood recovered from pipe, met the target nDNA concentration of 100 ng/µL for downstream STR analysis (black values on the ordinate). (b) Shown are the Avg and Max concentrations of mitochondrial DNA (mtDNA) from replicate extracts. Most extracts met the target mtDNA concentration of 10 copies/µL for downstream Sanger sequencing of the whole control region (black values on the ordinate). Notes—Results from substrates exposed to render-safe (backpack and suitcase) and soak conditions were pooled. Metal pipe exposed included cleaned and "dirty" pipe and unblasted and blasted pipe, whereas metal pipe controls included both cleaned and "dirty" pipe. Replicates ranged from 1 for cardboard exposed, wherein the Avg equals the Max, to 16 for metal pipe exposed. Some replicates did not quantify DNA, for example, 1 of 5 cardboard controls (nDNA and mtDNA), 3 of 16 (nDNA) and 1 of 16 (mtDNA) metal exposed, and 1 of 7 PVC pipe exposed (mtDNA). [Color figure can be viewed at wileyonlinelibrary.com]*

value. The adhesive side yielded fingerprints of value (quality score 2) for at least one sample for all tape types. Usable ridge detail was obtained on the nonadhesive side of all packing tape and regular duct tape samples, but not for electrical or for extra strength duct tape. The other substrates showed ridge detail, but usable ridge detail was developed only on galvanized metal flats.

Hairs and Fibers Following Exposure to AFC-380 Foam and Aqueous Gel

Few of the fibers, no carpet fibers, and no hairs were recovered following the render-safe exercise using foam (Table 1). Those fibers that were recovered, and all hairs and fibers in the

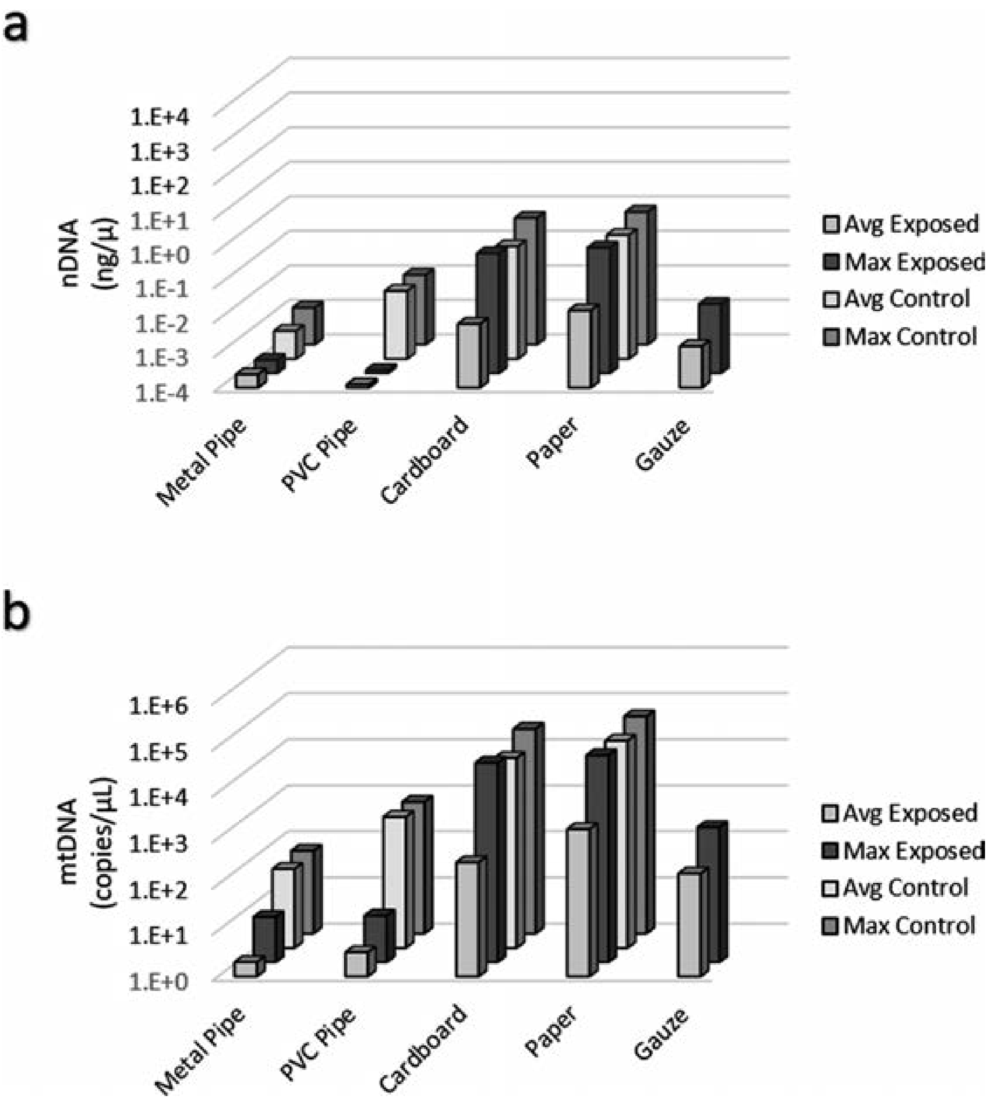


FIG. 5—*DNA recovered from blood deposited on items exposed to gel. (a) Shown are the geometric means (Avg) and maximal (Max) concentrations of nuclear DNA (nDNA) in DNA extracts of blood deposited on replicate items following exposure to gel, in comparison with the Avg and Max nDNA concentrations of extracts from blood deposited on unexposed, control replicate items. No extract met the target nDNA concentration of 100 ng/µL for downstream STR analysis (black values on the ordinate); however, this observation was unrelated to gel exposure since none of the controls met the nDNA target either. (b) Shown are the Avg and Max concentrations of mitochondrial DNA (mtDNA) from the same replicate extracts. In contrast to the nDNA results, several extracts met the target mtDNA concentration of 10 copies/µL for downstream Sanger sequencing of the mtDNA whole control region (black values on the ordinate). Notes—For both exposed and control items, two (2) replicates were tested for PVC pipe and gauze, while the remaining items had four (4). DNA was not extracted from blood on unexposed control gauze. [Color figure can be viewed at wileyonlinelibrary.com]*

foam soak condition, were positively associated with their respective controls via microscopy and spectroscopy. The same is true for all hairs and fibers exposed to the aqueous gel.

Documents Following Exposure to AFC-380 Foam and Aqueous Gel

All handwriting and machine printing by ink jet and laser was legible. No indented writing was recoverable from any document exposed to either foam or gel.

Discussion

DNA

Exposure to AFC-380 foam and gel resulted in a dilution effect on both mtDNA and nDNA for most items, and in general, mtDNA could be recovered from the experimental items even if nDNA was not recovered. The quality of nDNA and mtDNA recovered from items exposed to these agents was comparable to the untreated controls, that is, no inhibitors present and no or insignificant degradation. These results suggest that both nDNA and mtDNA can be recovered from bloodstains that have been exposed to AFC-380 foam and gel and from hair exposed to the former. It should be noted, however, that locating bloodstains, hairs, or other DNA evidence on items exposed to AFC-380 foam and gel may be difficult in a real-world scenario due to the aqueous nature of the foam. Many of the blood samples on the nonporous objects appeared to have been completely removed as a result of exposure to the foam and gel and were only located during processing in some instances because the location of the blood on each item had been marked during sample deposition. In addition, none of the hairs placed in the render-safe condition could be located in the persistent foam layer during sample recovery.

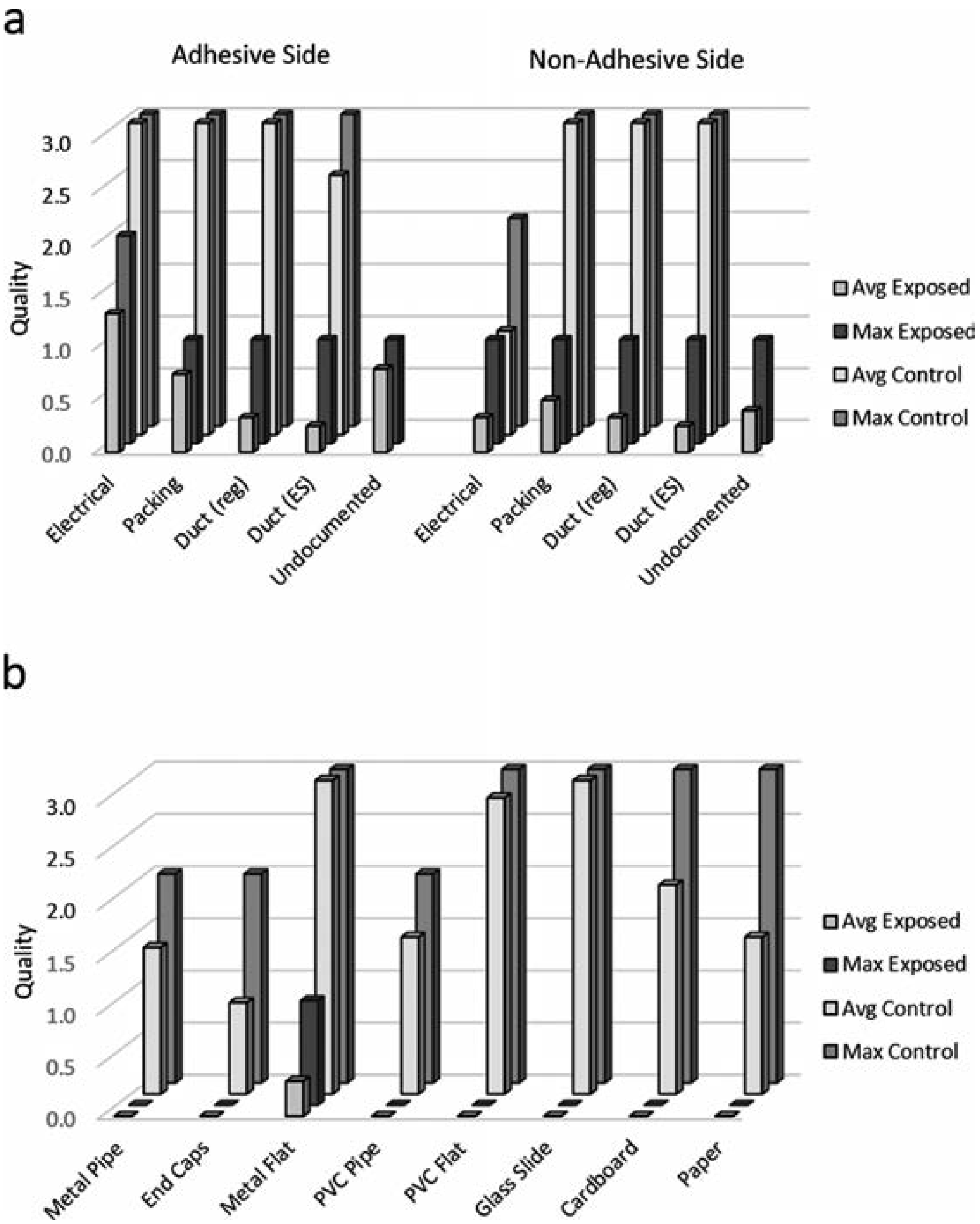


FIG. 6—*Average and maximum quality scores obtained after development of latent fingerprints on items exposed to AFC-380 foam. To be useful for comparison, quality must reach a threshold value of 2 (black values on the ordinate). (a) Shown are scores for adhesive and nonadhesive sides of various tapes. Only the adhesive side of foam-exposed electrical tape reached the quality threshold. Undocumented tapes were adventitious, therefore without controls. (b) Shown are additional nonporous and porous items. No foam-exposed items reached the quality threshold. Notes—Results from substrates exposed to render-safe (backpack and suitcase) and soak conditions were pooled. [Color figure can be viewed at wileyonlinelibrary.com]*

Despite the excellent results obtained from known marked samples, locating unknown DNA samples after exposure to AFC-380 foam or gel may be a limiting factor in forensic casework.

Fingerprints From AFC-380 Foam-Exposed Samples

In contrast to the results obtained from DNA samples, latent fingerprints on most porous and nonporous surfaces appear to be highly affected by exposure to AFC-380 blast suppression foam. Although ridge detail was recovered on both the adhesive and nonadhesive sides of several of the tape samples, ridge detail was only developed on one other experimental item, a metal flat in the soak condition. Originally, it was thought that the poor fingerprint recovery results for the nonporous items were due to a thick, mottled residue left behind by the foam that became particularly obscuring after cyanoacrylate fuming of the items. When the items were further processed using RAM and viewed under various light sources, the residue fluoresced and made it difficult to see any ridge detail that may have been present. This prompted further testing of several different residue removal methods and fingerprint processing methods for both porous and nonporous items. Experiments using alternative methods described in Appendix S1 also yielded very poor recovery results from supplemental items. Ridge detail was not developed on any of the exposed supplemental glass slides, regardless of rinse/soak time or processing method used. No ridge detail was observed on any of the porous supplemental items after processing with IND, physical developer, or WetPrint^{TM} (Lynn Peavy Co., Lenexa, KS).

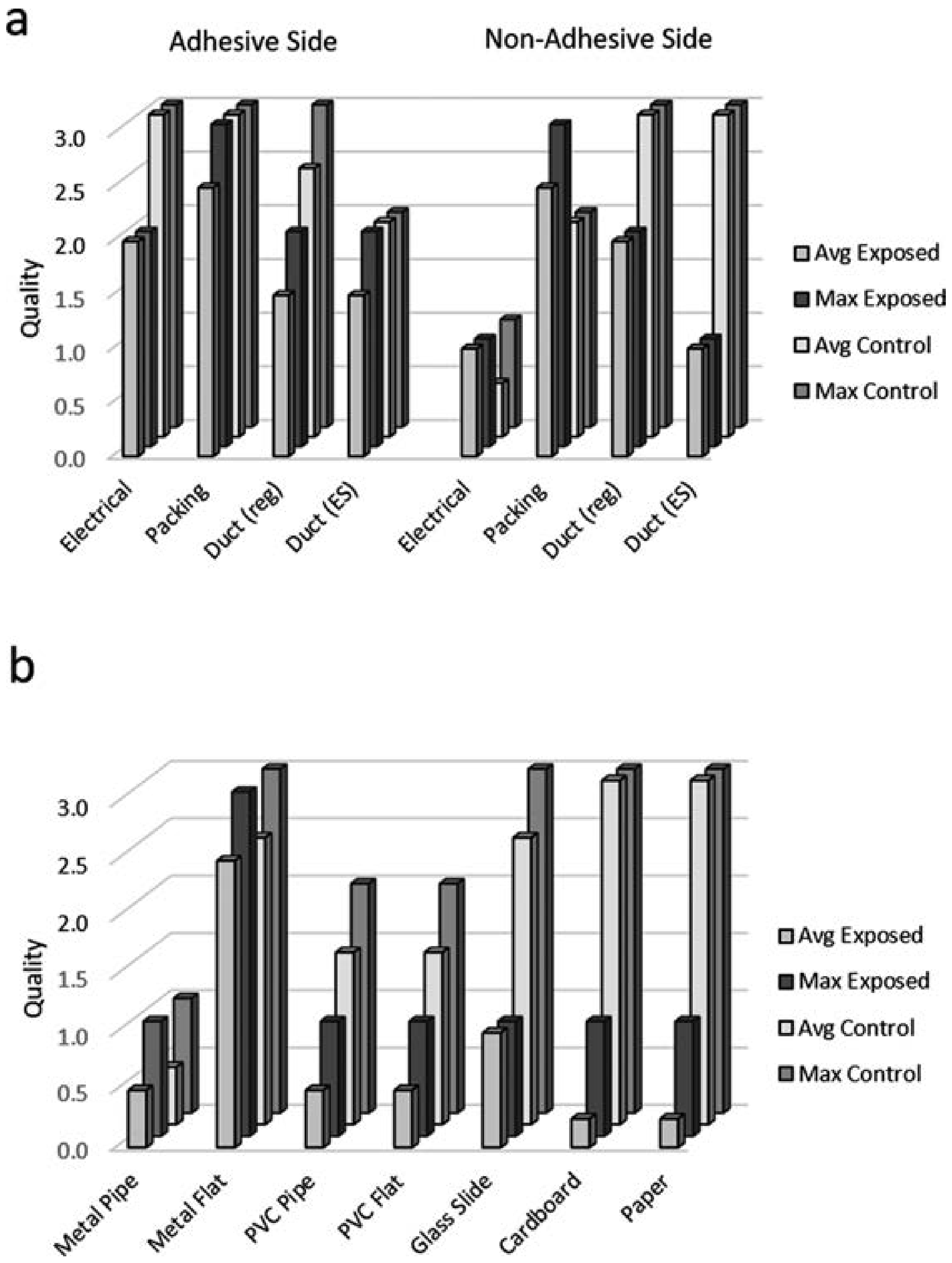


FIG. 7—*Average and maximum quality scores obtained after development of latent fingerprints on items exposed to aqueous gel. To be useful for comparison, quality must reach a threshold value of 2 (black values on the ordinate). (a) Shown are scores for adhesive and nonadhesive sides of various tapes. Except for the nonadhesive side of electrical and extra strength duct tape, at least one of all the other gel-exposed specimens reached the quality threshold. (b) Shown are additional nonporous and porous items. Only the metal flats reached the quality threshold. [Color figure can be viewed at wileyonlinelibrary.com]*

These results suggest that the foam itself may affect the preservation of fingerprints on most porous and nonporous surfaces, and attempts to process items exposed to AFC-380 foam for latent fingerprints will likely be unsuccessful. Of all substrates tested, tape specimens seem most likely to yield usable prints, particularly on the adhesive side.

Fingerprints From Aqueous Gel-Exposed Samples

Fingerprint development from various tapes and galvanized metal flats exposed to aqueous gel was usually successful, but not so for metal pipe, PVC, glass, cardboard, or paper substrates. For the metal items, these observations are contrary to those of Kuznetsov et al. (41), who reported greater postblast damage to fingerprints on light plates than on heavy plates. Successful development on tapes is plausible, due to their entrapment on the adhesive side or by residual adhesive on the other side, remaining when removed from the roll. Although the gel components are presumptively nonreactive in anhydrous state (37,90), solution behavior may at least partially explain the observed adverse effect on latent fingerprints on other substrates. One can visualize several potential interactions between sebaceous residues, which contain various polar lipid molecules (91), via solvation and ionic complexing by hydrated sodium polyacrylate

(92). When hydrated, sodium polyacrylate is an anionic polyelectrolyte. Sodium polyacrylate is used in detergents not only as a structurant (to give shape to fabric), but also as a substitute for phosphates, functioning as a builder and sequestrant (to capture and hold metal cations Ca^{2+} and Mg^{2+} found in hard water) (93). The outward-facing negative charges of the polyelectrolyte draw not only metallic cations, but also polar water molecules. This interaction with water explains polyabsorbency, but may cause attracted water molecules to orient themselves with respect to the polyacrylic acid chains with their oxygen atoms facing outward, creating an overall negatively charged surface.

Solvation could occur, as water-soluble polyelectrolyte gels will complex lipid surfactants stoichiometrically and in well-organized structures through ionic, hydrophobic, electrostatic, or hydrogen-bonding mechanisms (94-97). This interaction is roughly analogous to the way that fatty acid salts in soaps form micelles to surround nonpolar grease and oil molecules. Alternatively, the spatially oriented water molecules might function as a weak Brønsted–Lowry acid. The fractional charge of the polyacrylic acid molecule (i.e., degree of acid dissociation) is approximately 10-60% at pH 6 and pH 8, respectively, (98), that being our experimentally measured range of pH (precise pH measurement of a gel is difficult). By one or more of the proposed mechanisms, the aqueous gel may function as a pH-dependent, unprotonated polyacid (99) that could act to compromise the successful development of latent fingerprints. Further research would be required to elucidate the mechanisms involved.

Documents

As anticipated (81), indented writing exposed to either aqueous foam or gel could not be visualized by electrostatic detection apparatus. Handwriting and machine printing remained legible, although there was moderate lightening and diffusion of water-based pigments.

Hairs and Fibers

Microscopic appearance of hairs and fibers was unaffected by exposure to aqueous foam or gel. Under observation by polarized light microscopy and fluorescence microscopy, a small amount of debris was noted on the exterior of several fibers in gel samples that auto-fluoresced (even without exposure to specific wavelengths of light) and had a small amount of fluorescence. There was also a small amount of debris that fluoresced in the UV, violet, blue, and green light. In no case did this debris interfere with the fiber examination and assessment of fiber fluorescence as it was clearly exterior debris. Typically, during fiber examinations washing potential fibers prior to mounting is avoided to prevent the loss of any existing trace evidence that may be adhered to the exterior of the fiber (e.g., biological material). That being said, under these conditions it may be necessary to wash the fibers if the foam/gel obscure microscopic characteristics or interfere with the evaluation and assessment of optical properties and fluorescence.

Conclusions

Biological material exposed to AFC-380 foam and aqueous gel yielded nDNA and mtDNA that was not degraded or contained amplification inhibitors, demonstrating that neither agent reduced the quality of the DNA extracted. However, exposure to these agents resulted in washing away and/or dilution of some biological material, thereby reducing DNA yields, and this effect was more pronounced for hard impervious surfaces such as pipe. Nonetheless, nDNA and/or mtDNA analysis may be possible from foam- or gel-exposed items, but would be dependent on the amount of biological material, which was originally deposited and subsequently recovered in a forensic laboratory. In contrast, AFC-380 foam and aqueous gel blocks significantly affected latent fingerprints on nearly all porous and nonporous items tested, with marginally improved results from the adhesive and nonadhesive sides of several common types of tape. Exposed hairs and fibers were positively associated with controls via microscopy and spectroscopy. Indented writing was undetectable after exposure to either foam or gel, but inks and toners were unaffected. In any incident involving explosives and the use of AFC-380 foam and aqueous gel blocks, locating usable evidentiary samples will always present its own challenge.

Acknowledgments

We thank Brian Eckenrode for helpful discussions regarding the hypothesis that hydrated gel may function as a pH-dependent polyacid. We thank J. Stephen Veyera (Vickers and Nolan Enterprises), Keith Frakes (Sandia National Laboratory), and Bradley Koons, Byron Cody, and Marvin Garcia (FBI Albuquerque) for their assistance in logistics, render-safe procedures, and postblast sample recovery.

References

1. Medalia J."Dirty bombs": technical background, attack prevention and response, issues for Congress (CRS Report R41891). Congressional Research Service. 2011. https://fas.org/sgp/crs/nuke/R41890.pdf (accessed July 9, 2020).
2. Senate US. Dirty bombs and basement nukes: the terrorist nuclear threat (S. HRG. 107–575). 2nd edn. Washington, D.C.: U.S. Government Printing Office, 2002. https://www.govinfo.gov/app/details/CHRG-107shrg80848 (accessed July 9, 2020).
3. Monson K, Ali S, Brandhagen M, Duff M, Fisher C, Lowe K, et al. Effects of ionizing radiation on the evidentiary value of DNA, latent fingerprints, hair, and fibers: a comprehensive review and new results. Forensic Sci Int 2018;284:204–18. https://doi.org/10.1016/j.forsciint.2018.01.012
4. Connell LW.Dirty bomb risk and impact (SAND2017-9121R). Sandia National Laboratories. 2017. https://www.osti.gov/servlets/purl/1378173 (accessed July 9, 2020).
5. DHS. Planning guidance for protection and recovery following radiological dispersal device (RDD) and improvised nuclear device (IND) incidents (E8–17645). Fed Regist 2008;73(149):45029–48.
6. Kaminski MD, Lee SD, Magnuson M. Wide-area decontamination in an urban environment after radiological dispersion: a review and perspectives. J Hazard Mater 2016;305:67–86. https://doi.org/10.1016/j.jhazmat.2015.11.014
7. Wente WB.Unconventional Nuclear Warfare Defense (UNWD) containment and mitigation subtask (SAND2005-2859). Sandia National Laboratories. 2005. https://www.osti.gov/scitech/servlets/purl/923086 (accessed July 9, 2020).
8. Review NATO. The dirty bomb: low cost, high risk. 2010 https://www.nato.int/docu/review/2010/Nuclear_Proliferation/dirty_bomb/EN/index.htm
9. Burns WJ, Slovic P. The diffusion of fear: modeling community response to a terrorist strike. J Defense Model Simu 2007;4(4):298–317. https://doi.org/10.1016/j.jenvrad.2009.07.006
10. Giesecke JA, Burns WJ, Barrett A, Bayrak E, Rose A, Slovic P, et al. Assessment of the regional economic impacts of catastrophic events: CGE analysis of resource loss and behavioral effects of an RDD attack scenario. Risk Anal 2012;32(4):583–600. https://doi.org/10.1111/j.1539-6924.2010.01567.x
11. Tucker MD, Gao H, inventors. Sandia National Laboratory, assignee. Highly concentrated foam formulation for blast mitigation. US patent 7,850,865. 2010.

12. Hoffman DM, Mitchell AR.Development of defoamers for confinement foam (LLNL-02-010). Lawrence Livermore National Laboratory. 2003. https://www.osti.gov/scitech/servlets/purl/878603 (accessed July 9, 2020).
13. McRoberts VM, Martell M-A, Jones JA. Equipment compatibility and logistics assessment for containment foam deployment (SAND2005-5793). Sandia National Laboratories. 2005. https://www.osti.gov/scitech/servlets/purl/876288 (accessed July 9, 2020).
14. Rand PB, inventor. United States Department of Energy, assignee. Stabilized aqueous foam systems and concentrate and method for making them. US patent 4,442,018A. 1984.
15. Baer M, Cooper P, Kipp M. Investigations of emergency destruction methods for recovered, explosively configured, chemical warfare munitions: interim emergency destruction methods-evaluation report (SAND95-8248). Sandia National Laboratories. 1995. https://www.osti.gov/servlets/purl/86893 (accessed July 9, 2020).
16. Britan A, Shapiro H, Liverts M, Ben-Dor G, Chinnayya A, Hadjadj A. Macro-mechanical modelling of blast wave mitigation in foams. Part I: review of available experiments and models. Shock Waves 2013;23 (1):5–23. https://doi.org/10.1007/s00193-012-0417-4
17. Jones JA, McRoberts VM, Martell M-A. Equipment compatibility and logistics assessment for containment foam deployment (SAND2005-5793). Sandia National Laboratories. 2005. https://www.osti.gov/scitech/servlets/purl/876288 (accessed July 9, 2020).
18. Winfield F, Hill D. Preliminary results on the physical properties of aqueous foams and their blast attenuating characteristics. Defence Research Establishment Suffield Ralston (Alberta), 1977. http://www.dtic.mil/get-tr-doc/pdf?AD=ADA045650 (accessed July 9, 2020).
19. Clark CJ, Bennett EM, inventors. Wormald US Inc, assignee. Method for explosive blast control using expanded foam. US patent 4,589,341. 1986.
20. Hartman WF, Larsen ME, Boughton BA.Blast mitigation capabilities of aqueous foam (SAND2006-0533). Sandia National Laboratories, 2006. https://www.osti.gov/scitech/servlets/purl/877732 (accessed July 9, 2020).
21. Ballanger F, Counilh D, Rambert N, Lefrançois A, Haas J, Chinnayya A. Study of small scale experiments of detonations in an aqueous foam confinement. Proceedings of the 25th International Colloquium on the Dynamics of Explosions and Reactive Systems (ICDERS). 2015; Leeds, U.K. http://www.icders.org/ICDERS2015/abstracts/ICDERS2015-077.pdf (accessed July 9, 2020).
22. Jourdan G, Mariani C, Houas L, Chinnayya A, Hadjadj A, Del Prete E, et al. Analysis of shock-wave propagation in aqueous foams using shock tube experiments. Phys Fluids 2015;27(5):056101. https://doi.org/10.1063/1.4919905
23. Panczak TD, Krier H, Butler PB. Shock propagation and blast attenuation through aqueous foams. J Hazard Mater 1987;14(3):321–36. https://doi.org/10.1016/0304-3894(87)85004-5
24. Aubert JH, Kraynik AM, Rand PB. Aqueous foams. Sci Am 1986;254 (5):74–83.
25. Igra O, Falcovitz J, Houas L, Jourdan G. Review of methods to attenuate shock/blast waves. Prog Aerospace Sci 2013;58:1–35. https://doi.org/10.1016/j.paerosci.2012.08.003
26. Larsen ME. Aqueous foam mitigation of confined blasts. Int J Mechan Sci 1992;34(6):409–18. https://doi.org/10.1016/0020-7403(92)90008-5
27. Peregino PJ, Bowman D, Maulbetsch R, Saunders D, Vande Kieft L. Blast and fragmentation suppression with aqueous foam and a Kevlar tent. Proceedings of the 28th Department of Defense Explosives Safety Seminar; 1998 Aug 18–20; Orlando, FL. 1998. https://apps.dtic.mil/sti/pdfs/ADA513200.pdf (accessed July 9, 2020).
28. Weaire DL, Hutzler S. The physics of foams. Oxford, U.K.: Oxford University Press, 1999.
29. Parra R, Medina V, Conca J. The use of fixatives for response to a radiation dispersal devise attack–a review of the current (2009) state-of-the-art. J Environ Radioactiv 2009;100(11):923–34. https://doi.org/10.1016/j.jenvrad.2009.07.006
30. Wilkinson DA, Sweet D, Fairley D. Recovery of DNA from exhibits contaminated with chemical warfare agents: a preliminary study of the effect of decontamination agents and chemical warfare agents on DNA. Can Soc Forensic Sci J 2007;40(1):15–22. https://doi.org/10.1080/00085030.2007.10757148
31. Zuidberg M, van Woerkom T, de Bruin K, Stoel R, de Puit M. Effects of CBRN decontaminants in common use by first responders on the recovery of latent fingerprints—assessment of the loss of ridge detail on glass. J Forensic Sci 2014;59(1):61–9. https://doi.org/10.1111/1556-4029.12281
32. Wilkinson D, Hancock J, Lecavalier P, McDiarmid C. The recovery of fingerprint evidence from crime scenes contaminated with chemical warfare agents. J Forensic Identif 2005;55(3):326–61.
33. Noonan BJ, Steves HK. The performance of small arms ammunition when fired into water (NOLTR 70–174). Naval Ordnance Lab. 1970. http://www.dtic.mil/dtic/tr/fulltext/u2/713445.pdf (accessed July 9, 2020).
34. Nicholas N, Welsch J. Institute for non-lethal defense technologies report: ballistic gelatin. Pennsylvania State University Applied Research Lab. 2004. http://www.dtic.mil/get-tr-doc/pdf?AD=ADA446543 (accessed July 9, 2020).
35. WMDTech. Water gel system. 2018. https://www.wmdtech.com/products/containment/water-gel-system (accessed July 9, 2020).
36. Global Safety Management. Safety data sheet: sodium polyacrylate. 2015https://beta-static.fishersci.com/content/dam/fishersci/en_US/documents/programs/education/regulatory-documents/sds/chemicals/chemicals-s/S25565.pdf (accessed July 9, 2020).
37. Fisher Science Education. Safety data sheet: sodium polyacrylate. 2015. https://beta-static.fishersci.com/content/dam/fishersci/en_US/documents/programs/education/regulatory-documents/sds/chemicals/chemicals-s/S25565.pdf (accessed July 9, 2020).
38. Morton HS. Scaling the effects of air blast on typical targets. Johns Hopkins University Applied Physics Lab. 1966. http://www.dtic.mil/get-tr-doc/pdf?AD=AD0481144 (accessed July 9, 2020).
39. Ramasamy S, Houspian A, Knott F. Recovery of DNA and fingermarks following deployment of render-safe tools for vehicle-borne improvised explosive devices (VBIED). Forensic Sci Int 2011;210(1):182–7. https://doi.org/10.1016/j.forsciint.2011.03.006
40. McCarthy D. Latent fingerprint recovery from simulated vehicle-borne improvised explosive devices. J Forensic Identif 2012;62(5):488–516.
41. Kuznetsov VA, Sunde J, Thomas M. Explosive blast effects on latent fingerprints. Proceedings of the First International Conference on Forensic Applications and Techniques in Telecommunications, Information, and Multimedia and Workshop. Adelaide, Australia. Brussels, Belgium: ICST, 2008;1–4. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.683.3398&rep=rep1&type=pdf (accessed July 9, 2020).
42. Bille TW, Cromartie C, Farr M. Effects of cyanoacrylate fuming, time after recovery, and location of biological material on the recovery and analysis of DNA from post-blast pipe bomb fragments. J Forensic Sci 2009;54(5):1059–67. https://doi.org/10.1111/j.1556-4029.2009.01128.x
43. Berti A, Barni F, Virgili A, Colozza C, Maiorino F, Tocca M. The recovery of DNA profiles from saliva and touch evidences after postal bomb explosion. Forensic Sci Int-Gen 2011;3(1):e471–2. https://doi.org/10.1016/j.fsigss.2011.09.097
44. Phetpeng S, Kitpipit T, Thanakiatkrai P. Systematic study for DNA recovery and profiling from common IED substrates: from laboratory to casework. Forensic Sci Int Genet 2015;17:53–60. https://doi.org/10.1016/j.fsigen.2015.03.007
45. Hoffmann SG, Stallworth SE, Foran DR. Investigative studies into the recovery of DNA from improvised explosive device containers. J Forensic Sci 2012;57(3):602–9. https://doi.org/10.1111/j.1556-4029.2011.01982.x
46. Foran DR, Gehring ME, Stallworth SE. The recovery and analysis of mitochondrial DNA from exploded pipe bombs. J Forensic Sci 2008;54 (1):90–4. https://doi.org/10.1111/j.1556-4029.2008.00901.x
47. Esslinger KJ, Siegel JA, Spillane H, Stallworth S. Using STR analysis to detect human DNA from exploded pipe bomb devices. J Forensic Sci 2004;49(3):481–4. https://doi.org/10.1520/JFS2003127
48. Chargaff E. Isolation and composition of the deoxypentose nucleic acids and of the corresponding nucleoproteins. In: Chargaff E, Davidson JN, editors. The nucleic acids. New York, NY: Academic Press, 1955;307–407.
49. Nasiri H, Forouzandeh M, Rasaee MJ, Rahbarizadeh F. Modified salting-out method: high-yield, high-quality genomic DNA extraction from whole blood using laundry detergent. J Clin Lab Anal 2005;19(6):229–32. https://doi.org/10.1002/jcla.20083
50. Kulstein G, Wiegand P. Comprehensive examination of conventional and innovative body fluid identification approaches and DNA profiling of laundered blood- and saliva-stained pieces of cloths. Int J Legal Med 2018;132(1):67–81. https://doi.org/10.1007/s00414-017-1691-6
51. Edler C, Gehl A, Kohwagner J, Walther M, Krebs O, Augustin C, et al. Blood trace evidence on washed textiles – a systematic approach. Int J Legal Med 2017;131(4):1179–89. https://doi.org/10.1007/s00414-017-1549-y
52. Beresford AL, Brown RM, Hillman AR, Bond JW. Comparative study of electrochromic enhancement of latent fingerprints with existing development techniques. J Forensic Sci 2012;57(1):93–102. https://doi.org/10.1111/j.1556-4029.2011.01908.x
53. Salahuddin Z, Yasir Zahoor M, Kalsoom S, Rakha A. You can't hide encoded evidence: DNA recovery from different fabrics after washing. Austral J Forensic Sci 2018;50(4):355–60. https://doi.org/10.1080/00450618.2016.1237545

54. Bogas V, Carvalho M, Anjos MJ, Corte-Real F. Genetic identification of degraded and/or inhibited DNA samples. Austral J Forensic Sci 2016;48(4):381–406. https://doi.org/10.1080/00450618.2015.1069894
55. Kulstein G, Wiegand P. DNA/RNA co-analysis of seminal fluid-stained fabrics after water immersion for up to seven days. Forensic Sci Int Genic 2017;6:e27–8. https://doi.org/10.1016/j.fsigss.2017.09.015
56. Maslanka DS. Latent fingerprints on a nonporous surface exposed to everyday liquids. J Forensic Identif 2016;66(2):137–54.
57. Soltyszewski I, Moszczynski J, Pepinski W, Jastrzebowska S, Makulec W, Zbiec R, et al. Fingerprint detection and DNA typing on objects recovered from water. J Forensic Identif 2007;57(5):681–7.
58. Tontarski KL, Hoskins KA, Watkins TG, Brun-Conti L, Michaud AL. Chemical enhancement techniques of bloodstain patterns and DNA recovery after fire exposure. J Forensic Sci 2009;54(1):37–48. https://doi.org/10.1111/j.1556-4029.2008.00904.x
59. De Paoli G, Lewis SA Sr, Schuette EL, Lewis LA, Connatser RM, Farkas T. Photo- and thermal-degradation studies of select eccrine fingerprint constituents. J Forensic Sci 2010;55(4):962–9. https://doi.org/10.1111/j.1556-4029.2010.01420.x
60. Smyth A, Sims MR, Holt J. Detection of fingermarks from post-blast debris: a review. J Forensic Identif 2018;68(3):369–78.
61. Gardner SJ, Cordingley TH, Francis SC. An investigation into effective methodologies for latent fingerprint enhancement on items recovered from fire. Sci Justice 2016;56(4):241–6. https://doi.org/10.1016/j.scijus.2016.02.003
62. Bradshaw G, Bleay S, Deans J, NicDaeid N. Recovery of fingerprints from arson scenes: part 1-latent fingerprints. J Forensic Identif 2008;56(1):18–23.
63. Sutton R, Grenci C, Hrubesova L. A comparison on the longevity of submerged marks in field and laboratory conditions. J Forensic Identif 2014;64(2):143–56.
64. Fieldhouse S, Needham M. Latent fingermark longevity on non-porous surfaces in tap and salt water environments. Fingerprint Whorld 2016;41 (159):6–19.
65. Bleay S, Sears V, Downham R, Bandey H, Gibson A, Bowman V, et al. Fingerprint source book, 2nd edn. St. Albans, U.K.: Home Office Centre for Applied Science and Technology (CAST), 2017;396–653. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/678466/fingerprint-source-book-v2.pdf (accessed July 9, 2020).
66. Simmons RK, Deacon P, Farrugia KJ. Water-soaked porous evidence: a comparison of processing methods. J Forensic Identif 2014;64(2):157–73.
67. Jasuja OP, Kumar P, Singh G. Development of latent fingermarks on surfaces submerged in water: optimization studies for phase transfer catalyst (PTC) based reagents. Sci Justice 2015;55(5):335–42. https://doi.org/10.1016/j.scijus.2015.03.001
68. Dhall JK, Kapoor AK. Development of latent prints exposed to destructive crime scene conditions using wet powder suspensions. Egypt J Forensic Sci 2016;6(4):396–404. https://doi.org/10.1016/j.ejfs.2016.06.003
69. Trapecar M. Finger marks on glass and metal surfaces recovered from stagnant water. Egypt J Forensic Sci 2012;2(2):48–53. https://doi.org/10.1016/j.ejfs.2012.04.002
70. Wood M, James T. Latent fingerprint persistence and development techniques on wet surfaces. Fingerprint Whorld 2009;35(135):90–100.
71. Wilkinson D. Friction ridge detection from challenging crime scenes. In: Ramotowski R, editor. Lee and Gaensslen's advances in fingerprint technology, 3rd edn. Boca Raton, FL: CRC Press, 2012;381–408.
72. Cohen D, Cohen EH. A significant improvement to the SPR process: more latent prints were revealed after thorough wiping of small particle reagent-treated surface. J Forensic Identif 2010;60(2):152–62.
73. Steiner R, Bécue A. Effect of water immersion on multi-and monometallic VMD. Forensic Sci Int 2018;283:118–27. https://doi.org/10.1016/j.forsciint.2017.12.020
74. Champod C, Lennard CJ, Margot P, Stoilovic M. Fingerprint detection techniques. In: Fingerprints and other ridge skin impressions. Boca Raton, FL: CRC Press, 2016;105–79.
75. Madkour S, Abeer S, El Dine FB, Elwakeel Y, AbdAllah N. Development of latent fingerprints on non-porous surfaces recovered from fresh and sea water. Egypt J Forensic Sci 2017;7(1):3. https://doi.org/10.1186/s41935-017-0008-8
76. Cohen Y, Azoury M, Elad ML. Survivability of latent fingerprints part II: the effect of cleaning agents on the survivability of latent fingerprints. J Forensic Identif 2012;62(1):54–61.
77. Chadwick S, Neskoski M, Spindler X, Lennard C, Roux C. Effect of hand sanitizer on the performance of fingermark detection techniques. Forensic Sci Int 2017;273:153–60. https://doi.org/10.1016/j.forsciint.2017.02.018
78. Lepot L, Vanden Driessche T, Lunstroot K, Gason F, De Wael K. Fibre persistence on immersed garment – influence of knitted recipient fabrics. Sci Justice 2015;55(4):248–53. https://doi.org/10.1016/j.scijus.2015.02.006
79. Lepot L, Vanden Driessche T. Fibre persistence on immersed garment – influence of water flow and stay in running water. Sci Justice 2015;55 (6):431–6. https://doi.org/10.1016/j.scijus.2015.09.003
80. Santacroce G. The forensic examination of fire and water-damaged documents. Int J Forensic Doc Exam 1999;5(1):76–82.
81. SWGDOC. SWGDOC standard for indentation examinations. Scientific Working Group for Forensic Document Examination, 2015. https://www.swgdoc.org/documents/SWGDOC%20Standard%20for%20Indentation%20Examinations.pdf (accessed July 9, 2020).
82. Riebeling IJ, Kobus HJ. Some parameters affecting the quality of ESDA results. J Forensic Sci 1994;39(1):15–20. https://doi.org/10.1520/JFS13566J
83. Sears VG, Bleay SM, Bandey HL, Bowman VJ. A methodology for finger mark research. Sci Justice 2012;52(3):145–60. https://doi.org/10.1016/j.scijus.2011.10.006
84. Ellen DM, Foster DJ, Morantz DJ. The use of electrostatic imaging in the detection of indented impressions. Forensic Sci Int 1980;15(1):53–60. https://doi.org/10.1016/0379-0738(80)90195-4
85. Kavlick MF. Development of a triplex mtDNA qPCR assay to assess quantification, degradation, inhibition, and amplification target copy numbers. Mitochondrion 2019;46:41–50. https://doi.org/10.1016/j.mito.2018.09.007
86. Trozzi T, Schwartz R, Hollars M. Processing guide for developing latent prints. Washington, D.C.: FBI Laboratory, 2000. http://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/jan2001/lpu.pdf (accessed July 9, 2020).
87. Bécue A, Scoundrianos A, Moret S. Detection of fingermarks by colloidal gold (MMD/SMD)–beyond the pH 3 limit. Forensic Sci Int 2012;219(1–3):39–49. https://doi.org/10.1016/j.forsciint.2011.11.024
88. SWGDOC. SWGDOC standard for preservation of liquid soaked documents. Scientific Working Group for Forensic Document Examination, 2013. https://www.swgdoc.org/documents/SWGDOC%20Standard%20for%20Preservation%20of%20Liquid%20Soaked%20Documents.pdf (accessed July 9, 2020).
89. Satoh M, Kuroiwa T. Organization of multiple nucleoids and DNA molecules in mitochondria of a human cell. Exp Cell Res 1991;196 (1):137–40. https://doi.org/10.1016/0014-4827(91)90467-9
90. Corp Cabot.Safety data sheet: CAB-O-SIL® TS-622 fumed silica, 2018. www.cabotcorp.com/~/media/files/msds/fmo/thai/en/ts622-thai-en.pdf (accessed July 9, 2020).
91. Ramotowski R. Composition of latent print residue. In: Lee H, Gaensslen R, editors. Advances in fingerprint technology, 2nd edn. Boca Raton, FL: CRC Press, 2001;63–104.
92. Lankalapalli S, Kolapalli VRM. Polyelectrolyte complexes: a review of their applicability in drug delivery technology. Indian J Pharm Sci 2009;71(5):481–7. https://doi.org/10.4103/0250-474X.58165
93. Hollingsworth MW. Role of detergent builders in fabric washing formulations. J Am Oil Chem Soc 1978;55(1):49–51. https://doi.org/10.1007/bf02673389
94. Kwon HJ, Osada Y, Gong JP. Polyelectrolyte gels-fundamentals and applications. Polymer J 2006;38(12):1211–9. https://doi.org/10.1295/polymj.PJ2006125
95. Gole A, Phadtare S, Sastry M, Langevin D. Studies on interaction between similarly charged polyelectrolyte: fatty acid system. Langmuir 2003;19(22):9321–7. https://doi.org/10.1021/la0352063
96. Diamant H, Andelman D. Self-assembly in mixtures of polymers and small associating molecules. Macromolecules 2000;33(21):8050–61. https://doi.org/10.1021/ma991021k
97. Wei Y-C, Hudson SM. The interaction between polyelectrolytes and surfactants of opposite charge. J Macromol Sci Pol Rev 1995;35(1):15–45. https://doi.org/10.1080/15321799508014588
98. Swift T, Swanson L, Geoghegan M, Rimmer S. The pH-responsive behaviour of poly (acrylic acid) in aqueous solution is dependent on molar mass. Soft Matter 2016;12(9):2542–9. https://doi.org/10.1039/C5SM02693H
99. Mafé S. Estimation of pKa shifts in weak polyacids using a simple molecular model: effects of strong polybases, hydrogen bonding and divalent counterion binding. Chem Phys 2004;296(1):29–35. https://doi.org/10.1016/j.chemphys.2003.09.033

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Supplementary Experiments.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14501
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Katie M. White,[1] *M.S.; and Christopher S. Palenik*,[1] *Ph.D.*

Toner Particles as Forensic Evidence: Microanalytical Characterization of Known Toner and Recognition of Toner in Environmental Samples*'†'‡

ABSTRACT: Modern printing toners represent a prime example of subvisible particles that can be easily transferred to hands, clothing, and other surfaces. To explore the potential evidentiary value of toner particles, toner samples were collected from known printer cartridges and characterized by various microanalytical techniques to establish the properties most useful for recognition, identification, and comparison. Environmental samples (i.e., dust) were then collected from various locations at varying distances from toner-based printers, using both tape lifts and carbon adhesive stubs, to assess the possibility of detecting toner. By light microscopy, toner can be recognized on the basis of particle size and shape, as well as color. Further examination of the micromorphology in the field emission scanning electron microscope reveals characteristic morphologies and differences in surface texture and shape among toner sources. Raman spectroscopy provides chemical identification of the pigment (or pigment class) and, in some cases, also permits identification of the polymer component. While black and blue pigment chemistry remained constant among toner varieties that were studied (copper phthalocyanine and carbon black), variation in yellow and magenta pigments was observed. Analysis of dust samples collected from various environments demonstrated that while toner is consistently detectable in close proximity to printers (within 2 feet), it also can be detected in dust collected in nearby rooms. This research demonstrates that toner particles can be located, characterized, and discriminated, using a suite of microanalytical methods that are applicable to forensic casework.

KEYWORDS: laser printers and photocopiers, toner, trace evidence, nanoparticles, dust analysis, Raman spectroscopy

Although abundant and nearly ubiquitous, the wide varieties of subvisible particles and nanoparticles (100 µm–10 nm) that may be encountered in ordinary dust are rarely utilized as forensic evidence. The majority of forensic trace evidence laboratories generally limit their focus to larger particles, often ignoring these smaller particles and features (with the notable exception of GSR evidence). While not typically exploited as forensic evidence, subvisible particles are readily found in nature, generated as dusts in various anthropogenic processes, and engineered for a growing variety of applications in consumer and industrial products. Exploitation of such particles represents a natural extension of current trace evidence examinations.

Toner-based laser printers and photocopiers are found in homes and offices around the world. Toner is a free-flowing powder used in these printers to form printed text and images on media. Modern printing typically uses a black (K) toner, either alone or in combination with cyan (C), magenta (M), and yellow (Y) toners. Through strategic overlapping of these four colors, in a process referred to as CMYK printing, a significant region of color space can be reproduced. The fine toner particles used in printing processes are not transferred to paper with perfect efficiency, and as such, a trace, but certainly nontrivial, quantity of these particles are deposited loose onto printed materials or dispersed into the surrounding environment. In turn, these particles can be picked up and transferred by secondary and higher-order processes (1).

Traditionally, toner powder was made by dry mixing resin particles with pigments and other additives, which were heated and kneaded to fuse the components into a slab (2,3). This mass was then pulverized and air jet milled into a superfine powder, producing toner particles with irregular shapes and size ranges. The majority of toner manufacturers now prepare toners using more advanced chemical synthesis methods, which result in toner particles with more uniform shapes, a narrower size range, and carefully engineered nanostratigraphies and nanotexture. For example, Fig. 1 shows an ion milled cross section of a toner particle in which a sheath-core nanostructure is apparent. This level of control over the manufacturing process provides improved print resolution and color reproduction as well as particles with

[1] Microtrace, LLC, 790 Fletcher Drive, Suite 106, Elgin, IL, 60123.

Corresponding author: Katie M. White, M.S. E-mail: kwhite@microtrace.com

*Presented in whole or in part at the Inter/Micro 68th Annual Meeting, June 6-10, 2016, in Chicago, IL; the Inter/Micro 69th Annual Meeting, June 12-16, 2017, in Chicago, IL; and at PittCon, February 26-March 1, 2018, in Orlando, FL.

†Supported in part by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice.

‡The opinions, findings, and conclusions or recommendations expressed in the publication/program/exhibition are those of the author(s) and do not necessarily reflect those of the Department of Justice.

Received 14 April 2020; and in revised form 9 June 2020; accepted 11 June 2020.

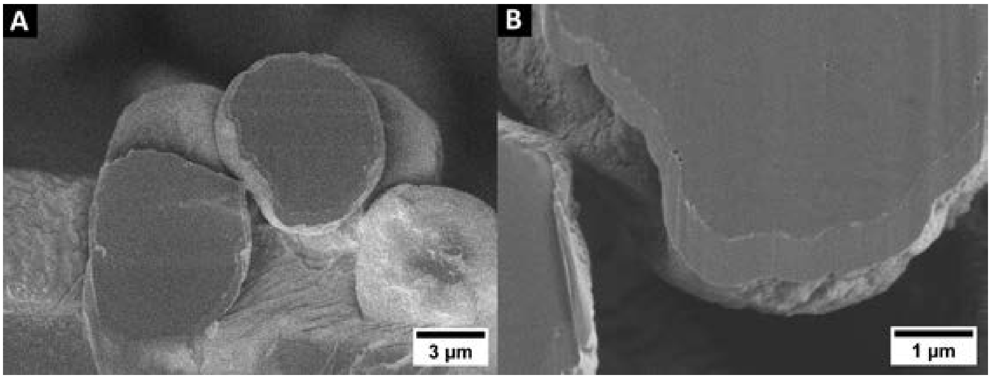


FIG. 1—*Ion milled cross section of a T010 toner particle.*

well-defined characteristics that might be exploited as trace evidence.

Toner powders are composed of two main ingredients: a colorant, dispersed in the toner to impart color (typically 4%–20% of the formulation), and a polymeric resin to transport the pigment during printing and fuse it to the paper (typically between 40–95% of the formulation) (4). Some of the more common polymers used in toners are styrene copolymers (*e.g.*, acrylic and butadiene modified) and polyesters. Pigments (as opposed to dyes) are primarily utilized to impart color to toners, with most manufacturers using azo or polycyclic pigments (5). The balance of a toner formulation consists of surface additives and binding agents, which are added at lower concentrations to achieve other performance-related properties. Such additives may include fumed silica, fumed titanium dioxide, metal stearates, fluoropolymer powders, magnetite, and cerium oxide (4).

Over the past 30 years, there have been many studies on the analysis of toner, including an in-depth three-part series examining the forensic discrimination of different toner samples (6–8). Researchers have focused on the use of scanning electron microscopy with energy-dispersive X-ray spectroscopy (SEM-EDS) (8–10); pyrolysis gas chromatography (py-GC) (8,11–15); infrared spectroscopy, both diffuse reflectance and infrared-reflection absorption (6,7,13–22); and thin layer chromatography (TLC) (19,23,24) for the analysis and comparison of toners, among others (9,15,25–27). In more recent years, Raman has additionally been suggested as a technique for the analysis of toner (27–30). However, the existing literature on toner analysis is geared toward document examination and focuses on the analysis of toner that has been printed on paper. Those that do study toner powder directly have performed analysis on bulk samples rather than individual particles that may be encountered as trace evidence. While the environmental sciences have examined potential hazards of airborne toner particles as respiratory hazards (31–33), we are not aware of any applications of toner dust as a forensic indicator or tracer of particular environments or printers.

The goal of this research is to explore the value of finding and characterizing subvisible particles (such as toner) by demonstrating the micro- and nano-morphological and chemical information that can be obtained from single particles that approach the resolution limits of light microscopy and whose engineered features can exceed this resolving power. It is also intended to assist the forensic practitioner who may encounter potential toner particles through a microanalytical study of toner particle morphology and the chemical properties of 53 known toner reference samples with the goal of establishing a basis by which toner particles can be (i) identified, (ii) discriminated, and (iii) used as an investigative probe for comparing or sourcing dust evidence in forensic investigation (34).

Materials and Methods

Known Sample Collection

Fifty-three (53) known toner samples were collected from the toner cartridges of laser printers and photocopiers in both home and office environments, as well as in a commercial printing facility (Table 1 and Fig. 2). Using a dry cotton swab, the opening of the cartridge was gently wiped to collect loose particles. For analysis, these reference particles were released from a swab by teasing it with a tungsten needle over a sample substrate.

Polarized Light Microscopy

Toner particles were mounted on a microscope slide in refractive index oil (n = 1.540) for examination with a Zeiss Axio Imager A2m polarized light microscope (PLM) using transmitted light. The microscopical features of each toner, including the particle shape, size and density of pigment granules (within the polymer matrix), and the presence of any anisotropic components, were observed and recorded. Photomicrographs were collected from each sample, at a lower (200×) and higher (1000×) magnification, using a Leica DFC540 camera. Particle size ranges for each toner were calculated from these photos using ImageJ (35).

Field Emission Scanning Electron Microscopy

Using a tungsten needle, loose toner particles from each reference sample were transferred to a conductive carbon adhesive tab mounted on an aluminum SEM stub. Samples were examined without any conductive coating in a JEOL 7100FT field emission scanning electron microscope (FESEM), typically at an accelerating voltage of 1.0 kV and a probe current of 3. Secondary electron images were collected at relatively low magnification (~1000×) to document the intra-sample variation in particle

TABLE 1—*Summary of the measured characteristics of reference toner samples.*

Sample			PLM observations		Size (μm)			FESEM observations		Raman identifications	
#	Make	Color	Pigment	Shape	Min	Max	Avg.	Morphology	Texture	Polymer	Pigment
T001	Dell	Cyan	Dense, even	Rounded	3.9	8.7	6.1	Rounded	Rough	NONE	PB15
T002	Dell	Magenta	Dense, even	Rounded	4.0	12.5	7.6	Rounded	Rough	Polystyrene	Naphthol AS
T003	Dell	Yellow	Dense, even	Rounded	4.2	9.3	6.0	Rounded	Rough	Polystyrene	PY74
T004	Dell	Black	Dense, even	Rounded	3.5	10.3	6.4	Rounded	Rough	Polystyrene	Carbon black
T005	HP	Cyan	Dense, even	Spherical	4.2	9.7	6.2	Spherical	Moderate	NONE	PB15
T006	HP	Magenta	Sparse, even	Spherical	4.7	9.9	6.4	Spherical	Moderate	Polystyrene	Naphthol AS
T007	HP	Yellow	Sparse, even	Spherical	4.3	13.2	7.8	Spherical	Moderate	Polystyrene	Diarylide
T008	HP	Black	Sparse, even	Spherical	3.7	10.0	5.8	Spherical	Moderate	Polystyrene	Carbon black
T009	HP	Cyan	Dense, even	Spherical	5.5	10.0	7.1	Spherical	Smooth	NONE	PB15
T010	HP	Magenta	Dense, even	Spherical	4.6	9.6	7.3	Spherical	Smooth	Polystyrene	Naphthol AS
T011	HP	Yellow	Sparse, even	Spherical	4.1	12.6	8.4	Spherical	Smooth	Polystyrene	Diarlyide
T012	HP	Black	Dense, even	Spherical	4.1	7.0	5.4	Spherical	Moderate	Polystyrene	Carbon black
T013	Brother	Black	Sparse, even	Spherical	6.7	17.8	11.1	Spherical	Moderate	NONE	Carbon black
T014	Brother	Cyan	Sparse, clumped	Rounded	4.7	15.7	10.0	Spherical	Moderate	NONE	PB15
T015	Brother	Magenta	Sparse, clumped	Rounded	6.8	16.3	10.9	Irregular	Rough	Polystyrene	Quinacridone (PR122)
T016	Brother	Yellow	Sparse, even	Spherical	7.4	16.8	13.1	Irregular	Rough	Polystyrene	Disazo Condens. Grp. 1
T017	Brother	Black	Sparse, even	Rounded	5.1	12.0	8.3	Irregular	Rough	Polystyrene	Carbon black
T018	Brother	Black	Sparse, even	Spherical	8.4	37.1	23.8	Spherical	Moderate	NONE	Carbon black
T019	HP	Cyan	Dense, even	Spherical	3.4	7.6	5.0	Spherical	Smooth	NONE	PB15
T020	HP	Magenta	Dense, even	Spherical	3.6	7.7	5.9	Spherical	Smooth	Polystyrene	Naphthol AS
T021	HP	Yellow	Dense, even	Spherical	3.5	7.2	5.3	Spherical	Moderate	Polystyrene	Diarylide
T022	HP	Black	Dense, even	Spherical	3.9	7.2	5.7	Spherical	Moderate	NONE	Carbon black
T023	HP	Black	Dense, even	Irregular	5.8	22.5	11.6	Irregular	Smooth	Polystyrene	Carbon black
T024	Xerox	Cyan	Dense, even	Rounded/irregular	4.5	12.9	7.9	Rounded	Rough	NONE	PB15
T025	Xerox	Magenta	Dense, even	Rounded/irregular	4.4	12.7	7.5	Rounded	Rough	FLUORESCENCE	FLUORESCENCE
T026	Xerox	Yellow	Dense, even	Rounded	4.4	11.0	7.2	Rounded	Rough	NONE	PY74
T027	Xerox	Black	Dense, even	Rounded	3.2	10.5	5.6	Rounded	Rough	NONE	Carbon black
T028	RICOH	Cyan	Dense, even	Rounded	4.3	11.2	6.9	Rounded	Moderate	NONE	PB15
T029	RICOH	Magenta	Dense, even	Rounded	4.1	11.7	7.1	Spherical	Moderate	NONE	Naphthol AS
T030	RICOH	Yellow	Dense, even	Rounded	5.3	10.7	7.8	Spherical	Moderate	Polystyrene	Diarylide
T031	RICOH	Black	Dense, even	Rounded	3.2	11.4	6.1	Spherical	Moderate	Polystyrene	Carbon black
T032	Konica Minolta	Cyan	Dense, even	Rounded/irregular	5.5	14.5	9.6	Rounded	Moderate	NONE	PB15
T033	Konica Minolta	Magenta	Dense, even	Rounded/irregular	5.4	17.8	11.2	Rounded	Rough	Polystyrene	BONA (PR48.2/48.3)
T034	Konica Minolta	Yellow	Dense, even	Rounded/irregular	6.4	14.0	9.2	Rounded	Moderate	Polystyrene	PY74
T035	Konica Minolta	Black	Dense, even	Irregular	5.9	15.2	9.8	Rounded	Moderate	Polystyrene	Carbon black
T036	Toshiba	Cyan	Dense, even	Rounded/irregular	9.0	22.9	14.0	Irregular	Smooth	NONE	PB15
T037	Toshiba	Magenta	Dense, even	Rounded/irregular	8.0	18.1	12.6	Rounded	Moderate	NONE	Naphthol AS
T038	Toshiba	Yellow	Dense, even	Irregular	6.0	24.6	13.4	Irregular	Moderate	PET	Benzimidazole (PY180)
T039	Toshiba	Black	Dense, even	Rounded/irregular	6.0	13.1	8.8	Irregular	Rough	NONE	Carbon black
T040	HP	Cyan	Dense, clumped	Rounded/irregular	4.5	8.8	6.1	Spherical	Moderate	NONE	PB15
T041	HP	Magenta	Dense, clumped	Rounded	3.8	11.0	7.1	Spherical	Moderate	Polystyrene	Naphthol AS
T042	HP	Yellow	Dense, even	Irregular	3.9	16.7	7.7	Spherical	Moderate	Polystyrene	Diarylide
T043	HP	Black	Dense, even	Irregular	5.9	38.9	14.1	Rounded	Moderate	Polystyrene	Carbon black
T044	HP	Cyan	Dense, even	Rounded	3.3	9.3	6.0	Spherical	Moderate	NONE	PB15
T045	HP	Magenta	Sparse, even	Rounded	4.2	9.3	6.9	Rounded	Smooth	Polystyrene	Naphthol AS
T046	HP	Yellow	Sparse, even	Rounded	6.6	20.1	12.7	Rounded	Smooth	Polystyrene	PY74
T047	HP	Black	Sparse, even	Rounded	6.9	18.0	11.6	Rounded	Smooth	Polystyrene	Carbon black
T048	Xerox	Black	Dense, even	Rounded	8.3	24.9	14.1	Spherical	Moderate	NONE	Carbon black
T049	HP	Black	Dense, even	Irregular	6.4	21.6	14.1	Rounded	Smooth	NONE	Carbon black
T050	Canon	Cyan	Dense, even	Rounded/irregular	4.3	11.7	7.1	Rounded	Rough	NONE	PB15
T051	Canon	Magenta	Dense, even	Rounded/irregular	4.6	10.0	6.7	Rounded	Rough	Polystyrene	BONA (PR57.1)
T052	Canon	Yellow	Dense, even	Rounded/irregular	3.0	11.0	6.7	Rounded	Rough	Polystyrene	Benzimidazole (PY180)
T053	Canon	Black	Dense, even	Rounded/irregular	4.0	10.5	6.6	Rounded	Rough	Polystyrene	Carbon black

size and morphology, and at higher magnification (~8000×), to capture the surface detail of individual toner particles, though samples were studied over a greater range of magnification.

The toner particle cross section image shown in Fig. 1 was prepared using a JEOL IB-9010CB cross section polisher, which utilized a 5 kV argon ion beam. Imaging of the sample without any conductive coating was carried out using the Through-The-Lens (TTL) detection system at an accelerating voltage of 1.0 kV.

Raman Microspectroscopy

For Raman microspectroscopy, loose toner powder was teased from each sample swab and transferred to an aluminum slide using a tungsten needle. Analysis was conducted using a Renishaw inVia Raman Microscope system with a 785 nm (red) laser. Confocal spectra (using a pinhole) were collected from individual toner particles, using the 100× dry objective and a laser power that was varied between 0.1 and 100 percent (measured laser energy at sample of 0.83–142 mV). Sample bleaching was not necessary. Spectral processing (baseline correction and smoothing) was completed in Renishaw's WIRE 3.3 software. Data interpretation was conducted using a library of known pigment and polymer spectra, both collected by Microtrace, and the pigment identification scheme described by Palenik, *et al*. (5).

Environmental Sample Collection

Dust samples were collected from various environments, which were then searched for the presence of toner particles. Locations were selected both where one would anticipate traces of toner to be present (*e.g.*, within 1 ft of a printer, on the hands of persons who recently changed toner cartridges, areas where printed papers are handled) and where the secondary transfer of toner particles might be deposited (*i.e.*, rooms near but without a printer). Both clear tape lifts (3M Magic Tape) and adhesive carbon tabs on SEM stubs were evaluated as potential sample collection substrates, as these are both commonly used by forensic investigators for the collection of trace evidence (*e.g.*, fibers, hair, gunshot residue). For each sample, a total of ten combined lifts were taken (Table 2).

Environmental Sample Analysis

Using an approach similar to the one utilized for analysis of the reference samples, tape lifts were initially screened for the presence of toner using PLM at a range of lower magnifications (~200×). Lifts were placed face up on a slide in acetone (to dissolve the cellulose acetate backing of the tape) and mounted in Cargille refractive index liquid (n = 1.47). [It should be noted that while some polymers exhibit a degree of solubility in acetone, the toner particles observed in these preparations appeared to remain intact. Though known polystyrene toner particles mounted in acetone did exhibit evidence of softening (*i.e.*, some smeared when the coverslip was removed), it seems that individual toner particles already trapped in the tape lift adhesive were not noticeably affected.] Potential toner particles, identified in the tape lifts on the basis of size, color, and morphology, were photographed and then analyzed *in situ* by Raman spectroscopy. Similarly, the samples collected on SEM stubs were screened by reflected light microscopy and particles of interest were analyzed by confocal Raman microspectroscopy. The stubs were also manually screened for particles with recognizable features.

Results

Known Toner Samples

PLM Examination

Each of the 53 toner samples was examined by PLM (Table 1). Individual toner particles typically range from 3 to 25 μm, with the range of minimum to maximum size for a given toner sample typically within 15 μm (87% are within 15 μm). Figure 3 shows photomicrographs of toner particles that illustrate the range of particles sizes observed, with each sample population representing 50 measurements. It should be noted that particle sizes can be altered in environmental samples (for example, by crushing).

Besides particle size, the most defining morphological features observed by light microscopy are the particle shape and pigment density. Most particles exhibited a spherical morphology; however, some particles showed an oblong morphology or irregular morphology. Examples of various toner particle shapes are presented in Fig. 4. At higher magnifications (100× oil immersion objective), the pigment distribution within an individual particle could be observed (Fig. 5), and the samples showed some differences, which have been described as dense or even, and fine or clumped. Examples of each pigment density classification are shown in Fig. 6.

It is important to note that these assignments are based upon a visual categorization rather than quantitative metrics. While a more objective approach could certainly be developed, a visual classification is anticipated to be both sufficiently accurate and of more practical benefit in classifying toner particles in casework samples. To illustrate the potential discriminating power of these classifications, the results are graphically presented in Fig. 7. While the reference pigments studied in this work do not necessarily represent the full range of toner particles available to the market, they do show that these morphological variables alone can be used to provide some degree of discrimination, and they provide a framework by which additional toner samples could be classified.

FESEM Examination

The appearance of the toner particles was observed in the FESEM, which provides the benefit of visualizing not only a higher magnification view of particle shape, but also the surface characteristics of the toner particle. Particle shape has been assigned to one of three categories, which are shown in Fig. 8: (A) spherical, (B) round, and (C) irregular. Although the data have been classified into a discrete group of shapes, the actual particle shapes observed (both intra- and inter-sample) can span a continuum. In the classification conducted in this research, the categorization was manually assigned; however, should this prove to be a commonly referenced characteristic, it could be possible to objectively assign these values based upon a quantitative calculation of particle circularity.

The surface texture of individual particles was generally consistent within a given toner sample and could therefore be a useful classification feature. The surface textures could be broadly grouped into one of three categories based on a qualitative evaluation of surface roughness. As shown in Fig. 9, the toner particle surfaces were classified as (A) rough (70–200 nm surface particles); (B), moderate (15–70 nm surface particles); or (C) smooth (<15 nm surface particles).

Shape and texture characteristics for each of the samples have been summarized in Table 1, while a plot of the distribution of various features has been summarized in Fig. 10. Spherical and rounded toner particles represent nearly equal populations in the sample set, with irregular particles being less common. The moderate texture was seen most often in conjunction with the spherical particle morphology, and a rough surface texture was typically associated with the rounded particles. While different colored toners from the

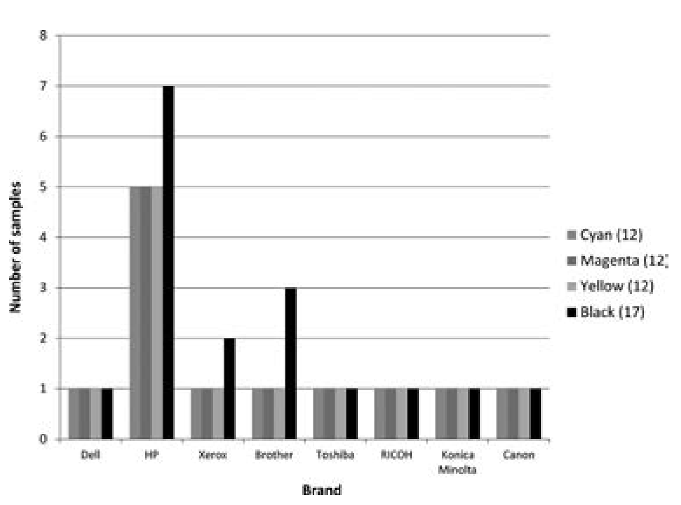


FIG. 2—*Known toner samples collected by brand. [Color figure can be viewed at wileyonlinelibrary.com]*

TABLE 2—*Summary of environmental sample results.*

#	Lift type	Location of lift	Closest toner sources	Approx. dist. to source	Methods	Toner observed?	Approx. quantity
M001	Tape	Copy room, printer tray	T001–T004	0 ft	PLM	Yes	>200
M002	Tape	Copy room, counter	T001–T004	1–2 ft	PLM, Raman	Yes	>200
M003	Tape	Copy room, near mailboxes	T001–T004	~10 ft	PLM	Yes	~50
M004	Tape	Office K, desk	T005–T008	80 ft	PLM	Yes	~30
M005	Tape	Office C, light switch	T001–T004 T018	20 ft 10 ft	PLM	Yes	~10
M006	Tape	Break room	T005–T008	>100 ft	PLM	No	0
S001	Carbon tab	Right hand, after toners replaced	T001–T004	–	FESEM, Raman	Yes	>200
S002	Carbon tab	Copy room, printer tray	T001–T004	0 ft	FESEM	Yes	~70
S003	Carbon tab	Copy room, counter	T001–T004	1–2 ft	FESEM	Yes	~50
S004	Carbon tab	Office S, printer tray	T001–T004	0 ft	FESEM	Yes	~50
S005	Carbon tab	Lab M, counter	T005–T008	1–2 ft	FESEM	Yes	~50

same source (*i.e.*, the same printer) often exhibited similar morphologies and textures, this was not always the case. An example is illustrated in a comparison of the four colors of Toshiba toner (T036 – T039; Fig. 11).

Raman Microspectroscopy

Raman analysis was effective for the identification (or classification) of pigments in toner particles. In all but one sample, the colorant could be constrained to a specific pigment or pigment class on the basis of the collected Raman spectrum. The exception was a magenta Xerox toner (T025), which exhibited fluorescence that could not be avoided under the available instrumental conditions.

The results, detailed by sample, are shown in Table 1 and summarized by color in Fig. 12. As expected, carbon black was the black pigment identified in all of the black toners and Pigment Blue 15 (PB15) was the blue pigment identified in all of the cyan toners. Variety was observed among the pigments detected in the known magenta toners. Of the eleven magenta toners in which pigment was identified, eight were classified as Naphthol AS group pigments, while three others (a Brother, a Konica Minolta, and a Canon) were each determined to be from different classes of pigment. These other pigments were identified as two Beta OxyNaphthoic Acid (BONA) lake pigments (C.I. Pigment Red 48:2/48:3 and C.I. Pigment Red 57:1) and a quinacridone pigment (C.I. Pigment Red 122).

The most variety was observed in the yellow toner pigments. Of the twelve yellow toners evaluated, four were identified as Pigment Yellow 74 (a Monoazo class pigment) (C.I. 11741), five were classified as Diarylides, two classed as Benzimidazoles, and one was classified as a Disazo Condensation Group 1 pigment. These identifications were made and assigned using the pigment classification scheme developed in Palenik, *et al.* (5). The HP toners were of particular interest: of the five yellow HP

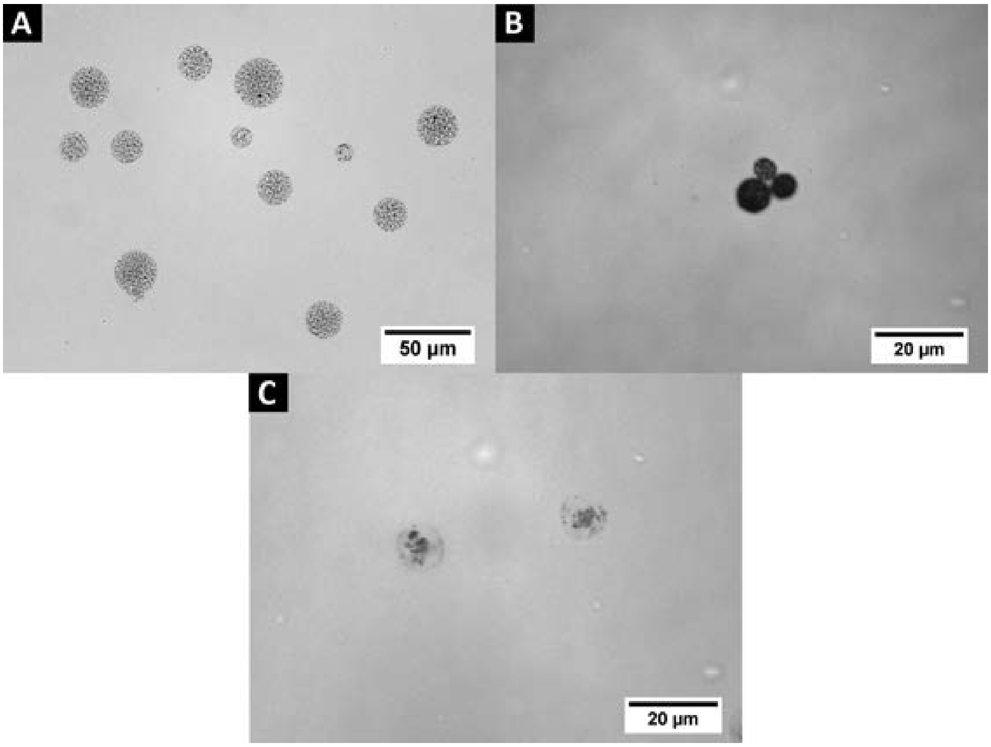


FIG. 3—*Light micrographs which illustrate the range of particle sizes observed in the known toner set: (A) average size of ~ 24 μm (T018); (B) average size of ~ 11 μm (T015); and (C) average size of ~ 6 μm (T022). [Color figure can be viewed at wileyonlinelibrary.com]*

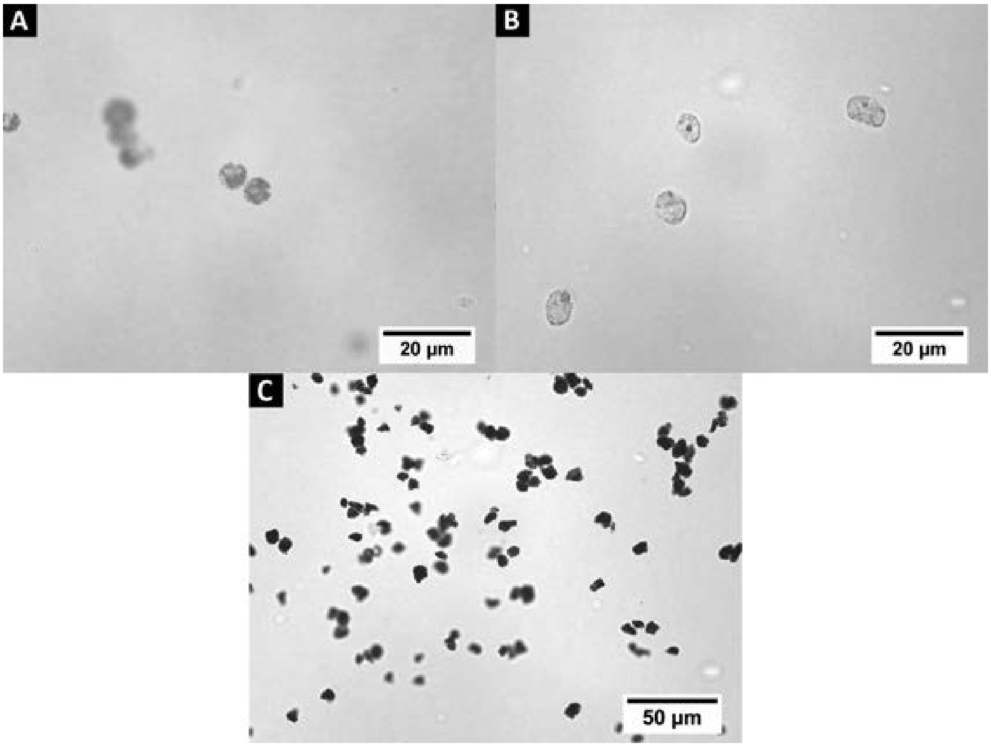


FIG. 4—*Light micrographs which illustrate the range of particle morphologies observed in the known toner set: (A) spherical (T019); (B) rounded (T030); and (C) irregular (T039). [Color figure can be viewed at wileyonlinelibrary.com]*

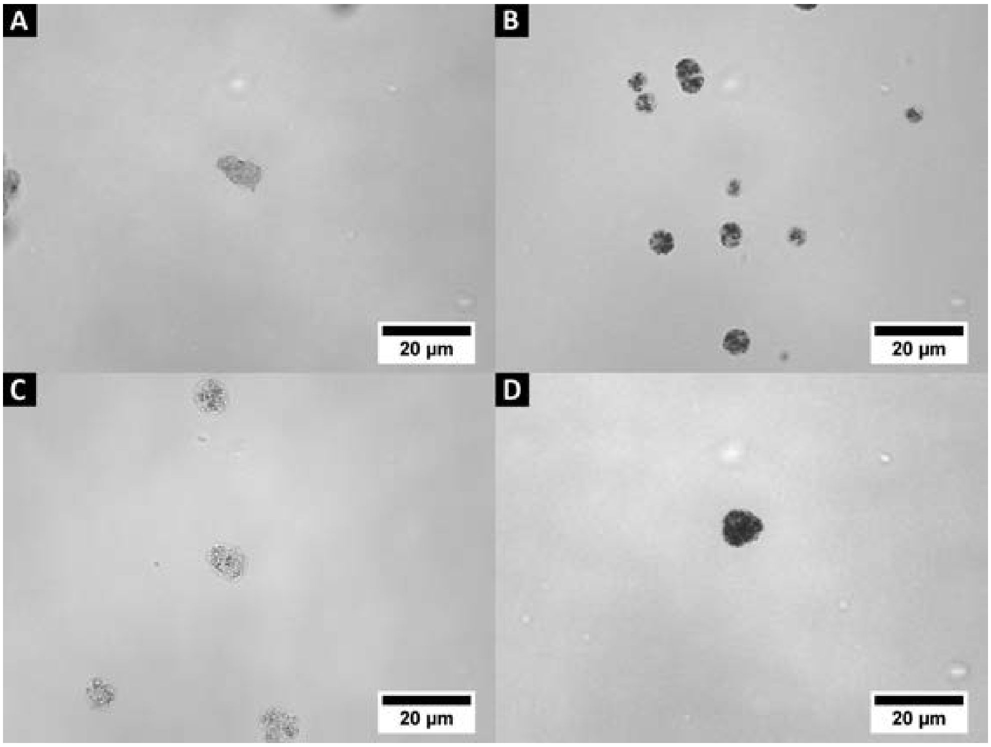


FIG. 5—*Light micrographs showing individual pigment particles present in known toner particles viewed using oil immersion microscopy at high magnification: (A) cyan (T032); (B) magenta (T020); (C) yellow (T046); (D) black (T027). [Color figure can be viewed at wileyonlinelibrary.com]*

toners, all but one contained a Diarylide class pigment (the exception used PY74, a Monoazo General class pigment (Palenik, *et al*., 2011)).

In addition to the pigment signal in the Raman spectra, additional peaks often provided an identification of the polymer binder (Table 1). While the literature suggests that many polymers

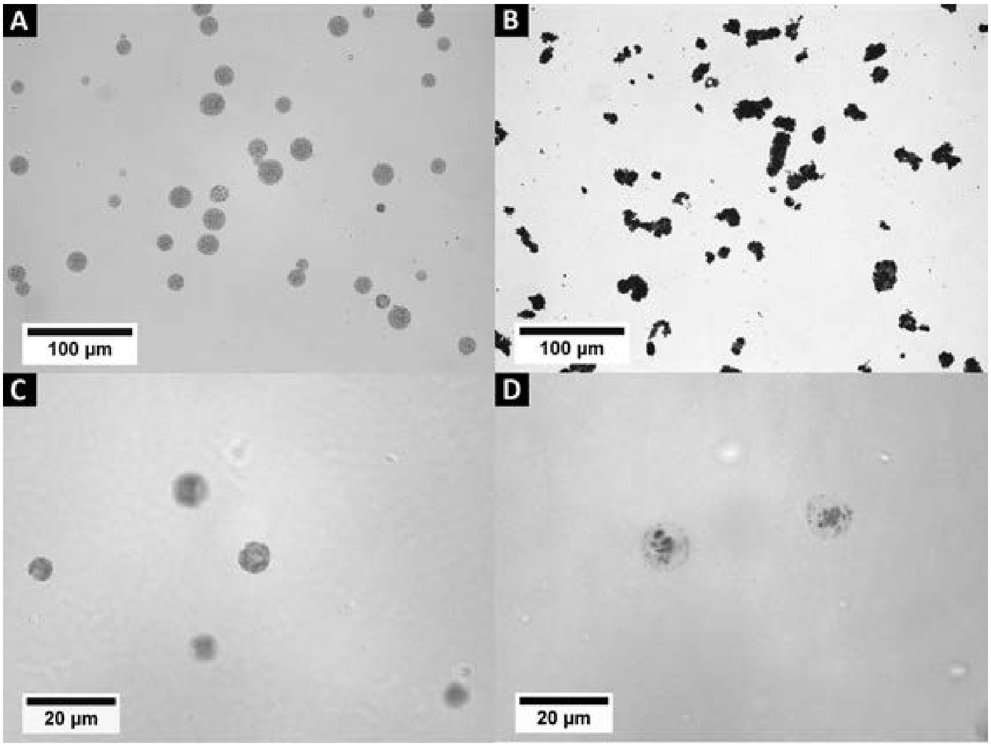


FIG. 6—*Light micrographs showing extreme differences in pigment density, from sparse (A, T018) to dense (B, T043), and in pigment distribution, from even (C, T044) to clumped (D, T015). [Color figure can be viewed at wileyonlinelibrary.com]*

(*e.g.,* polystyrene, styrene copolymers, polyethylene terephthalate, polyvinyl chloride, acrylic resin) are used as toner resins, it is unclear how commonly these polymers are used in regularly occurring commercial products (36). One example, shown in Fig. 13 depicts a spectrum collected from a known magenta toner compared to known reference spectra of polystyrene and Pigment Red 57.1 (PR57.1). Though the stronger scatter of the pigments could overpower the signal from the polymer component, an indication of the polymer resin was present in the spectra of more than half the known samples (30 of the 53). This identification was based largely on the presence of the strongest polystyrene peak at 1000 cm^{-1}. It is notable that no peaks attributable to a polymer were detected in the cyan toner reference samples. This is almost certainly due to the relatively strong resonance signal from the PB15 pigment, which appears to have obscured any signal that may have been produced by the polymer component.

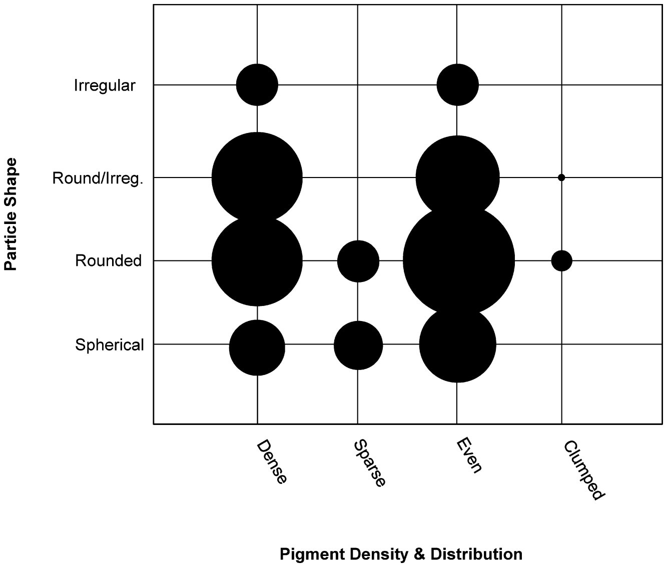


FIG. 7—*A plot of morphological features (particle shape versus pigment density and distribution) observed in the reference toner samples that were studied. The size of each circle represents the relative number of toner samples represented in a given category.*

In cases where a resin was identified, it was almost always identified as polystyrene. In only one instance was another resin polymer detected. Polyethylene terephthalate (PET) was identified as the polymer resin in a yellow Toshiba toner (T038). From these results, it appears that the resin composition will rarely be a discriminating factor; however, identification of the resin does serve as additional corroboration for the identification of a suspected toner particle. This feature becomes more significant when other identifying characteristics, such as color in a black particle or morphology in a crushed or distorted particle, are less obvious.

Environmental Dust Samples

When screening the dust samples listed in Table 2 for the presence of toner particles, color was established as the most reliable feature for initial recognition. Because of this, scanning tape lifts for toner by transmitted PLM were faster and more efficient than manually searching for toner morphology by electron microscopy.

In a study of PLM preparations of environmental samples, black toner particles were more commonly encountered than colored toner particles. This is not surprising, given that black

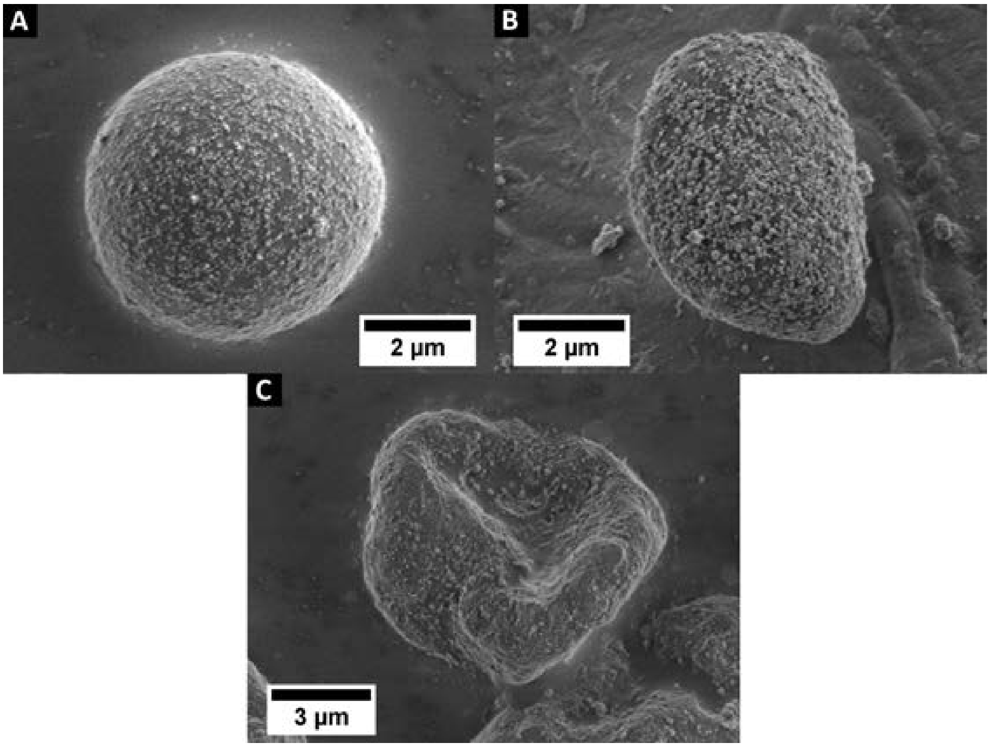


FIG. 8—*Secondary electron images which illustrate the range of particle morphologies observed in the known toner set: (A) spherical (T042); (B) rounded (T051); and (C) irregular (T015).*

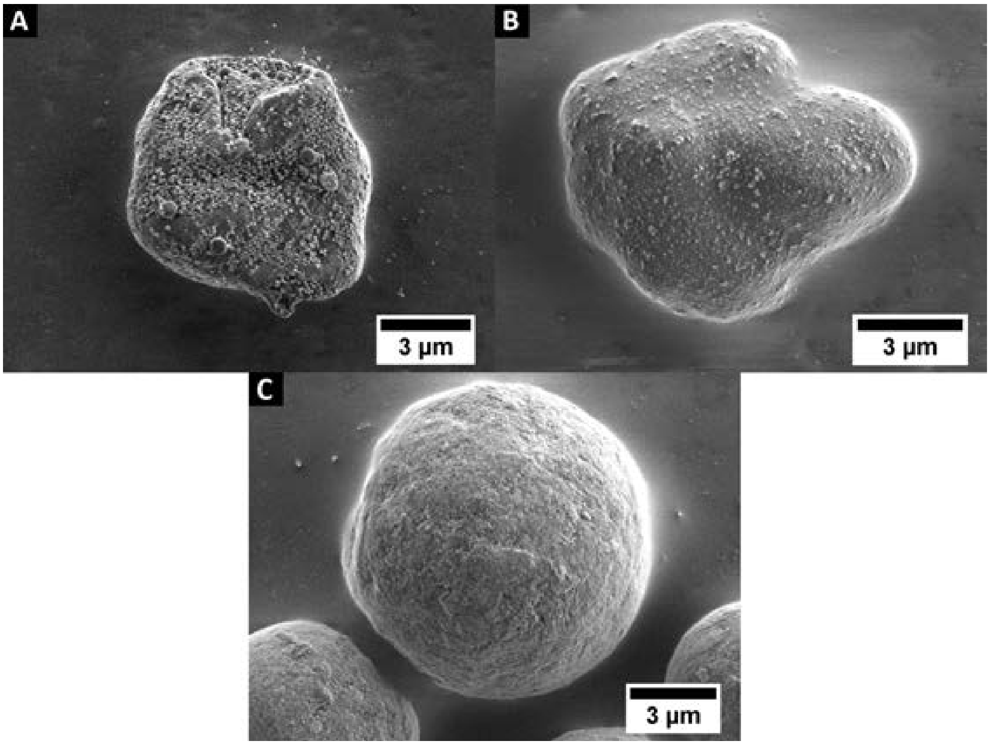


FIG. 9—*Secondary electron images showing the different surface textures observed in the known toner set: (A) rough (T024); (B) moderate (T035); and (C) smooth (T011).*

toner is more commonly utilized than colored toners in our sampled environment. However, not only are black toner particles more difficult to recognize (as other colors are more notable against the general background of an environmental dust sample), but carbon black pigment is more commonly encountered in other particles such as char, rubber, soot, and

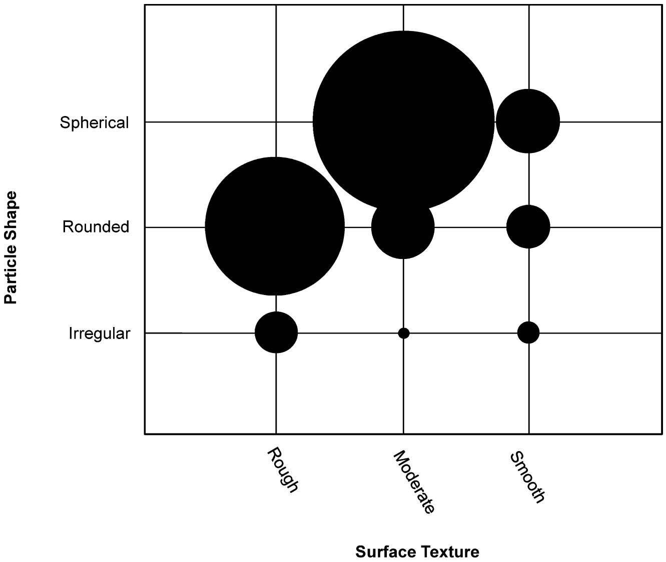


FIG. 10—*A plot of particle shape and surface texture observed in the reference toner samples by FESEM. The size of each circle represents the relative number of toner samples represented in a given category.*

paint (to name a few). As a result, the identification of carbon black pigment alone is not a sufficient basis upon which to identify a particle as toner.

Fortunately, many of the possible toner particles encountered in the environmental samples retained some aspect of their rounded morphology (Fig. 14*A*). In other cases, a distorted (*i.e.*, flattened) morphology was noted (Fig. 14*B*) (possibly deformed by fuser heat, shed from printed paper, or imparted by sample collection). It was also demonstrated that the same chemical information that could be obtained from known toner particles could be collected from suspected toner particles observed in these environmental samples. Figure 15 shows a Raman spectrum obtained from a magenta particle on a tape lift, in which the pigment was successfully identified as belonging to the Naphthol AS group.

Toner was present and readily observed in environmental samples collected in the near vicinity of printers (same room), from tape lifts (Fig. 16) and carbon adhesive stubs (Fig. 17). In samples collected near the printer, particles of toner were often found in a cluster (rather than as lone particles) and sometimes more than one color of particle was present in the group. Moving further away from the printing source, the relative amounts of toner were seen to decrease; however, individual particles of toner were still observed in samples taken several feet away from a printer (same room), on the hands of someone who handled toner cartridges, in a room without a printer (where paper from the printer is regularly handled), and on a light switch in a room with a printer. Toner was not observed in lifts collected from environments (without a printer) where handling of printed paper is infrequent, such as a break room. The observations of toner quantity in an environmental sample are based on visual estimates (a commonly accepted practice in the environmental examination of samples by light microscopy [37]). While more quantitative approaches could certainly be applied, the results presented in Table 2 illustrate the general trends in the relative toner particle concentration of the various samples studied. In these environmental samples, the particles are generally present at trace levels, with respect to the other background dust particles in the preparations.

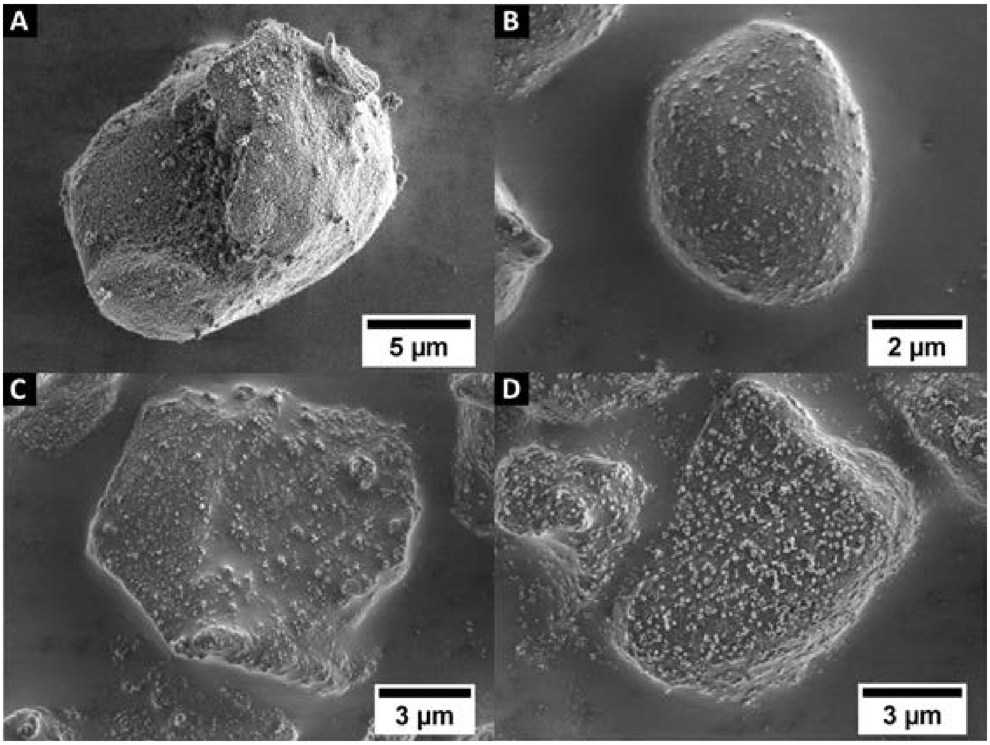


FIG. 11—*Secondary electron images showing differences in shape and texture among the four different colored toners from the same Toshiba printer: (A) cyan (T036); (B) magenta (T037); (C) yellow (T038); and (D) black (T039).*

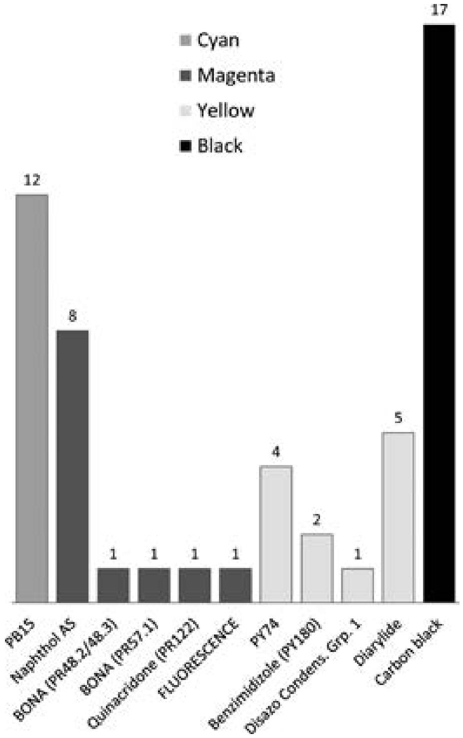


FIG. 12—*Summary of pigments and pigment classes identified in known toners by Raman. [Color figure can be viewed at wileyonlinelibrary.com]*

In addition to the distance from the printing source, many other factors could influence the amount and prevalence of toner in a given environment. For example, the type and frequency of printing, the time since the surface/hands were cleaned, and room traffic represent a few of the primary variables anticipated to impact transfer and persistence. This research does not attempt to study the effects of these variables, as the complexity of such processes suggests that a study would be of little general applicability to any given case. Instead, the research presented here is intended to demonstrate that toner particles can be found and identified in environmental samples where they are anticipated to be present. The environmental samples studied here also indicate that toner particles are not ubiquitous in all areas of an office environment.

Discussion

The techniques of PLM and FESEM provide complementary information about toner particle morphology. While transmitted light permits the observation of color, which is arguably the most recognizable feature of toner (and thus most useful for detection in unknown dust samples), it also provides information about pigment distribution. Engineered features of toner that exceed the resolution of light microscopy can be studied with the higher resolution afforded by a low voltage electron microscope, which permits the study of features and textures well under 50 nm in size.

As with the data obtained from optical microscopy, observation of microscopical features in the scanning electron microscope provides additional support for the positive identification of an unknown particle as toner. For example, a review of particles in the Particle Atlas (34) shows that pigment particles with the characteristics observed by light and electron microscopy can be confidently distinguished from all other listed particle types. The inter-sample differences noted by these methods could also be used to provide some comparative and sourcing information.

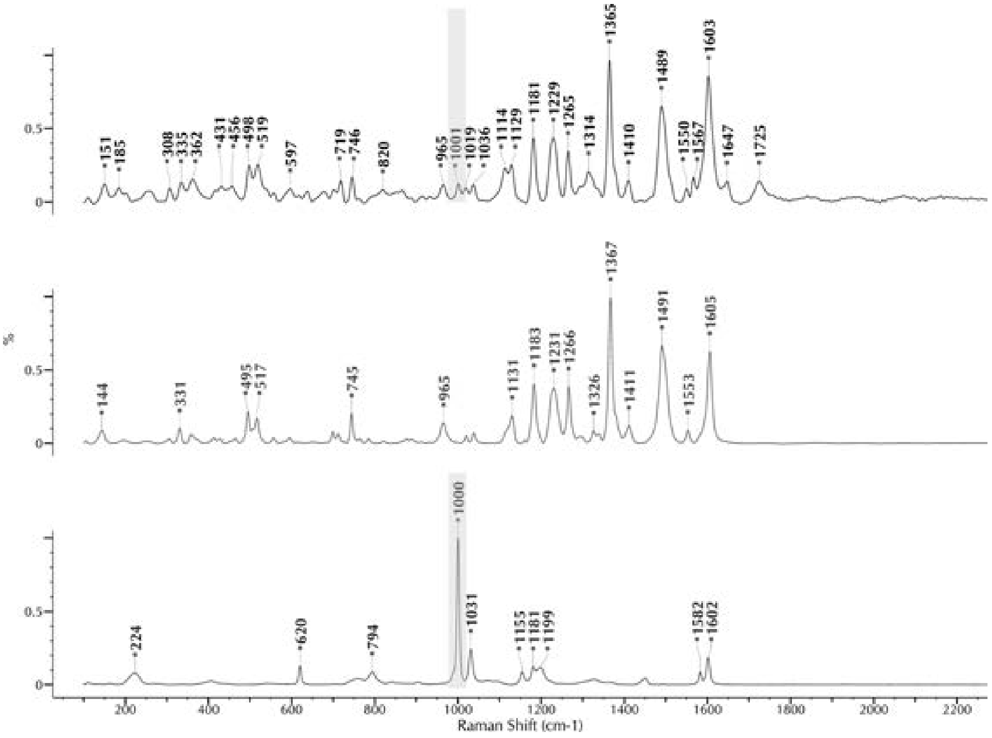


FIG. 13—*Raman spectrum of a known toner particle (T051, top), showing the pigment (PR57.1, middle) and polymer (polystyrene, bottom) identified. [Color figure can be viewed at wileyonlinelibrary.com]*

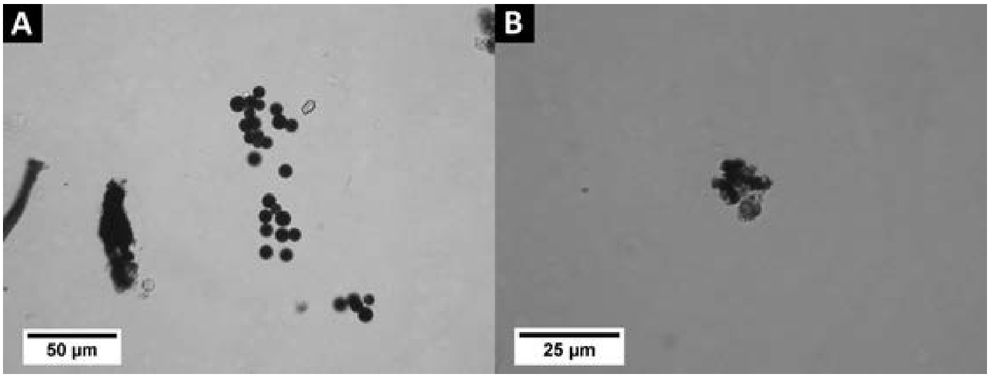


FIG. 14—*PLM images of environmental tape lift samples collected near a laser printer, showing the (A) rounded (M002) and (B) deformed particles encountered (M003). [Color figure can be viewed at wileyonlinelibrary.com]*

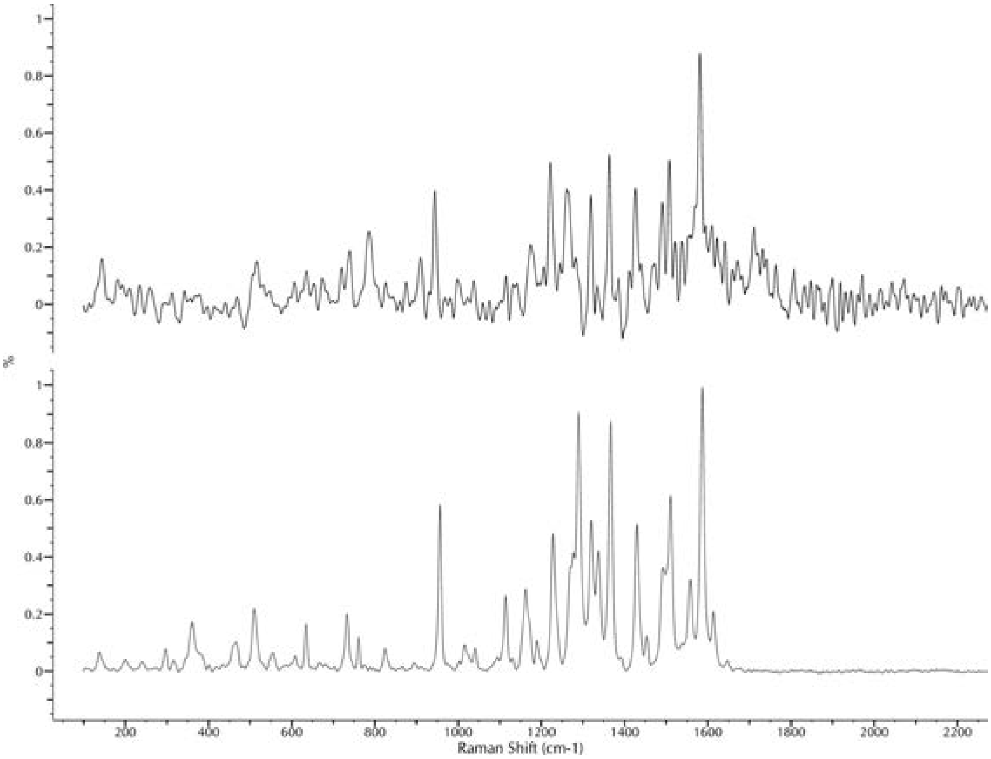


FIG. 15—*In situ Raman spectrum of a toner particle from an environmental tape lift (top), showing the Naphthol AS pigment identified (bottom). [Color figure can be viewed at wileyonlinelibrary.com]*

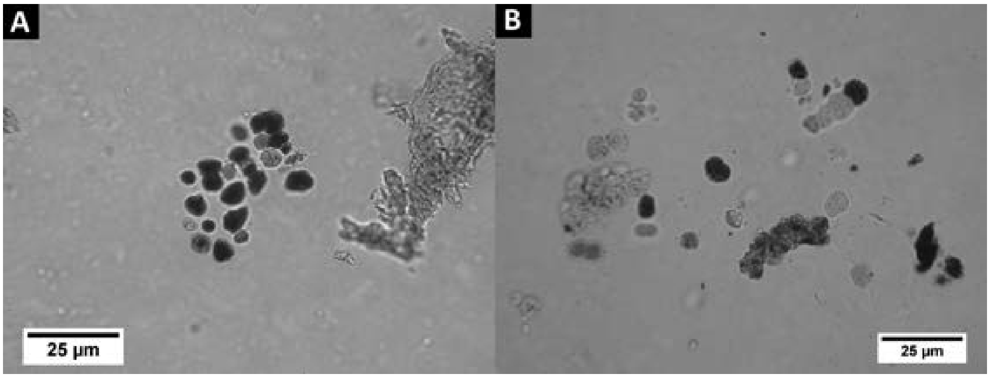


FIG. 16—*PLM images of environmental tape lift samples collected near a laser printer (M002). [Color figure can be viewed at wileyonlinelibrary.com]*

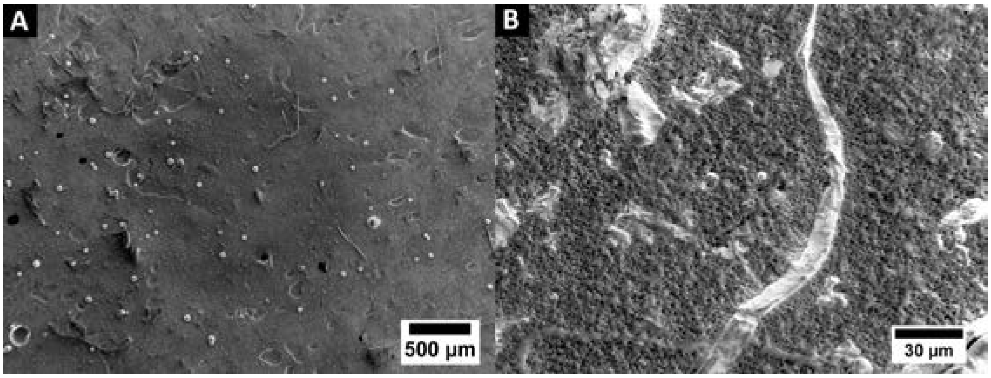


FIG. 17—*Secondary electron images showing toner present in environmental lifts (carbon adhesive tab) collected from: (A) the hands of someone who changed a toner cartridge (S001); and (B) the tray of a printer (S003).*

While some toners of the same color could be discriminated from one another using Raman spectroscopy to identify their pigments (specifically yellow and magenta toners), there was no strong correlation between the manufacturer or printer make/model. However, if different, the pigment identification could exclude a particular known toner as a source. Although the literature cites various polymeric resins are used in the production of toner (4,9) when the polymer was detected in the known set, most toners could not be differentiated based on the polymer peaks observed in their spectra. Though there was one exception (a toner containing PET instead of polystyrene), routinely relying on this property for discrimination would require that different resins are commonly used, which seems unlikely given this population of data.

In the end, morphology is a large part of recognition for toner. Looking for the distinctive color, size, and shape of the particles made them easy to spot by light microscopy. But the characteristics that were most useful for discrimination were determined by chemical analysis and electron microscopy. Working in confocal mode with the Raman microscope and examining the particle morphology with the high resolution of the FESEM permitted the analysis of individual toner particles and the nanoscale features present on their surfaces. Unlike methods such as infrared spectroscopy and TLC, which are suited for bulk analysis of printed toner, these nondestructive approaches are appropriate for examination of suspected toner traces, but may also be useful for other potential subvisible and nanoparticle evidence present in unknown dusts, depending on their scattering properties and morphologies. Analyzing particles directly on tape lifts and SEM stubs is another advantage to these techniques, as they require little to no sample preparation and also do not require familiarity with particle manipulation.

Environmental dust samples studied here have demonstrated that (i) traces of toner are present in environments that contain a printer and can also be transferred to new environments where direct sources of toner are absent; (ii) toner particles can be recovered from these environments using collection methods already in use; and (iii) existing analytical techniques can be employed to characterize and identify individual toner particles. Depending on the features observed, comparative analysis may allow potential known toners to be excluded or included as sources for the recovered toner evidence. Even without known sources for comparison, detection of toner could provide a valuable investigative lead, depending on the characteristics of those particles.

Conclusions

First, this research illustrates various morphological and chemical features upon which single toner particles can be positively identified. These features include color, size, shape, pigment distribution, surface texture, internal structure, pigment composition, and resin composition. Based upon a comparison of these features to the range of particle types characterized in the Particle Atlas (38), these properties (or even a subset of these properties) permit toner particles to be positively identified as such.

Secondly, the research shows that the same factors used for toner particle identification can provide some level of discrimination between toner particles of different make. While it is not possible to individualize all toners to a single manufacturer, there are certainly discriminating characteristics that permit some toners to be discriminated and classified. While this study provides some indications of significance, a larger sample set would need to be characterized to determine which variables are most distinctive and whether a certain combination of features is sufficient to ascribe a toner particle to a particular manufacturer.

Finally, this research demonstrates that toner particles can be recognized and identified in environmental dust samples. In particular, toner particles can be recognized in samples collected on substrates (tape, GSR-type stubs) already used for the collection of forensic evidence. A study of samples from within an office environment suggests that toner particles generally decrease in concentration as a function of distance from a printer and that toner particles are not ubiquitous in office dust.

In sum, these results suggest that toner particles hold potential as a form of sub-visible trace evidence with nanoscale features that can provide more than what is typically utilized in a trace evidence examination. The applications of this form of evidence range from investigative to comparative.

Acknowledgments

The authors would like to acknowledge the following individuals and businesses who contributed known toner samples for analysis: Mary McKeon; Mike Martin; Maria R. Hernandez, MBA; Corie White; and CPR Printing, Inc. (Geneva, IL).

References

1. Palenik CS, Brinsko-Beckert K, Insana J, Palenik SJ. Analytical and transfer characteristics of a fluorescent detection spray: implications for

subvisible and nanotrace particle transfers. Forensic Sci Int 2018;286:96–105. https://doi.org/10.1016/j.forsciint.2018.03.007.

2. Galliford GJ.Chemically prepared toner (CPT). http://gallifordconsulting.com/wp-content/uploads/2016/10/Chemically-Prpared-Toner-Basics.pdf (accessed May 24, 2017).
3. Galliford GJ.Particle shape of toners and the advantage of using chemical toner manufacturing methods. http://gallifordconsulting.com/wp-content/uploads/2016/10/Particle-Shape-of-Toners.pdf (accessed May 24, 2017).
4. Galliford GJ.The anatomy of a toner. http://gallifordconsulting.com/wp-content/uploads/2016/10/The-Anatomy-of-a-Toner.pdf (accessed May 24, 2017).
5. Palenik CS, Palenik SJ, Herb J, Groves E. Fundamentals of forensic pigment identification by Raman microspectroscopy: a practical identification guide and spectral library for forensic science laboratories. Rockville, MD: National Institute of Justice, 2011; NIJ Award No. 2010-DN-BX-K236.
6. Merrill RA, Bartick EG, Taylor JH III. Forensic discrimination of photocopy and printer toners: I. The development of an infrared spectral library. Anal Bioanal Chem 2003;376(8):1272–8. https://doi.org/10.1007/s00216-003-2073-0.
7. Egan WJ, Morgan SL, Bartick EG, Merrill RA, Taylor JH III. Forensic discrimination of photocopy and printer toners: II. Discriminant analysis applied to infrared reflection-absorption spectroscopy. Anal Bioanal Chem 2003;376(8):1279–85. https://doi.org/10.1007/s00216-003-2074-z
8. Egan WJ, Galipo RC, Kochanowski BK, Morgan SL, Bartick EG, Miller ML, et al. Forensic discrimination of photocopy and printer toners: III. Multivariate statistics applied to scanning electron microscopy and pyrolysis gas chromatography/mass spectrometry. Anal Bioanal Chem 2003;376(8):1286–97. https://doi.org/10.1007/s00216-003-2099-3.
9. Zappa G, Carconi P, Gatti R, D'Alessio A, Di Bonito R, Mosiello L, et al. Feasibility study for the development of a toner reference material. Measurement 2009;42(10):1492–6. https://doi.org/10.1016/j.measurement.2009.08.006.
10. Shaffer DK. Forensic document analysis using scanning microscopy. In: Postek MT, Newbury DE, Platek SF, Joy DC, editors. Proceedings of the SPIE: Scanning Microscopy 2009. Vol. 7378; 2009 May 4–7; Monterey, CA. Bellingham, WA: SPIE, 2009. https://doi.org/10.1117/12.825186.
11. Levy EJ, Wampler TP. Applications of pyrolysis gas chromatography/mass spectrometry to toner materials from photocopiers. J Forensic Sci 1986;31(1):258–71. https://doi.org/10.1520/JFS11880J.
12. Chang W, Huang C, Giang Y. An improvement on pyrolysis gas chromatography for the differentiation of photocopy toners. J Forensic Sci 1993;38(4):843–63. https://doi.org/10.1520/JFS13482J.
13. Totty RN. Analysis and differentiation of photocopy toner. Forensic Sci Rev 1990;2(1):1–23.
14. Zimmerman J, Mooney D. Preliminary examination of machine copier toners by infrared spectrophotometry and pyrolysis gas chromatography. J Forensic Sci 1986;31(2):489–93. https://doi.org/10.1520/JFS12279J.
15. Totty RN. The examination of photocopy documents. Forensic Sci Int 1990;46(1–2):121–6. https://doi.org/10.1016/0379-0738(90)90148-R.
16. Almeida Assis AC, Barbosa MF, Valente Nabais JM, Custodio AF, Tropecelo P. Diamond cell Fourier transform infrared spectroscopy transmittance analysis of black toners on questioned documents. Forensic Sci Int 2012;214(1–3):59–66. https://doi.org/10.1016/j.forsciint.2011.07.019.
17. Mazzella WD, Lennard CJ, Margot PA. Classification and identification of photocopying toners by diffuse reflectance infrared Fourier transform spectroscopy (DRIFTS): I. Preliminary results. J Forensic Sci 1991;36(2):449–65. https://doi.org/10.1520/JFS13047J.
18. Mazzella WD, Lennard CJ, Margot PA. Classification and identification of photocopying toners by diffuse reflectance infrared Fourier transform spectroscopy (DRIFTS): II. Final report. J Forensic Sci 1991;36(3):820–37. https://doi.org/10.1520/JFS13092J.
19. Saini K, Saroa J. Differentiation of color photocopy toners using TLC, UV, and FTIR techniques. J Forensic Identif 2011;61(6):561–80.
20. Williams RL. Analysis of photocopying toners by infrared spectroscopy. Forensic Sci Int 1983;22(1):85–95. https://doi.org/10.1016/0379-0738(83)90122-6.
21. Kemp GS, Totty RM. The differentiation of toners used in photocopy processes by infrared spectroscopy. Forensic Sci Int 1983;22(1):75–83. https://doi.org/10.1016/0379-0738(83)90121-4.
22. Tandon G, Jasuja OP, Sehgal VN. Characterization of photocopy toners using Fourier transform infrared spectroscopy: a structural diagnosis of chemical constituents in black photocopier toners used in India. Int J Forensic Doc Exam 1997;3(2):119–26.
23. Thakur J, Jasuja O, Singla A. Thin-layer chromatography of photocopy toners. J Forensic Identif 2004;54(1):53–63.
24. Tandon G, Jasuja OP, Sehgal VN. Thin layer chromatography analysis of photocopy toners. Forensic Sci Int 1995;73(2):149–54. https://doi.org/10.1016/0379-0738(95)01746-6.
25. Subedi K, Trejos T, Almirall J. Forensic analysis of printing inks using tandem laser induced breakdown spectroscopy and laser ablation inductively coupled plasma mass spectrometry. Spectrochim Acta Part B: Atom Spec 2015;103–104:76–83. https://doi.org/10.1016/j.sab.2014.11.011.
26. Szynkowska MI, Czerski K, Paryjczak T, Parczewski A. Ablative analysis of black and colored toners using LA-ICP-TOF-MS for the forensic discrimination of photocopy and printer toners. Surf Interface Anal 2010;42(5):429–37. https://doi.org/10.1002/sia.3194.
27. Heudta L, Deboisa D, Zimmerman TA, Köhlerd L, Banoc F, Partouchee F, et al. Raman spectroscopy and laser desorption mass spectrometry for minimal destructive forensic analysis of black and color inkjet printed documents. Forensic Sci Int 2012;219(1–3):64–75. https://doi.org/10.1016/j.forsciint.2011.12.00.
28. Feldmann JM. Discrimination of color copier/laser printer toners by Raman spectroscopy and subsequent chemometric analysis [dissertation]. West Lafayette, IN: Purdue University, 2013.
29. Božičevića MS, Gajovićb A, Zjakić I. Identifying a common origin of toner printed counterfeit banknotes by micro-Raman spectroscopy. Forensic Sci Int 2012;223(1–3):314–20. https://doi.org/10.1016/j.forsciint.2012.10.007.
30. Udriştioiu EG, Bunaciu AA, Aboul-Enein HY, Tănase IGh. Forensic analysis of color toners by Raman spectroscopy. Instrum Sci Technol 2009;37(1):23–9. https://doi.org/10.1080/10739140802584707.
31. Morawska L, He C, Johnson G, Jayaratne R, Salthammer T, Wang H, et al. An investigation into the characteristics and formation mechanisms of particles originating from the operation of laser printers. Environ Sci Technol 2009;43(4):1015–22. https://doi.org/10.1021/es802193n.
32. Kagi N, Fujii S, Horiba Y, Namiki N, Ohtani Y, Emi H, et al. Indoor air quality for chemical and ultrafine particle contaminants from printers. Build Environ 2007;42(5):1949–54. https://doi.org/10.1016/j.buildenv.2006.04.008.
33. He C, Morawska L, Taplin L. Particle emission characteristics of office printers. Environ Sci Technol 2007;41(17):6039–45. https://doi.org/10.1021/es063049z.
34. Palenik SJ. The determination of geographical origin of dust samples. In: McCrone WC, Delly JG, Palenik SJ, editors. The particle atlas. vol. V. Ann Arbor, MI: Ann Arbor Science Publishers, 1979;1347–61.
35. ImageJ [computer program]. Windows version. Bethesda, MD: U. S. National Institutes of Health, 1997.
36. Brunelle RL, Crawford KR. Advances in the forensic analysis and dating of writing ink. Springfield, IL: Charles C Thomas Publisher Ltd, 2003.
37. ASTM D6602-03be1. Standard practice for sampling and testing of possible carbon black fugitive emissions or other environmental particulate, or both. West Conshohocken, PA: ASTM International, 2009. https://doi.org/10.1520/D6602-03be01.
38. McCrone WC, Draftz RG, Delly JG. The particle atlas: a photomicrographic reference for the microscopical identification of particulate substances. Ann Arbor, MI: Ann Arbor Science Publishers, 1967.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14491
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Samara Alves Testoni,[1] *Ph.D.; Vander Freitas Melo*,[1] *Ph.D.; Lorna Anne Dawson*,[2] *Ph.D.; Joice Malakoski*,[3] *M.Sc.; Edimar Cunico*,[3] *M.Sc.; and Jorge Andrade Junqueira Neto*,[3] *M.Sc.*

The Use of a Sequential Extraction Technique to Characterize Soil Trace Evidence Recovered from a Spade in a Murder Case in Brazil*

ABSTRACT: Soil trace evidence can be useful in criminal investigations. A homicide which had occurred in South Brazil been concluded through the courts with a guilty conviction. A spade with soil traces adhering to it was seized from the confessed killer's house, it having been established that it had been used to bury parts of the victim's body. In the context of this confession, it provided an opportunity to test a protocol of analysis and verify the potential of discriminate soil sample analysis in such case works. This allowed us to test the practice of sequential analysis which had been developed for forensic case works in Brazil, with three sequential extractions: (i) 0.2 mol/L pH 3.0 ammonium oxalate; (ii) dithionite–citrate–bicarbonate; and (iii) 0.5 mol/L NaOH. It was possible to predict the sequence of events related to the homicide by using the sequential extraction technique and to conclude that: (i) the A horizon soil from the burial location of the torso was found to be very similar to the soil samples which had been recovered from the spade, which was able to be established despite there only being a small amount of soil adhering to the spade; (ii) the location where the legs were buried contributed a low amount of soil adhering to the spade. Therefore, it is suggested that, where possible, sequential extractions should be prioritized from a questioned sample to best provide information about the likely sequence of contact places and this test likely scenarios and criminal events.

KEYWORDS: forensic geology, soil evidence, closed case, spade, transfer, trace evidence, granitic rocks, iron oxides, kaolinite

A range of physical evidence types have been used to help test and establish what had happened at crime scenes, such as DNA, hair, paint, glass, fibers, and plant and soil traces. Although it has recently been more thoroughly investigated (1–3), the use of soil to assist criminal investigations dates back around 150 years ago (4) and even before that to the Roman time when they were reported to examine the hooves of enemies horses to work out where they had travelled. The use of soils in forensic investigations can be a valuable source of information due to its variable nature and the ability to ascertain the likely origin of soil traces adherent to a wide range of objects and people, including shoes, tires, clothes, and hair (5–9). Soil variability derives from pedological processes such as the underlying geological parent material, position in the landscape, predominant local climate, human action, vegetation, living organisms in the soil, and time (10,11). Each of these factors in turn defines the intensity of weathering, which will result in the formation of specific soil types. When soil is transferred to different types of surface and persists to enable detection and recovery, the questioned sample can be examined and analyzed to ascertain likely origin.

Chemical extractions have been frequently used in forensic studies: for example, total extraction and analysis by ICP-OES/AES-MS (1,3,12); organic extractions (2,13,14). All of these chemical methods were applied to an isolated matrix of the soil. The advantage of a chemical sequential analysis is to dissolve mineral phases in a preestablished sequence, where the separation and the characterization of each mineral phase improve the quality of the soil analysis (15) and provide data on a range of complementary analyses on a limited size of sample.

Some work in Brazil has concentrated on sequential chemical extraction of short-range order materials (SOR) and crystalline minerals from the clay fraction in forensic cases (16–22). The sequence of analyses adopted mixed methodologies for extraction of short-range order phase (SOR) and crystalline Fe oxides in nonforensic studies (23–25) (with modifications): (i) 0.2 mol/L pH 3.0 ammonium oxalate (AO); (ii) dithionite–citrate–bicarbonate (DCB); and (iii) 0.5 mol/L NaOH. The main modification used first for (16) in a simulated forensic case in relation to the method proposed by (24) was the inclusion of the extraction step with DCB. Sodium dithionite, as a strong reducing agent,

[1]Department of Soil Science and Agricultural Engineering, Federal University of Paraná, Funcionários St. 1540, Curitiba, State of Paraná 80035-050, Brazil.

[2]Environmental and Biochemical Sciences Group, The James Hutton Institute, Craigiebuckler, Aberdeen, Scotland AB15 8QH, U.K.

[3]Criminalistics Institute of Paraná, Visconde de Guarapuava Avenue 2652, Curitiba, State of Paraná 80010-100, Brazil.

Corresponding author: Samara Alves Testoni, Ph.D. E-mail: testonisamara@gmail.com

*Financial support provided by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*—Capes, Brazil (reference Capes Forense 25/2014).

Received 30 Dec. 2019; and in revised form 12 May 2020; accepted 28 May 2020.

solubilizes hematite and goethite from the clay fraction of the soil (26).

The acid ammonium oxalate (AAO) method extracts the poorly crystalline forms of inorganic and organic Fe and Al from soils. It attacks most silicate minerals and goethite and hematite slightly, and it dissolves magnetite and finely divided easily weathered silicates such as olivine to a considerable extent. Through complexation and reduction of the pH, which promote the protonation to extract poorly ordered Fe and Al oxides, using the acid ammonium oxalate (pH 3), such as in the following reaction:

$$Fe_2O_{3(s)} + 2H_3O+_{(aq)} \leftrightarrow 2Fe(OH)_2 +_{(aq)} + H_2O_{(l)}$$

In the DCB extraction, sodium dithionite, as a strong reducing agent, solubilizes hematite and goethite from the soil clay fraction. The extraction of SOR by AO and NaOH allows access to the elements with intermediate mobility and solubility in relation to soluble, exchangeable, and crystalline forms. The inclusion of DCB (crystalline minerals) in sequential extraction was due to the importance of pedogenetic Fe oxides (hematite and goethite) in the differentiation of pedogenetic and geochemical environments such as in Brazil (27,28). It is important to emphasize that the potential of soil sample discrimination with these extractions is greatly increased with the use of the sequential protocol to improve the separation of the soil mineral phases, relevant for such an environment with highly weathered soils such as in this study.

Elements differ in form and in the phase they are in (26,29,30). Elements in the soluble fraction and adsorbed to the negative and positive charges of mineral and organic colloids should not be used in forensic science (16), as they are the most available and mobile forms of the elements, being easily affected by temporal changes such as fertilization and rainfall. If, for example, one of these management interferences on the soil surface occurs after the suspect's presence at the crime scene, the trace sample may lose its comparability and traceability to that location thus invalidating any comparisons made.

To avoid these transitory variations, several authors (16,18–21) used more aggressive extraction methods to analyze different and less transient mineral phases. The extraction of SOR by AO and NaOH allows access to the elements with intermediate mobility and solubility in relation to soluble and exchangeable and crystalline forms (20,25,26,29–32). This perspective is essential in the discrimination of soils which are highly similar pedologically (i.e., the same soil class and parent material) but separated by short distances (i.e., 10 to hundreds of meters). Under such a situation of similarity, the most appropriate chemical form to ascertain local chemical changes is the assessment of the low crystallinity materials (33).

One of the first forensic soil surveys reported in Brazil proposed significant changes in traditional analytical methods are to reduce the amount of soil traces necessary for analysis (16). These authors worked with less than 1.0 g of soil and used only the silt + clay fraction in the sequential analyses. The sand fraction compromises the homogenization of a small amount of sample (1.0 g) and has mainly primary minerals (quartz), which usually present low geochemical response to environmental variations (16). From this initial work of methodological definitions, other authors have successfully used sequential chemical extractions in simulated forensic cases (17–22). Another common procedure in all these simulated forensic cases was the treatment of the data by PCA and Bray–Curtis analysis, which were successfully applied and helped in the interpretation of forensic data including soil traces (14,16,17,19–22).

An opportunity arose to apply and test the sequential analysis protocol in a known murder case (femicide) in the State of Paraná, Brazil (19). Sequential mineralogical and chemical analyses as itemized above were used to chemically assess a limited amount of soil traces (less than 0.5 g in total was available on the questioned item) recovered forensically from a suspect's vehicle (adhering to the outside rear view mirror and to the left front fender of the suspect vehicle). All results were analyzed by PCA and Bray–Curtis and show moderate comparability between the questioned samples and some of the scene samples, and suggest that the soil recovered from the suspect's vehicle could have originated from the edge of the Graciosa Road (State of Paraná, South Brazil), close to the place where the victim's body was located. It could not be excluded as having originated from that location. However, in this case the suspect did not confess to having been to or near to this location, and in this homicide case, the soil data were used as one of a set of analyses in a combination of evidence types (19).

Another homicide occurred in the State of Paraná, South Brazil. In this case, the murderer confessed to the crime to the police. A spade with soil traces was subsequently seized from the murderer's house. Under this context, the hypothesis of this investigation was that the application of the protocol of sequential analysis to soil samples collected from the spade will test the hypothesis that the murderer used this tool at the crime scene. In order to maintain the alignment with all research that had used the sequential analysis, the data of the present work were analyzed by chemometric techniques, PCA and Bray–Curtis.

Material and Methods

Case Background and Soil Sampling

In 2016, a man was murdered, and his body was cut up into four parts, with two parts buried in a rural area at Colombo municipality, Curitiba Metropolitan Region (CMR), State of Paraná, South Brazil (Fig. 1*a*). The victim's ex-wife confessed to the crime and indicated where the parts of the body had been buried. The torso was found buried at 0–0.40 m depth in an agricultural area (Inceptisol, under agricultural cropping), and the legs were found 10.6 km away from the place where the torso was found, also buried at 0–0.40 m, in an area under native vegetation (Inceptisol, under forest).

The CMR is composed of four main cities (Fig. 1*a*): Curitiba (capital of the state of Paraná), Araucária (location where the parts of the body were buried), Colombo, and São José dos Pinhais. The predominant soil class in these four cities is Inceptisol, with variation of the underlying parent material: Curitiba municipality—claystone; Araucária municipality—granite/gneiss, Colombo municipality—limestone, and São José dos Pinhais municipality—granite/gneiss. Therefore, soil sampling in the municipalities of Curitiba and Colombo aimed to validate the discriminatory power of the sequential analyses of soil samples collected from different parent materials. Soils from São José dos Pinhais municipality would allow a test of discrimination of soils from the same pedological unit (Inceptisol) and parent material (granite/gneiss) in relation to the crime scene. The design of sampling locations is shown in Fig. 1.

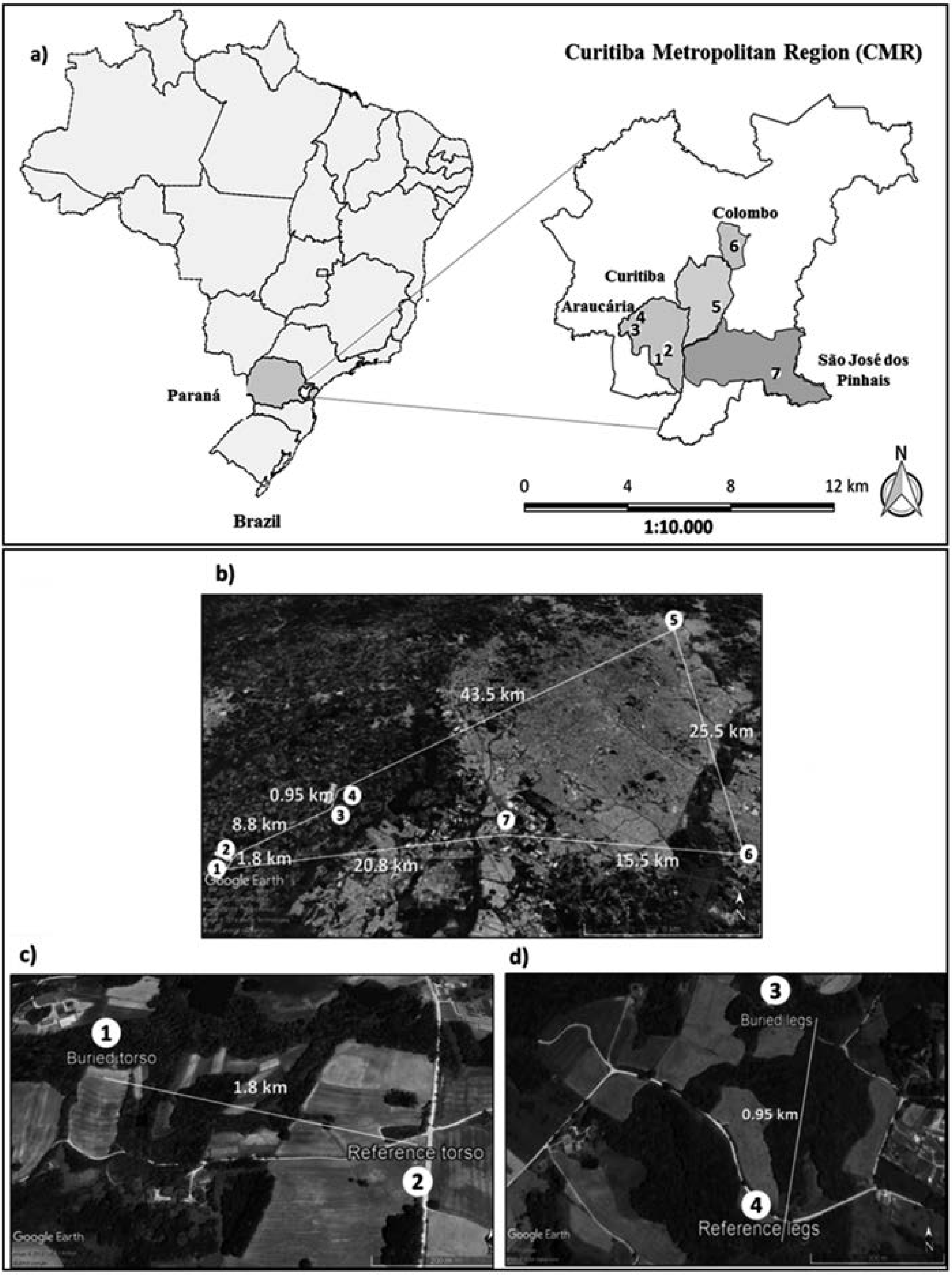


FIG. 1—*(a) Map of Brazil showing the State of Paraná and sites where the samples were collected (Araucária municipality, Colombo municipality, Curitiba municipality, and São José dos Pinhais municipality); (b) sampling locations and their relative distances: 1) Location 1 (samples from burial site of torso—Araucária municipality); 2) Location 2 (reference samples from site with torso—Araucária municipality); 3) Location 3 (samples from burial site of legs—Araucária municipality); 4) Location 4 (reference samples from site with legs—Araucária municipality); 5) Location 5 (claystone samples—Curitiba municipality); 6) Location 6 (limestone samples—Colombo municipality); 7) Location 7 (granite samples—São José dos Pinhais municipality); (c) position of location 1 in relation to the location 2; and d) position of the location 3 in relation to the location 4. [Color figure can be viewed at wileyonlinelibrary.com]*

- *Location 1—(Araucária municipality)* where the victim's torso was buried (coordinates −25°33′36″S, −49°28′31″W);
- *Location 2—(Araucária municipality)* reference samples, 1.8 km from location 1 (coordinates −25°33′43″S, −49°28′12″W). Soils at both locations 1 and 2 are Inceptisols, formed by granite/gneiss and are under the same land use (agricultural cropping). Even with the similarity and proximity among soils, use of the sequential analysis for discrimination between locations 1 and 2 was tested;
- *Location 3—(Araucária municipality)* where the victim's legs were buried, 8.8 km from location 2 (coordinates −25°31′28″ S, −49°25′39″W);
- *Location 4—(Araucária municipality)* reference samples, 0.95 km from location 3 (coordinates −25°31′46″S, −49°25′42″W). Locations 3 and 4—Inceptisols from granite/gneiss under native vegetation (pine forest).
- *Location 5—(Curitiba municipality)* reference samples formed from claystone, 43.5 km from location 4 (coordinates - 25°22′20.55″S, - 49°11′23.97″W) and under old pasture;
- *Location 6—(Colombo municipality)* reference samples, formed from limestone, 25.5 km from location 5 (coordinates - 25°33′37.56″S, - 49°12′30.54″W) and under old pasture;
- *Location 7—(São José dos Pinhais municipality)* reference samples formed from granite/gneiss, 15.5 km from location 6, 20.8 km from location 1 (coordinates −25°32′45.02″S, −49°19′52.15″W) and under old pasture.

Samples collected at locations 5, 6, and 7 are representative reference soils from the Metropolitan Region of Curitiba (MRC), which covers the municipality where the body parts were buried in Araucária (Fig. 1). In the MRC, there are three parent materials: claystone (location 5), limestone (location 6), and granite (location 7). Although location 7 has the same parent material as the locations where the torso and legs were recovered from, sampling at this point cannot be used as a reference for locations 1 and 3 due to its considerable distance from the crime scene. As the crime was confessed to, the actual locations of the burial sites of the torso and legs were found, and it was therefore possible to choose reference locations 2 and 4 which were under the same parent material, land use and also close to locations 1 and 3. The collection of reference soil samples at locations 5, 6, and 7 aimed to test the sequential analysis in a wider geographical area, with soils under different parent materials. Locations 5, 6, and 7 form a group in the principal component analysis (PCA). If the sequential analysis is efficient in forensic studies, replicate samples collected from location 7 (granite/gneiss) should form an isolated group in the PCA from the other groups from location 1 to 4 (also granite/gneiss) due to the distance from location 7 in relation to the crime scene. This discrimination based on the distance of soils from the same parent material is fundamental in forensic studies (19–21)

The sampling procedure was carried out using a standard operating procedure (SOP) for forensic soils developed in Brazil (20). At each location, 4 replicate soil samples were collected from the corners of a quadrant (1.5 m apart from each other), in both the A and B horizons (0–0.1 m (surface) and 0.1–0.4 m (middle depth), respectively; Fig. 2). The sampling of two horizons aimed to simulate the crime conditions, where the spade used to bury the parts of the body had potential contact with both the A and B horizons (0–0.4 m) at some time during the digging process. The sampling separation allowed the assessment as to which horizons had most soil adherent to the spade when recovered. Sampling equipment was carefully cleaned between uses to avoid any potential cross contamination between sampling positions.

The Scientific Police of Paraná State sized a spade which was suspected of being used to bury parts of the victim's body. Samples adhering to the spade were collected both from the front and from the rear of the tool, at several positions (Fig. 3). Visually, samples from both front and rear of the spade showed great homogeneity in color and texture. Six samples (replicates) were collected from the spade with no color or texture differentiation. Due to color homogeneity, the front of the spade with the most soil adhering to it was divided into 4 equal-area parts and the back of the spade with the least soil adhering to it was divided into 2 equal-area parts. The smallest subsample of the spade (area of the subsample 5) determined the amount of soil sample for all collections (3 g).

Sample Preparation and Soil Chemical Analyses

Prior to sample preparation and soil chemical analyses, a visual comparison of the soils was carried out in order to characterize the samples (Fig. 4). However, it was not possible to clearly differentiate the samples as they had a similar color.

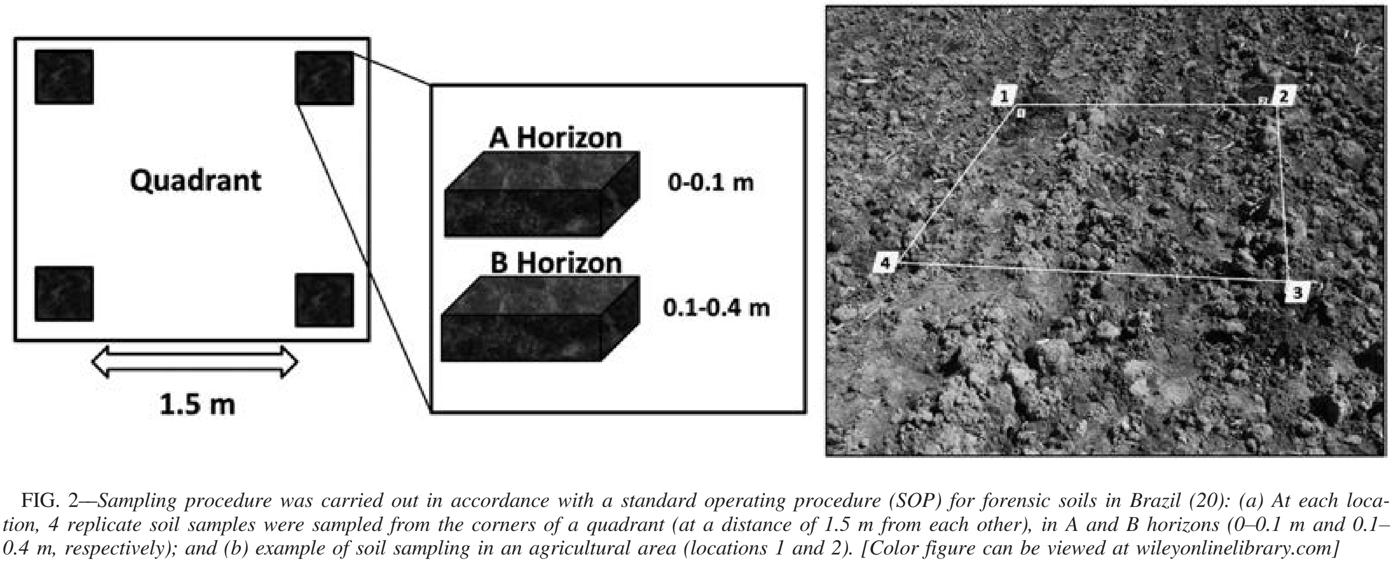


FIG. 2—*Sampling procedure was carried out in accordance with a standard operating procedure (SOP) for forensic soils in Brazil (20): (a) At each location, 4 replicate soil samples were sampled from the corners of a quadrant (at a distance of 1.5 m from each other), in A and B horizons (0–0.1 m and 0.1–0.4 m, respectively); and (b) example of soil sampling in an agricultural area (locations 1 and 2). [Color figure can be viewed at wileyonlinelibrary.com]*

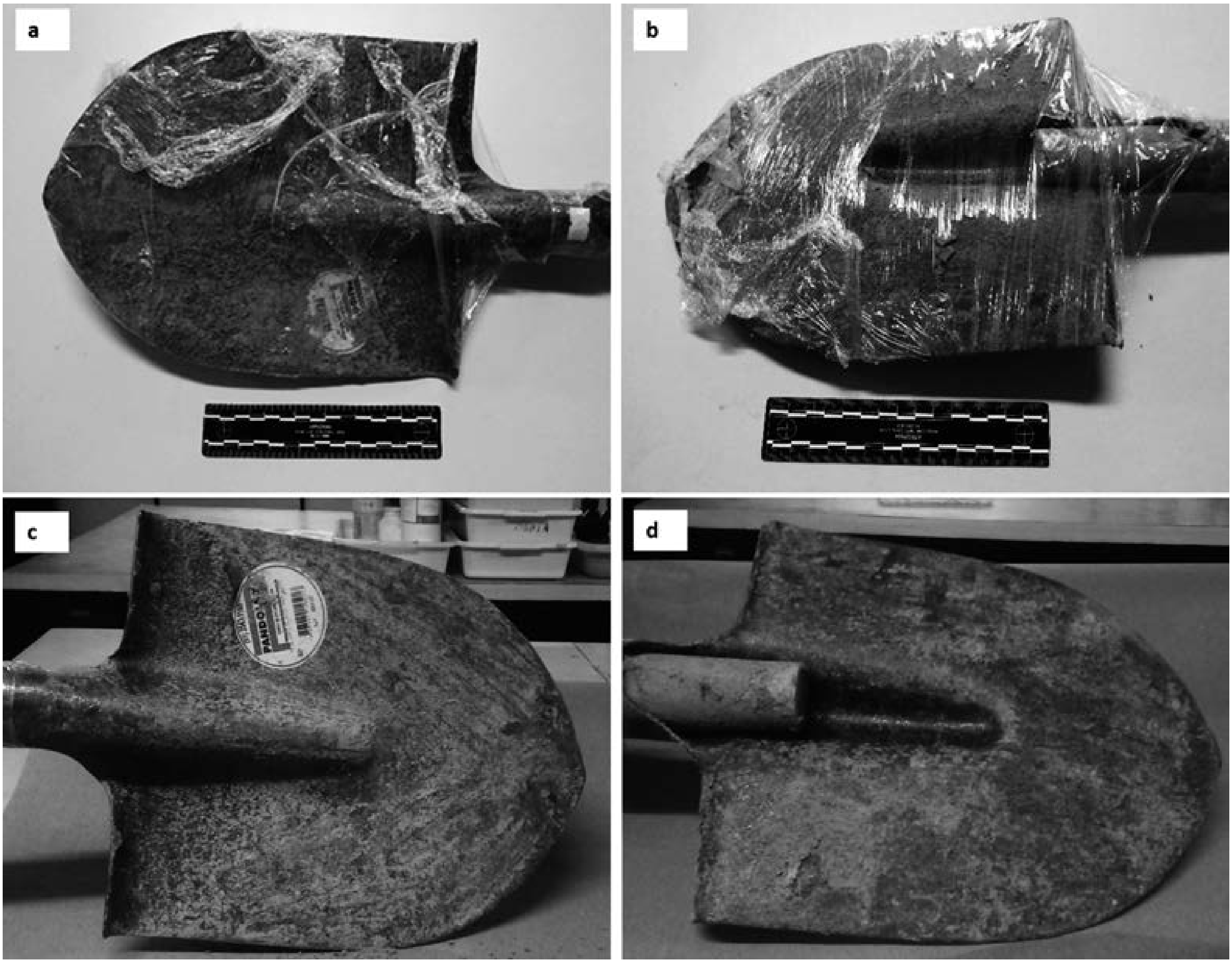


FIG. 3—*General image of the spade used to bury the body parts: before (a and b) and after (c and d) of the soil sampling. [Color figure can be viewed at wileyonlinelibrary.com]*

Soil samples were oven-dried (40°C, for 24 h) and sieved through a 2 mm sieve. A sequence of soil analytical procedures was carried out on the sampled soils, following the procedures previously developed for forensic investigations (16) (Fig. 5). The amount of dried soil for all samples used in this study was defined by the available amount in the subsample collected on the spade (Fig. 3). Three grams of soil was subjected to the organic matter removal with hydrogen peroxide (H_2O_2) 30% (v/v) in water bath under 70°C. The soil was dispersed via grinding with a rubber stick in the presence of pH 10 deionized water (1 g of Na_2CO_3 in 10 L of H_2O). The sand fraction was retained in a 0.053 mm mesh. Fractions smaller than 0.053 mm (silt + clay) were ground (with a pestle in an agate mortar) and sieved through a 0.2 mm mesh. Afterward, the silt and clay fractions were placed in an oven to remove the water. When dried, the silt + clay forms a thin crust, which in turn is lightly ground and then sieved through a 0.2 mm sieve, in order to homogenize the silt + clay fraction. This uniformity is intended to assist in the chemical extraction processes.

The most reactive fraction to the partial chemical extractions is the clay fraction. However, due to the reduced amount of soil traces available (0.8 g; Fig. 5), the silt fraction was too small for subsequent quantitative analysis. The finer particles (silt and clay) are preferentially retained on shoes, tyres, clothes, etc. (34). One gram of soil has been used previously in several simulated crime scenes (16,18–20). Similarly, in the application in an actual criminal case (19) 0.5 g of trace sample was used, and the full procedure was applied as described in Fig. 5. Therefore, this work followed the same procedures of previous work which applied the sequential analyses in the silt + clay fraction (16,18–20).

The amount of 0.8 g of silt + clay fraction was placed in centrifuge tubes covered with aluminum foil. After addition of 20 mL (weight ratio of 25:1) of 0.2 mol/L ammonium oxalate pH 3 solution (AO) (34), the tubes were agitated for 2 h. The suspension was centrifuged at 5000 rpm, and the supernatant collected to determine element contents.

The AO residue was treated with dithionite–citrate–bicarbonate (DCB) method (26). Samples of 0.65 g were placed in 100 mL tubes and subjected to the extraction three times with 10.4 mL (weight ratio of 16:1) of solution of sodium citrate 0.3 mol/L + 1.3 mL (weight ratio of 2:1) of a solution of sodium bicarbonate 1.0 mol/L + 0.26 g (weight ratio of 0.8:1) of sodium dithionite. Samples in solution were manually agitated while heated at 70°C in a water bath for a period of 30 min.

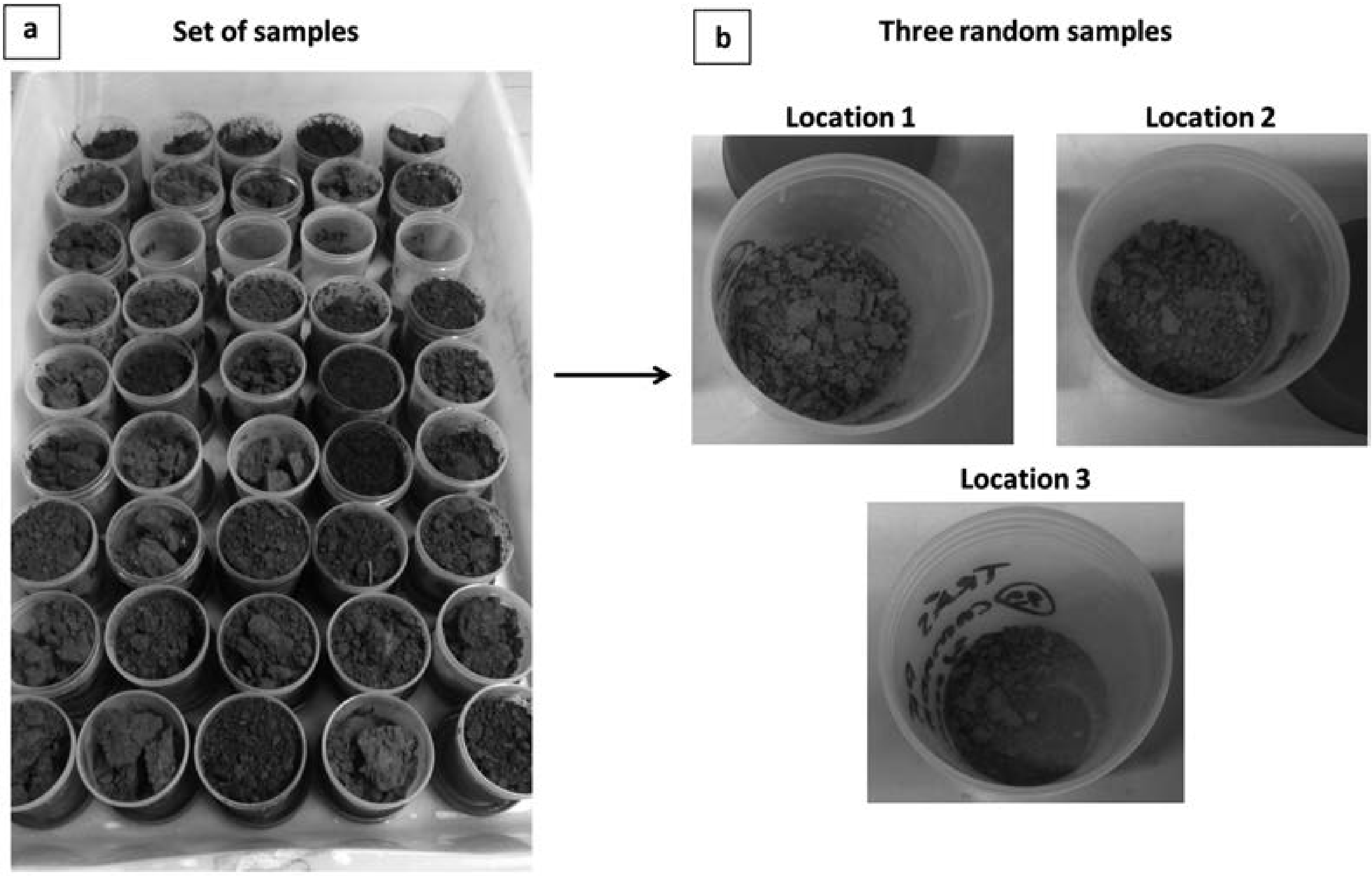


FIG. 4—*(a) Set of samples, and (b) visual comparison of samples collected from Araucária (Locations 1, 2, and 3), formed by granite/gneiss before they were dissolved by the sequential and chemical extractions. [Color figure can be viewed at wileyonlinelibrary.com]*

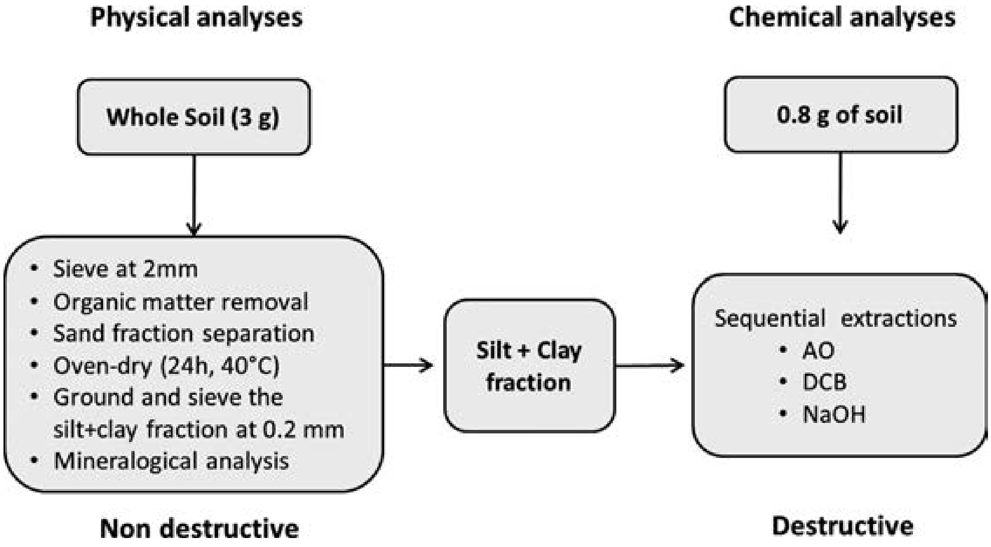


FIG. 5—*Scheme of physical and chemical procedures applied to the soil samples in this study. Note: AO, ammonium oxalate; DCB, dithionite–citrate–bicarbonate.*

The last step in the extraction process is the NaOH extraction. Around 0.40 g of DCB residue was placed in tubes with 2 mL of NaOH 0.5 mol/L without heating and homogenized. Teflon beakers containing 15 mL of the NaOH 0.5 mol/L solution were placed in a sand bath at 200°C until boiling. The total volume of solution was 17 mL (weight ratio of 42.5:1) which was then added to the tube with the samples. The boiling solution remained in contact with the sample for three minutes under constant manual agitation. Immediately after, samples were cooled in a recipient with cool water and then centrifuged.

For all sequential extractions, a series of common steps were established: (i) The extracting solutions and volumetric flasks were filled with ultrapure water (18.2 MΩ.cm at 25°C – Millipore Direct–Q System), and high purity acids were used in the analyses (Merck PA); (ii) before each extraction, the samples were dried for 24 hours in an oven at 40°C, ground with a pestle in an agate mortar, and sieved at 0.2 mm in order to improve the efficiency of the extractions; (iii) after each extraction, salt excess was removed by washing with 0.5 mol/L $(NH_4)_2CO_3$ and ultrapure water; (iv) after washing, the samples were oven-dried at 40°C for 24 h; (v) extracts were obtained by centrifugation (3000 rpm for 10 min), and all extracts were filtered in a slow filter paper (Macherey Nagel®); (vi) after dissolution, element concentrations were determined by optical emission spectrometry inductively coupled with plasma (Varian-720S) (ICP-OES). For the extracts obtained by AO and NaOH 0.5 mol/L, 0.6 and 1 mL of HNO_3 were respectively added before the analysis in the ICP-OES, to avoid the formation of precipitates in NaOH extracts and to better preserve the organic extractant AO.

The operational conditions of the ICP-OES with an axial configuration were the following: radiofrequency power—1200 W; replicate—3; plasma gas flow rate—15 L/min; auxiliary gas flow rate—1.5 L/min; sample uptake rate—1.0 mL/min; nebulizer gas flow rate—0.5 L/min; nebulizer type—spray chamber; spray chamber—double spray; injector tube diameter—1.2 mm; signal integration time—15 sec; and wavelength—203 nm. All procedures were performed including the blank analytical solutions.

Mineralogical Data

Mineralogical data from work previously carried out with samples from the same location (Araucária) (21) were included in this investigation to widen the discussion about the elementary data obtained by sequential chemical extractions. In this work, the following conditions to XRD data were used: diffraction patterns (random powder samples) were obtained in the equipment Panalytical X'Pert3, under 0.42 °2θ sec^{-1} speed, and analyzed in the range from 3 to 35 °2θ. The diffractometer was equipped with nickel filter, graphite monochromator, and CuKα radiation, and it was operated at 40 kV and 40 mA.

Multivariate Statistics

The application of multivariate statistics to soil data is a useful approach and sometimes crucial to help elucidate and evaluate patterns of data in criminal cases. This type of statistical approach offers a clear ability to test for grouping of samples as well as separation between samples, that is, inter- and intragroup comparisons. In addition, it is possible to verify the similarities and the dissimilarities and discrimination in a set of soil samples based on the analytical data obtained.

Data obtained from the silt + clay fraction were exported from a single matrix, square root transformed and statistically analyzed by principal component analysis (PCA) using the software Statistica (35) and Paleontological Statistics (PAST), through the application of clustering using Bray–Curtis similarities (36). From the matrix data, two PCAs were generated: PCA 1) elemental contents extracted from AO, DCB, and NaOH extractions for the A horizon; PCA 2) elemental contents extracted from AO, DCB, and NaOH extractions for the B horizon. The elements which presented values higher than the limit of detection (LD) in the ICP-OES for all samples and extraction methods were selected: Al (LD—0.01 mg/L), Fe (LD—0.009 mg/L), Mg (LD—0.006 mg/L), Zn (LD—0.003 mg/L), Cu (LD—0.01 mg/L), Mn (LD—0.006 mg/L), and Pb (LD—0.003 mg/L; values are available on Tables S1–S3).

The Bray–Curtis clustering approach has been successfully used in other forensic investigations (19,20) and presents the groups (clusters) and the index of maximum similarity among the groups of samples, particularly those which share common characteristics. Pearson correlations among the elemental concentrations of the AO, DCB, and NaOH extractions were carried out using the software Statistica (35).

Results and Discussion

Soils from the same sampling areas have been used in previous work carried out in South Brazil (19–21), with a very similar mineralogical profile among the samples, particularly among samples from Araucária (21). X-ray diffraction consistently revealed a predominance of kaolinite, quartz (explained by the presence of the silt in the samples analyzed), gibbsite, and hematite (Fig. 6).

Considering the high degree of similarity in the mineralogical assemblage, soils from the same parent material and located close by may be distinguished. Data obtained by sequential chemical extraction with AO (ammonium oxalate), DCB (dithionite–citrate–bicarbonate), and 0.5 mol L^{-1} NaOH were compiled in a matrix separated by horizons (Figs 7 and 8). The elemental profiles (contents of Al_2O_3, Fe_2O_3, Mg, Zn, Cu, Mn, and Pb) of the samples from the A horizon showed a lower variation among sites compared to the samples from the B horizon (please see Tables S1–S3): the larger variation among the sites was verified in the Mg content extracted by NaOH 0.5 mol/L (A horizon—1.0 mg/kg, B horizon—0.1 mg/kg, Claystone location); Mn extracted by AO (A horizon—321.8 mg/kg, B horizon—79.9 mg/kg, Reference torso location), and extracted by DCB (A horizon—172.8 mg/kg, B horizon—57.6 mg/kg, Reference torso location).

The highest Fe_2O_3 contents were obtained by DCB extraction (Figs 7 and 8). One of the advantages of sequential extraction is the possibility of isolating the SRO and crystalline mineral phases of the soil. This behavior follows the nature of sequential and selective extractions: AO—dissolution of allophanes and imogolites (short-range order material—SRO with Si/Al molar ratio greater than 2:1; and amorphous Fe minerals (ferrihydrite, fougerite, and schwertmannite), Mn and Al oxides (37,38); DCB—crystalline Fe oxides (hematite and goethite) (25); and NaOH—Al-hydroxide, Al–O–Si layers resistant to previous AO extraction and Si–O (opaline silica) (24). The greater extraction of Fe_2O_3 by DCB in relation to AO is due to the predominance of crystalline forms of Fe (hematite and goethite) over Fe-SRO (e.g., ferryhidrite) (39,40).

The highest Fe_2O_3-DCB contents were found in the soils of the external municipalities (used as reference locations—Curitiba, Colombo, and São José dos Pinhais) to the burial locations of the victim's body parts (Araucária municipality). The granite/gneiss from São José dos Pinhais municipality presents higher levels of primary iron–magnesian minerals in the sand fraction (16,18). The weathering of these minerals released more Fe for hematite and mainly goethite formation in the clay fraction.

DCB does not extract Al_2O_3 in an isolated phase. The incorporation of Al into the hematite and goethite structure is verified by isomorphic substitution (IS) of Fe^{3+} by Al^{3+}, both in octahedral coordination. The highest correlation coefficient ($p < 0.01$)

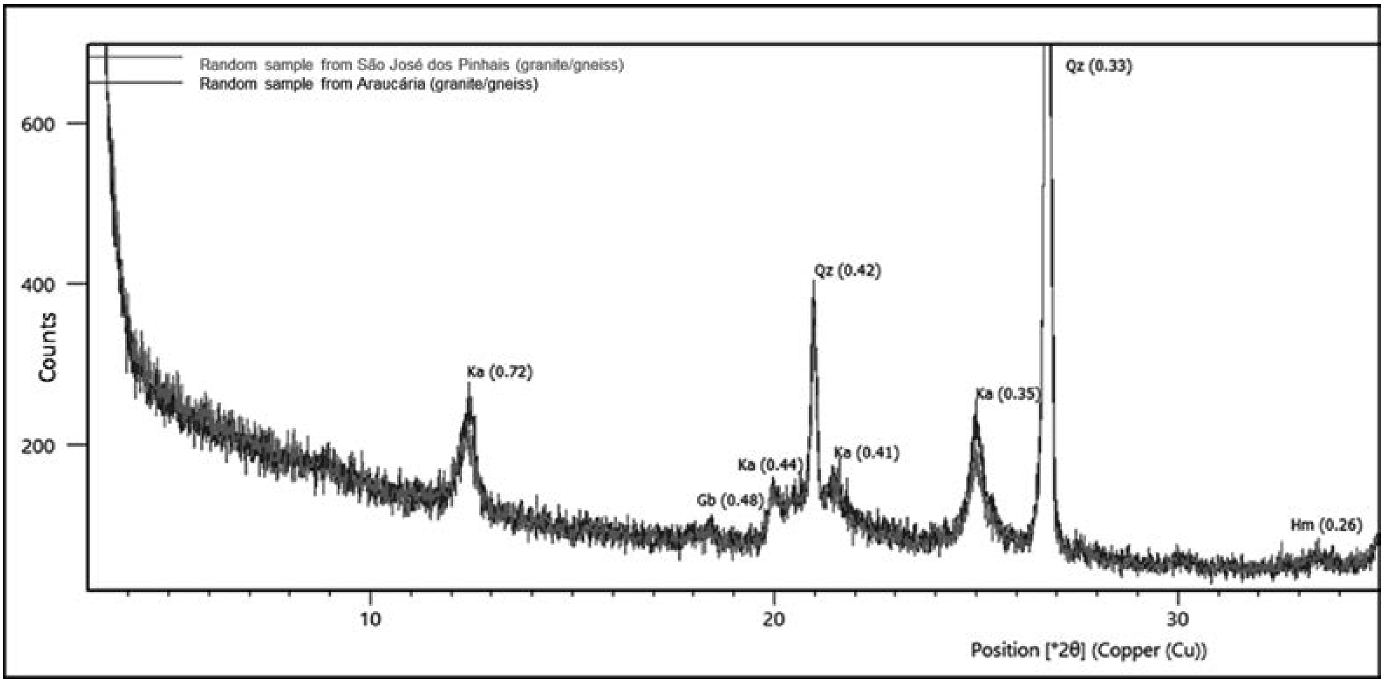


FIG. 6—*Mineralogical profile of ramdom samples from São José dos Pinhais and Araucária municipalities (granite/gneiss), State of Paraná, Brazil. (Adapted from [21]). Note: Ka, kaolinite; Gb, gibbsite; Qz, quartz; Hm, hematite. The interlayer distances in nanometers (nm) are presented in parenthesis. [Color figure can be viewed at wileyonlinelibrary.com]*

between the elements extracted by the DCB was observed between Fe and Al, indicating a similar source for both elements (Table 1). Hematite and goethite in the A horizon from the location where the torso was buried presented lower IS levels (smaller Al_2O_3-DCB contents; Fig. 7), a factor that also contributed to discrimination of this location from the other locations in the statistical analysis PCA and clustering by Bray–Curtis (Figs 9–12).

AO extracted more amorphous Al oxides than amorphous Fe oxides (Figs 7 and 8). This behavior has also been shown by other authors (24,25). However, samples that had more amorphous Al oxides also had higher contents of amorphous Fe oxides ($r = 0.77$, $p < 0.01$; Table 1). The conditions that favor the formation and stabilization of amorphous Fe and Al oxides are higher organic matter contents and lower redox potential conditions (excess of water) (41).

The content of Al_2O_3-NaOH was important to separate soil samples collected at the Curitiba (claystone) and São José dos Pinhais municipalities (granite/gneiss), and the reference locations in the set of A horizon samples (Fig. 7). The behavior for the B horizon samples was much more variable (Fig. 8), with greater variation of Al_2O_3-NaOH and, consequently, providing only a low power of forensic sample discrimination.

In general, for both the A and B soil horizons, the main source of Mn was the amorphous oxides extracted by AO (Figs 7 and 8). Another possibility is the association between Fe and Mn in the crystals of amorphous oxides. The correlation coefficient between Fe-AO and Mn-AO contents was high, positive, and significant ($r = 0.77$, $p < 0.01$; Table 1). The difference of the ionic radii between Fe^{3+} (0.067 nm) and Mn^{4+} (0.052 nm) is only 22%. Under room temperature and pressure conditions, the theoretical limit of ionic radius difference to enable isomorphic substitution (IS) is 35% (42). Mn-AO contents were high for the A horizon samples of the location where the legs had been buried and the reference samples from the location where the torso had been buried (800 and 300 mg/kg, respectively), being lower than where the legs had been buried for the B horizon (900 mg/kg). Mn-AO contents isolated these samples from the others.

The other elements (Mg, Cu, Zn, and Pb) are normally associated with the colloidal fraction of the soil by IS or by inner-sphere adsorption to the surface reactive groups of minerals. The highest Mg-DCB contents in the A and B horizons from the Colombo municipality soils are consistent with originating from the carbonate material (Figs 7 and 8). These sedimentary rocks are formed of carbonates of Ca (calcite), Mg (magnesite), and Ca-Mg (dolomite) (43). The Zn-DCB contents separated the reference samples group from the A and B horizons from the Curitiba municipality, Colombo municipality, and São José dos Pinhais municipality from the other samples belonging to the crime scene region (Araucária municipality). Higher Mg-AO contents of A horizon also separated the locations where the victim's torso and legs were buried (Fig. 7). The higher contents of Pb in the A horizon of the agricultural area (torso and reference torso; Fig. 7) are compatible with a long period of the addition of fertilizers to the soil, mainly phosphate-based ones (44,45).

The sequential chemical extractions carried out with ammonium oxalate (AO), dithionite–citrate–bicarbonate (DCB), and NaOH 0.5 mol/L (Fig. 4) were clearly efficient in separating samples by their sampling locations using PCA (Figs 9 and 10). A tendency for greater clustering was observed for the set of soil samples collected from the superficial soil layers (i.e., the A horizon). Even with the removal of organic matter prior to the sequential extractions, the effect of this colloidal fraction was seen in the chemical characteristics of the secondary minerals from the different sampling locations. The organic matter favors the formation and stabilization of SRO minerals in the clay fraction (46,47).

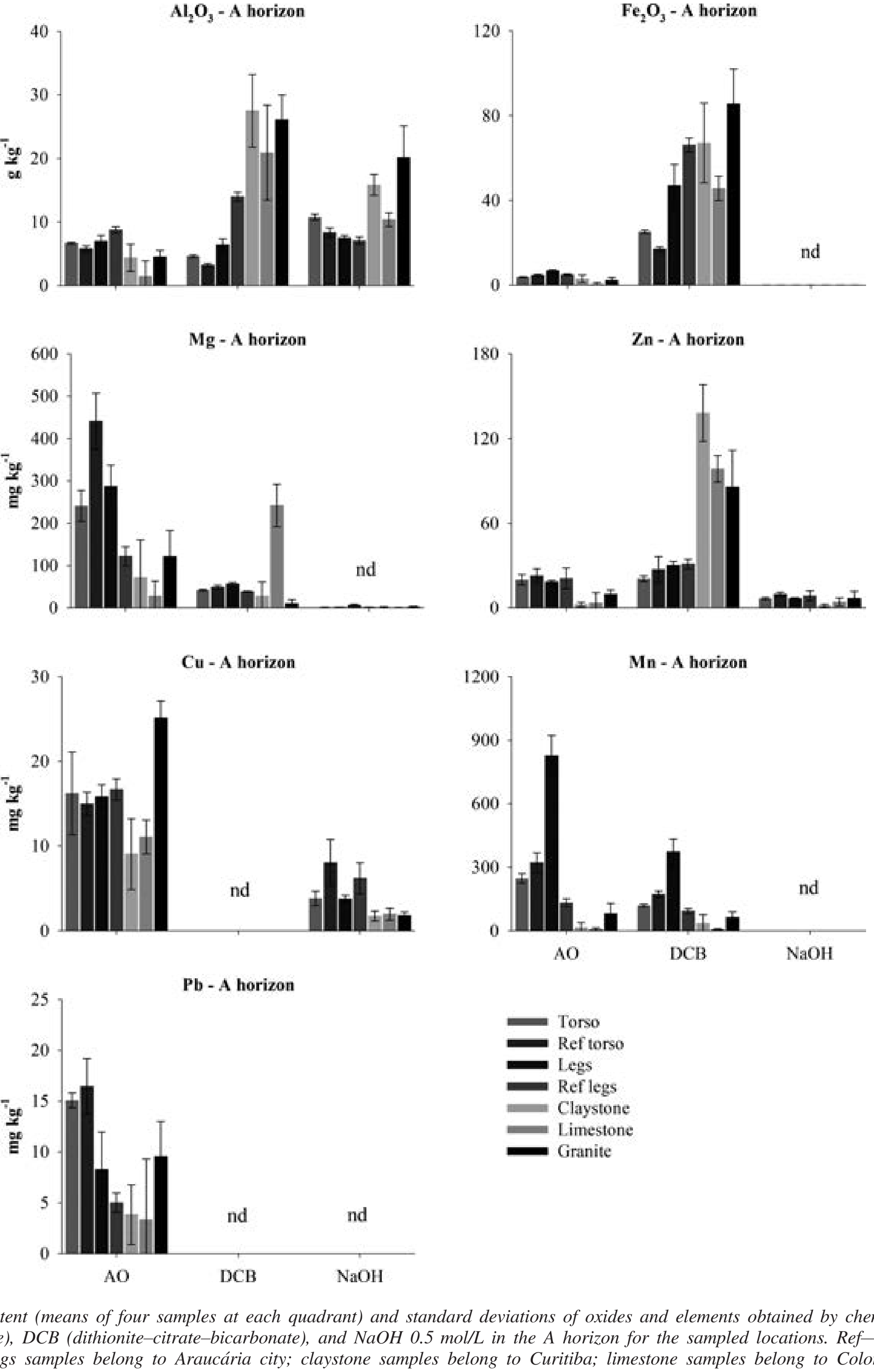


FIG. 7—*Average content (means of four samples at each quadrant) and standard deviations of oxides and elements obtained by chemical extractions with AO (ammonium oxalate), DCB (dithionite–citrate–bicarbonate), and NaOH 0.5 mol/L in the A horizon for the sampled locations. Ref—reference; Torso, Ref torso; Legs and Ref legs samples belong to Araucária city; claystone samples belong to Curitiba; limestone samples belong to Colombo; granite samples belong to São José dos Pinhais; and cities from Metropolitan Region of Curitiba (MRC), State of Paraná, South Brazil. nd—below the detection limit of ICP-OES. [Color figure can be viewed at wileyonlinelibrary.com]*

The cluster analysis allows a comparison of the samples collected at the same location (i.e., the replicates; Figs 11 and 12). In the A horizon, the following sequence of similarity among the sample groups was the following: (Fig. 11) (i) with 98% of similarity was a grouping of all six soil replicates collected from the spade. This shows the great similarity of the soil mass adhering to the spade likely during the burial of the victim's body parts. The four replicates collected at the torso burial location had a 97% similarity. There was also a clustering of all reference samples of the location where the legs were recovered from and two samples from the burial location where the legs were recovered from (97% similarity), and which also compared with the limestone and granite reference locations. It was expected that there would be a grouping of all replicates collected at the same

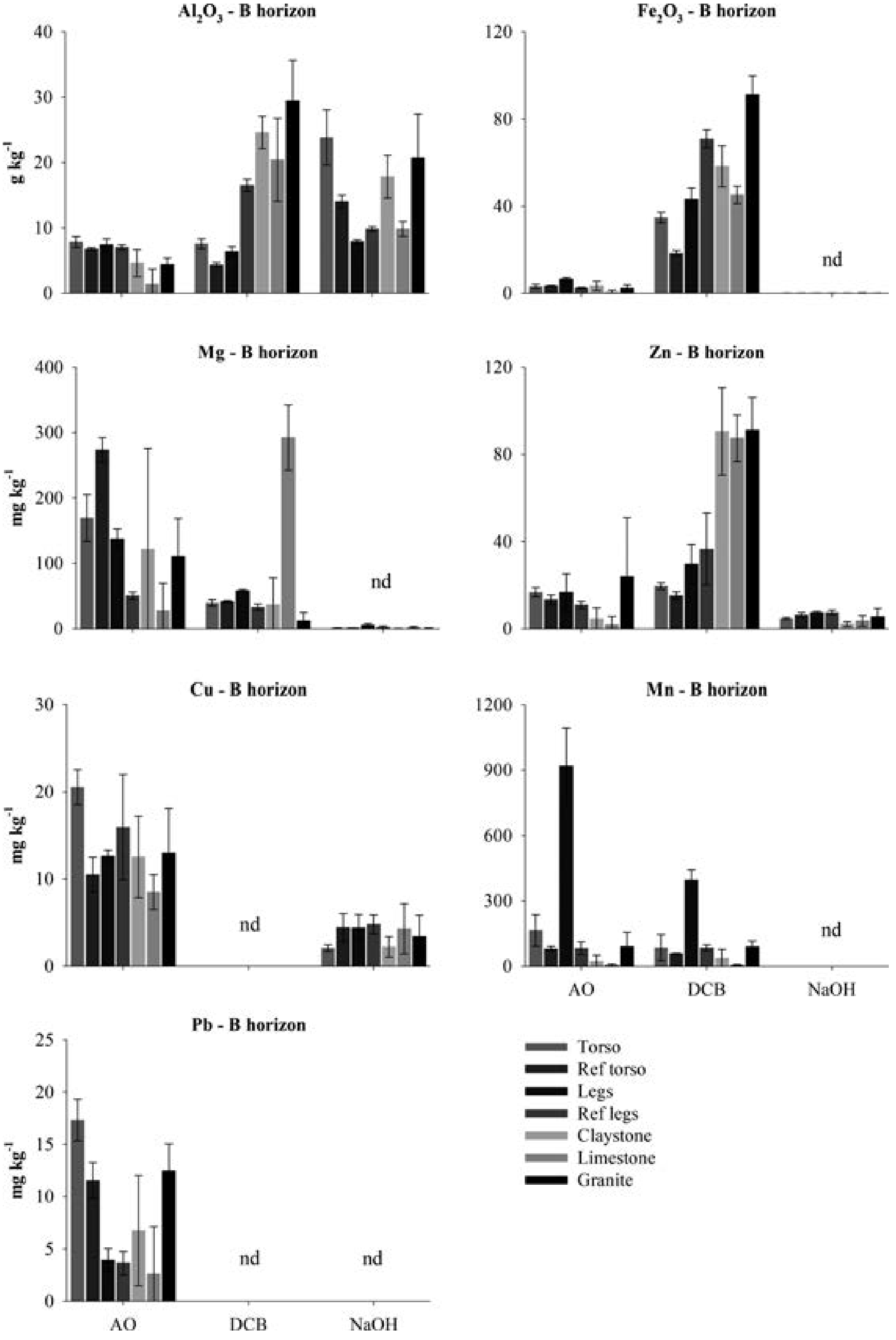


FIG. 8—*Average content (four samples at each quadrant) and standard deviations of oxides and elements obtained by chemical extractions with AO (ammonium oxalate), DCB (dithionite–citrate–bicarbonate), and NaOH 0.5 mol/L in the B horizon for the sampled locations. Ref—reference; Torso, Ref torso; Legs and Ref legs samples belong to Araucária city; claystone samples belong to Curitiba; limestone samples belong to Colombo; granite samples belong to São José dos Pinhais; and cities from Metropolitan Region of Curitiba (MRC), State of Paraná, South Brazil. nd—below the detection limit of ICP-OES. [Color figure can be viewed at wileyonlinelibrary.com]*

place with a high index of similarity. However, even if replicates are collected at a close distance to each other (i.e., 1.5 m from each other), horizontal differences in the chemical characteristics of the soil samples can occur; (ii) the best grouping was for the samples collected from the spade with the samples collected at the burial location of the torso (95% of similarity). As the murderer had confessed to having used the same spade to bury all the victim's parts, the clustering data show that there was a preferential adherence of the soil from the A horizon during the burial of the torso. This was the most important association found from this work. The initial contact of the clean spade with the first surface layer promoted considerable adherence of the A

TABLE 1—*Pearson correlations among the elements obtained by AO (ammonium oxalate), DCB (dithionite–citrate–bicarbonate), and NaOH 0.5 mol/L extractions for all samples of this study.*

	Cu	Fe	Mg	Mn	Pb	Zn
AO extraction						
Al	0.43*,**	0.77*,**	0.46*,**	0.45*,**	0.40*,**	0.49*,**
Cu		ns	ns	ns	0.29*,**	ns
Fe			0.63*,**	0.77*,**	0.27*,**	0.36*,**
Mg				0.43*,**	0.63*,**	0.39*,**
Mn					ns	0.33*,**
Pb						0.43*,**
DCB extraction						
Al	ns	0.80*,**	ns	0.54*,**	ns	0.76*,**
Cu		ns	ns	ns	ns	0.53*,**
Fe			ns	ns	ns	ns
Mg				ns	ns	−0.42*,**
Mn					ns	ns
Pb						ns
NaOH extraction						
Al	−0.47*,**	ns	−0.37*,**	ns	ns	−0.31*,**
Cu		ns	ns	ns	ns	0.60*,**
Fe			ns	ns	ns	ns
Mg				ns	ns	0.28*,**
Mn					ns	ns
Pb						ns

ns, not significant.
*,**Significant at 5 and 1 % probability, respectively.

horizon soil. When the spade reached the subsurface, the transfer of horizon B soil was less, possibly due the already adhering soil from the A horizon. More soil adherence on a metal surface (horizon A on the spade) is expected than soil–soil adherence (horizon B on top of horizon A soil). A similarity index of 95% was observed for the group of all reference samples from the location where the legs were recovered from and the grouping of three replicates of the samples collected at the location where the legs had been buried (iii) the large group of spade plus torso burial location + torso reference samples showed a 85% similarity. Even though these two sampling sites were positioned at a close horizontal distance from each other and were in a similar environment (agricultural soils under the same management and with the same underlying parent material; Fig. 1), data produced by chemical extractions were able to identify greater similarity between replicates (intragroup similarity—similarity of the isolated groups of the torso and reference torso equal to 97 and 95%, respectively). These data confirm the preferential adherence of soil on the spade at the torso burial location.

For the B horizon soils, the formation of the groups was more dispersed and presenting with lower similarity indexes (Fig. 12): (i) the group of all the samples of the location where the torso had been buried presented lower similarity (92%) compared to the A horizon (97%); (ii) the first similarity value between two groups occurred at the 83% similarity index (torso + reference torso). The second important group (torso + legs) showed a similarity value of 80%; (iii) at 77% of similarity index, the spade was placed at the burial location (large group legs + torso + reference torso + spade); and (iv) the sample groups of the different parent materials (São José dos Pinhais municipality—granite/gneiss, Curitiba municipality—claystone, and Colombo municipality—limestone) showed 87% of intragroup similarity.

Conclusions

The chemical sequential extractions were performed according to the previous studies (19,20). The soil from the A horizon at the burial torso site location had adhered to the spade. However,

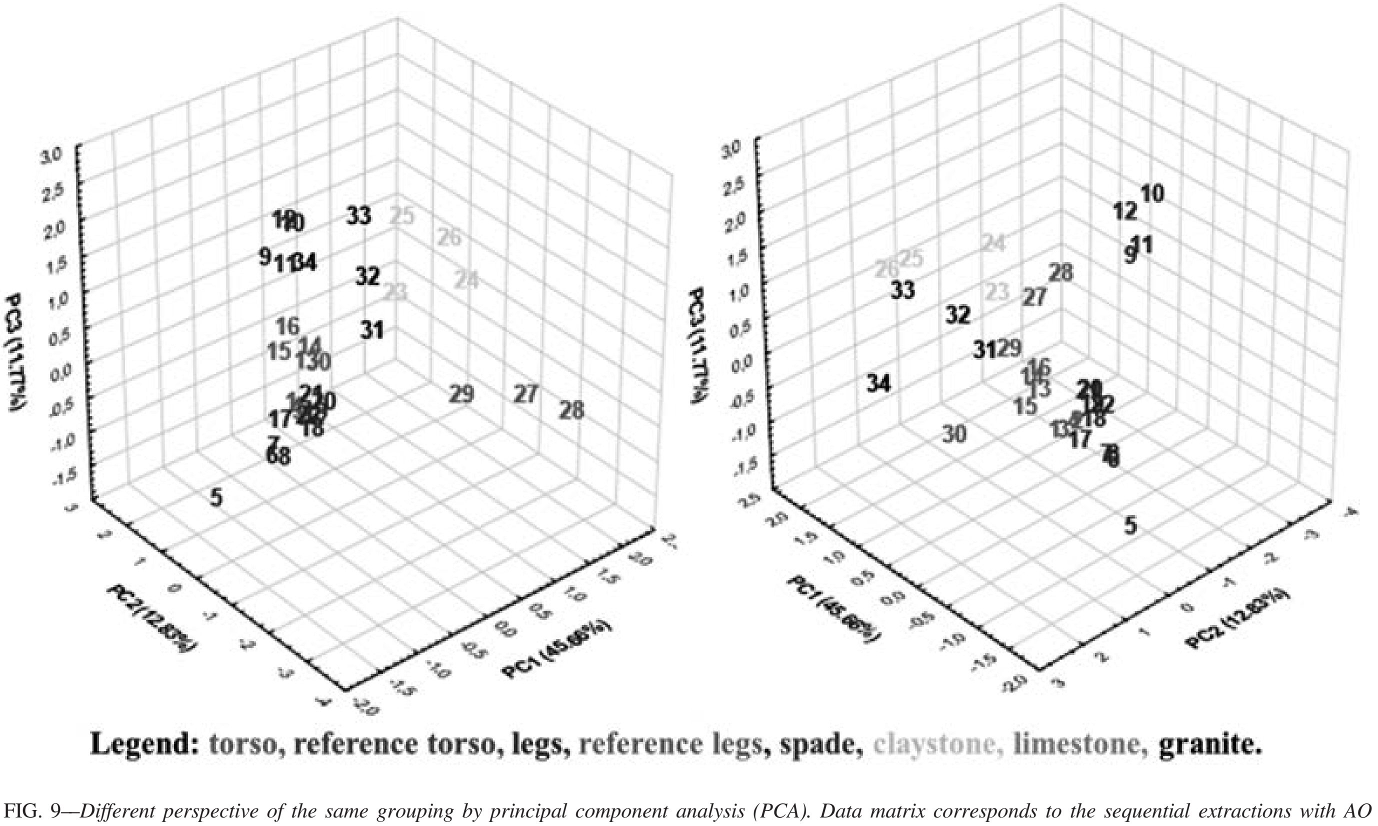


FIG. 9—*Different perspective of the same grouping by principal component analysis (PCA). Data matrix corresponds to the sequential extractions with AO (ammonium oxalate), DCB (dithionite–citrate–bicarbonate), and NaOH 0.5 mol/L obtained in the A horizon for all samples of this study (including spade samples). Ref, reference. [Color figure can be viewed at wileyonlinelibrary.com]*

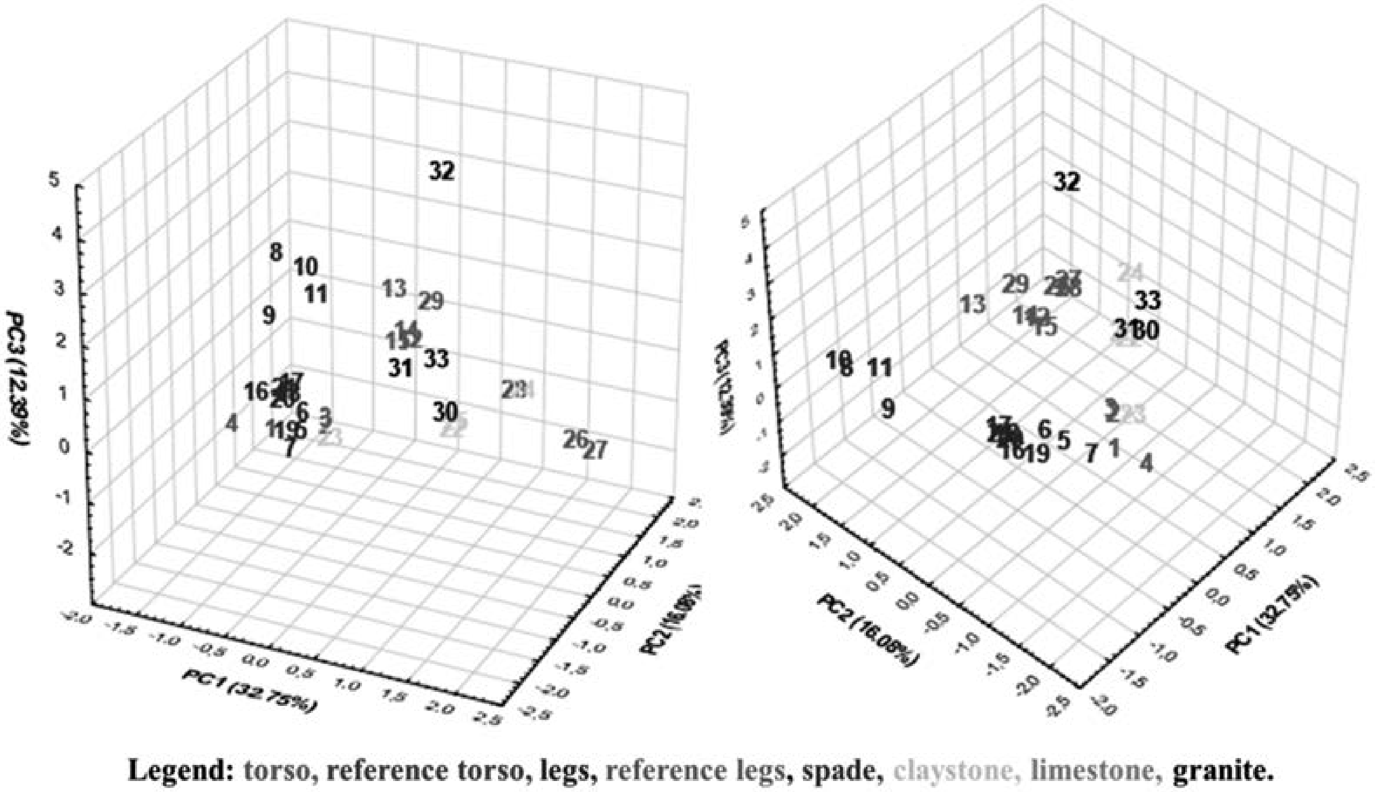


FIG. 10—*Different perspective of the same grouping by principal component analysis (PCA). Data matrix corresponds to the sequential extractions with AO (ammonium oxalate), DCB (dithionite–citrate–bicarbonate), and NaOH 0.5 mol/L obtained in the B horizon for all samples of this study (including spade samples). Ref, reference. [Color figure can be viewed at wileyonlinelibrary.com]*

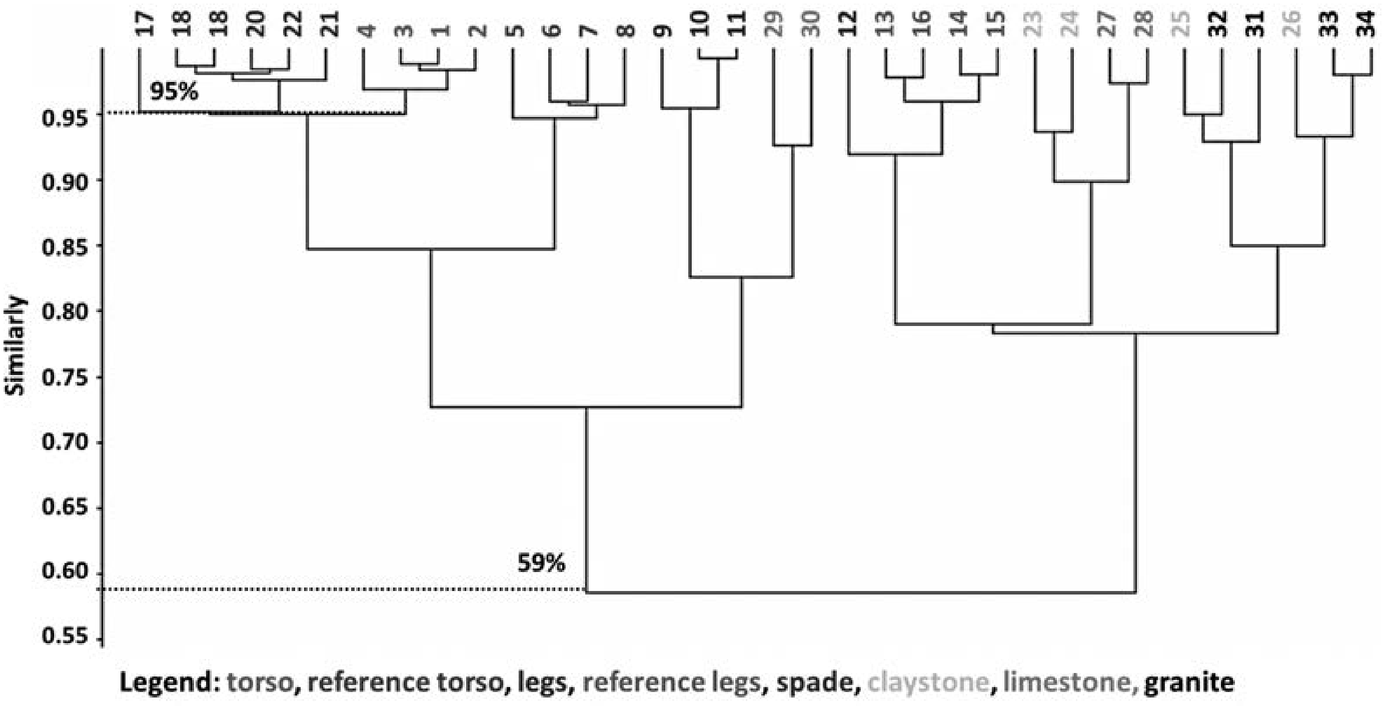


FIG. 11—*Clustering by Bray–Curtis. Data matrix corresponds to the sequential extractions with AO (ammonium oxalate), DCB (dithionite–citrate–bicarbonate), and NaOH 0.5 mol/L obtained in the A horizon for all samples of this study (including spade samples). Ref, references. [Color figure can be viewed at wileyonlinelibrary.com]*

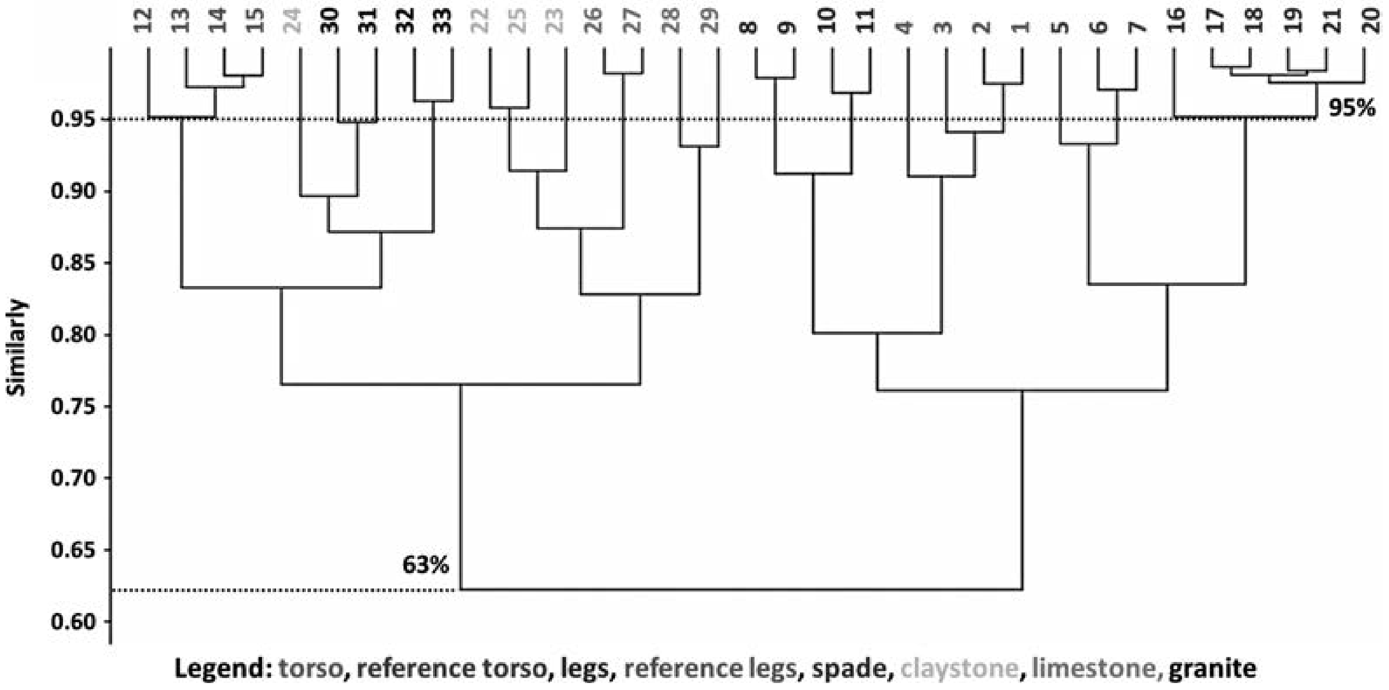


FIG. 12—*Clustering by Bray–Curtis. Data matrix corresponds to the sequential extractions with AO (ammonium oxalate), DCB (dithionite–citrate–bicarbonate), and NaOH 0.5 mol/L obtained in the B horizon for all samples of this study (including spade samples). Ref, references. [Color figure can be viewed at wileyonlinelibrary.com]*

the site where the legs were subsequently buried contributed much less trace soil material to the spade. Therefore, the analytical procedures proposed in previous studies (16) presented a promising approach for the analysis of soils collected at crime scenes in Brazil, provided enough soil is available from the questioned item. The sequential chemical analysis should be considered and prioritized instead of isolated extractions to improve discrimination and comparability between trace soil samples from a questioned item and a crime scene. It also demonstrates the importance in considering aspects of activity level information, such as which location the tool had been used at first, in this case the site where the torso had been buried, and thus likely provides the stronger evidence of that contact.

References

1. Concheri G, Bertoldi D, Polone E, Otto S, Larcher R, Squartini A. Chemical elemental distribution and soil DNA fingerprints provide the critical evidence in murder case investigation. PLoS One 2011;6(6):4–8. https://doi.org/10.1371/journal.pone.0020222.
2. Dawson LA. Soil organic characterisation in forensic case work. J Int Geosci 2017;40(2):157–65. https://doi.org/10.18814/epiiugs/2017/v40i2/017018.
3. Reidy L, Bu K, Godfrey M, Cizdziel JV. Elemental fingerprinting of soils using ICP-MS and multivariate statistics: a study for and by forensic chemistry majors. Forensic Sci Int 2013;233(1–3):37–44. https://doi.org/10.1016/j.forsciint.2013.08.019https://doi.org/10.1016/j.forsciint.2013.08.019.
4. Fitzpatrick R, Raven M, Self P. Clay mineralogy as significant evidence in 4 murder investigations involving a wide range of earth materials from Perth, Adelaide, Melbourne and Sydney. In: Proceedings of the Australian Clay Minerals Society Conference; 2014 Feb 3–5; Perth, Australia. Melbourne, Australia: Australian Clay Minerals Society, 2014;23–6.
5. Farrugia KJ, Bandey H, Dawson L, Daéid N. Chemical enhancement of soil based footwear impressions on fabric. Forensic Sci Int 2012;219(1–3):12–28. https://doi.org/10.1016/j.forsciint.2011.11.011.
6. Mayes RW, Macdonald LM, Ross JM, Dawson LA. Discrimination of domestic garden soils using plant wax compounds as markers. In: Ritz K, Dawson L, Miller D, editors. Criminal and environmental soil forensics. Scotland, U.K: Springer, 2009;463–76.
7. Finley SJ, Benbow ME, Javan GT. Potential applications of soil microbial ecology and next-generation sequencing in criminal investigations. Appl Soil Ecol 2015;88:69–78. https://doi.org/10.1016/j.apsoil.2015.01.001.
8. Woods B, Lennard C, Kirkbride KP, Robertson J. Soil examination for a forensic trace evidence laboratory – Part 3: a proposed protocol for the effective triage and management of soil examinations. Forensic Sci Int 2016;262:46–55. https://doi.org/10.1016/j.forsciint.2014.08.009.
9. Demanèche S, Schauser L, Dawson L, Franqueville L, Simonet P. Microbial soil community analyses for forensic science: application to a blind test. Forensic Sci Int 2017;270:153–8. . https://doi.org/10.1016/j.forsciint.2016.12.004.
10. Fontana A, Chagas CS, Donagemma GK, Menezes AR, Filho BC. Soils developed on geomorphic surfaces in the mountain region of the State of Rio de Janeiro. Rev Bras Ciência do Solo 2017;41:e016057. https://doi.org/10.1590/18069657rbcs20160574.
11. Ramos PV, Dalmolin RSD, Marques Júnior J, Siqueira DS, de Almeida JA, Moura-Bueno JM, et al. Magnetic susceptibility of soil to differentiate soil environments in southern Brazil. Rev Bras Ciência do Solo 2017;41:e0160189. https://doi.org/10.1590/18069657rbcs20160189.
12. Pye K, Croft D. Forensic analysis of soil and sediment traces by scanning electron microscopy and energy-dispersive X-ray analysis: an experimental investigation. Forensic Sci Int 2007;165(1):52–63. https://doi.org/10.1016/j.forsciint.2006.03.001.
13. Carvalho Á, Ribeiro H, Mayes R, Guedes A, Abreu I, Noronha F, et al. Organic matter characterization of sediments in two river beaches from northern Portugal for forensic application. Forensic Sci Int 2013;233(1–3):403–15. https://doi.org/10.1016/j.forsciint.2013.10.019.
14. Melo VF, Mazzetto JML, Dieckow J, Bonfleur EJ. Factor analysis of organic soils for site discrimination in a forensic setting. Forensic Sci Int 2018;290:244–50. https://doi.org/10.1016/j.forsciint.2018.07.005.
15. Melo VF, Schaefer CEGR, Novais RF, Singh B, Fontes MPF. Potassium and magnesium in clay minerals of some Brazilian soils as indicated by a sequential extraction procedure. Commun Soil Sci Plant Anal 2002;33(13–14):2203–25. https://doi.org/10.1590/S0100-06832000000200004.
16. Melo VF, Barbar LC, Zamora PGP, Schaefer CE, Cordeiro GA. Chemical, physical and mineralogical characterization of soils from the

Curitiba Metropolitan Region for forensic purpose. Forensic Sci Int 2008;179:123–34. https://doi.org/10.1016/j.forsciint.2008.04.028.
17. Corrêa RS, Melo VF, Abreu GGF, Sousa MH, Chaker JA, Gomes JA. Soil forensics: how far can soil clay analysis distinguish between soil vestiges? Sci Justice 2018;58(2):138–44. https://doi.org/10.1016/j.scijus.2017.09.003.
18. Prandel LV, Melo VF, Brinatti AM, Saab SC, Salvador FAS. X-ray diffraction and rietveld refinement in deferrified clays for forensic science. J Forensic Sci 2018;63(1):251–7. https://doi.org/10.1111/1556-4029.13476.
19. Melo VF, Testoni SA, Dawson L, Lara AG, Salvador FAS. Can analysis of a small clod of soil help to solve a murder case? Sci Justice 2019;59(6):667–77. https://doi.org/10.1016/j.scijus.2019.06.008.
20. Testoni SA, Melo VF, Dawson LA, Salvador FAS, Kunii PA. Validation of a standard operating procedure (SOP) for forensic soils investigation in Brazil. Rev Bras Ciência do Solo 2019;43:e0190010. https://doi.org/10.1590/18069657rbcs20190010.
21. Testoni SA, Melo VF, Dawson LA, Salvador FAS. Prandel LV. Evaluation of forensic soil traces from a crime scene: robbery of a safety deposit box in Brazil. Geol Soc London Spec Publ 2019;492:1–31. https://doi.org/10.1144/SP492-2019-35.
22. Testoni AS. Pedologia e mineralogia do solo aplicadas às ciências forenses [Pedology and soil mineralogy applied to forensic sciences] [thesis]. Curitiba, Brazil: Federal University of Paraná, 2019.
23. Simas FNB, Schaefer CEGR, Melo VF, Guerra MBB, Saunders M, Gilkes RJ. Clay-sized minerals in permafrost-affected soils (Cryosols) from King George Island, Antarctica. Clays Clay Miner 2006;54(6):721–36. https://doi.org/10.1346/CCMN.2006.0540607.
24. Mendonça T, Melo VF, Schaefer CEGR, Simas FNB, Michel RFM. Clay mineralogy of gellic soils from the Fildes Peninsula, Maritime Antarctica. Soil Sci Soc Am J 2013;77:1842–51. https://doi.org/10.2136/sssaj2012.0135.
25. Poggere GC, Melo VF, Francelino MR, Schaefer CE, Simas FN. Characterization of products of the early stages of pedogenesis in ornithogenic soil from Maritime Antarctica. Eur J Soil Sci 2016;67(1):70–8. https://doi.org/10.1111/ejss.12307.
26. Mehra OP, Jackson ML. Iron oxide removal from soils and clays by a dithionite-citrate system buffered with sodium bicarbonate. Clays Clay Miner 1960;7(1):317–27. https://doi.org/10.1346/CCMN.1958.0070122.
27. Muggler CC. Polygenetic oxisoil on tertiary surfaces, Minas Gerais, Brazil: soil genesis and landscape development [thesis]. Wageningen, Netherlands: Wageningen University & Research, 1998.
28. Inda Junior AV, Kämpf N. Avaliação de procedimentos de extração dos óxidos de ferro pedogênicos com ditionito-citrato-bicarbonato de sódio [Evaluation of pedogenic iron oxide extraction procedures with sodium dithionite-citrate-bicarbonate]. Rev Bras Ciência do Solo 2003;27(6):1139–47.
29. Correa MM, Ker JC, Barrón V, Fontes MPF, Torrent J, Curi N. Caracterização de óxidos de ferro de solos do ambiente tabuleiros costeiros [Characterising iron oxides from coastal and central plain soils]. Rev Bras Cienc do Solo 2008;32(3):1017–31. https://doi.org/10.1590/S01006832008000300011.
30. Pereira TTC, Ker JC, Schaefer CEGR, Barros NF, Neves JCL, Almeida CC. Genesis of latosols and cambisols developed from pelitic rocks of the Bambui Group, Minas Gerais State – Brazil. Rev Bras Cienc do Solo 2010;34(4):1283–95. https://doi.org/10.1590/S01006832010000400026.
31. Medeiros PSC, Nascimento PC, Inda AV, Silva DS. Caracterização e classificação de solos graníticos em topossequência na região Sul do Brasil [Soil characterisation and classification of granitic soils in toposequence in Southern Brazil]. Ciência Rural 2013;43(7):1210–7. https://doi.org/10.1590/S010384782013000700011.
32. Melo V, Singh B, Schaefer CEGR, Novais RF, Fontes MPF. Chemical and mineralogical properties of kaolinite-rich Brazilian soils. Soil Sci Soc Am J 2001;65(4):1324–33. https://doi.org/10.2136/sssaj2001.6541324x.
33. Oliveira JC, Melo VF, Souza LCP, Rocha HO. Terrain attributes and spatial distribution of soil mineralogical attributes. Geoderma 2014;213:214–25. https://doi.org/10.1016/j.geoderma.2013.08.020.
34. Fitzpatrick RW, Raven MD. How pedology and mineralogy helped solve a double murder case: using forensics to inspire future generations of soil scientists. Soil Horizons 2012;53(5):1–16. https://doi.org/10.2136/sh12-05-0016.
35. StatSoft Inc. STATISTICA (data analysis software system), 2011. Tulsa, OK: StatSoft Inc, 2011.
36. Hammer Ø, Harper DAT, Ryan PD. PAST: paleontological statistics software package for education and data analysis. Palaeontol Electron 2001;4(1):1–9.
37. Jackson ML, Lim CH, Zelazny LW. Oxides, hydroxides, and aluminosilicates. In: Klute A, editor. Methods of soil analysis. Madison, WI: American Society of Agronomy, 1986;101–50.
38. Singh B, Gilkes RJ. Properties of soil kaolinites from south-western Australia. J Soil Sci 1992;43:645–67. https://doi.org/10.1111/j.1365-2389.1992.tb00165.x.
39. Kämpf N, Schwertmann U. Avaliação da estimativa de substituição de Fe por Al em hematitas de solos [Evaluation of Al for Fe substitution in soil hematites]. Rev Bras Ci Solo 1998;22(2):209–13. https://doi.org/10.1590/S0100006831998000200005.
40. Fontes MPF, Weed SB. Iron oxides in selected Brazilian oxisols. Soil Sci Soc Am J 1991;75:1143–99. https://doi.org/10.2136/sssaj1991.03615995005500040040x.
41. Borer P, Hug SJ. Phytoextration of bromine from contaminated soil. J Geochemical Explor 2017;174:21–8. https://doi.org/10.1016/j.gexplo.2016.03.012.
42. Sambatti JA, Costa ACS, Muniz AS, Sengik E, Souza Junior IG, Bigham JM. Relações entre a substituição isomórfica de Fe por Al e as características químicas e mineralógicas de hematitas sintéticas [Relationships between the isomorphic substitution of Fe for Al and the chemical and mineralogical characteristics of synthetic hematites]. Rev Bras Ciência do Solo 2002;26(1):117–24. https://doi.org/10.1590/S0100-06832002000100011.
43. Bigarella J, Salamuni R. Características texturais dos sedimentos da Bacia de Curitiba. [Textural characters of sediments from the Curitiba Basin] Bol da UFPR 1962;7:1–164.
44. García-Rodeja E, Nóvoa JC, Pontevedra X, Martínez-Cortizas A, Buurman P. Aluminium fractionation of European volcanic soils by selective dissolution techniques. Catena 2004;56:155–83 https://doi.org/10.1007/978-3-540-48711-1_25..
45. Valle LAR, Rodrigues SL, Ramos SV, Pereira HS, Amaral DC, et al. Beneficial use of a by-product from the phosphate fertilizer industry in tropical soils: effects on soil properties and maize and soybean growth. J Clean Prod 2016;112:113–20. https://doi.org/10.1016/j.jclepro.2015.07.037.
46. Schöning I, Knicker H, Kögel-knabner I. Intimate association between O/N-alkyl carbon and iron oxides in clay fractions of forest soils. Org Geochem 2005;36:1378–90. https://doi.org/10.1016/j.orggeochem.2005.06.005.
47. Hanke D. Gênese, interação organo-mineral e estabilidade de agregados de solos desenvolvidos de basalto [Genesis, organo-mineral interaction and stability of soil aggregates developed from basalt] [thesis]. Curitiba, Brazil: Federal University of Paraná, 2012.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Elements extracted by ammonium oxalate (AO) in the sequential extraction of the clay + silt fraction.

Table S2. Elements extracted by ditionite-citrate-bicarbonate (DCB) in the sequential extraction of the clay + silt fraction.

Table S3. Elements extracted by NaOH 0.5 mol L^{-1} in the sequential extraction of the clay + silt fraction.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14548
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Colby E. Ott,[1] M.S.; Kourtney A. Dalzell,[1] B.S.; Pedro José Calderón-Arce,[2] M.S.; Ana Lorena Alvarado-Gámez,[2] Ph.D.; Tatiana Trejos,[1] Ph.D.; and Luis E. Arroyo,[1] Ph.D.

Evaluation of the Simultaneous Analysis of Organic and Inorganic Gunshot Residues Within a Large Population Data Set Using Electrochemical Sensors*,†

ABSTRACT: The increasing demand for rapid methods to identify both inorganic and organic gunshot residues (IGSR and OGSR) makes electrochemical methods, an attractive screening tool to modernize current practice. Our research group has previously demonstrated that electrochemical screening of GSR samples delivers a simple, inexpensive, and sensitive analytical solution that is capable of detecting IGSR and OGSR in less than 10 min per sample. In this study, we expand our previous work by increasing the number of GSR markers and applying machine learning classifiers to the interpretation of a larger population data set. Utilizing bare screen-printed carbon electrodes, the detection and resolution of seven markers (IGSR; lead, antimony, and copper, and OGSR; nitroglycerin, 2,4-dinitrotoluene, diphenylamine, and ethyl centralite) was achieved with limits of detection (LODs) below 1 μg/mL. A large population data set was obtained from 395 authentic shooter samples and 350 background samples. Various statistical methods and machine learning algorithms, including critical thresholds (CT), naïve Bayes (NB), logistic regression (LR), and neural networks (NN), were utilized to calculate the performance and error rates. Neural networks proved to be the best predictor when assessing the dichotomous question of detection of GSR on the hands of shooter versus nonshooter groups. Accuracies for the studied population were 81.8 % (CT), 88.1% (NB), 94.7% (LR), and 95.4% (NN), respectively. The ability to detect both IGSR and OGSR simultaneously provides a selective testing platform for gunshot residues that can provide a powerful field-testing technique and assist with decisions in case management.

KEYWORDS: gunshot residues, GSR, electrochemistry, machine learning algorithms, simultaneous detection, screen-printed carbon electrodes, SPCE, square-wave anodic stripping voltammetry, inorganic, organic

Gun violence in America continues to devastate communities across the United States, resulting in approximately 100 firearm-related deaths each day, totaling approximately 39,000 deaths in 2019 (1,2). The forensic science and law enforcement communities require rapid and efficient methods when dealing with firearm-related incidents. The analysis of gunshot residues (GSRs) can provide investigative leads, identify potential shooters, and refute or corroborate statements (3–5).

The discharge of a firearm results in the ejection of a cloud of compounds out the barrel, chamber, and other openings of the firearm. This gaseous cloud contains various inorganic and organic compounds typical of GSR (3–5). Inorganic gunshot residues (IGSR) include elements such as lead (Pb), antimony (Sb), barium (Ba), copper (Cu), and aluminum (Al) within the primer, projectile, and cartridge casings (5–7).

The current standard practice for the identification of IGSR uses scanning electron microscopy energy-dispersive X-ray spectroscopy (SEM-EDS) (7–9). The ASTM E1588-17 method assesses morphology followed by elemental analysis (7). This method is the gold standard, providing confirmatory analysis of distinctive morphological features of GSR particles ranging from 0.5 μm to 10 μm and elemental composition on a single particle. Nonetheless, SEM-EDS is bulky and time-consuming, limiting its use to the forensic laboratory setting. The management of cases from crime scene to the courtroom could benefit from the implementation of reliable screening tests that are capable of quick on-site detection for better selection of samples prior to their submission to the laboratory and use as evidence for preliminary hearings to speed up the legal process.

Screening methods capable of dual detection of inorganic and organic markers benefit from improved selectivity despite the absence of morphology identification. Common organic gunshot residues (OGSR) include nitroglycerin (NG), diphenylamine (DPA), 2,4-dinitrotoluene (DNT), and ethyl centralite (EC). OGSR compounds are found in the gunpowder as stabilizers and explosive compounds (5–7). Current methods of testing are

[1] Department of Forensic and Investigative Science, West Virginia University, Morgantown, WV, 26506.

[2] Centro de Electroquímica y Energía Química, CELEQ, Universidad de Costa Rica, San Pedro de Montes de Oca, San José, 11501-2060, Costa Rica.

Corresponding author: Luis E. Arroyo, Ph.D. E-mail: luis.arroyo@mail.wvu.edu

*Presented in part at the 72nd Annual Scientific Meeting of the American Academy of Forensic Sciencesin, February 17-22, 2020, in Anaheim, CA.

†This project is sponsored by Award No. 2018-DU-BX-0186 from the National Institute of Justice to West Virginia University. The opinions, findings, and conclusions are those of the authors and do not necessarily reflect those of the Department of Justice.

Received 8 May 2020; and in revised form 19 June 2020; accepted 28 July 2020.

limited to either inorganic markers or organic markers and generally use instrumental techniques such as mass spectrometry and gas chromatography, which are expensive, time-consuming, and difficult to make portable (10). Therefore, the development and assessment of rapid, portable, and sensitive techniques providing simultaneous analysis of IGSR and OGSR are of analytical importance to the forensic science community.

Electrochemistry is a mature analytical technique offering sensitive and versatile testing platforms for a variety of materials and compounds in fields such as biochemistry, thermodynamics, and environmental chemistry (11). The analysis of inorganic metals has been demonstrated previously, allowing for the detection of IGSR (5,12–17). In addition, many organic components are electroactive species. It has been previously demonstrated by several research groups that the simultaneous detection of IGSR and OGSR via electrochemical methods is a viable solution for screening GSR samples (5,16–18) as Goudsmits et. al stated that simultaneous detection via other methods has been a challenge for the field (19). Its speed of analysis and portability makes electrochemistry ideal for field testing and strategic triage approaches at the laboratory. Moreover, the nondestructive nature of electrochemical methods makes it feasible for further confirmation by other techniques when needed.

This work aims to demonstrate the importance of a simultaneous screening of GSR samples for both IGSR and OGSR through the analysis of a large population study to provide statistical interpretations of the strength of the electrochemical method to answer the dichotomous question of presence or absence of GSR and its implications with recent firing activities. Comparison of several populations including samples collected from the hands of shooters and nonshooters is presented herein along with statistical analysis.

Materials and Methods

Reagents and Standards

Sodium acetate anhydrous, glacial acetic acid (HPLC grade), and acetonitrile (Optima®) were purchased from Fisher Scientific (Fair Lawn, NJ). 1,3-Diethyl-1,3-diphenylurea 99% (ethyl centralite) and Antimony (III) Oxide (Reagent Plus®) powder 5-micron 99% were purchased from Sigma-Aldrich (St. Louis, MO). Diphenylamine standard was acquired from SPEX Certiprep® (Metuchen, NJ). Lead, copper, and antimony standards were obtained from Ultra Scientific® (Kingstown, RI). Nitroglycerin and 2,4-dinitrotoluene standards were purchased from AccuStandard® (New Haven, CT). Ultrapure, 18.2 MΩ water was obtained using a Millipore Direct-Q® UV water purification system (Billerica, MA). Nitrogen was obtained from Matheson Tri-Gas, Inc. (Irving, TX).

Electrodes and Instrumentation

Screen-printed carbon electrodes (SPCEs) model type DRP-110 were purchased from Metrohm DropSens, USA. Electrochemical measurements were carried out using an Autolab PGSTAT128N potentiostat along with the NOVA software, version 2.1.4, from Metrohm USA, Inc. A Metler Toledo FiveEasy (Columbus, OH) pH meter was used for determining pH values.

Sample Collection

Collection of shooter samples along with background samples was performed following Institutional Review Board (IRB) procedure # 1506706336 for the collection of samples from the general public. Background hand samples were collected from volunteers who stated that they had not been in contact with firearms or fireworks in the past 24 h. Collection was performed on the WVU campus and at the World Scout Jamboree at a station located several miles away from the shooting ranges. A total of 350 background samples were collected.

Authentic shooter samples consisted of samples collected from the hands of individuals who had recently fired a gun at one of two locations: the West Virginia University (WVU) Ballistics Laboratory or the World Scout Jamboree at the Summit Bechtel Reserve, WV. The shooter's hands were sampled either immediately after shooting (indoor range) or less than one hour after the discharge of the firearm (outdoor ranges). In both locations, the sampling was conducted using aluminum SEM-EDS stubs coated with adhesive carbon tape purchased from Ted Pella, Inc. (Redding, CA). The back and palms of both hands were sampled as described previously (5) following the standard practice sampling procedure for GSR collection in the United States. This involves the use of four collection stubs, where one stub was used for collection from each of the following: right palm, right back, left palm, and left back of the hand. Several firearms and ammunitions were utilized in the study. Table 1 provides a summary of the firearms and ammunitions used in this study. A total of 395 authentic shooter samples were collected.

Sample Preparation

All measurements were conducted in 0.1 M acetate buffer pH 4.5 on the surface of the bare screen-printed carbon electrodes. The left palm shooter samples were utilized for analysis. Figure 1 demonstrates the sample preparation procedure. The adhesive stub surface was first washed by placing a 50 μL drop of acetate buffer on a portion of the stub surface and agitated by pipetting the drop several times onto the surface of the carbon stub to improve extraction. This drop was then removed and placed in a centrifuge tube. A 50 μL drop of acetonitrile was then used to wash the stub surface in the same spot as the buffer wash following the same procedure. The drop of acetonitrile was removed and placed in a separate microfuge tube. The organic fraction was then evaporated under a constant stream of nitrogen in a ductless hood. The buffer fraction was then used to reconstitute the dried organic fraction and vortexed prior to analysis through placing the 50 μL portion on the electrode, covering the working, auxiliary, and reference electrode, and measured via square-wave anodic stripping voltammetry. The extraction procedure from a portion of the stub allows for the collection stub to undergo future analysis via SEM-EDS and ensures a representative sample as a result of the collection procedure and the random distribution of GSR particles.

Square-Wave Anodic Stripping Voltammetric (SWASV) Method

Analysis of samples was achieved first through the application of −0.95 V potential, used as a preconcentration step to deposit the analytes in their reduced form at the surface of the working electrode. Following this preconcentration step, a square-wave procedure was used to sweep the potential between −1.0 and +1.2 V in order to strip analytes from the surface and obtain the oxidation peaks of the analytes of interest. The parameters used for the SWASV analysis can be found in Table 2. This procedure was utilized for the construction of calibration curves for each analyte of interest in similar fashion using drop analysis. Several quality control (QC) samples were analyzed at the start

TABLE 1—*Summary of firearms and ammunitions utilized in collection of authentic GSR samples.*

Firearm	Type	Ammunition
Springfield XD9, 9 mm pistol	Leaded	Blazer factory standard 115 grain copper full metal jacket
		Winchester 231 (4 grains) 115 or 124 grain total metal jacket, Remington primer
		Winchester 231 (4 grains) 115 or 124 grain total metal jacket, Winchester primer
		Manufacturer loaded Remington Luger 115 grain full metal jacket
		Alliant Unique powder
		Alliant Bullseye powder
		Accurate No. 2 powder
		Hodgdon HS-6 powder
		Hodgdon HP-38 powder
		IMR 700-X powder
		IMR PB powder
		Vectan Ba-9 powder
		Winchester WSF powder
		Alliant Blue Dot powder (gray and blue)
Sig Sauer P320, 9 mm pistol	Lead-free primer	Federal Syntech training 9 mm match 124 and 127 grain total synthetic jacket
Ruger Mark IV, 22 LR pistol	Leaded	Federal .22 caliber long rifle 40 grain solid
Taurus Model 905, 9-mm revolver	Leaded	Winchester 231 (4 grains) 115 or 124 grain total metal jacket, Remington primer
Taurus Model 608, .357 Magnum Revolver	Leaded	American Eagle Federal, 38 special 130 grain full metal jacket

and end of each analysis period, including blank buffer, negative stub control (a carbon stub not used or exposed during sampling), IGSR standard mix, and OGSR standard mix.

Data Analysis

Sample response was assessed as the current area of each peak at potentials corresponding to the analytes of interest. Critical thresholds were calculated for each analyte based on the average current response for the analysis of the background (nonshooter) samples. Critical thresholds (CT) were calculated as the average current response plus three times the standard deviation of the current response as seen in Equation 1.

$$\mathrm{CT} = \mathrm{I}_{\text{background skin avg.}} + (3*\mathrm{sd}) \qquad (1)$$

Outliers based on the generation of box plots from the background data set were removed prior to calculation, and all samples demonstrating a zero response were replaced by the lowest quartile value for each analyte. An analyte was considered present if the measured response was above the critical threshold, and absent if at or below that experimental limit. Threshold values were used for the assessment of shooter samples to determine performance measures including true positives (sensitivity), true negatives (specificity), false positives, and false negatives as outlined previously (5). Further, machine learning algorithms including naïve Bayes, logistic regression, and neural networks were used to assess the performance of the method as detailed by our group (20) for classification of shooter from nonshooter samples. To this end, known shooter and known nonshooter samples were identified as such and a random splitting of the data was performed to use 60% of the data for training, 20% for validation, and 20% for testing. JMP Pro statistical software version 14.0.0 was used for statistical analysis of the data using the above-mentioned machine learning algorithms. The neural network consisted of a single hidden layer with a TanH activation sigmoid identity radial of 3 with a learning rate of 0.1 and a squared penalty method.

Results and Discussion

While forensic science fields such as the analysis of seized drugs, explosives detection, and trace evidence boast powerful field-testing tools, the discipline of gunshot residue analysis currently lacks screening techniques for use in the field. However, incorporation of reliable screening methods can result in

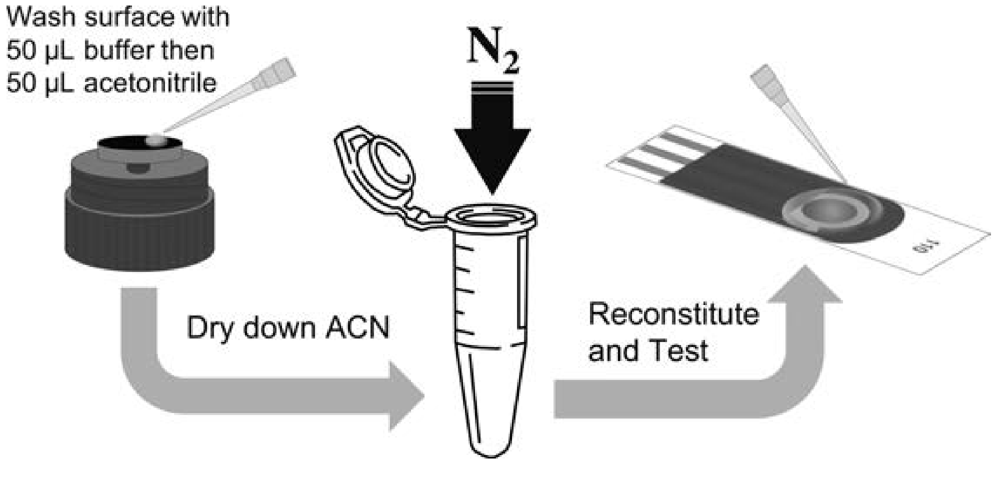


FIG. 1—*GSR extraction procedure from the sampling stub. [Color figure can be viewed at wileyonlinelibrary.com]*

TABLE 2—*SWASV parameters for analysis of GSR with SPCEs.*

Bare Carbon	
Parameter	Value
Technique	SWASV
Deposition potential	−0.95 V
Deposition time	120 s
Start potential	−1.0 V
End potential	1.2 V
Step	0.004 V
Modulation amplitude	0.025 V
Frequency	8 Hz
Interval time	0.125 s

opportunities for more efficient case management. For instance, law enforcement agencies from over ninety cities within the United States utilize acoustic detection methods to allow officers to arrive at the scene of a firearm-related incident in a matter of minutes. Despite this technology, current detection methods cannot match this response speed causing the investigators to wait weeks to months for laboratory results. In some cases, this also implies that suspects may be spending unnecessary time in jail.

As a result, the main purpose of this study was to evaluate the feasibility of modern disposable electrochemical sensors as fast and reliable screening tools for field detection of gunshot residues. The utility of the method was assessed in two stages. First, the optimization and evaluation of analytical performance for the simultaneous identification of IGSR and OGSR was accomplished and is discussed in the *Analytical measures* section. Second, the assessment of the occurrence of the monitored GSR markers within the background populations relative to the shooter populations was conducted. This second stage involved the validation of the method by estimation of error rates and accuracy of the electrochemical method and is discussed in the *Critical threshold* and *Machine learning* sections.

The adoption of this method for use in the field was evaluated through three main aspects: (i) practicality of the method and speed of analysis in less than 10 min, (ii) demonstrable sensitivity and selectivity in large sets of casework-like authentic samples to demonstrate fit-for-purpose, and (iii) overall selectivity and accuracy greater than 80% to provide reliable screening testing.

Analytical Measures of the Method

The electroanalytical method proposed for the assessment of GSR markers required an initial evaluation of critical variables that play a role in the identification of the individual elements and compounds. For instance, peak potential is probably the main variable that provides qualitative information, the current obtained under specific buffer conditions, and pH is representative of the analytical response of the materials in solution. Therefore, individual assessment of each analyte was conducted to determine the electrochemical potential of interest, linear range, limits of detection, and limits of quantitation. Independent calibration solutions were prepared via serial dilution from a stock and measured as individual 50 μL drops deposited onto the electrode surface. These solutions were used to build calibration curves of the target analytes. The limits of detection were calculated as three times the standard deviation of the lowest calibrator response divided by the average slope. Table 3 demonstrates the calculated analytical measures for each analyte of interest measured on the carbon electrode.

Calibration curves demonstrated acceptable linearity and limits of detection below 1 μg/mL. In general, the limits of detection achieved by this electrochemical method are comparable to other methods used for detection of IGSR and OGSR. For instance, mass spectrometric methods typically report OGSR compounds in the low nanogram to microgram range or low part-per-billion to part-per-million range depending on instrumentation (21). IGSR detection limits for methods such as SEM-EDS are typically low micrograms, while mass spectrometry methods range from low nanogram/part-per-billion to microgram/low part-per-million, and in the sub-to-low nanogram level for laser-induced breakdown spectroscopy (LIBS) (21). Of these, only SEM-EDS and TOF-SIMS are capable of single particle analysis and morphology, which is a desirable feature. Although electrochemistry cannot offer morphological information, the simultaneous detection of IGSR and OGSR adds significant confidence to the results. Further, despite inferior OGSR detection limits by electrochemistry than LC/MS, this method has been shown to be effective in the detection of GSR at levels typically encountered in casework-like samples, as supported by the true positive rates in this study and previous work. Table 3 shows copper as the most sensitive element under the current measuring conditions, but overall the square-wave method demonstrated good performance for metal detection and for the organic compounds. The initial deposition step was beneficial for the reduction of all metals to improve the sensitivity of the method. However, due to possible differences in the electrochemical mechanisms of some of the organic species, the sensitivities varied. This can be seen in the case of ethyl centralite (EC) and diphenylamine (DPA) that showed LODs in the range of 0.4 μg/mL in contrast to NG and 2,4-DNT that presented lower values. However, in general, the low detection capability obtained for all OGSRs is considered an advantage and a feature that is highly desirable due to the low concentration of these compounds that may be present in the authentic samples. Furthermore, this work demonstrates an expanded panel of analytes for GSR analysis with improved limits of detection compared to our previous work (5), which will allow for increased reliability in the analysis when multiple GSR markers are present.

Figure 2 demonstrates standard mixtures of the target GSR markers along with the signal from the buffer blank. Peak potential separation was observed for the majority of analytes; however, resolution problems were observed for NG and DPA, as well as Cu and Sb, depending on their relative concentration in the sample. When present together in a mixture, the NG and DPA oxidation peaks presented a slight overlap. Despite this, the assignment of the peak shape and their position differ enough offering an opportunity for their detectability. More importantly, peak potential is a key indicator of the presence of nitroglycerin as the peak will be greater than 0.50 V.

The copper and antimony potentials were also of interest. As mentioned before, the method was extremely sensitive for the analysis of copper. As such, the copper peak can obscure the oxidation peak of antimony when the concentration of antimony is small, but the peak of copper was still easily observed when the concentration of antimony was 40 times that of copper. However, the analysis of antimony proved difficult as the standard was made from a stock solution containing trace nitric acid and trace tartaric acid. Tartaric acid is commonly used in stock commercial solutions to add stability to antimony. It was shown that the addition of tartaric acid hinders the electro-oxidation of antimony (Fig. 3). This can also be considered when observing the lower values for linearity and LOD compared to the other metals.

TABLE 3—*Analytical performance for the detection of IGSR and OGSR analytes.*

	Potential (V)	Linear Range (μg/mL)	R^2	Repeatability (%RSD, $n = 3$)	LOD (μg/mL)
IGSR					
Lead	−0.784 ± 0.035	0.10– 2.0	0.999	4.4	0.055 ± 0.01
Antimony	−0.401 ± 0.027	0.75–7.5	0.986	10	0.183 ± 0.07
Copper	−0.292 ± 0.053	0.05–1.0	0.990	2.3	0.012 ± 0.001
OGSR					
2,4-Dinitrotoluene	−0.132 ± 0.032	1.0–20	0.982	5.6	0.200 ± 0.03
Diphenylamine	0.406 ± 0.018	1.0–8.0	0.987	6.2	0.462 ± 0.06
Nitroglycerin	0.509 ± 0.010	0.50–8.0	0.998	10	0.147 ± 0.08
Ethyl centralite	1.03 ± 0.045	0.50–8.0	0.998	8.0	0.450 ± 0.09

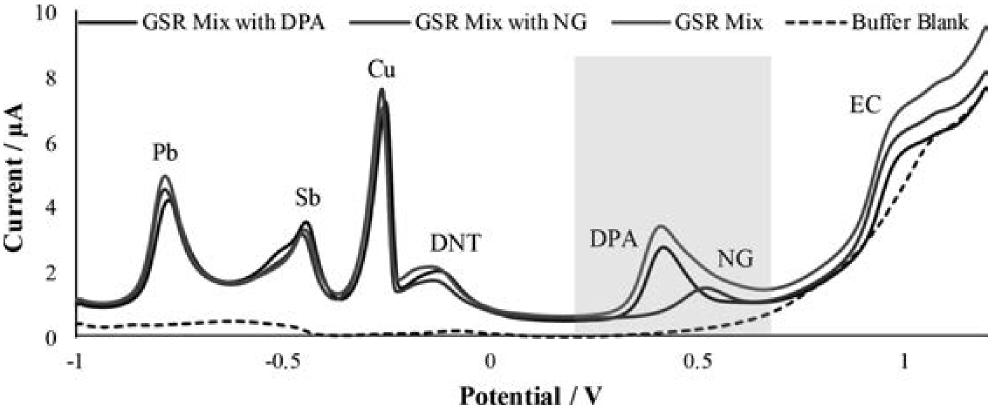


FIG. 2—*Standard mixtures of GSR markers analyzed using SWASV on bare carbon SPCEs. [Color figure can be viewed at wileyonlinelibrary.com]*

It was also found that the presence of chloride ion in the solution resulted in shifting of the potential peaks due to influence with the pseudo-silver reference electrode (Fig. 4), causing issues when attempting to utilize powdered forms of antimony due to solubility issues except when using hydrochloric acid.

The presence of the chloride ion demonstrated potential shifts not only on the antimony oxidation peak but also on the peaks for the other GSR markers. When preparing standard mixtures of GSR using antimony dissolved in hydrochloric acid, the presence of the chloride ion shifted the peak potentials more anodic by approximately 100 mV despite a large dilution of the stock antimony that was originally at 10,000 μg/mL. These potential shifts can be seen in Fig. 5. For this reason, the antimony standard containing trace nitric and tartaric acid was used for reporting the analytical measures, despite higher LOD. It is worth noting that the presence of chloride ion (in high concentrations like those tested) and tartaric acid are not anticipated on skin samples, and therefore, it can cause false-negative identifications if the examiner is not aware of the possible matrix effects caused by the medium of some of the standards. It is important for the analyst to carefully select reagents and standards that do not cause interference problems during electrochemical measurements.

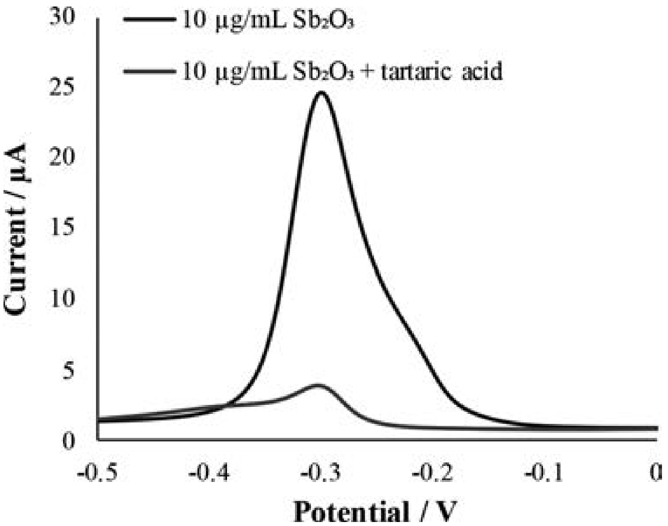


FIG. 3—*Effect of tartaric acid on the current response of 10 μg/mL antimony in acetate buffer using SWASV on bare SPCEs. [Color figure can be viewed at wileyonlinelibrary.com]*

Critical Thresholds in Authentic Samples

Authentic samples were split into three sample sets for analysis and comparison: samples collected at WVU (150 background and 175 shooter), samples collected at the World Scout Jamboree (200 background and 220 shooter), and combined samples from both locations for a total of 350 background and 395 shooter samples. Critical threshold values were calculated using the backgrounds from each sample set individually. Performance measures were then calculated for each sample set. The WVU and WSJ sample sets were then reassessed using the critical thresholds calculated for the combined sample set to determine any difference from the critical thresholds of each individual set. Separation of the sample sets was performed due to collection environment, where larger amounts of dust and dirt were suspected within the WSJ population. Also, the ammunition used at the WVU location was standard leaded, while the Jamboree set consisted of a mixture of leaded and lead-free. Therefore, this provided an opportunity to assess any differences that could be

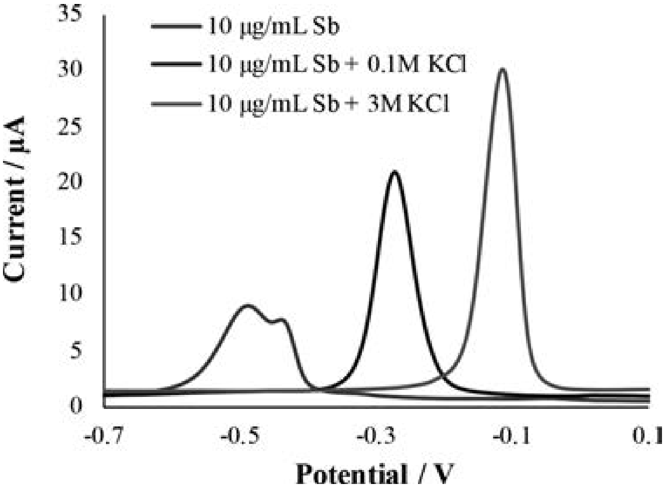


FIG. 4—*Effect of chloride ion on the oxidation potential of 10 µg/mL antimony using SWASV on bare SPCEs. [Color figure can be viewed at wileyonlinelibrary.com]*

present. Analysis of authentic samples revealed that the most prevalent residues detected in this data set using bare carbon were lead, copper, and nitroglycerin. Some samples demonstrated the presence of antimony, ethyl centralite, diphenylamine, and 2,4-dinitrotoluene; however, these were not used for statistical analysis of the authentic samples due to their limited presence in this data set (less than approximately 5% of samples). The importance of this method for the detection of a large panel of GSR markers cannot be underestimated as the combination of both inorganic and organic GSR markers adds increased reliability to the analysis. The sole presence of inorganic markers such as lead, barium, antimony, and copper can result in false-positive cases due to their prevalence in materials in certain industries including automotive, construction, electrical, paint, and plumbing (6,17,19). However, organic markers generally represent lower levels of false positives as their prevalence within communities is limited. Nitroglycerin is especially useful as there have been no reports of widely finding nitroglycerin naturally within the environment, except particular circumstances where exposure to explosives and certain pharmaceutical products is expected (6). The same is true of ethyl centralite and both are considered to be highly correlated with GSR (19). Moreover, the combined incidence of IGSR and OGSR markers was not observed on the background samples, stressing the increased confidence when both inorganic and organic constituents are considered in evidence interpretation.

Therefore, the unique capability of electrochemical analysis for simultaneous detection of both inorganic and organic GSR markers is especially desirable. Moreover, although three of the more prevalent markers in this population were chosen to study the performance rates of the method, this does not prevent the use of the other markers detected using this methodology for other population sets, as the addition of any of the OGSR markers in this study provides increased reliability and lower false-positive results. This is particularly important when assessed along with the infrequency of observing combined IGSR/OGSR markers in the background population. Several authentic shooter samples along with an example background population sample are displayed in Fig. 6 with the buffer blank and negative control.

Critical threshold values for lead, copper, and nitroglycerin differed between the sample sets. A clear difference was noticeable between the WVU sample set and WSJ sample set as seen in Table 4 for the calculated threshold values. As predicted, critical threshold values for the WSJ set were higher than WVU, representing higher levels of some compounds in the background samples. This could have resulted from several factors including increased debris on the hands of participants and their participation in many different activities at the camp ranging from robotics to archery to swimming. The background population for WSJ may have also been higher for nitroglycerin due to potential use of fireworks at the camp or high level of interaction between camp participants resulting in unintentional transfer between shooters and nonshooters. However, it is important to note that the variation could also be due to the WSJ sample set being larger than the WVU sample set.

Critical thresholds were also displayed graphically to demonstrate the difference between the background samples and the authentic shooter samples. Each compound demonstrated large differences in the peak current for shooter samples compared to that of the background samples as evidenced by the large distance between the critical threshold line and the peak current of each shooter sample. An example of this exploratory data analysis can be seen in Fig. 7 for a subset of the data obtained from the WVU data set. A subset was displayed for visual purposes, and the full graphical thresholds can be found in the Figures S1-S9.

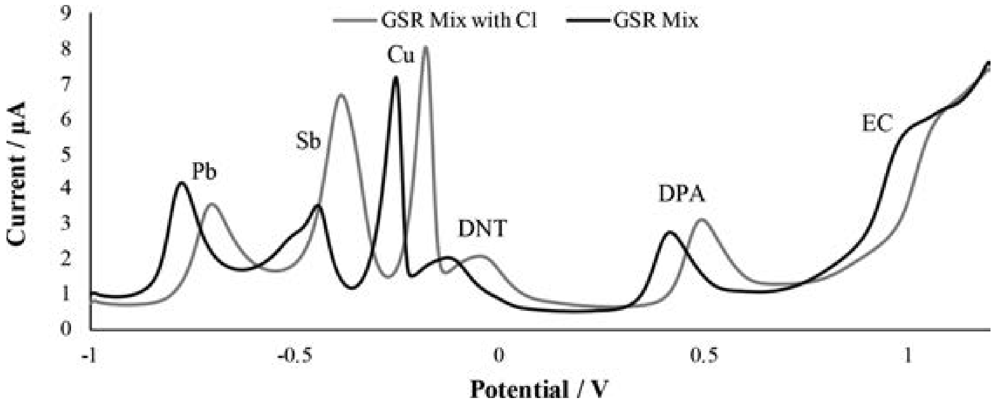


FIG. 5—*Potential shift of GSR marker oxidation peaks in the presence of chloride ion from the diluted hydrochloric acid. [Color figure can be viewed at wileyonlinelibrary.com]*

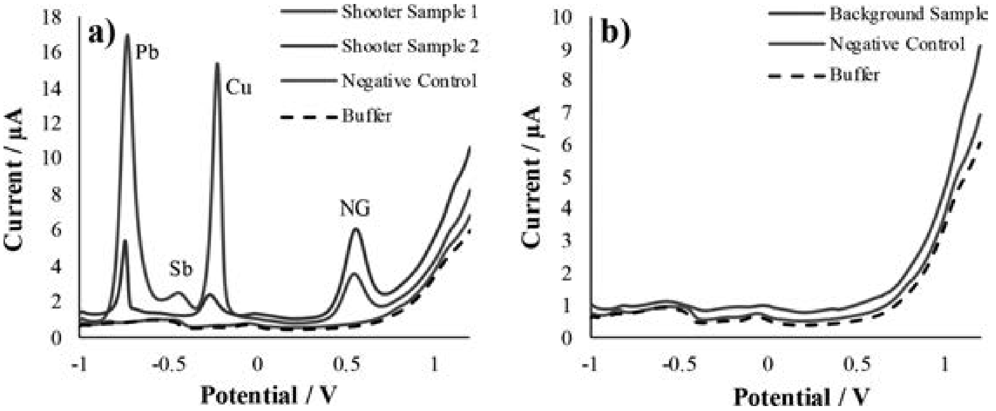


FIG. 6—*Voltammograms demonstrating (a) authentic shooter sample current response and (b) authentic population background current response along with controls. [Color figure can be viewed at wileyonlinelibrary.com]*

Table 5 demonstrates the performance measures for the sample sets that were calculated as previously described by our group for false positives, false negatives, true positives, true negatives, and overall accuracy (5). For this analysis, samples were considered to be positive for gunshot residue when at least two of the markers had currents above their respective critical threshold. If shooter samples demonstrated only a single marker or no markers above the threshold, this sample was considered to be a false negative. Performance was improved for the WSJ sample set (83.8% accuracy) compared to the WVU sample set (79.7%). Using the combined set critical thresholds, the performance was intermediate to the other two sample sets as expected. In all cases, the false-negative rate was lower than 38%. Also, of great interest to forensic science is the false-positive rate, as this rate could result in the incarceration of an innocent person. All sample sets demonstrated false-positive rates lower than 2% when using the critical thresholds. Furthermore, using the critical threshold method, Pb, Cu, and NG were seen in the background population in 4.9%, 6.3%, and 4.3% of the samples, respectively. Of particular interest, although false positives were low, NG was never seen in conjunction with another GSR marker, and therefore did not lead to a false-positive result (the false positives were a result of IGSR only). This is in contrast to the shooter samples where Pb, Cu, and NG were seen in 93.7%, 57.0%, and 43.5% of the samples, respectively. In other words, the chance of detecting NG or the other markers was more than 10 times higher for authentic shooter samples compared to the background samples. It is also important to note that in authentic samples, NG and Cu were rarely found without lead (3.0%). The performance of the critical threshold method was shown to provide excellent results for a screening method. However, machine learning algorithms have better capabilities to identify new patterns in data than the exploratory method of critical threshold and, therefore, were utilized to improve data classification.

TABLE 4—*Critical threshold values for each sample set for the GSR compounds of interest.*

	Critical Threshold Value/A × V		
	Lead	Copper	Nitroglycerin
WVU	1.80×10^{-8}	3.84×10^{-8}	2.71×10^{-9}
WSJ	$1.35 \times \times 10^{-8}$	2.87×10^{-8}	4.83×10^{-9}
All Samples	1.59×10^{-8}	3.33×10^{-8}	4.28×10^{-9}

Machine Learning Outcomes

Several machine learning algorithms were used including naïve Bayes, logistic regression, and neural networks as previously detailed and described by our group (20). The naïve Bayes approach assumes that the variables used for comparison are independent, which is rarely true. This algorithm compares the binary classification of shooter sample versus background sample to the current response for each GSR marker to determine the probability of an unknown sample falling into one of the two classification categories (20,22,23). The logistic regression approach seeks to classify data into two distinct outcomes. Again, those outcomes are simply shooter versus nonshooter. Logistic regression determines how likely it is to observe a particular current response given that the sample is a shooter or a nonshooter class. From this point, a decision boundary is created for analysis of unknown samples (20,24). Lastly, neural networks provide a powerful method for classification due to the ability of the algorithm to adapt to the data set and provide improved discrimination between groups without making general statistical assumptions common in other methods. Neural networks seek to emulate biological networks and are generally said to "learn," meaning that the process must train on known data to determine the optimal values for the algorithm (20). The neural network utilized for this work was a single hidden layer with 3 nodes and 3 input variables (the peak areas for Pb, Cu, and NG), which can be seen in Figure S10.

Method performance rates generally improved from naïve Bayes, to logistic regression, to neural networks, which offered the best accuracy of the tested methods. Comparison of the performance measures for the various machine learning algorithms across all three sample sets can be seen in Table 6. Performance for the WSJ sample set appeared to be more successful than for the WVU sample set. This trend was true throughout the different machine learning algorithms and for the critical threshold method described above. This was expected due to the environment in which the WSJ samples were collected. Since the participants were present in an outdoor shooting range with many different stations, it was suspected that the signals from the hands of shooters in the sample set would present higher amounts than those collected in a more controlled environment. This is clearly evident upon analysis of the performance rates across all methods. Nevertheless, the WSJ samples were also expected to have more possible contamination from the background environment, which did not appear to interfere with

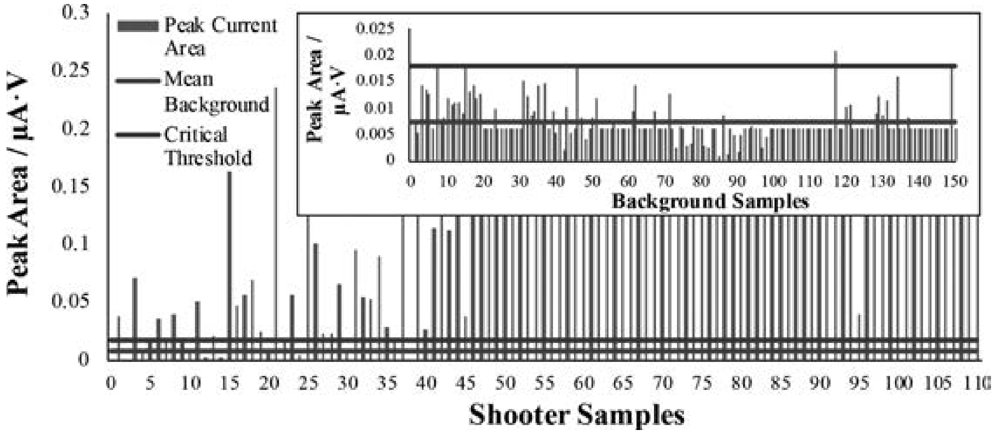


FIG. 7—*Demonstration of the calculated critical threshold showing the current response for shooter samples compared to the mean background and critical threshold for nitroglycerin for a subset of the background samples and the shooter samples. [Color figure can be viewed at wileyonlinelibrary.com]*

TABLE 5—*Performance rates for method assessment utilizing calculated critical threshold values for the various sample sets.*

Performance Measure	WVU Sample Set/%	WSJ Sample Set/%	Combined Sample Sets/%
False Positive	0	1.5	0.6
False Negative	37.7	29.6	33.7
True Negative	100	98.5	99.4
True Positive	62.3	70.5	66.3
Accuracy	79.7	83.8	81.8

TABLE 6—*Performance rates using machine learning algorithms for the sample sets.*

	WVU Sample Set/%	WSJ Sample Set/%	Combined Sample Sets/%
Naïve Bayes			
Performance Measure			
False Positive	6.45	0	2.67
False Negative	41.9	13.6	21.1
True Negative	93.6	100	97.3
True Positive	58.1	86.4	79.0
Accuracy	75.8	93.3	88.1
Logistic Regression			
Performance Measure			
False Positive	3.23	0	2.67
False Negative	41.9	6.82	7.89
True Negative	96.8	100	97.3
True Positive	58.1	93.2	92.1
Accuracy	77.4	96.6	94.7
Neural Network			
Performance Measure			
False Positive	0	6.67	2.67
False Negative	16.1	2.27	6.58
True Negative	100	93.3	97.3
True Positive	83.9	97.7	93.4
Accuracy	91.9	95.5	95.4

current classification. However, it is interesting to note that the use of critical thresholds provided slightly better classification than naïve Bayes and logistic regression for the WVU sample set, 79.7% accuracy compared with 75.8% and 77.4%, respectively. Logistic regression was demonstrated to be a strong classification method when assessed in comparison with neural networks, as the performance rates were similar for the WSJ and combined sample sets, approximately 95% for each, but logistic regression was superior for the WSJ set since it produced no false positives. Despite these differences, it was clear that neural networks present an improvement in the ability to differentiate between shooters and nonshooters. The accuracy of the method was greater than 90% with very low false-positive and false-negative rates, making this a desirable outcome for screening of samples from an individual of interest in an investigation. Further, the large size of this population data set provides statistical strength to the results offered. The reliability and power of the test was estimated through use of 10-fold cross-validation. The 10-fold cross-validation was repeated 10 times on the combined data set, and the performance measures were compared. No statistical difference was seen between the training and validation sets, demonstrating accuracy of the fit model and that a sufficiently large population for statistical analysis was present.

Conclusion

The use of electrochemical methods for the analysis of gunshot residues proved to be a simple, rapid, and sensitive approach for detecting both inorganic and organic GSR markers. One advantage of this analysis method is that it is not exhaustive and nondestructive, allowing further confirmation by SEM-EDS, as reported by our group (5). Simultaneous analysis of both OGSR and IGSR was achieved using a simple extraction protocol with electrochemical analysis in less than 10 minutes. An expanded testing panel consisting of lead, antimony, copper, nitroglycerin, ethyl centralite, 2,4-dinitrotoluene, and diphenylamine was presented. This provides the ability to increase the reliability of the analysis due to the presence of both inorganic and organic markers. While GSR-like inorganic components can be common in some occupational environments, the prevalence of organic markers is more unusual. Furthermore, the lower prevalence of combined IGSR and OGSR in the background populations improves confidence in the results.

This work represents the largest GSR population study conducted to the authors' knowledge on authentic shooter and background samples using electrochemistry. The confidence and usefulness of trace materials rely on relevant population studies to provide information about the rarity of evidence within the relevant population. As such, large population studies and databases are typically required for interpretation of the probative value of finding, or not finding, GSR on an individual of interest. This study provides experimental foundations to formulate conclusions based on identification of IGSR and OGSR markers and their relative occurrence versus the background population, rather than solely analyte identification. Further, machine learning algorithms serve as useful tools for the evaluation of evidence with limited input from the analyst. The outputs of the predictive models can be used to provide probabilistic interpretation of the evidence and improved discrimination and accuracy when compared to traditional analyst-based methods.

As demonstrated by the large sample sets consisting of 395 authentic shooter samples and 350 background samples, the electrochemical method is capable of providing high accuracy screening results using a simple critical threshold method (>80% accuracy), as well as machine learning algorithms including neural networks (>95% accuracy). Assessment of this promising screening application against a large population set provides a basis for the future application of this method in the field of forensic firearm-related investigations. Faster and more sensitive approaches will improve investigative response and justify the use of more costly and time-consuming methods on positive samples. Also, electrochemistry can be used as a triage tool at crime scenes and in the laboratory to help reduce backlogs and to open opportunities for processing alternative matrices, other than hands, that otherwise are not typically tested because of time constraints. Further, better informed decisions can be made on-site and potential use of evidence for preliminary hearings can reduce unnecessary jail time. Finally, the detection of OGSR will provide another layer of reliability to the test that will aid in preventing false-positive results.

This simple electrochemical approach provides a GSR screening method with high accuracy for use in the laboratory or in the field for on-site response. Future work will focus on the incorporation of more GSR markers in the testing and analysis of samples, as well as high-risk background samples and nonstandard ammunitions that lack some of the characteristic GSR markers. Expansion of the population study to include these sample groups will allow for the assessment of electrochemistry to its full potential.

Acknowledgments

Thank you to Ph.D. student Korina Menking-Hoggatt for her help with sample collection, organization, and statistical machine learning algorithms. Thank you to undergraduate research student Sara L. Kuberski for her help with this project. The authors would like to thank Dr. Keith Morris and his research group for their assistance in loading of ammunition and collection of samples at the WVU Ballistics laboratory. The World Scout Jamboree is acknowledged for their support of this project and coordination in the collection of samples from the outdoor shooting range, as well as support from Sean McIntosh at Foothold Labs.

References

1. Giffords Law Center to Prevent Gun Violence. Gun violence statistics. 2019. https://lawcenter.giffords.org/facts/gun-violence-statistics/ (accessed February 2, 2020).
2. Archive GV.Past summary ledgers. 2020. https://www.gunviolencearchive.org/past-tolls (accessed February 2, 2020).
3. Meng H, Caddy B. Gunshot residue analysis–a review. J Forensic Sci 1997;42(4):553–70. https://doi.org/10.1520/JFS14167J.
4. Blakey LS, Sharples GP, Chana K, Birkett JW. Fate and behavior of gunshot residue—a review. J Forensic Sci 2018;63(1):9–19. https://doi.org/10.1111/1556-4029.13555.
5. Trejos T, Vander Pyl C, Menking-Hoggatt K, Alvarado AL, Arroyo LE. Fast identification of inorganic and organic gunshot residues by LIBS and electrochemical methods. Forensic Chem 2018;8:146–56. https://doi.org/10.1016/j.forc.2018.02.006.
6. Maitre M, Kirkbride KP, Horder M, Roux C, Beavis A. Current perspectives in the interpretation of gunshot residues in forensic science: a review. Forensic Sci Int 2017;270:1–11. https://doi.org/10.1016/j.forsciint.2016.09.003.
7. Dalby O, Butler D, Birkett JW. Analysis of gunshot residue and associated materials–a review. J Forensic Sci 2010;55(4):924–43. https://doi.org/10.1111/j.1556-4029.2010.01370.x.
8. ASTM International. ASTM E1588–17: standard guide for gunshot residue analysis by scanning electron microscopy/energy dispersive X-ray spectrometry. West Conshohocken, PA: ASTM. International, 2017. https://doi.org/10.1520/E1588-17.
9. Harris A. Analysis of primer residue from CCI Blazer® Lead Free ammunition by scanning electron microscopy/energy dispersive X-ray. J Forensic Sci 1995;40(1):27–30. https://doi.org/10.1520/jfs13755j.
10. Goudsmits E, Sharples GP, Birkett JW. Recent trends in organic gunshot residue analysis. Trends Analyt Chem 2015;74:46–57. https://doi.org/10.1016/j.trac.2015.05.010.
11. Wang J, Tian B, Wang J, Lu J, Olsen C, Yarnitzky C, et al. Stripping analysis into the 21st century: faster, smaller, cheaper, simpler and better. Anal Chim Acta 1999;385(1–3):429–35. https://doi.org/10.1016/S0003-2670(98)00664-3.
12. Salles MO, Naozuka J, Bertotti M. A forensic study: lead determination in gunshot residues. Microchem J 2012;101:49–53. https://doi.org/10.1016/j.microc.2011.10.004.
13. Erden S, Durmus Z, Kiliç E. Simultaneous determination of antimony and lead in gunshot residue by cathodic adsorptive stripping voltammetric methods. Electroanalysis 2011;23(8):1967–74. https://doi.org/10.1002/elan.201000612.
14. Omahony AM, Samek IA, Sattayasamitsathit S, Wang J. Orthogonal identification of gunshot residue with complementary detection principles of voltammetry, scanning electron microscopy, and energy-dispersive X-ray spectroscopy: sample, screen, and confirm. Anal Chem 2014;86(16):8031–6. https://doi.org/10.1021/ac5016112.
15. Woolever CA, Dewald HD. Differential pulse anodic stripping voltammetry of barium and lead in gunshot residues. Forensic Sci Int 2001;117(3):185–90. https://doi.org/10.1016/S0379-0738(00)00402-3.
16. O'Mahony AM, Wang J. Electrochemical detection of gunshot residue for forensic analysis: a review. Electroanalysis 2013;25(6):1341–58. https://doi.org/10.1002/elan.201300054.
17. Vuki M, Shiu KK, Galik M, O'Mahony AM, Wang J. Simultaneous electrochemical measurement of metal and organic propellant constituents of gunshot residues. Analyst 2012;137(14):3265–70. https://doi.org/10.1039/c2an35379b.
18. Olson EJ, Isley WC, Brennan JE, Cramer CJ, Bühlmann P. Electrochemical reduction of 2,4-dinitrotoluene in aprotic and pH-buffered media. J Phys Chem C 2015;119(23):13088–97. https://doi.org/10.1021/acs.jpcc.5b02840.
19. Goudsmits E, Sharples GP, Birkett JW. Preliminary classification of characteristic organic gunshot residue compounds. Sci Justice 2016;56(6):421–5. https://doi.org/10.1016/j.scijus.2016.06.007.
20. Menking-Hoggatt K, Arroyo L, Curran J, Trejos T. Novel LIBS method for micro-spatial chemical analysis of inorganic gunshot residues. J Chemom 2019;1–13. https://doi.org/10.1002/cem.3208 Epub 2019 Dec 09.
21. Feeney W, Vander Pyl C, Bell S, Trejos T. Trends in composition, collection, persistence, and analysis of IGSR and OGSR: a review. Forensic Chem 2020;19:100250. https://doi.org/10.1016/j.forc.2020.100250.
22. Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Mach Learn 1997;29(2–3):131–63. https://doi.org/10.1002/9780470400531.eorms0099.
23. Zhang H. In: Barr V, Markov Z, editors. The optimality of naive Bayes. In: Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference 2004 May 12–14, Miami Beach, FL. Menlo Park, CA: American Association for Artificial Intelligence, 2004;562–7.
24. Bewick V, Cheek L, Ball J. Statistics review 14: logistic regression. Crit Care 2005;9(1):112–8. https://doi.org/10.1186/cc3045.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Critical threshold plot for lead within the background population set (350 samples).

Figure S2. Critical threshold plot for lead within the shooter population set (395 samples) showing the maximum current on the y-axis

Figure S3. Critical threshold plot for lead within the shooter population set (395 samples) with expanded y-axis to demonstrate the critical threshold and mean value

Figure S4. Critical threshold plot for copper within the background population set (350 samples).

Figure S5. Critical threshold plot for copper within the shooter population set (395 samples) showing the maximum current on the y-axis

Figure S6. Critical threshold plot for copper within the shooter population set (395 samples) with expanded y-axis to demonstrate the critical threshold and mean value.

Figure S7. Critical threshold plot for nitroglycerin within the background population set (350 samples).

Figure S8. Critical threshold plot for nitroglycerin within the shooter population set (395 samples) showing the maximum current on the y-axis.

Figure S9. Critical threshold plot for nitroglycerin within the shooter population set (395 samples) with expanded y-axis to demonstrate the critical threshold and mean value.

Figure S10. GSR Neural Network arrangement for the three input nodes for the peak area currents of lead, copper, and nitroglycerin with a 3-node, single hidden layer, and classification output.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14549
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Eric F. Law,[1] *M.S.; and Keith B. Morris*,[1] *Ph.D.*

Three-Dimensional Analysis of Cartridge Case Double-Casts

ABSTRACT: Due to the shot-to-shot variability in tool mark reproduction on fired cartridge cases, a method of replication is needed for the creation of training and testing sets. Double-casting is one method that has been used for this application, but the accuracy and variability of this method needs to be characterized. Three firearms were used to fire 25 cartridges each to create the master cartridge cases. The double-casting method consists of creating a silicone mold of the master cartridge case. A plastic resin mix is then poured into the mold to create the double-cast reproduction. Fifteen double-casts of each of the 75 fired cartridge cases were created across different silicone molds to analyze within- and between-mold variability. The master cartridge cases and double-casts were scanned with a confocal microscope (Sensofar® S neox) to create three-dimensional representations of the surfaces. Two similarity metrics were used for the objective comparison of the double-casts to their master cartridge cases: the areal correlation coefficient ($ACCF_{MAX}$) and the number of congruent matching cells (CMC). The $ACCF_{MAX}$ and CMC data, along with visual examinations, showed that the double-casting method produces accurate reproductions. Within-mold variability was found to be minimal, and between-mold variability was low. These results illustrate that double-casting can be applied for training and testing purposes.

KEYWORDS: firearm tool mark, firearm identification, double-casting, breech face, congruent matching cells, cartridge case variability

Breech face and firing pin impressions on fired cartridge cases appear differently from shot to shot due to variability in reproduction through the firing process. Due to this variability, a method of replication is necessary to create training and testing sets. The National Institute of Standards and Technology (NIST) maintains a standard cartridge case, standard reference material (SRM) 2461, that is a reproduction of a master cartridge case created through electroforming (1). NIST has used confocal microscopy to acquire 3D scans of their reproduced standard cartridge case to compare with the master cartridge case (2).

Plastic replicas have been used, and the methods published on, as far back as 1956 with high detail resolution (3). More recently, this method has been called double-casting, where silicone mold negatives are made from master cartridge cases, and plastic replicas are created using the molds (4). The European Network of Forensic Science Institutes (ENFSI) has been using double-cast samples in proficiency tests for the purpose of having all participants examine the same samples (5). Visual comparisons between casts and their master cartridge cases have verified that the method produces high-quality reproductions (5). Recently, NIST created the SRM 2460a Standard Bullet that is a polyurethane cast coated in a fine layer of metal to increase surface reflectivity and better represent an actual fired bullet (6).

However, there is expected variability in the reproduction of the fine detail from the master cartridge case when creating multiple double-casts from a single mold, or multiple double-casts of the same cartridge case from different molds. This variability needs to be characterized to ensure each double-cast is representative of the master cartridge case. An IBIS® Heritage^{TM} System has been used for this purpose in a previous study (7). Those results were based on 2D grayscale images, and therefore, lighting differences between the master cartridge cases (metal) and double-casts (black plastic), due to differences in surface reflectivity, may have affected similarity scores.

To further evaluate the accuracy of double-cast tool mark reproduction compared with their master cartridge cases, confocal microscopy was used to measure the surfaces. This allowed for the depths of the impressions and striations to be directly considered. Confocal microscopy is also less dependent on lighting than conventional 2D imaging, and therefore, effects of surface reflectivity differences will be reduced. The goal of this study was to analyze the variability in the level of detail reproduced through the double-casting process using the areal correlation coefficient ($ACCF_{MAX}$) and the congruent matching cell (CMC) algorithm developed by NIST. The breech face impression area will be the focus of this study because CMC is currently optimized for the breech face area. The overall similarity between surfaces using the $ACCF_{MAX}$ and the number of congruent cells were used to provide an objective measure of similarity between double-casts and their master cartridge cases.

Materials and Methods

Twenty-five cartridge cases were fired from each of the three firearms to produce master cartridge cases for double-casting: SCCY® CPX-2, Hi-Point® C9, and Smith & Wesson® SD9VE. These firearms were selected based on the types of tool marks

[1] Department of Forensic and Investigative Science, West Virginia University, P.O. Box 6121, Morgantown, WV 26506.

Corresponding author: Keith B. Morris, Ph.D. E-mail: keith.morris@mail.wvu.edu

Received 7 May 2020; and in revised form 4 June 2020; accepted 29 July 2020.

they produce on the breech faces of fired cartridge cases. The CPX-2 produces prominent aperture shear, the C9 leaves parallel impressions, and the SD9VE imparts distinct granular impressions. One cartridge case from each of these firearms is shown in Fig. 1.

Fifteen double-casts were created from each of the 25 cartridge cases fired by each of the three firearms. This resulted in 1125 double-casts. To make the silicone molds for each firearm, the 25 master cartridge cases were arranged on a custom 3D printed holder (Fig. 2). The cartridge cases were then covered with a two-part liquid silicone mix, Smooth-On® Mold Star^{TM} 30. The mix was placed into a pressure pot for curing at 45 psi to remove air bubbles from the mixture at room temperature for approximately 6 h. The result was a single cylindrical piece of silicone that was approximately 11.5 cm in diameter and 4.0 cm in thickness (Fig. 2). This was repeated three times for the master cartridge cases from each firearm for a total of three molds. After the mold had cured and the master cartridge cases were removed, a two-part plastic resin mix, Smooth-On® Smooth Cast® 327, was dyed black with Smooth-On® SO-Strong® black colorant and poured into the mold. This plastic resin mix was also cured in the pressure pot at 35 psi at room temperature for approximately 6 h. The molds and casts were cured at different pressure levels. The thought was that if the casts were cured at higher pressure than the mold, then the structural integrity of the mold may change under the higher pressure. To be safe, the mold was cured at higher pressure to ensure that it could withstand the lower pressure at which the casts were cured. After the plastic resin finished curing, each double-cast was removed from the mold. This process was repeated to create five double-casts of each master cartridge case from each of the three molds per firearm, for a total of 15 double-casts per master cartridge case. Figure 2 displays five reproductions of master cartridge case #1 created from Mold 1 for one firearm. This allowed for the analysis of within- and between-mold variability. This method has also been described in a previous study where an IBIS® Heritage^{TM} System was used to evaluate the reproducibility of the same firearms, molds, and double-casts as those used in this study (7).

A Sensofar® S neox optical profiler was used to acquire three-dimensional scans of the fired master cartridge cases and all the reproductions using the "Confocal Fusion" scan mode. This is a proprietary method developed by Sensofar® that primarily measures a surface using confocal microscopy; however, if a confocal data point is not measured at a given location, then focus variation is used instead (e.g., areas with steep slopes). Generally, this method works well. It is difficult to measure the edges of the firing pin impression even with this method, although this area is not typically of interest in cartridge case comparisons. The scans were all converted to the *.x3p file format for software interoperability and ease of use (9). All files were named in a way that allowed for the specific mold number, cast set number, and master cartridge case number to be known. An example filename is "DAN-FA-UNK-0203-0001," where each part is defined as:

- DAN: unique letter combination that identifies these as double-casts of the CPX-2 master cartridge cases.
- FA: Federal American Eagle manufactured ammunition was used.
- UNK: not applicable here because manufactured ammunition was used but can be used to refer to the primer and powder of a reloaded cartridge.
- 0203: identifies this double-cast was created from the second mold and was made in the third set of double-casts created from that mold.
- 0001: refers to the first cartridge case of the 25 masters fired from this firearm.

Two similarity metrics were used for the objective comparison of the double-casts to their master cartridge cases: the areal correlation coefficient ($ACCF_{MAX}$) and the number of CMCs. The comparison software was developed by NIST and was provided to the authors for research purposes as part of an ongoing collaboration. The CMC method has been described in detail elsewhere (10–12) but will be explained in brief here. To prepare the *.x3p files for CMC analysis, the scans were cropped so that only the breech face area remains. Care was taken to select the same breech face areas on the casts as selected on their respective master cartridge cases to reduce comparison effects due to differences in sample domain. To attenuate noise, form, and waviness, a Gaussian regression filter was applied to each measured surface with λ_S = 25 μm and λ_C = 400 μm cutoff wavelengths. Furthermore, the images were downsampled from their original resolution of 1.38 μm/pixel by a factor of two to 2.76 μm/pixel. This resulted in faster comparisons while not sacrificing accuracy. One cartridge case scan was then set as the reference surface, and another was set as the comparison surface. A grid of 64 (500 × 500 μm) cells was defined on the reference surface. For each of the cells, a search was performed over all positions of the comparison surface to find the cell registration location that yielded the highest value for the respective cell pair similarity value ($ACCF_{MAX}$). Each cell was also allowed 360°

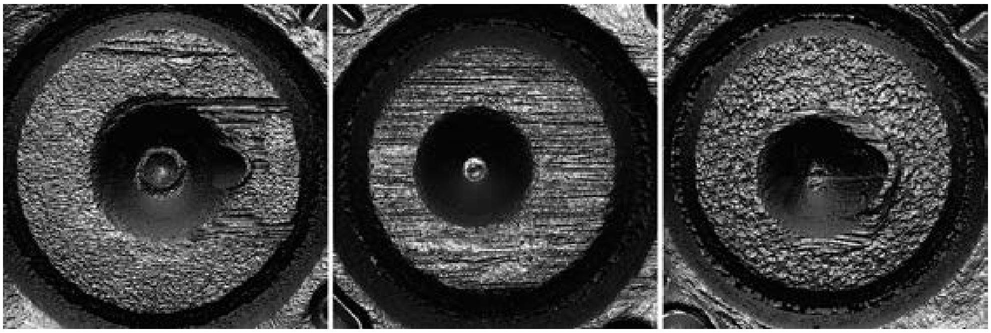


FIG. 1—*Examples of cartridge cases from each of the three firearms. Left is from the SCCY® CPX-2, middle is from the Hi-Point® C9, and right is from the Smith & Wesson® SD9VE. Images saved from the Cadre X3P viewer (8) with the enhanced contrast option enabled.*

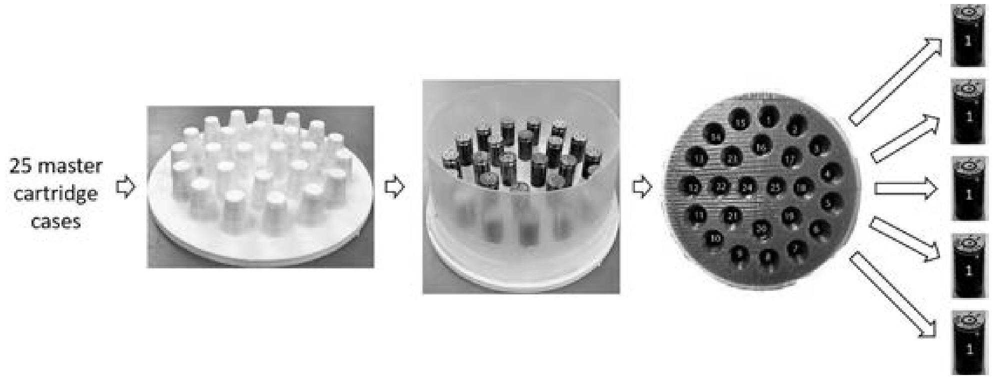


FIG. 2—*Illustration of the double-casting method. [Color figure can be viewed at wileyonlinelibrary.com]*

rotation during this search. The comparison metric of the CMC method is the number of congruent matching cell pairs, that is, the number of cell pairs that have both a sufficient similarity and a congruent registration location. In this study, to be considered a congruent matching cell, the cell similarity value had to be at least 20% and the errors in registration position and orientation cannot exceed 125 μm and 3°, respectively. These criteria are based on research done by NIST (10–12), as well as in-house testing of the algorithms. The output included the number of congruent matching cells, as well as an $\mathrm{ACCF_{MAX}}$ value corresponding to the overall similarity of the two surfaces.

Figure 3 illustrates the comparison of two cartridge cases fired by the same CPX-2 pistol. The grid of cells, color-coded to show congruency, was overlaid onto the reference surface (Fig. 3, left). The black cells represent a CMC with the comparison surface (Fig. 3, right), while the red cells represent non-CMCs. In the corresponding comparison surface, the cells are displayed based on their best-fit locations, as well as color-coded the same as in the reference surface. In this example, the non-congruent cells fall in the aperture shear area of the breech face. Aperture shear is typically present in cartridge cases fired by this particular CPX-2; however, it is not reproducible. The results of this comparison illustrate the effectiveness of utilizing the congruent cell approach. In this comparison, an $\mathrm{ACCF_{MAX}}$ of 33.0% and 27 congruent cells resulted. This $\mathrm{ACCF_{MAX}}$ may not be any larger than nonmatching comparisons, but 27 congruent cells would lead an examiner to believe these cartridge cases were fired by the same firearm. The highest known nonmatching CMC score has never exceeded five in NIST's research (10–12).

Figure 4 illustrates a CMC comparison with a double-cast of the reference surface used in Fig. 3 (left). As with all double-cast comparisons performed in this study, the master cartridge case was set as the reference surface (Fig. 4, left) and the double-cast was set as the comparison surface (Fig. 4, right). In this instance, all 39 cells from the reference surface were found to be congruent in the comparison surface.

The tolerances for congruent cells may appear to be low considering the goal here is to analyze the accuracy of reproductions. It may be expected that individual cells would have greater than 20% similarity, and less than the allowable 125 μm in x and y spatial positioning and 3° in rotation if the double-casts are accurately reproducing the fine detail from the master. However, these criteria were selected to be consistent with comparisons that have previously been run. Using these same criteria allowed for the similarity of the casts to their master cartridge cases versus the comparison of multiple cartridge cases fired by the same firearm to be analyzed, such as in Figs 3 and 4. Had more strict tolerances been selected for the double-casts, the difference in comparing the reproductions and multiple cartridge cases fired by the same firearm may not have been clear.

Once all the comparisons of the reproductions to their master cartridge cases had been run, the CMC and $\mathrm{ACCF_{MAX}}$ data were organized into Microsoft Excel® spreadsheets. The naming convention was columns corresponding to Cast Set 1.1, Cast Set 1.2, Cast Set 1.3, up to Cast Set 3.5, where the first number was the mold number (1, 2, or 3) and the second number was the cast set number from that mold (1, 2, 3, 4, or 5). There were 25 values (rows in the data files) within each cast set that represented the similarity of the double-cast compared with its master cartridge case. The Friedman test was then used to compare the cast set groups (the columns in the data files) for each of the firearms to determine whether any of the CMC or $\mathrm{ACCF_{MAX}}$ values were significantly different for each of the three firearms. All plots and statistical analyses were performed using R and RStudio (13,14).

Results and Discussion

Overall Results

Comparisons were set up with the master cartridge cases as the reference surfaces and the double-cast scans as the comparison surfaces. The algorithm outputs the overall similarity of the surfaces represented by the $\mathrm{ACCF_{MAX}}$ and the number of congruent cells. Using the raw number of congruent cells with double-casts to investigate the accuracy of the reproductions was slightly misleading. For example, master cartridge case #1 may have had a larger breech face area than master cartridge case #2, potentially due to a smaller firing pin impression or a different degree of primer surface contact with the breech face of the firearm. The grid of cells overlaid on master cartridge case #1 will have more cells than the grid on master cartridge case #2. During comparison, there are more potential cells to be found congruent in the double-casts of master cartridge case #1 than with master cartridge case #2. For this reason, the CMC results displayed are represented as percent values where the number of congruent cells was divided by the total number of cells possible. Descriptive statistics for the number of potential cells for each of the three firearms are shown in Table 1. On average, the CPX-2 had the most potential cells, followed by the SD9VE and C9.

Scatterplots of the results for each firearm are displayed in Fig. 5. The x-axis shows the percent of congruent cells, and the

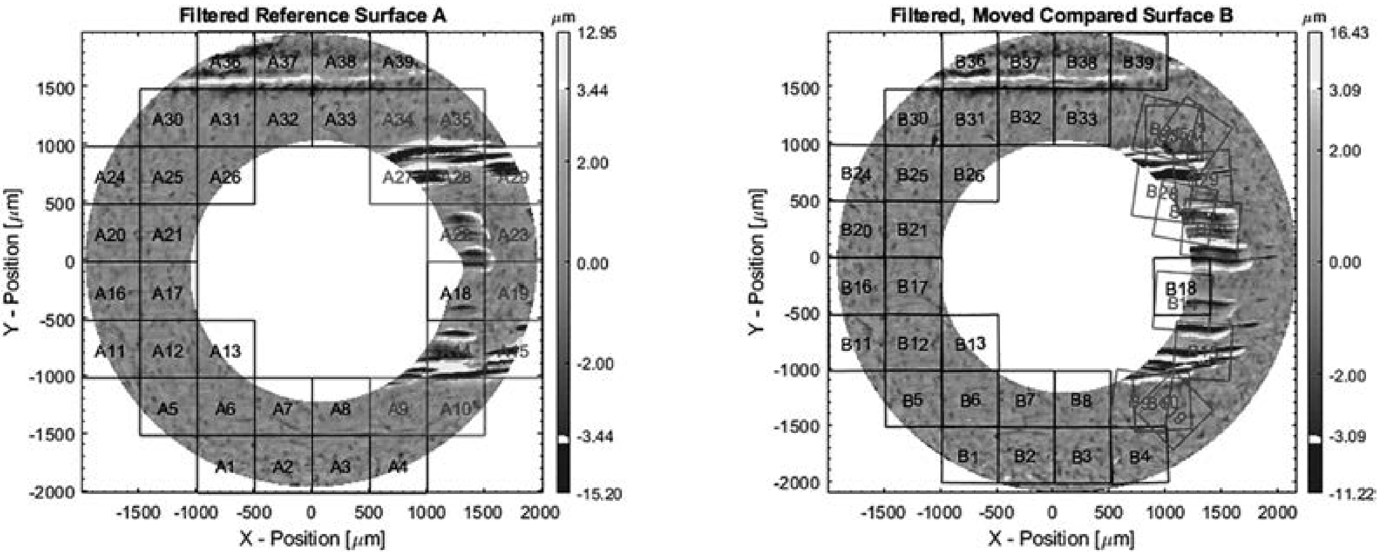


FIG. 3—*CMC example for a CPX-2 known match comparison. The overall ACCF$_{MAX}$ was 33.0%, and there were 27 of 39 congruent cells. [Color figure can be viewed at wileyonlinelibrary.com]*

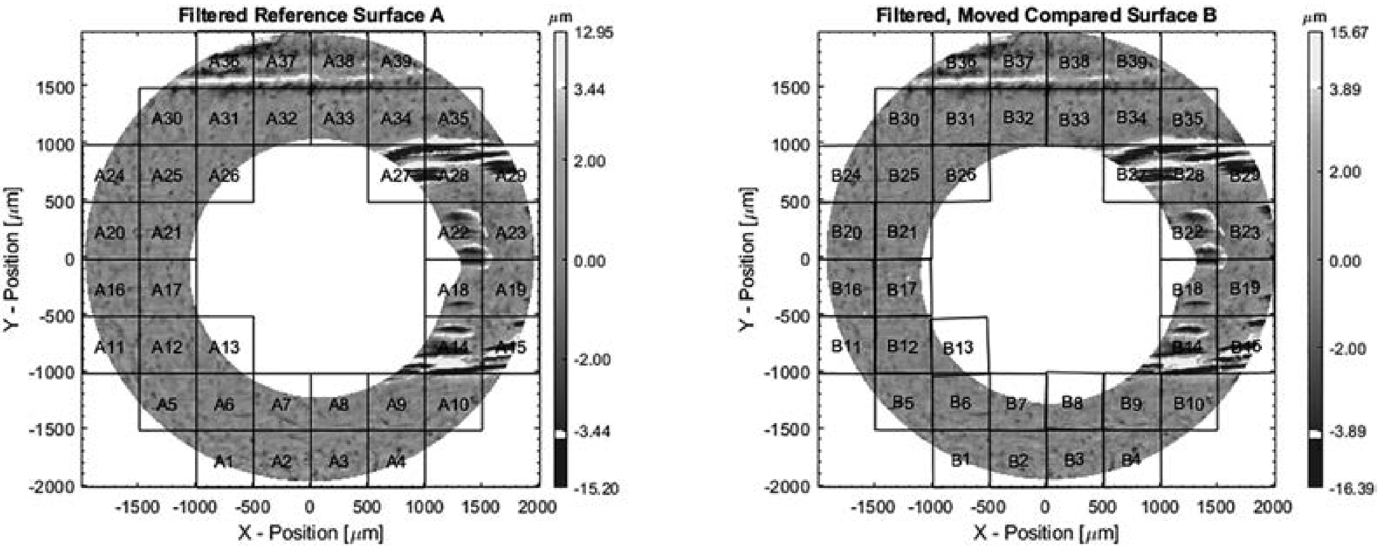


FIG. 4—*CMC example for a CPX-2 master cartridge case compared with one of its double-casts. The overall ACCF$_{MAX}$ was 93.0%, and there were 39 of 39 congruent cells. [Color figure can be viewed at wileyonlinelibrary.com]*

y-axis shows the overall similarity (ACCF$_{\text{MAX}}$). The red squares are the comparison results of the intercomparisons of the master cartridge cases for the respective firearms. Each blue circle represents a single comparison of a double-cast to its master cartridge case. Each comparison is made up of n cells, each of which has their own similarity, position, and rotation values. The master cartridge case intercomparisons exhibited larger variability in CMC and ACCF$_{\text{MAX}}$ percentages, with the C9 clearly performing the best of the three firearms. Furthermore, the double-casts all displayed higher similarity than the master cartridge case intercomparisons, which was expected. The double-casts generally had ACCF$_{\text{MAX}}$ values larger than 80% and a CMC percent above 95% with all three firearms. Lateral scale differences influence the similarity metrics, more so the ACCF$_{\text{MAX}}$ due to the position tolerance with the CMC metric. Lateral scale differences may be introduced because of shrinkage of the casts during curing; however, linear shrinkage with the casting material used is 0.0075 in/in, limiting this effect (15). Even so, the values are indicative of the high level of detail reproduced through the double-casting process.

TABLE 1—*Descriptive statistics for the number of potential cells.*

Firearm	Mean	Standard Deviation
SCCY CPX-2	40	2
Hi-Point C9	33	1
Smith & Wesson SD9VE	39	2

To further investigate the reproduction accuracy, the differences in the number of potential cells on the reference surface (the master cartridge case) and the number of congruent cells in the comparison surface (the double-cast) were plotted and are shown in Fig. 6. It would be expected that all data points should

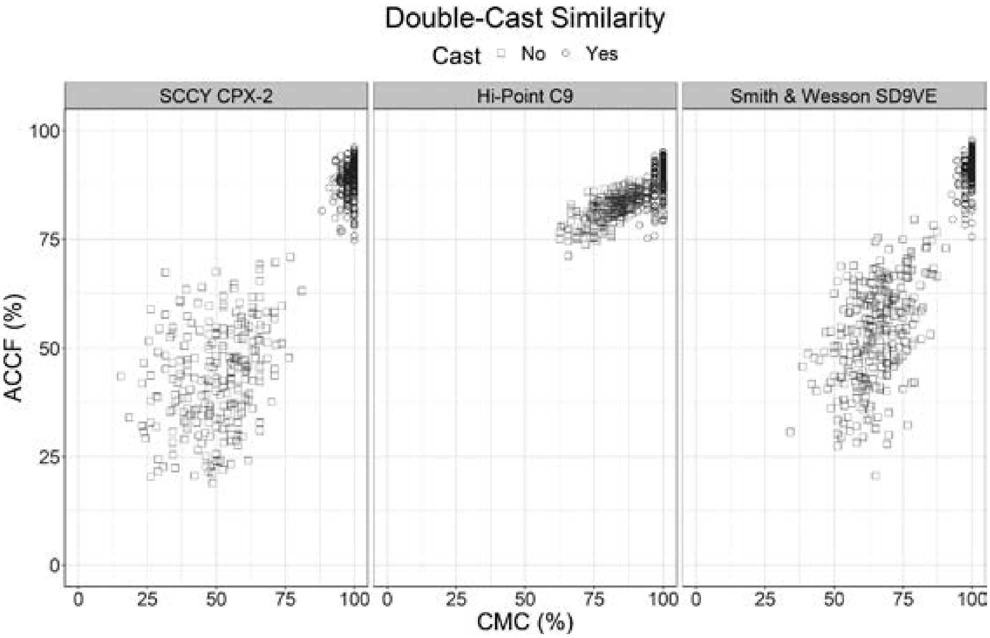


FIG. 5—*CMC and ACCF$_{MAX}$ results for the master cartridge cases compared with their reproductions. [Color figure can be viewed at wileyonlinelibrary.com]*

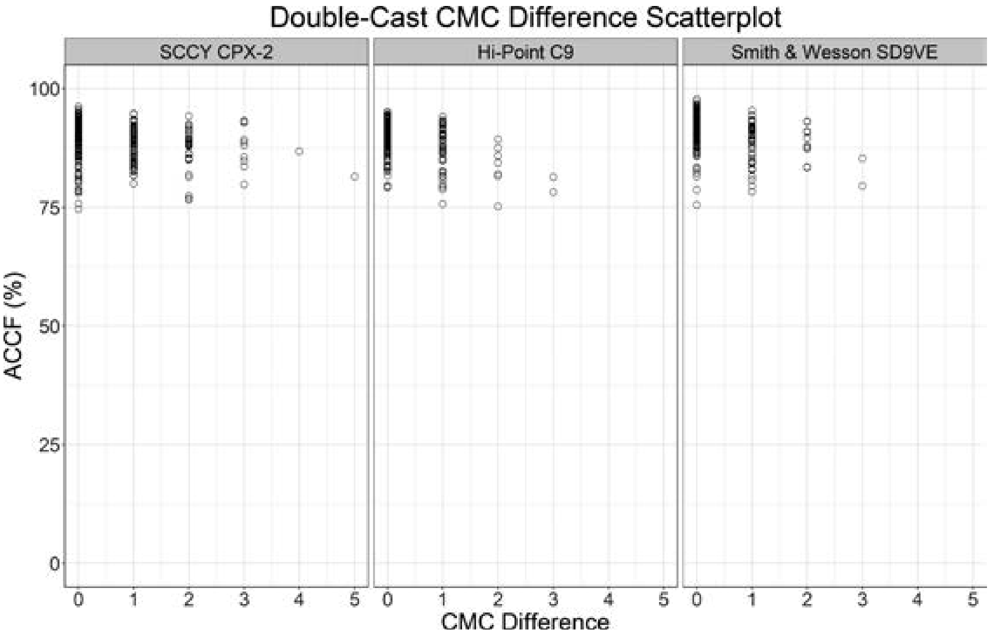


FIG. 6—*Scatterplot displaying the number of congruent cells less than the number of potential cells. Any difference in the number of congruent cells means that there were cells that fell outside of the similarity, position, and rotation thresholds.*

be in the “0” category because the double-casts should be identical to the master cartridge cases. However, due to variation in the casting process there are areas in some of the casts that did not reproduce as well as others and therefore fall outside of the congruency criteria (20% similarity, 125 μm in position, 3° rotation). The CPX-2 had the most cases where the number of CMCs was less than the number of potential cells. The SD9VE had the least, and the C9 results were in between the other two firearms.

The Friedman test was used to compare across the mold and set numbers (the columns in the Excel® data files) to determine whether there was evidence of significant differences at the 0.05 level of significance (16). The Friedman test is a global test meaning it does not provide information on which groups are different, only if any of the groups are different from any of the others. The *p*-values from this test are shown in Table 2 for both the CMC and ACCF$_{MAX}$ data. Significant differences, where the *p*-value was less than 0.05, were found in all cases except for the C9 CMC results. For all three firearms, there were smaller *p*-values for the ACCF$_{MAX}$ data, indicating more evidence for group differences than with the congruent cell data.

TABLE 2—*Friedman test* p-*values for the three firearms based on both CMC and* $ACCF_{MAX}$ *data.*

Firearm	CMC	$ACCF_{MAX}$
SCCY CPX-2	0.0239	<0.0001
Hi-Point C9	0.7613	<0.0001
Smith & Wesson SD9VE	0.035	<0.0001

Significant differences are highlighted in gray.

$ACCF_{MAX}$ Data

Figure 7 shows the distributions of $ACCF_{MAX}$ values for the different mold and set numbers (*x*-axis) for each of the three firearms. For example, Cast Set 2.3 refers to Mold 2 and the third set of casts made from that mold. Within each firearm, the boxplots all overlap with each other, so based on visual examination, it does not appear that any of the molds produce better or worse reproductions than any others.

Because the Friedman test was significant for the $ACCF_{MAX}$ data for all three firearms (Table 2), multiple comparison procedures (17) were performed to determine which individual groups were significantly different. Tables 3, 4, and 5 display the *p*-values for the individual comparisons of the groups that are displayed in the boxplots with significant differences (*p*-values less than 0.05) highlighted in gray. Based on the $ACCF_{MAX}$ data, out of the 105 total comparisons per firearm, for the CPX-2, there were 17 groups that were significantly different, for the C9, there were nine groups that were significantly different, and for the SD9VE, there were four groups that were significantly different. Out of these significant differences, only two were within-mold differences (Cast Sets 1.1–1.5 and 2.1–2.5 with the C9). All other differences were between-mold comparisons. This indicates that within-mold variability is minimal, and between-mold variability is low. It is also important to note that no decreasing trend was observed from the first cast set from each mold to the fifth. This shows that there were no significant changes in the mold causing a decrease in the cast similarity scores over the five sets.

Figures 8, 9, and 10 each display a master cartridge case and reproduction that resulted in one of the lowest overall similarity values from one of the three firearms. Due to the allowable thresholds for CMC, the low $ACCF_{MAX}$ comparisons were still able to produce a large number of congruent cells: 37 of 39 for

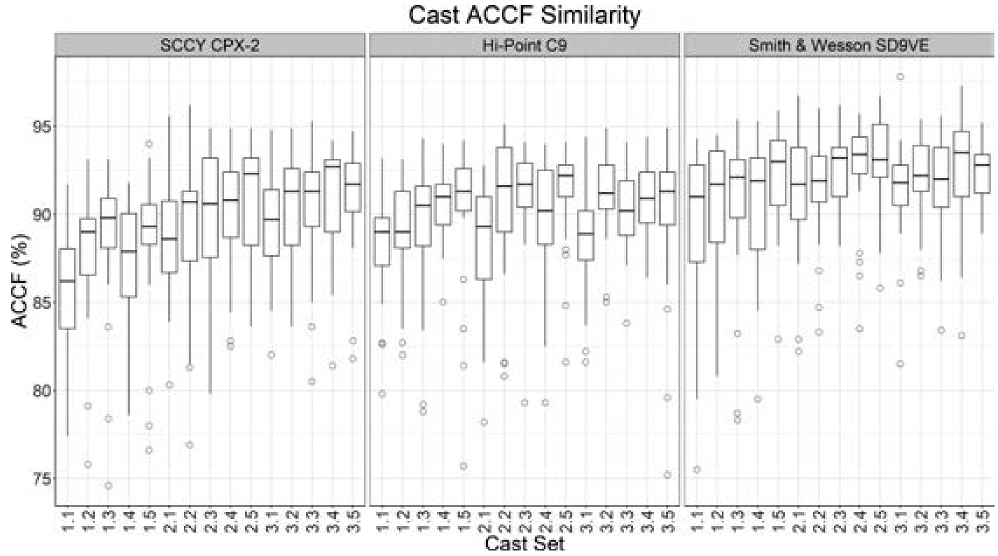


FIG. 7—*Boxplot of the* $ACCF_{MAX}$ *scores of all casts compared with their master cartridge cases separated by the different mold and set numbers.*

TABLE 3—*Friedman test multiple comparison procedure for the CPX-2* $ACCF_{MAX}$ *values.*

	1.1	1.2	1.3	1.4	1.5	2.1	2.2	2.3	2.4	2.5	3.1	3.2	3.3	3.4
1.2	0.9675	–	–	–	–	–	–	–	–	–	–	–	–	–
1.3	0.3199	0.9987	–	–	–	–	–	–	–	–	–	–	–	–
1.4	1.0000	0.9999	0.7826	–	–	–	–	–	–	–	–	–	–	–
1.5	0.3408	0.9991	1.0000	0.8021	–	–	–	–	–	–	–	–	–	–
2.1	0.917	1.0000	0.9999	0.9989	0.9999	–	–	–	–	–	–	–	–	–
2.2	0.0158	0.6846	0.9994	0.124	0.9991	0.8116	–	–	–	–	–	–	–	–
2.3	0.003	0.3734	0.9797	0.0349	0.9754	0.5148	1.0000	–	–	–	–	–	–	–
2.4	0.0086	0.564	0.9969	0.0786	0.996	0.7076	1.0000	1.0000	–	–	–	–	–	–
2.5	0.0002	0.0992	0.7724	0.0043	0.7516	0.1663	0.9997	1.0000	1.0000	–	–	–	–	–
3.1	0.2899	0.998	1.0000	0.7516	1.0000	0.9997	0.9996	0.9851	0.998	0.8021	–	–	–	–
3.2	0.0008	0.1949	0.9051	0.0117	0.8922	0.2997	1.0000	1.0000	1.0000	1.0000	0.9225	–	–	–
3.3	0.0002	0.0992	0.7724	0.0043	0.7516	0.1663	0.9997	1.0000	1.0000	1.0000	0.8021	1.0000	–	–
3.4	<0.0001	0.033	0.5148	0.0009	0.4904	0.0616	0.9914	0.9999	0.9977	1.0000	0.5517	1.0000	1.0000	–
3.5	<0.0001	0.0098	0.2803	0.0002	0.2616	0.0200	0.9376	0.9954	0.9729	1.0000	0.3097	0.9998	1.0000	1.0000

Significant differences at the 0.05 level are highlighted in gray.

TABLE 4—*Friedman test multiple comparison procedure for the C9 $ACCF_{MAX}$ values.*

	1.1	1.2	1.3	1.4	1.5	2.1	2.2	2.3	2.4	2.5	3.1	3.2	3.3	3.4
1.2	0.9616	–	–	–	–	–	–	–	–	–	–	–	–	–
1.3	0.8832	1.0000	–	–	–	–	–	–	–	–	–	–	–	–
1.4	0.1286	0.9793	0.9965	–	–	–	–	–	–	–	–	–	–	–
1.5	0.0227	0.7746	0.9025	1.0000	–	–	–	–	–	–	–	–	–	–
2.1	0.9998	1.0000	0.9999	0.734	0.3283	–	–	–	–	–	–	–	–	–
2.2	0.0038	0.4354	0.6225	0.9995	1.0000	0.1086	–	–	–	–	–	–	–	–
2.3	0.0026	0.3695	0.5519	0.9987	1.0000	0.0834	1.0000	–	–	–	–	–	–	–
2.4	0.3184	0.9991	1.0000	1.0000	0.9998	0.9298	0.986	0.975	–	–	–	–	–	–
2.5	0.0003	0.1286	0.2368	0.9616	0.9995	0.0182	1.0000	1.0000	0.8121	–	–	–	–	–
3.1	1.0000	1.0000	0.9992	0.5873	0.2129	1.0000	0.0602	0.0449	0.8461	0.0086	–	–	–	–
3.2	0.0053	0.493	0.6797	0.9998	1.0000	0.134	1.0000	1.0000	0.9919	1.0000	0.0761	–	–	–
3.3	0.3910	0.9997	1.0000	1.0000	0.9993	0.9583	0.9727	0.9549	1.0000	0.7444	0.8963	0.9829	–	–
3.4	0.0109	0.6341	0.8031	1.0000	1.0000	0.2129	1.0000	1.0000	0.9983	1.0000	0.1286	1.0000	0.9955	–
3.5	0.0834	0.9512	0.9887	1.0000	1.0000	0.6225	0.9999	0.9998	1.0000	0.9845	0.4697	1.0000	1.0000	1.0000

Significant differences at the 0.05 level are highlighted in gray.

TABLE 5—*Friedman test multiple comparison procedure for the SD9VE $ACCF_{MAX}$ values.*

	1.1	1.2	1.3	1.4	1.5	2.1	2.2	2.3	2.4	2.5	3.1	3.2	3.3	3.4
1.2	0.9191	–	–	–	–	–	–	–	–	–	–	–	–	–
1.3	0.7537	1.0000	–	–	–	–	–	–	–	–	–	–	–	–
1.4	0.8533	1.0000	1.0000	–	–	–	–	–	–	–	–	–	–	–
1.5	0.0791	0.9771	0.998	0.9918	–	–	–	–	–	–	–	–	–	–
2.1	0.8533	1.0000	1.0000	1.0000	0.9918	–	–	–	–	–	–	–	–	–
2.2	0.5387	1.0000	1.0000	1.0000	0.9999	1.0000	–	–	–	–	–	–	–	–
2.3	0.0345	0.9136	0.9844	0.9581	1.0000	0.9581	0.9985	–	–	–	–	–	–	–
2.4	0.0011	0.3371	0.5742	0.4454	0.9983	0.4454	0.7834	0.9999	–	–	–	–	–	–
2.5	0.0009	0.3074	0.5387	0.4117	0.9973	0.4117	0.7537	0.9998	1.0000	–	–	–	–	–
3.1	0.9942	1.0000	1.0000	1.0000	0.8288	1.0000	0.9992	0.6559	0.1176	0.1034	–	–	–	–
3.2	0.6559	1.0000	1.0000	1.0000	0.9995	1.0000	1.0000	0.9942	0.6785	0.6444	0.9999	–	–	–
3.3	0.9918	1.0000	1.0000	1.0000	0.8533	1.0000	0.9995	0.6897	0.1332	0.1176	1.0000	0.9999	–	–
3.4	0.0071	0.6673	0.8684	0.7737	1.0000	0.7737	0.9644	1.0000	1.0000	1.0000	0.3371	0.9244	0.3682	–
3.5	0.0721	0.9725	0.9973	0.9898	1.0000	0.9898	0.9999	1.0000	0.9987	0.998	0.8113	0.9993	0.8372	1.0000

Significant differences at the 0.05 level are highlighted in gray.

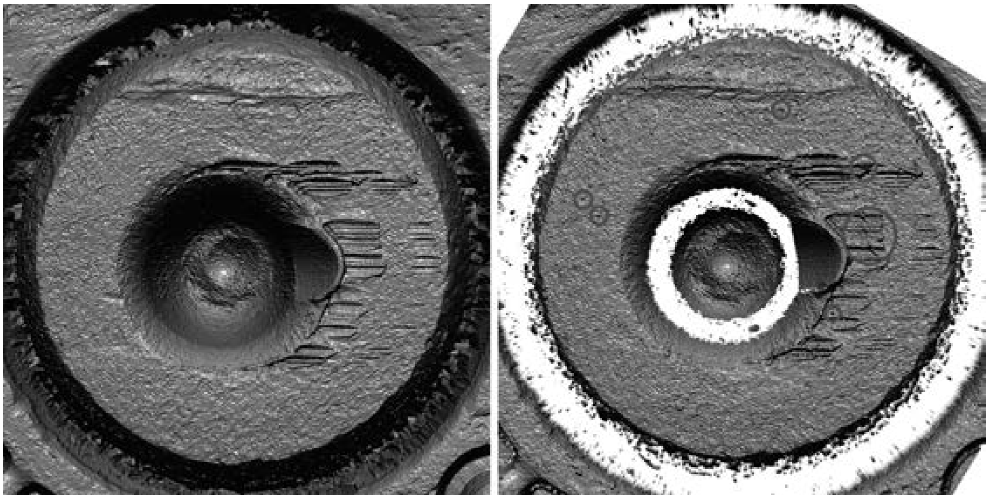


FIG. 8—*One of the lowest $ACCF_{MAX}$ comparisons for the CPX-2, resulting in an $ACCF_{MAX}$ of 76.6% and 37 of 39 congruent cells. The master cartridge case is on the left and the reproduction from Mold 1, Set 5 is on the right. Artifact examples from casting are circled in red on the double-cast. Images saved using the Cadre X3P viewer (8). [Color figure can be viewed at wileyonlinelibrary.com]*

the CPX-2, 34 of 36 for the C9, and 39 of 39 for the SD9VE. Although some differences may be found, the reproductions are still visually accurate representations of the features from the master cartridge cases.

While examining Figs 8, 9, and 10, there are features on the double-casts that do not appear on the master cartridge cases. Some examples of these are circled in red on the images, and additional marks can also be found. These marks do not carry

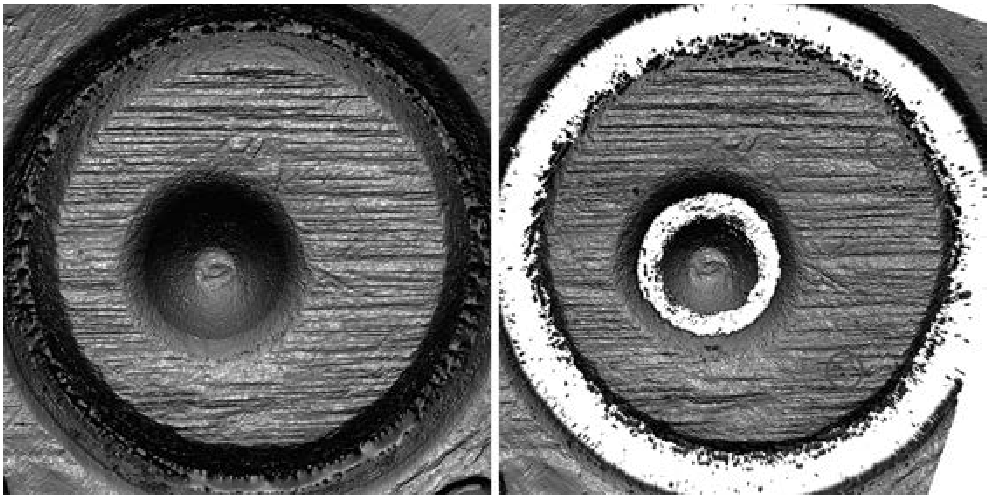


FIG. 9—*One of the lowest ACCF$_{MAX}$ comparisons for the C9, resulting in an ACCF$_{MAX}$ of 78.8% and 34 of 35 congruent cells. The master cartridge case is on the left and the reproduction from Mold 1, Set 3 is on the right. Artifact examples from casting are circled in red on the double-cast. Images saved using the Cadre X3P viewer (8). [Color figure can be viewed at wileyonlinelibrary.com]*

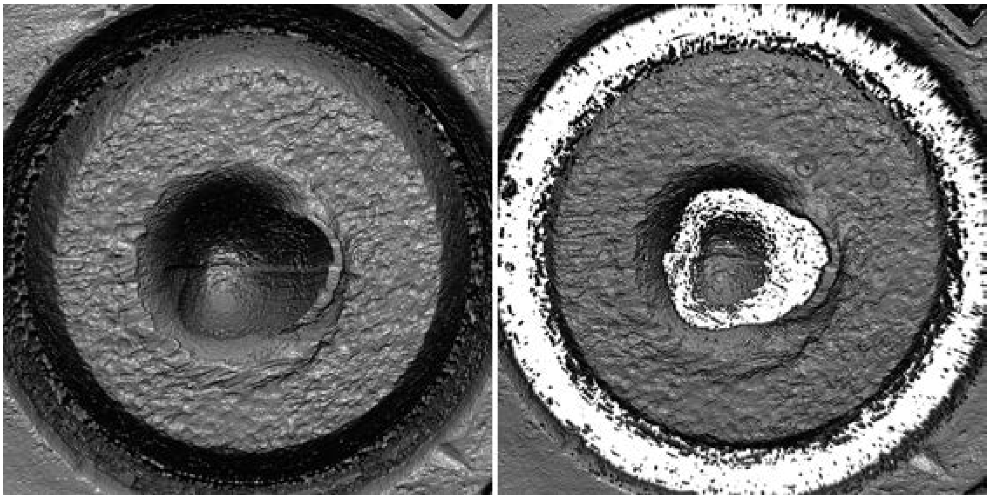


FIG. 10—*One of the lowest ACCF$_{MAX}$ comparisons for the SD9VE, resulting in an ACCF$_{MAX}$ of 75.5% and 39 of 39 congruent cells. The master cartridge case is on the left and the reproduction from Mold 1, Set 1 is on the right. Artifact examples from casting are circled in red on the double-cast. Images saved using the Cadre X3P viewer (8). [Color figure can be viewed at wileyonlinelibrary.com]*

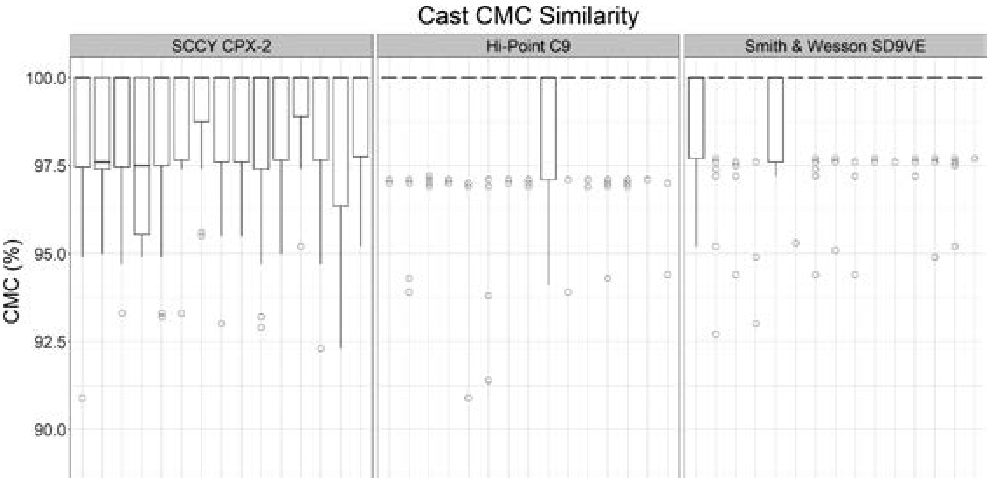


FIG. 11—*Boxplot of the CMC percent values of all casts compared with their master cartridge cases separated by the different mold and set numbers.*

across all casts from the same mold indicating these features are not from the molding process. Debris that either fell into the cured mold or contaminated the casting material during mixing may have been the cause of these additional marks leading to lower ACCF_{MAX} values. While the additional marks may not appear to be major differences between the two surfaces, the ACCF_{MAX} is sensitive to these differences as they would be highlighted as features when applying the Gaussian regression filter. It is important to note that there was no quality control process used. Double-casts created for training or testing purposes that included additional marks such as those shown in Figs 8, 9, and 10 would not be used, and new double-casts would be created.

CMC Data

The ACCF_{MAX} may be the more appropriate metric to use for comparing the accuracy of double-casting because the overall similarity is of interest rather than individual areas covered by cells. However, the congruent cell data are also important because many comparisons not involving reproductions to their master cartridge cases will not result in 100% of potential cells being congruent due to shot-to-shot variability. Refer to Fig. 3 for a typical known match example.

As discussed earlier, the CMC data were presented as a percent value by dividing the number of congruent cells by the total number of potential cells for each comparison. Recall, the number of potential cells was overlaid onto the reference surface (the master cartridge case) and searched over the comparison surface (the double-cast). Any cells that were within the similarity, positioning, and angular thresholds were considered congruent. The distributions of the percentages of congruent cells are displayed in Fig. 11. For the CPX-2, 62% of comparisons resulted in 100% congruent cells, 83% were 100% congruent for the C9, and 82% were 100% congruent for the SD9VE.

The Friedman test results for the congruent cell data were significant for the CPX-2 and SD9VE, so multiple comparison procedures were used for those two firearms to determine which comparisons were significant. The minimum p-values for the CPX-2 and SD9VE were 0.0850 and 0.1500, respectively, indicating no individual significant differences, and therefore, the p-values are not all being included here as with the ACCF_{MAX} data. There may have been no individual significant differences found with the multiple comparison test for a couple of reasons. The Friedman test p-values were not much below the significance level of 0.05. Because of this, the multiple comparison procedure may not have had enough statistical power to detect any individual group differences at those Friedman test p-value magnitudes.

Double-casts of cartridge cases from the CPX-2 seemed to produce lower overall ACCF_{MAX} values, a lower number of congruent matching cells, and more significant differences (p-values less than 0.05). Further research is required to determine why the CPX-2 showed this difference compared with the other two firearms.

Conclusions

Three-dimensional analysis utilizing the ACCF_{MAX} and CMC data shows that the double-casting process creates reproductions that are representative of the fine detail present in the surfaces of the master cartridge cases. The double-casts from this study with the lowest overall similarity percentages still visually produced the detail from the master cartridge cases. Furthermore, implementation of a quality control process would lead to removal of any reproductions that are below the desired quality, whether measured through objective or visual comparisons. Double-casting has importance in the forensic science community for easily creating exemplars for database imaging, proficiency testing, and error rate analysis. Based on the results presented, double-cast sets could be created for these purposes with little concern for being inaccurate cartridge case representations.

Acknowledgments

The authors would like to acknowledge Dr. Casey Jelsema for his support with the statistical analysis and Mr. Barrett Towns for his assistance with creating the molds and casts. The authors would also like to acknowledge the firearm and tool mark analysis team at the National Institute of Standards and Technology (NIST) for providing access to their tool mark research software. The software was applied in this paper to calculate the ACCF and CMC comparison results.

References

1. Song J, Whitenton E, Kelley D, Clary R, Ma L, Ballou S, et al. SRM 2460/2461 standard bullets and casings project. J Res Natl Inst Stand Technol 2004;109(6):533–42. https://doi.org/10.6028/jres.109.040.
2. Seiler DF, Watters RL Jr. Certificate: standard reference material 2461 standard cartridge case. Gaithersburg, MD: National Institute of Standards and Technology, 2012.
3. Biasotti AA. Plastic replicas in firearms and tool mark identifications. J Crim Law Criminol 1956;47(1):110–7.
4. Firearms Programme General Secretariat. INTERPOL Ballistics Information Network: Handbook on the collection and sharing of ballistics data, 3rd edn. Lyon, France: INTERPOL, 2014.
5. Pauw-Vugts P, Walters A, Øren L, Pfoser L. FAID2009: proficiency test and workshop. AFTE J 2013;45(2):115–27.
6. Gundlach D, Choquette SJ. Certificate: standard reference material 2460a standard bullet replica. Gaithersburg, MD: National Institute of Standards and Technology, 2018.
7. Law E, Morris K. The utility of double-casting for creating cartridge case reproductions. AFTE J 2020;52(1):26–39.
8. Cadre Forensics. X3P viewer software. https://www.cadreforensics.com (accessed July 27, 2020).
9. Open Forensic Metrology Consortium. OpenFMC & X3P. https://www.cadreforensics.com/x3p.html (accessed July 27, 2020).
10. Song J. Proposed NIST ballistics identification system (NBIS) using 3D topography measurements on correlation cells. AFTE J 2013;45(2):184–9.
11. Chu W, Tong M, Song J. Validation tests for the congruent matching cells (CMC) method using cartridge cases fired with consecutively manufactured pistol slides. AFTE J 2013;45(4):361–6.
12. Song J, Vorburger TV, Chu W, Yen J, Soons JA, Ott DB, et al. Estimating error rates for firearm evidence identifications in forensic science. Forensic Sci Int 2018;284:15–32. https://doi.org/10.1016/j.forsciint.2017.12.013.
13. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2019.
14. RStudio, Inc. About RStudio. http://www.rstudio.com/about (accessed July 27, 2020).
15. Smooth-Cast™ ColorMatch™ 325 Series. Technical bulletin. https://www.smooth-on.com/tb/files/SMOOTHCAST_325_326_327_COMBO_TB.pdf (accessed July 27, 2020).
16. Conover WJ. Practical nonparametric statistics, 2nd edn. Hoboken, NJ: John Wiley and Sons, 1980.
17. Pohlert T. PMCMRplus: calculate pairwise multiple comparisons of mean rank sums extended. R package version 1.4.4. 2020. https://cran.r-project.org/web/packages/PMCMRplus/PMCMRplus.pdf (accessed July 30, 2020).

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14538
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Emily E. Fairbanks,[1] *M.S.; Jennifer Turner,*[2] *M.S.; Junkun Ma,*[3] *Ph.D.; and Jorn Yu,*[1] *Ph.D.*

Development of a Novel Finger-Trigger Interface for Trigger Pull Measurement

ABSTRACT: Trigger pull is the force that needs to be exerted on the trigger to discharge a firearm. The measurement of trigger pull can assist in the evaluation of the safety, function, and manufacturing characteristics associated with a firearm during the forensic firearm examination process. Nonetheless, the accuracy and uncertainty of trigger pull measurements may be affected by the measuring device, test procedure, and environmental conditions. In this work, an innovative finger-trigger interface device was developed to facilitate accurate trigger pull measurements. The idea was to reduce the variation related to the position of the measurement device on the trigger in existing measuring methods and devices. Three force sensors based on different technologies were initially evaluated. While two of the three sensors failed to produce data, the miniature capacitive plate sensor exhibited high precision and a linear response over the range of typical trigger pulls. To examine the effects of the finger-trigger interface on trigger pull measurement, different sensor housing prototypes were designed *in silico* and 3D printed for the construction of three finger-trigger interface devices. The performance of each finger-trigger interface device was evaluated by measuring the trigger pulls of several selected firearms and comparing the data to a previously published study. Our preliminary results demonstrated the novel finger-trigger interface device offered a new way to measure trigger pull *in situ* with acceptable accuracy and precision.

KEYWORDS: firearms examination, trigger pull, finger-trigger interface, force sensor, 3D printed device, unintentional discharge

Trigger pull is the amount of force that must be applied to the trigger to release the sear causing a firearm to discharge. Trigger pull varies based on the type of trigger action. Single action firearms typically have trigger pulls in the range of 3.5–5.0 lbf (pounds of force), whereas the trigger pull of double action firearms can be in the range of 5.0–12.0 lbf (1). Note that single action and double action trigger pull measurements are not necessarily confined to these ranges and can fall outside these ranges; that is, some single action trigger pull measurements are greater than double action trigger pull measurements for the same firearm, and trigger pull measurements for some cylinder positions of double action revolvers may be 12.0 lbf. Firearms examiners measure trigger pull to help determine the operating condition of a firearm, expose possible alterations in the firing mechanism, characterize the design of specific makes and models of firearms, and serve as an aid in criminal cases involving unintentional or accidental discharge (2). The first three uses of trigger pull deal with recognizing abnormalities that can aid in examining the function of a firearm, but the fourth use is applied in court to provide weight to existing evidence in cases involving unintentional discharge. The common argument is that firearms with heavier trigger pulls require greater intent to pull the trigger than those with lighter trigger pulls (3).

Although trigger pull is used in court cases involving unintentional discharge, research of the statistical analysis of trigger pull and its relation to unintentional discharge has not been completed. Most research concerning unintentional discharge focuses on the situational characteristics of past incidents (4,5) and examines how involuntary muscle contractions from sympathetic movements, loss of balance, and startle reactions can result in pulling the trigger unintentionally (5–8).

Currently, four methods of testing are available to the firearms examiner for the measurement of trigger pull: the use of dead weights, spring gauges, force gauges, and automated trigger pull devices (9,10). Dead weights and spring gauges were the first two methods considered for the determination of trigger pull and have been adopted for a long time. The use of dead weights involves placing a hook over the trigger, while the firearm is held vertically and adding calibrated weights to the other end of the hook until the trigger releases the sear. For the spring gauge method, a spring is attached to the trigger via a hook, while the firearm is immobilized and in the horizontal position. The spring is pulled rearward, and the gauge attached to the spring is observed until the trigger releases the sear. Comparison of these two methods has revealed that the dead weight method is more accurate but is also susceptible to variations from outside factors (11). Note that the study was preliminary and had a limited scope of four firearms with ten measurements using each device. The greatest variation in both methods has been observed when varying the position of the hook or spring on the trigger (11–13). Since the dead weight method is more accurate, it became the default method used by laboratories to collect trigger pull data for reference databases (14–16).

[1] Department of Forensic Science, College of Criminal Justice, Sam Houston State University, Huntsville, TX, 77340.

[2] Firearms Identification Laboratory, Harris County Institute of Forensic Sciences, Houston, TX, 77054.

[3] Department of Engineering Technology, Sam Houston State University, Huntsville, TX, 77340.

Corresponding author: Jorn Yu, Ph.D. E-mail: jornyu@shsu.edu

Received 5 May 2020; and in revised form 25 June 2020; accepted 16 July 2020.

The rise of technology introduced digital force gauges as another option for measuring trigger pull. These force gauges are operated by placing the end of a rod extending from the measuring device against the trigger of a firearm that is mounted horizontally in a vice. The measuring device is pulled rearward, exerting pressure on the trigger. The peak force is displayed on the digital display once the sear is released. Accuracy of the digital force gauge method can be established by using certified weights (12). A recent study examining the uncertainty associated with a digital force gauge showed significant differences in trigger pull means obtained between participants when measuring the same firearm (17).

At the present time, the only computer-operated automated trigger pull device available to firearms examiners is the TriggerScan™ System by Dvorak Instruments (Tulsa, OK). The system consists of a pair of moveable and fixed arms that sit between the trigger and trigger guard. A firearm is mounted on the system's adjustable mounting rail. The fixed arm rests against the forward portion of the trigger guard, while the moveable arm is compressed against the trigger by a stepper motor. As the trigger is depressed, a force sensor measures the force applied to the trigger every five ten-thousandths of an inch it travels and a microcontroller collects and sends the data to a computer. A graphical representation of force versus trigger travel distance is displayed by the instrument's software. An evaluation of the TriggerScan™ system described the instrument as capable of measuring trigger pull with a tolerance of +/− of 0.1 pounds and a resolution capability of 0.0007 pounds (2). This is only true for measurements made on the same firearm that has not been repositioned between tests; accuracy testing has revealed that TriggerScan™ remains susceptible to the same issue as the traditional testing methods regarding the position of force on the trigger (10,13,18). A standardized device that offers more accurate and precise trigger pull measurements should be developed first so the statistical analysis of trigger pull and its relation to unintentional discharge can be investigated.

In this project, three types of miniature force sensors were assessed for precision, accuracy, and functional design. The sensors examined were a capacitive plate sensor, strain gauge load cell, and a piezoelectric ceramic disk. The goal was to create a finger-trigger interface device incorporated with a miniature force sensor that is compact, computer-based and can maintain the desired precision and accuracy regardless of its position on the trigger of any given firearm. We hypothesized that the precision and accuracy of trigger pull measurements could be improved by reducing the issue of force placement on the trigger.

Materials and Methods

Force Sensors

Three different types of force sensors were examined, where each force sensor's mode of input was based on a different electrical property. The first type of sensor was a capacitive plate force sensor (S8-100N; SingleTact by Pressure Profile Systems, Glasgow, U.K.). Capacitive plate force sensors contain two conductive plates that are separated by a dielectric material. Force applied to the sensor causes the distance between the plates to decrease resulting in an increased capacitance. The SingleTact sensor used in this study was 0.35 mm thick with a diameter of 8.0 mm and a reported force measurement capacity of 22 pounds with an error of <1%. The second type of sensor examined was a thin beam load cell (LCL-020; OMEGA, Norwalk, CT) capable of measuring up to 40 pounds with a reported error of 0.25%. The load cell has deformable metal beams with a series of electrical resistors called strain gauges attached to their surfaces. As a force is applied to the load cell, the metal beams deform putting tension or compression on the attached strain gauges. The deformation of the strain gauges cause an increase or decrease in electrical resistance depending on the type and magnitude of the force applied. The third type of sensor was a piezoelectric ceramic disk (SMD05T04R111WL; STEMiNC, Davenport, FL). Piezoelectric materials generate an electrical voltage when an external force is applied due to electron displacement in the material. Ceramics such as barium titanate and lead zirconate Titanate are considered strong piezoelectric materials because they have an asymmetric polycrystalline structure that exhibits polarization at normal operating temperatures (19,20). The piezoelectric disk used in this project has a piezoelectric coefficient of 320 picocoulombs per Newton. Table 1 lists the advantages and disadvantages for each of the sensor types.

Sensor Evaluation

Each sensor was evaluated using different weight plate combinations between 2.8 and 19.8 pounds. The voltage difference for each weight combination was measured using a digital multimeter (HT118A; Kaiweets, Shenzhen, China). The voltage differences were recorded in Excel (Office 360; Microsoft, Redmond, WA), and the correlation between the electrical voltage and the applied load was determined by means of linear regression. The voltage produced by each weight was measured six times, and the precision and bias of the sensor were calculated.

Finger-Trigger Interface Design

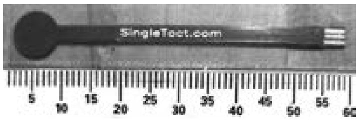
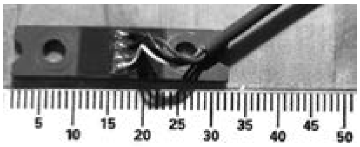
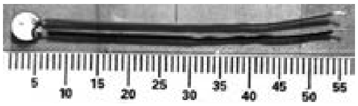
Three finger-trigger sensor housings were designed to interface with the miniature force sensors selected for this project. The design of the finger-trigger sensor housings focused on compatibility with different trigger shapes and sizes, user interaction, sensor function, and repeatability of placement on the trigger. The purpose of the sensor housing is to facilitate the sensor-finger and sensor-trigger interactions. Computer-aided 3D design software (Inventor 2018; Autodesk, San Rafael, CA) was used to create 3D models of components and their assemblies. The prototype of these components was printed using a Form 2 3D printer (PKG-F2; Formlabs, Somerville, MA). The printing material was Tough 2000 (RS-F2-TO20; Formlabs). This material was chosen for its comparable strength and stiffness of ABS plastic material.

Device Evaluation

Once the sensor interfaces were prototyped, the assembled devices were used to measure the trigger pull of various firearms. Measurements were recorded by connecting the chosen force sensor to an Arduino Uno (Uno R3; LAFVIN, Shenzhen, China) microcontroller that was connected to a computer as shown in Fig. 1. A simple program was developed to read the analog value of the voltage produced by the force sensor every 100 ms, and the data were subsequently transferred to an Excel worksheet using a Data Acquisition Macro for Excel (PLX-DAQ; Parallax Inc, Rocklin, CA). The calibration curve in Fig. 2 was used to calculate the force in pounds using the maximum voltage at the time the trigger was pulled.

The firearms used in the study are outlined in Table 2 and were borrowed from the Harris County Institute of Forensic

TABLE 1—*Summary of sensor evaluation.*

Sensor	Advantages	Disadvantages	Results	Conclusions
Capacitive Plate Sensor  (SingleTact, S8-100N)	• Low profile • Inexpensive • Well-developed and supported	• Area dependent • Plates must remain parallel across whole surface area	• Linear relationship between applied force and output voltage was observed	• Met the desired specifications for precision and reliability
Thin Beam Load Cell  (OMEGA, LCL-020)	• Used in weight balances • High precision • Easy to use	• Size limits interface design	• The sensor that was received had a zero-mass reading of 2.5 V which was the maximum output voltage	• Sensor was faulty, opted against a replacement due to size limitations
Piezoelectric Ceramic Disk  (STEMiNC, SMD05T04R111WL)	• First order relationship • Ceramic is inexpensive	• Ceramic is brittle	• Sensors cracked when exceeding 10 pounds	• Ceramic is too brittle, quartz or an enclosed unit sensor may work better

Sciences. This was done so that the trigger pull measurements collected in this study could be compared to the results of the Alvarez Bacha study (17) to determine accuracy. In the Alvarez Bacha study, a total of 15 firearms (5 pistols, 3 revolvers, and 7 long guns) were used. For this preliminary study, 9 firearms (4 pistols, 2 revolvers, and 3 long guns) were available and selected based on the types of firearms seen regularly in firearms identification laboratories. However, during testing, the trigger pulls of 4 firearms could not be measured due to the limited space between the trigger guard and trigger.

For each firearm, five measurements were collected by each device for each firearm by one user. This differed from the Alvarez Bacha study which collected ten measurements for each firearm by multiple users. Five measurements were chosen since this is a preliminary study, and unpaired *t*-tests were used for comparison of the average trigger pulls from the two methods since the number of measurements was not kept the same. Additionally, because the Alvarez Bacha study had multiple individuals conducting trigger pull measurements, higher variations of the standard deviation and %RSD were expected and these parameters could not be directly compared to the results of this study. For the purpose of device evaluation, the Alvarez Bacha data were used primarily to assist with the evaluation of the accuracy of the finger-trigger interface devices.

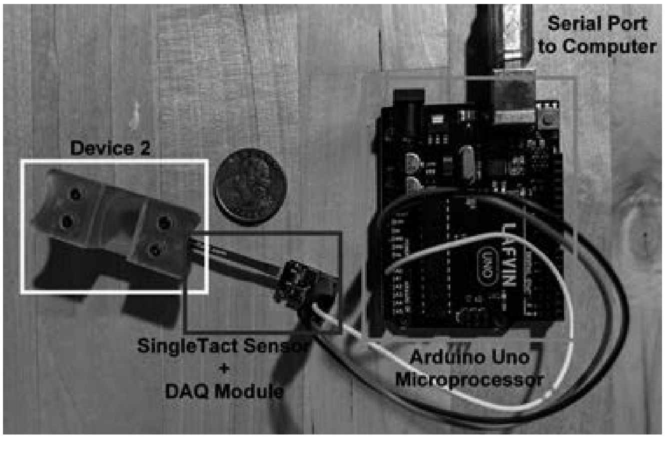


FIG. 1—*Setup for device evaluation. SingleTact sensor was connected to the microprocessor via an I2C board as detailed in the sensor's user manual (21). [Color figure can be viewed at wileyonlinelibrary.com]*

Results and Discussion

Sensor Evaluation

Of the three sensors tested, the capacitive plate sensor performed the best for its accuracy and precision. The thin beam load cell and piezoelectric ceramic disk sensors failed during

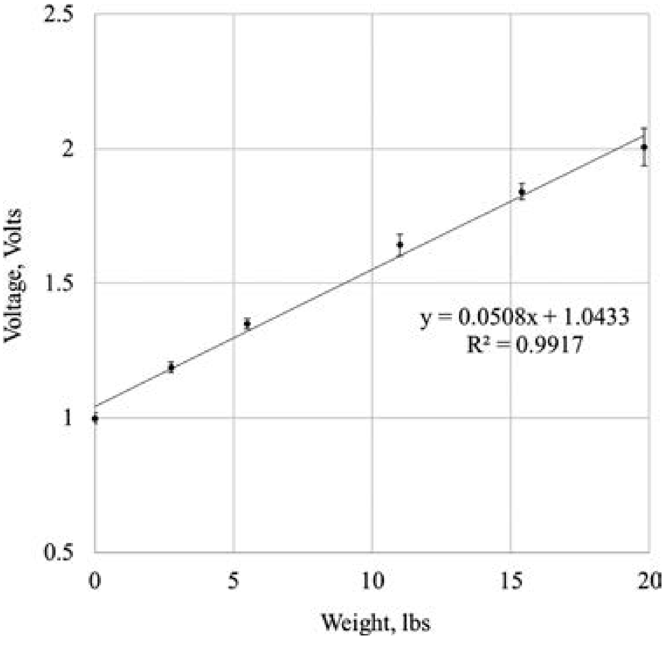


FIG. 2—*Linear regression model for capacitive plate sensor* (n = 6).

TABLE 2—*Firearms of different calibers, makes, models, and actions selected for testing*.*

Caliber	Make	Model	Type	Serial Number	Action
223 Rem	Colt	AR15-SPI	Rifle	SP176984	SA
20 Gauge	Winchester	1300 XTR	Shotgun	GS5584	SA
9mm Luger	Glock	17	Pistol	ATE442US	DA
9mm Luger	Ruger	P89	Pistol	309-77865	SA/DA
45 Auto	Sig Sauer	P220	Pistol	G241628	SA/DA

*Firearms were selected from the reference collection at Harris County Institute of Forensic Sciences.

TABLE 3—*Precision and bias for capacitive plate sensor (*n = *6).*

Weight, lbs	Precision	Bias					
0.00	2%	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6
2.76	2%	0%	5%	3%	5%	2%	4%
5.5	1%	16%	6%	6%	12%	10%	7%
11	2%	4%	10%	9%	8%	8%	5%
15.4	2%	2%	4%	1%	3%	4%	-1%
19.8	4%	-4%	-6%	-4%	-5%	-5%	-1%

the evaluation. Specifically, the thin beam load cell had a zero-mass reading of 2.5 V, which was the maximum output voltage, making the dynamic range of this sensor incompatible with the range for the trigger pull measurements. Further, the thin beam load cell had a size larger than three centimeters, which was too large for the construction of a reasonably sized finger-trigger interface device. As for the piezoelectric ceramic disk sensor, it was too brittle and cracked when exceeding ten pounds. Table 1 provides a summary of the observations made for each sensor. Evaluation of the capacitive plate sensor revealed that the correlation between the applied weight and resulting voltage was linear with a regression coefficient of 0.9917. The calibration curve and calibration range are shown in Fig. 2. After the calibration curve was constructed, each standard weight was remeasured six times for the evaluation of precision and bias. The precision was calculated using the relative standard deviation (RSD), which is the percentage of the value of standard deviation divided by the mean. The bias was calculated by subtracting the expected value of the standard weight by the measured value. A positive bias means the measured weight was higher than the expected weight (true weight). Likewise, a negative bias means the measured weight was lower than the expected weight. The precision and bias for several measurements are shown in Table 3. The precision for each weight used over a period of 6 measurements was <10%, and the bias for each weight used over a period of 6 measurements was <20%.

Finger-Trigger Interface Design

Sensor interface design focused on three goals. The first goal was to ensure the repeatability of placement on the trigger for a wide range of firearms. The second goal was to take into consideration limitations of the sensor. For example, in the case of the SingleTact capacitive plate sensor, the sensor needed to be seated between two parallel flat surfaces that have the same diameter as the sensor. The final goal was to keep the device compact so that it could be used while pulling the trigger. The devices were modeled after trigger shoes that were traditionally attached to triggers to distribute the force and minimize the weight of the trigger pull experienced by shooters.

As shown in Fig. 3*a*, Device 1 provided a consistent trigger sensor interaction. This was achieved by using spring-supported rods to push against the trigger so that the full trigger would be in contact with the device. The intention of this design was also to have the force evenly distributed across the whole trigger to eliminate the issue of device placement affecting the trigger pull. The bottom image of Fig. 3*a* shows the device while being used to measure trigger pull. As shown in Fig. 3*b*, Device 2 provided more consistent finger-sensor interaction. A curved finger guide was designed for the finger to rest on, while the trigger was

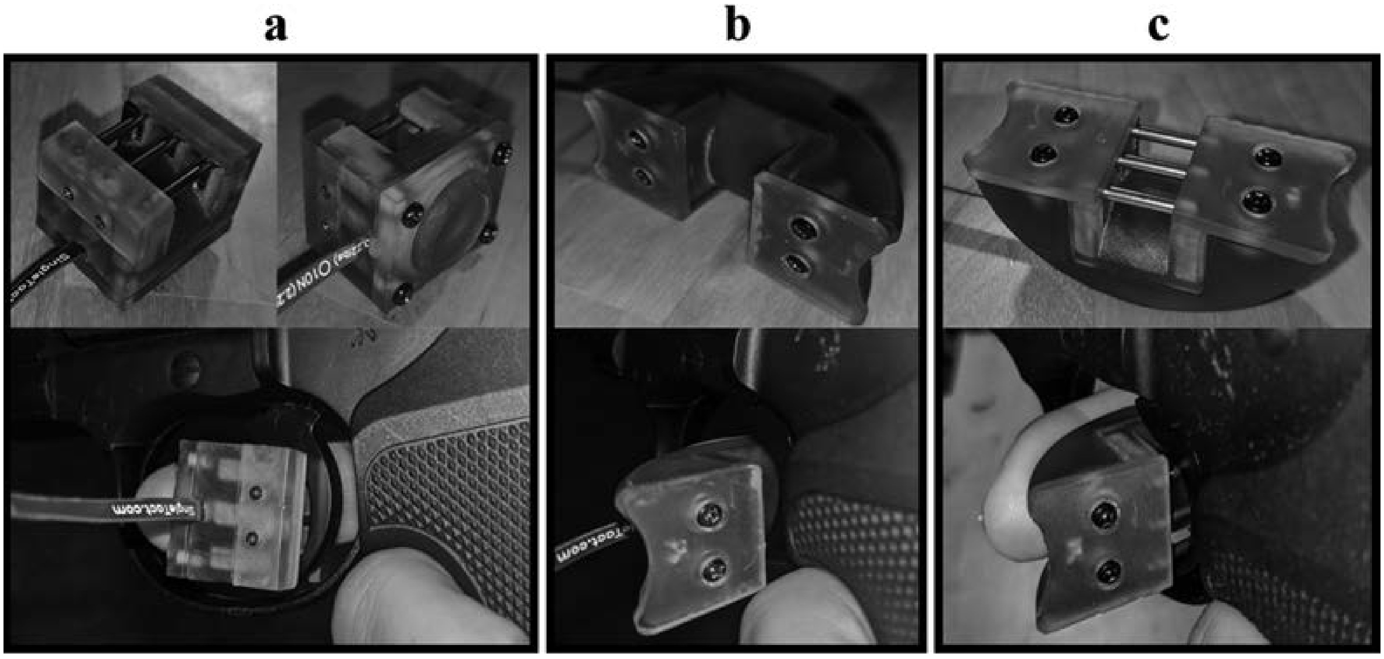


FIG. 3—*Final device designs fully assembled (top) and in use (bottom). Device 1 (a), Device 2 (b), Device 3 (c). [Color figure can be viewed at wileyonline library.com]*

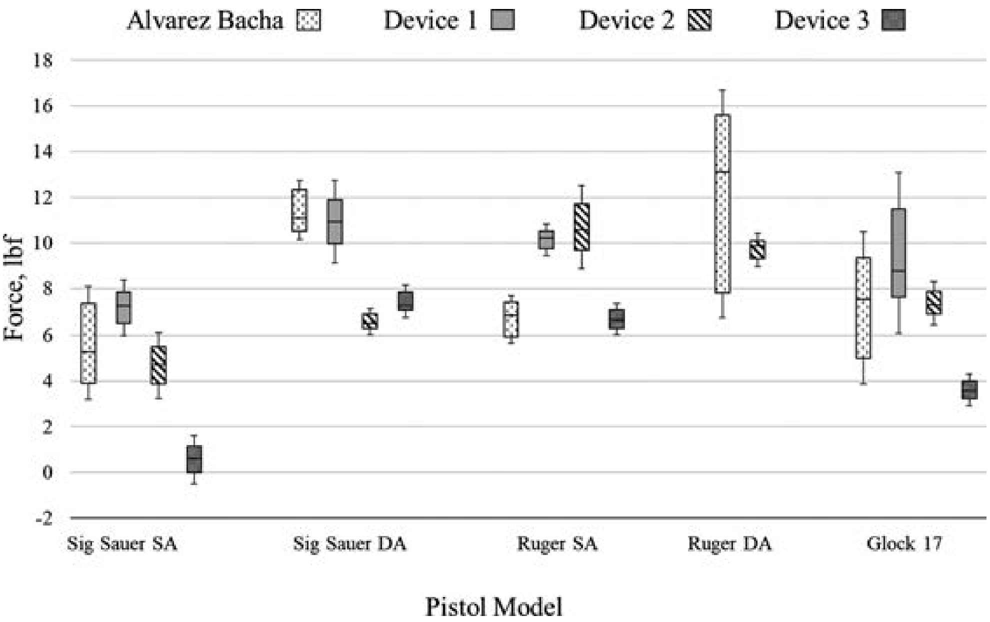


FIG. 4—*Comparison of trigger pulls measured by each device to trigger pulls measured from the Alvarez Bacha study for several pistols* (n = 5)*. Each vertical box is the maximum and minimum with the mean represented by the line dividing the box. The bars extending from the boxes are the standard deviations. Data for the Ruger DA using Device 1 and Device 3 are not included. Device 1 broke during measurement, and Device 3 was not used for the trigger pull measurement of the Ruger DA to prevent damage to the device since it shared a similar design to Device 1.*

TABLE 4—*Trigger pull data for each device compared to data from the Alvarez Bacha study*.*

		Alvarez Bacha Study	Device 1	Device 2	Device 3
Sig Sauer SA	Average	5.29	7.25	4.72	0.57
	STDEV	0.72	0.53	0.64	0.46
	%RSD	14%	7%	14%	80%
Sig Sauer DA	Average	11.1	10.94	6.53	7.26
	STDEV	0.38	0.84	0.26	0.33
	%RSD	3%	8%	4%	5%
Ruger SA	Average	6.83	10.23	10.58	6.65
	STDEV	0.28	0.32	0.81	0.29
	%RSD	4%	3%	8%	4%
Ruger DA	Average	13.12	N/A	9.86	N/A
	STDEV	1.08	N/A	0.33	N/A
	%RSD	8%	N/A	3%	N/A
Glock 17	Average	7.54	8.80	7.29	3.57
	STDEV	1.13	1.57	0.43	0.32
	%RSD	15%	18%	6%	9%
AR15-SPI	Average	6.55	N/A	12.07	11.23
	STDEV	0.44	N/A	0.39	0.22
	%RSD	7%	N/A	3%	2%
20 Gauge Winchester	Average	7.26	N/A	8.78	7.73
	STDEV	0.61	N/A	0.44	0.29
	%RSD	8%	N/A	5%	4%

*Averages and standard deviations have the units of force pounds, lbf.

depressed. The trigger shoe portion of the device acted as a button that the trigger would depress, transferring the force to the sensor. As shown in Fig. 3*c*, Device 3 combined the features of Device 1 and Device 2, so that both the trigger sensor and finger-sensor interactions would be consistent.

Device Evaluation

To evaluate the performance of each finger-trigger interface device, a box plot was constructed for the data obtained from the Alvarez Bacha study and all trigger pull measurements obtained using the three prototype devices in this project. As shown in Fig. 4, the measurements obtained by Device 1 were consistent with the measurements from the Alvarez Bacha study for three of the four pistols tested. The average readings from Device 1 were slightly higher than the readings from the Alvarez Bacha study (per Table 4—except for the double action of the Sig Sauer pistol), suggesting that the design of Device 1 could distribute the applied force across the trigger. Note that in the Alvarez Bacha study, the applied force was centralized to one location on the trigger where the shooter's index finger would typically rest. The %RSD (Percent Relative Standard Deviation) for Device 1, shown in Table 4, is below 10% for all measurements except for the Glock 17. For the single action mode of the Sig Sauer pistol, the %RSD is the lowest out of all the modes of measurement, making it the most precise for that pistol (per Table 4, the single action Sig Sauer %RSD for Device 1 is 7% versus 14% and 80% for Devices 2 and 3, respectively). During testing, Device 1 broke while measuring the double action trigger pull of the Ruger P89. This could have been an issue in the design of the spring retainers. Since Device 3 and Device 1 shared a similar design, Device 3 was not used for the double action trigger pull measurement of the Ruger P89 to prevent damage. The trigger pull measurement for long guns was not able to be completed by using Device 1 since it broke before the trigger pulls for the long guns were measured. Design improvements for Device 1 would include better reinforced spring retainers and decreasing the size of the sensor housing.

As shown in Figs 4 and 5, the trigger pull measurements obtained by Device 2 are consistent with the data obtained from the Alvarez Bacha study for four of the seven firearms tested. The average readings for Device 2 did not show systematic error when compared to the trigger pull averages obtained from the Alvarez Bacha study. The %RSD for Device 2, shown in Table 4, is below 10% for all measurements except for the single action mode of the Sig Sauer pistol.

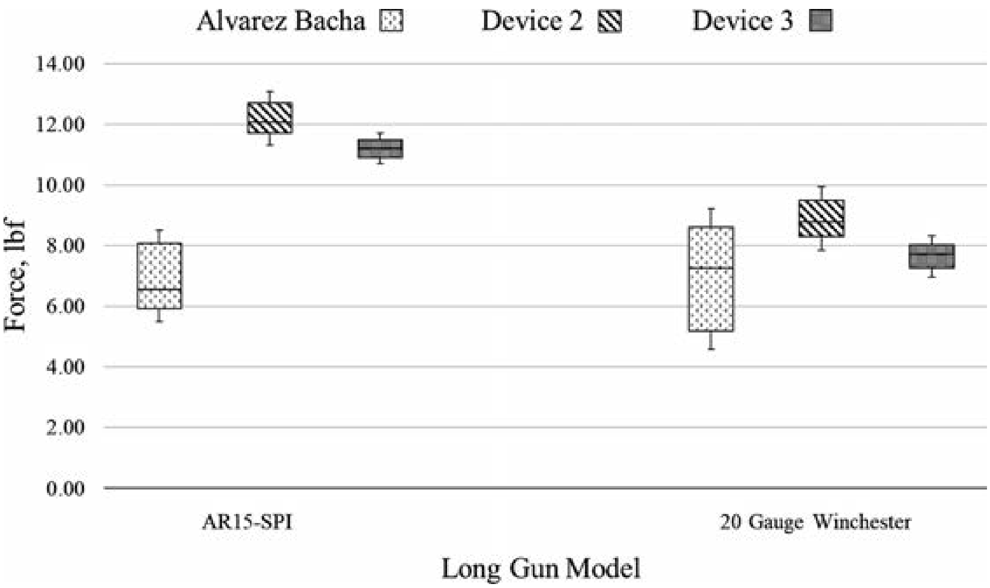


FIG. 5—*Comparison of trigger pulls measured by each device to trigger pulls measured from the Alvarez Bacha study for a rifle and shotgun (*n = 5*). Each vertical box is the maximum and minimum with the mean represented by the line dividing the box. The bars extending from the boxes are the standard deviations.*

As for Device 3, the trigger pull measurements are consistent with the data obtained from the Alvarez Bacha study for only two out of the six firearms tested. The average trigger pull measurements obtained for Device 3 were lower than those obtained from the Alvarez Bacha study (per Table 4, except for the two long guns). The results suggested that there might be a systematic error associated with the accuracy of Device 3, which was outperformed by the other two devices in terms of the precision of the trigger pull measurement.

For Device 1, the sensor was affixed to the portion of the housing that interacted with the trigger, while the button compressing the sensor was on the side of the finger interaction. For Devices 2 and 3, the sensor was affixed to the portion of the sensor that the finger rested on and the button compressing the sensor was on the side of the trigger. To determine whether this influences measurement, another device could be designed that is the same as Device 1 or 2 but changes the way the button interacts with the sensor. Additionally, Device 1 could be simplified by removing the spring-supported rods and this could be compared to Devices 1 and 2 as well. By continuing to test additional designs, the best combination of design features can be used to create a device that has the ideal accuracy and precision.

It is important to note that all testing in this study was done by one person, and a future step would be to compare the measurements between different users. Measuring devices that allow the trigger pull to be measured while actively pulling the trigger could show that the force required to pull the trigger may differ between individuals. The novel finger-trigger interface may be beneficial to investigators to measure the trigger pull, while a suspect pulls the trigger of a given firearm. This could provide an improved interpretation of trigger pull in court.

Conclusion

In this project, several miniature force sensors were evaluated for the construction of a finger-trigger interface device to measure trigger pull. Of the force sensors evaluated, the capacitive plate force sensor maintained a linear relationship between applied force and output voltage for the range of zero to twenty pounds. A housing for the sensor allowing for the measurement of trigger pull *in situ* was designed, and prototypes were constructed. The three prototype devices were used to measure the trigger pull of several firearms, and the results were compared to a previous trigger pull study (17). Results showed that the precision for all three devices was improved when compared to that of the digital force gauge used in the Alvarez Bacha study. The results have also demonstrated that it is possible to design a device that allows for the measurement of trigger pull while actively pulling the trigger with the user's finger. Future work would increase the number of firearms tested and address some of design issues encountered during testing.

Acknowledgments

The support of the advanced 3D printing devices by the Department of Engineering Technology at Sam Houston State University is acknowledged. The summer internship opportunity provided by the Firearms Identification Laboratory at Harris County Institute of Forensic Sciences is greatly appreciated. The authors appreciate Lucio Avellaneda for his assistance in data collection.

References

1. Paradis P, Hendrick HW. Accidental shootings: gun design and training issues. In: *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting*. Los Angeles, CA: SAGE Publications, 2001;823–7. https://doi.org/10.1177/154193120104501102
2. Dillon JH. The triggerscan system – microprocessor technology applied to precision trigger pull analysis. AFTE J 1999;31(2):123–30.
3. Hendrick HW, Paradis P, Hornick RJ. Human factors issues in handgun safety and forensics. Boca Raton, FL: CRC Press, 2008;61–2.
4. Petersson U, Bertilsson J, Fredriksson P, Magnusson M, Fransson P-A. Police officer involved shootings – retrospective study of situational characteristics. Police Pract Res 2017;18(3):306–21. https://doi.org/10.1080/15614263.2017.1291592
5. O'Neill J, Hartman ME, O'Neill DA, Lewinski WJ. Further analysis of the unintentional discharge of firearms in law enforcement.

Appl Ergon 2018;68:267–72. https://doi.org/10.1016/j.apergo.2017.12.004

6. Enoka RM. Involuntary muscle contractions and the unintentional discharge of a firearm. Law Enforc Exec Forum 2003;3(2):27–39.
7. Heim C, Schmidtbleicher D, Niebergall E. The risk of involuntary firearms discharge. Hum Factors J Hum Factors Ergon Soc 2006;48(3):413–21. https://doi.org/10.1518/001872006778606813
8. Kee S, Jing Sok Y, Alaric CW. Evaluating the likelihood of unintentional discharge of handguns due to sympathetic contractions. AFTE J 2019;51(1):30–3.
9. SWGGUN. Guidelines for trigger pull analysis. AFTE J 2008;40(2):219–20.
10. Firearms & Toolmarks Subcommittee of the Organization of Scientific Area Committees (OSAC) for Forensic Science. Best practice recommendation for measuring trigger pull of a firearm and estimating its uncertainty. Gaithersburg, MD: NIST, 2019;1–16.
11. Rios F, Thornton J. Static vs. dynamic determination of trigger pull. AFTE J 1984;16(3):84–7.
12. Lawrence GR, Lee H. The effect of hand grip angle on the measurement of trigger pull forces. AFTE J 2011;43(2):154–61.
13. Koffman A, Argaman U, Silverwater H, Hocherman G, Shoshani E. Triggerscan computerized trigger pull system. AFTE J 1999;31(4):449–56.
14. Lomoro VJ. A statistical analysis of trigger pulls. AFTE J 1986;18(1):35–48.
15. Krylo J, Slonina S. Trigger pull statistics. AFTE J 1985;17(1):79–92.
16. Wilson WH, Turbok RD. Trigger pull data. AFTE J 2003;35(4):400–30.
17. Alvarez Bacha C, Cavelier D. Imada DS2-44 digital force gauge: trigger pull measurements and associated uncertainty of measurements. AFTE J 2019;51(2):68–91.
18. Cunningham J. Accuracy testing on dvorak instruments' trigger-scan system. AFTE J 2000;32(4):364–6.
19. Garcia RE, Carter WC, Langer SA. The effect of texture and microstructure on the macroscopic properties of polycrystalline piezoelectrics: application to barium titanate and PZN-PT. J Am Ceram Society 2005;88(3):750–7. https://doi.org/10.1111/j.1551-2916.2005.00109.x
20. Araujo EB, Lima EC, Bdikin IK, Kholkin AL. Thickness dependence of structure and piezoelectric properties at nanoscale of polycrystalline lead zirconate titanate thin films. J Appl Phys 2013;113:187206. https://doi.org/10.1063/1.4801961
21. Pressure Profile Systems. SingleTact miniature force sensors user manual. Version 2.3. Glasgow, U.K.: Pressure Profile Systems, 2017. https://www.singletact.com/SingleTact_Manual.pdf (accessed November 24, 2019).

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14522
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Cassandra Kapsa,[1] *H.B.Sc.; Michael Ho*,[1,2] *H.B.Sc.; and Meadow Libby*,[1,2] *H.B.Sc.*

The Use of Liquid Latex to Recover Latent Fingerprints that are Covered in Debris from Exterior Glass Surfaces of Vehicles

ABSTRACT: The purpose of this research is to determine if latent fingerprints deposited on the exterior glass surfaces of vehicles, then covered in debris, can be recovered. Past research used liquid latex to lift soot to recover trace evidence. Recently, liquid latex has been used to recover latent fingerprints along the bottom of vehicles. In this study, a total of 216 latent fingerprints were deposited on the exterior windows of three vehicles. Three control and three experimental latent fingerprints were placed on each side window. The vehicles collected debris for either 2, 3, or 4 weeks. After debris collection, liquid latex was applied to the experimental sections. The underlying fingerprints were developed with white granular powder. Control fingerprints were developed directly with white granular powder. A chi-square test revealed a significant difference in fingerprint recovery between the control and liquid latex method ($X^2 = 9.026$, d.f. = 1, $p = 0.003$). An odds ratio determined that the control method increases the probability of latent fingerprint recovery by 2.68. Fisher's exact test indicated that there is no statistically significant difference between the detail of the recovered control and experimental fingerprints ($p = 0.065$). This study demonstrates that recovery of fingerprints is possible using the liquid latex method; however, the control method recovers more fingerprints on the glass exterior of vehicles. If latent fingerprints are thought to be present on the exterior glass surfaces of vehicles, the control method should be used to improve vehicle processing by investigators.

KEYWORDS: forensic identification, Bandey scale, debris, latent fingerprints, liquid latex, vehicles

The purpose of this research is to determine if latent fingerprints deposited on exterior glass surfaces of vehicles, then subsequently covered in debris, can be recovered. This research was accomplished using liquid latex to lift the debris, and white granular powder (a fine fingerprint powder) to develop the fingerprint(s) underneath. The quality of the recovered fingerprints was then analyzed based on ridge detail. This research is significant because if successful, this technique will aid in identifying the possible person(s) of interest at a crime scene involving a vehicle. The exterior glass surfaces of vehicles are of specific interest because they are a common area of the vehicle that a suspect can be in contact with as it is near the doors of the vehicles. This research contributes to an expanding literature of the application of liquid latex in a forensic context.

To test the feasibility of using liquid latex to remove debris that is covering fingerprints deposited on exterior vehicle glass, latent fingerprints were deposited on the exterior windows of a vehicle that was naturally covered in debris through exposure to various environmental conditions (e.g., rain and snow) (1). In this case, debris is considered "loose natural material" (2). Some researchers have stated that the debris that collects on the vehicle's surface covers and protects the fingerprints underneath (1).

At arson scenes, soot was an issue that impacted evidence recovery as the soot concealed trace evidence (3). Many methods were developed to try to solve this problem, such as cleaning the soot with water; however, they tended to disrupt or destroy the underlying evidence (3). Larkin et al. (4) were the first to successfully use liquid latex to lift soot to recover the fingerprint evidence from a room that had been set on fire in a homicide investigation. Since this first publication, other researchers have also tested liquid latex to lift soot to recover fingerprints and other evidence on surfaces exposed in a fire. Clutter et al. (5), for example, found that liquid latex was not successful in recovering the sample fingerprints that were covered in soot, but was successful in recovering blood spatter patterns covered in soot. The success of liquid latex to remove overlying material suggests that it may also prove useful in removing other types of debris.

Recently, Ho and George (1) conducted a study involving the use of liquid latex on vehicle exteriors to lift debris and recover the underlying latent fingerprints. They determined liquid latex to be successful in lifting the debris to recover the fingerprints (1). That study, however, tested the use of liquid latex on the painted metal exterior of the vehicle instead of the glass surfaces (1). In contrast, Grossi et al. (6) conducted a study regarding a comparison of the liquid latex method and tape method to remove debris and recover fingerprints on glass windowpanes, not vehicles. The tape method is used to lift powdered fingerprint impressions by placing transparent tape onto the

[1] Forensic Science Department, University of Toronto - Mississauga, 3359 Mississauga Road, Mississauga, Ontario, L5L 1C6, Canada.

[2] Forensic Services Branch, Hamilton Police Association, 155 King William St, Hamilton, Ontario, L8R 1A7, Canada.

Corresponding author: Cassandra Kapsa, H.B.Sc. E-mail: cassandra.kapsa@mail.utoronto.ca

Received 11 May 2020; and in revised form 18 June 2020; accepted 8 July 2020.

impression, rubbing the tape to remove air bubbles, then lifting the tape and placing it on a backing card (7). That study had no success in recovering fingerprints with the liquid latex method (6). In this research, latent fingerprints were deposited on the exterior glass surfaces of vehicles to test latent fingerprint recovery using liquid latex in order to clarify the above discrepancies in the literature regarding liquid latex.

Materials and Methods

The total sample for this research was 216 latent fingerprints divided into the following (Table 1):

TABLE 1—*Fingerprint sample breakdown.*

Trial	Variable (Weeks)	Experimental	Control
1	3	18	18
2	4	18	18
3	2	18	18
4	3	18	18
5	4	18	18
6	2	18	18

Three 2015 white Dodge Grand Caravans in similar wear conditions, with no damage to the windows, were provided by the Hamilton Police Service for the duration of the study. The vehicles are driven daily by officers in the execution of their duties. Prior use did not affect the fingerprint deposition or collection because the glass was cleaned before fingerprint deposition.

Each vehicle was given a trial number and each trial number was assigned a different length of time for debris collection, 2, 3, and 4 weeks, as these time frames allowed for appropriate debris collection during the research timeframe. The different lengths of time for debris collection exposed the fingerprints to various environmental situations and tested the liquid latex's ability to lift various amounts and types of debris. During the study period, the three vehicles were intentionally not washed (e.g., car wash or manual washing) but were exposed to rain or other precipitation that could potentially wash away the latent fingerprints. The vehicles were not protected from the elements in order to mimic real-life conditions.

The windshield and back window were not examined as part of this study because the window wipers may distort and/or wash away the fingerprints and fingerprints are not typically found on these windows. Only the three windows on each side of the vehicle were tested. Both the control and experimental fingerprints were placed on the left side (driver's side) and the right side (passenger's side) glass surfaces of each vehicle to control for the possible exposure differences of each side of the vehicle. Six fingerprints were deposited onto each window: three control and three experimental. Of the six fingerprints, three were on the top half of the window and three were on the bottom half. The placement of the three control and experimental fingerprints on either the top or the bottom of the window alternated and the placement of the fingerprints on each side of the vehicle is opposites (Fig. 1). Overall, 18 fingerprints were deposited on each side of the vehicle (nine control and nine experimental per vehicle) for a total of 36 fingerprints per vehicle. Six latent fingerprints per window were utilized because it is expected that there would be little variation between the debris collected on these windows. Even though there are six fingerprints placed on each window, all 18 fingerprints are essentially the same as they are all sebaceous fingerprints, placed by the same person and are on the same side of the vehicle. Each time period (2, 3, and 4 weeks) was tested twice to increase the likelihood of exposing the vehicles to various weather conditions and debris. It was expected that the weather conditions would vary between November 2019 and February 2020. Since there were three variables and each variable was completed twice (six trials in total), there were a total of 216 fingerprints deposited on to the exterior glass surfaces.

The experimental fingerprints were the fingerprints that were uncovered using white liquid latex (Amscan, Elmsford, NY) and then developed using white granular powder (Lynn Peavey Company, Lenexa, KS). The control fingerprints were the fingerprints that were not treated with liquid latex and were exclusively developed with white granular powder over the debris. The control fingerprints mimic the recovery method that is currently used by the Hamilton Police Service and other police services when collecting fingerprint evidence on the exterior glass surfaces of vehicles. White granular powder was used in this analysis instead of black and gray powder because it provided the best contrast on the glass. Cobalt Elite nitrile gloves (Maxill, St. Thomas, ON, Canada) were worn when developing and lifting all fingerprints.

The experimental procedure followed the protocol by Ho and George (1). Before placing the fingerprints, the vehicles were washed using water, Armor All® All Purpose Car Wash (Armored AutoGroup, Danbury, CT), and a brush to mimic everyday washing. The glass was cleaned before fingerprint deposition because this research focused on the debris collecting on top of the latent fingerprints, not latent fingerprints being deposited on debris. Likewise, the International Fingerprint Research Group (8) recommends using 1–3 clean substrates in Phase 1 pilot studies. The exterior surface glass was then examined with oblique lighting from a flashlight to ensure that no fingerprints were present on the surface after cleaning. Sebaceous fingerprints, also known as groomed fingerprints, were deposited onto the glass by one donor after they rubbed their finger(s) on oily areas of their face (e.g., nose and forehead) (1). The use of one donor and groomed fingerprints opposes the guidelines outlined by the International Fingerprint Research Group (8) in which it is recommended that Phase 1 Pilot Studies use 3–5 donors and natural instead of groomed fingerprints. One donor was selected to deposit fingerprints to control for the variation between individuals as this research focuses on the baseline effectiveness of the liquid latex method and not the effectiveness with a variety of donors (1). Sebaceous/groomed fingerprints were used to ensure that an adequate fingerprint was deposited on the glass surface. As a pilot study, sebaceous fingerprints were used because it is an initial starting point for this research as it is likely to produce a positive result when compared to natural fingerprints which would consist of a limited transfer of secretions. (8). Oblique lighting from a flashlight was used again to ensure the deposited fingerprints were adequate (1). In this study, an adequate fingerprint is defined as a fingerprint with visible ridge detail. The fingerprints were not marked once they were placed on the vehicle to replicate a real situation in which the location of the latent fingerprints is unknown. Each vehicle was parked outside and driven around the streets of Southern Ontario to be exposed to debris. The conditions the vehicles were exposed to—sun, rain, snow, wind, etc.—caused debris to collect on the surface of the vehicles. All vehicle drivers were notified of the research project to ensure they did not disrupt the

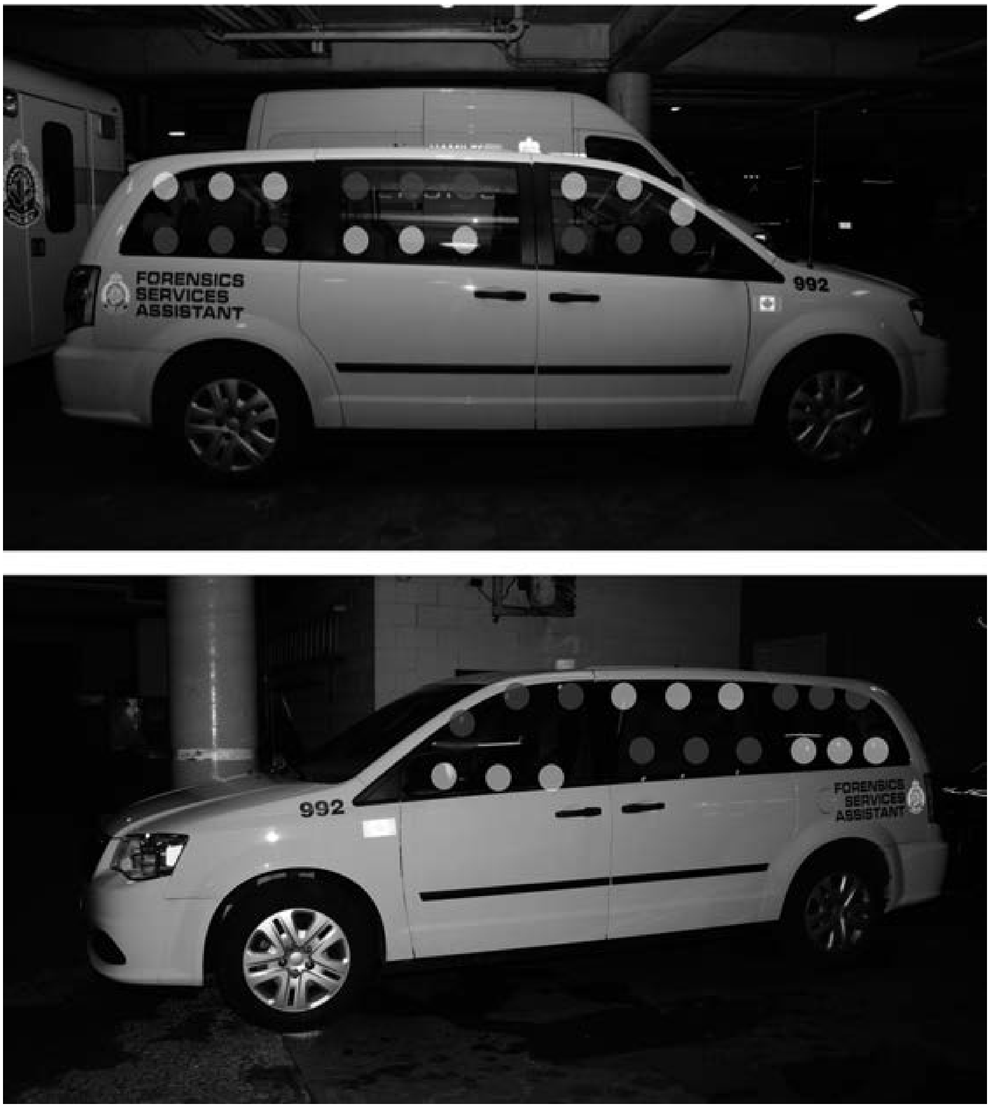


FIG. 1—*Location of fingerprints on the exterior glass windows. Each red circle represents one experimental latent fingerprint. Each yellow circle represents one control latent fingerprint. Experimental and control fingerprints alternate from the top of the window to the bottom of the window. Fingerprint placement on the top and bottom is the opposite on the other side of the vehicle. [Color figure can be viewed at wileyonlinelibrary.com]*

fingerprints while debris was being deposited. The distance the vehicles traveled and the specific conditions that vehicles were exposed to could not be controlled because vehicles are driven based on calls for service. To mitigate this difference between the vehicles, the drivers/officers alternated the vehicles they used when called to a crime scene, to expose all vehicles to debris collected on the road. The route and distance each vehicle traveled were variable as it depended on where the intended destination was located (Table 1).

Once each vehicle collected debris according to its designated time period, the control and experimental fingerprints were recovered. The vehicles were parked in the Hamilton Police garage during fingerprint recovery. Since this study took place in the winter months, each vehicle was left to acclimatize and dry in the parking garage for 1 h before fingerprint recovery began. The experimental fingerprints were treated with liquid latex. Three layers of liquid latex were applied to the surface with a 4″ 10 mm foam roller (Bennett Canada, Concord, ON, Canada) (1). The drying time for liquid latex is approximately 30 min per layer, thus the total drying time was approximately 1 h and 30 min (1). Once the liquid latex was dry, it was peeled off the glass surface. The experimental fingerprints were developed with white granular powder and a fiberglass brush (Lynn Peavey Company, Lenexa, KS). The control fingerprints were not treated with liquid latex. The white granular powder was applied directly to the control fingerprints covered in debris (1). All of the recovered fingerprints were photographed using a Nikon D7000 camera with a sigma lens and then lifted using 1.5″ fingerprint tape (Sirchie, Youngsville, NC). The lifted fingerprints were then placed on acetate sheets (Atlas Graphic Supply Inc., Markham, ON, Canada). The photographed fingerprints were given to a trained Forensic Identification Officer from the Hamilton Police Service to be scored. The lifted fingerprints were made available for reference in case the details in the photographs were unclear.

The Forensic Officer scored the fingerprints based on the Bandey scale and their own professional opinion. The Bandey scale is the most commonly used scale for assessing fingerprint quality (9). The Bandey scale scores fingerprints from 0–4 based on ridge detail (Table 2). Fingerprints with a Bandey scale score of

TABLE 2—*Bandey scale fingerprint scoring scheme.*

Score	Fingerprint Detail
0	No development
1	No continuous ridges; all discontinuous or dotty
2	One third of the mark comprised of continuous ridges; remainder either show no development or dotty
3	Two thirds of the mark comprised of continuous ridges; remainder either show no development or dotty
4	Full development; whole mark comprised of continuous ridges (9,10)

three or four are considered ideal for individual identification (10).

The authors classified "not recovered" fingerprints as fingerprints given a score of zero or the fingerprints that were not detected with the white granular powder. "Recovered" fingerprints were those given a score of 1–4 on the Bandey scale. A score of zero and an undetected fingerprint are classified as "not recovered" because in both cases, no ridge detail is present to confirm if it is, in fact, a fingerprint. However, fingerprints given a score of zero were able to be photographed and collected because there were outlines present in the area the initial fingerprint was deposited that indicated a fingerprint. Since the officers recovering the fingerprints were not responsible for scoring the fingerprints, the recovery of all noticeable and detectable traces of a fingerprint prevented bias from choosing specific fingerprints to recover based on visible detail.

Statistics and Results

Once data collection was complete and recovered fingerprints were scored, it was determined that 53 (25%) of the 216 latent fingerprints deposited were recovered. Of the 53 latent fingerprints recovered, 36 (68%) were recovered with the control method, while 17 (32%) were recovered with the liquid latex method. Figure 2 illustrates the difference between the quality of the fingerprints recovered by the control method and the experimental method using liquid latex. Overall, the control method had more fingerprints per score than the experimental method.

A chi-square test was performed on the recovered versus not recovered data for each method. Chi-square is used to compare the control and experimental method when categorical data are being used (11). The chi-square test revealed a significant difference in fingerprint recovery between the control and liquid latex method ($X^2 = 9.026$, d.f. = 1, $p = 0.003$). An odds ratio was used to compare the performance of the control and experimental method by comparing the recovered and not recovered fingerprints for each method (12). The odds of recovering latent fingerprints are 2.68 more likely using the control method compared to using liquid latex. To compare the scoring/quality of the recovered fingerprints for each method, a Fisher's exact test was performed because the assumptions for a chi-square test were not met (cell frequency <5) (13). There is no statistically significant difference between the detail of the recovered control and experimental fingerprints ($p = 0.065$).

The window location and the trial number (number of weeks to collect debris) were compared with the number of fingerprints recovered by each method, respectively (Figs 3 and 4).

Discussion and Conclusion

This study demonstrated the advantage of processing fingerprints with white granular powder over the debris that accumulates on the exterior windows of vehicles. The control method (processing with powder over debris) recovered more fingerprints than the experimental method of using liquid latex to remove debris. The quality of the recovered fingerprints, however, is not significantly different between methods when fingerprints can be recovered (Fig. 5).

The results from this study are similar to the conclusion made by Grossi et al. (6), as they also determined that the liquid latex method was not as effective on glass compared to other

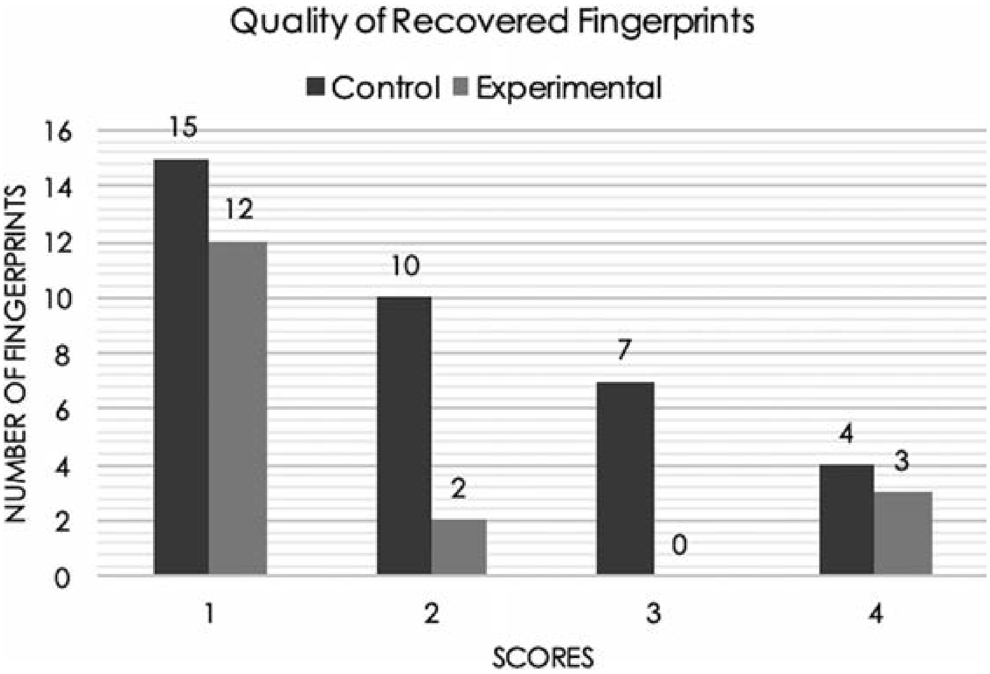


FIG. 2—*Bar graphs illustrating the comparison between the quality of fingerprints recovered by the control method and experimental/liquid latex method. The scores correspond with the Bandey scale criteria in Table 2. [Color figure can be viewed at wileyonlinelibrary.com]*

traditional methods. However, Grossi et al. (6) also found the liquid latex method to be 100% unsuccessful, which contradicts findings in this study and other studies. Although applying the granular powder directly over debris recovered more fingerprints than the liquid latex method, the liquid latex method was capable of recovering fingerprints of varying quality. The differences in liquid latex results between this study and Grossi's study are likely due to the differences in procedure and experimental setup. Firstly, Grossi et al. (6) deposited fingerprints on top of existing debris, unlike this study in which latent fingerprints were deposited on a clean surface. This difference would likely impact recovery because, in theory, the debris is supposed

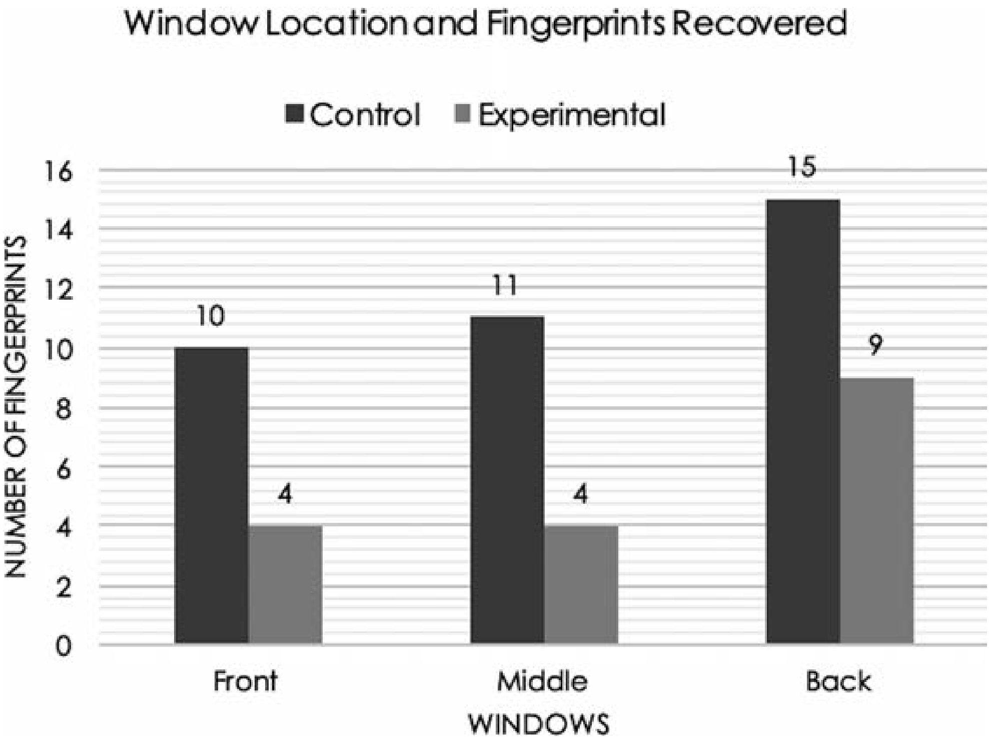


FIG. 3—*Bar graphs illustrating the comparison between the window location and the number of fingerprints recovered by the control method and experimental method using liquid latex. [Color figure can be viewed at wileyonlinelibrary.com]*

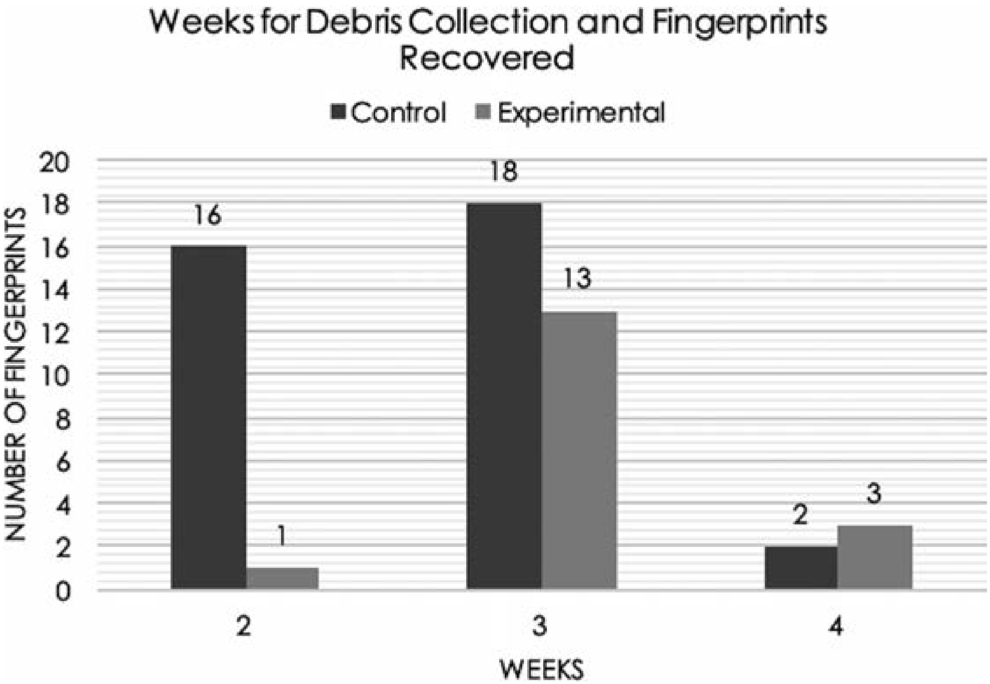


FIG. 4—*Bar graphs illustrating the comparison between the number of weeks to collect debris and the number of fingerprints recovered by the control method and experimental method using liquid latex. [Color figure can be viewed at wileyonlinelibrary.com]*

to protect the fingerprints so that liquid latex will solely lift the debris and leave the underlying fingerprint. Depositing the fingerprint on top of the debris and then applying liquid latex does not allow for debris protection. Secondly, in the study by Grossi et al. (6), the glass was in the form of stationary windowpanes that were put outdoors as well as indoors. An immobile windowpane in areas where individuals walk will result in a difference in the amount and types of debris collected on the surface when compared to a mobile vehicle, exposed to various environmental conditions.

The results of this study contradict the conclusion made by Ho and George (1), as they found the liquid latex pretreatment to perform better than the control method. They acknowledged that there needs to be a sufficient amount of debris on the surface for the liquid latex method to work (1). The Ho and George study (1) tested the ability to recover liquid latex on the bottom of the vehicles where large amounts of debris accumulate. In the present study, the height increase of the deposited latent fingerprints on the vehicle could account for the difference in the performance of liquid latex. Generally, the lower portion of vehicles collect more debris as this area is closer to the ground and the tires. Thus, in this study, with the latent fingerprints placed on the exterior windows, it was expected that less debris would accumulate on the surface. As seen in Fig. 3, both methods recovered more latent fingerprints on the back window as this is likely where most debris accumulated. Debris deposition levels are likely higher near the back of the vehicle because the debris from the vehicle's front tires travels to the back of the vehicle as it is being driven. Likewise, the middle and back windows are larger and can collect more debris than the front window. The front window is also partially blocked by the side view mirror which could have an impact on debris accumulation on the front window. Even though a difference was observed when comparing the three windows, the debris they collected (even the back window) was likely much less than the debris collected on the bottom of vehicles with Ho and George (1), which made the control method more successful than the liquid latex method.

It was hypothesized that the longer trials would have a greater number of fingerprints recovered because more debris would collect on top of the fingerprints due to longer exposure time. However, this was not the case. The results comparing the fingerprints recovered for each method in terms of the number of weeks for debris collection indicate that the 3-week trials recovered the most fingerprints for both methods (Fig. 4). The 4-week trial had the fewest number of fingerprints overall. An explanation for this

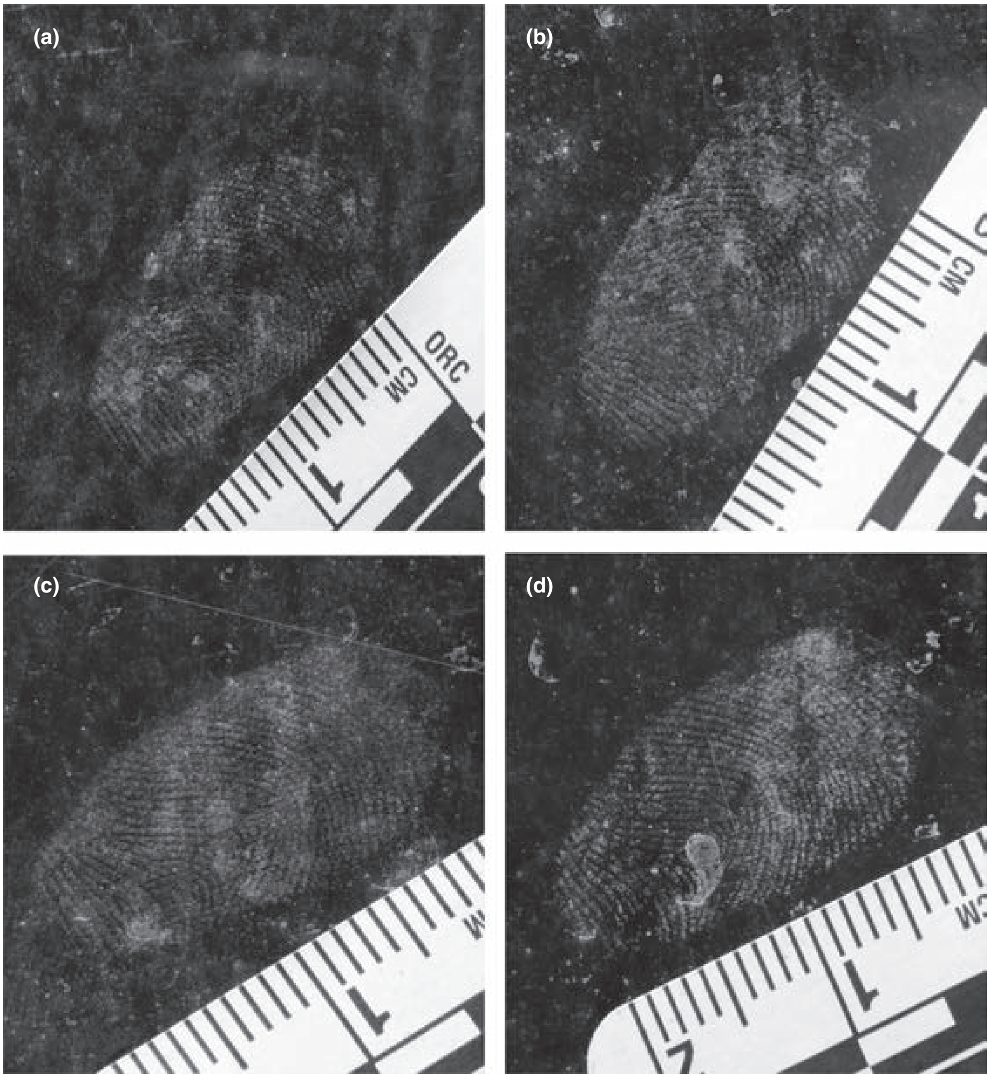


FIG. 5—*Varying quality of fingerprints after being processed with either the control method or liquid latex and then treated with white granular powder. The fingerprints above are fingerprints assigned a score of (a) 1, (b) 2, (c) 3, and (d) 4.*

observation is the environmental conditions experienced over the different trials. It is recommended that additional research is completed, specifically accounting for the weather conditions the vehicles are exposed to during the trials, to understand the relationship between debris accumulation and successful fingerprint recovery.

Ho and George (1) suggested processing the area containing possible fingerprints with granular powder before using liquid latex to determine if there is a sufficient amount of debris. If the fingerprints are successfully developed with granular powder, it is indicative that debris accumulation over the fingerprints is insufficient. If the fingerprints do not develop with the granular powder, it is possible that fingerprints are not present or that a significant amount of debris has covered the area. It is recommended by Ho and George (1), that if results are negative with granular powder, that liquid latex should be applied to the surface to remove existing debris and processed again with granular powder. This study tested the success of liquid latex on external windows regardless of the amount of debris.

Since the use of liquid latex on vehicles is a relatively new area of study, there are many future study recommendations. It is recommended that future studies following a similar protocol focus on following 2 specific aspects of the International Fingerprint Research Group guidelines for Phase 1 studies: the use of 3–5 donors (to test the variation between individuals) and natural instead of groomed fingerprints (to test the sensitivity of the method) (8). Similar to the recommendations made by Ho and George (1), future studies should focus on conducting trials at different times of the year. Both this study and the study by Ho and George (1) conducted the trials between November 2019 and February 2020. The amount and types of debris collected during this time of the year are likely different than the debris that would be collected during the spring, summer, or fall months. Ho and George (1) also recommended conducting longer trials in future studies. This change in methodology can be used to determine if the length of time for debris collection affects which method for fingerprint recovery should be used. Moreover, future studies should focus on other areas of the vehicle where latent fingerprints are likely deposited, specifically around the door handles, sides where the door opens, and the frame surrounding the door. These areas will be lower on the vehicle compared to the windows and could potentially collect more debris. The International Fingerprint Research Group (8) indicates that the liquid latex method requires optimization and validation before it can be applied to casework.

In this study, it was determined that fingerprint development over the debris will recover more fingerprints than the liquid latex method. However, the quality of the control fingerprints recovered according to the Bandey scale is similar to the fingerprints recovered using liquid latex. The results of this study indicate that it is advantageous to process the exterior glass surface of vehicles with white granular powder as opposed to the liquid latex method. Not only does the control method recover more fingerprints, but it is also much more cost-effective and time-efficient when compared to the liquid latex method. More research regarding the use of liquid latex to remove debris should be conducted to help investigators decide on methods to use when processing vehicles.

Acknowledgments

The authors wish to thank the Hamilton Police Services Forensic Services Branch for graciously providing the vehicles and materials for this research; and the Forensic Science Department at the University of Toronto Mississauga for their guidance and support.

References

1. Ho M, George M. The use of liquid latex as a pretreatment to remove debris off the exterior surface of vehicles for fingerprint recovery. J Forensic Identif 2019;69(3):329–37.
2. Pearshall J, editor. The concise Oxford dictionary, 10th edn. New York, NY: Oxford University Press, 1999;369.
3. Brodbeck SMC. The latex lifting method for the recovery of blood, DNA, and dermal ridge evidence in arson cases. J Bloodstain Pattern Anal 2011;27(4):3–7.
4. Larkin TPB, Marsh NP, Larrigan PM. Using liquid latex to remove soot to facilitate fingerprint and bloodstain examinations: a case study. J Forensic Identif 2008;58(5):540–50.
5. Clutter SW, Bailey R, Everly JC, Mercer K. The use of liquid latex for soot removal from fire scenes and attempted fingerprint development with ninhydrin. J Forensic Sci 2009;54(6):1332–5. https://doi.org/10.1111/j.1556-4029.2009.01143.x
6. Grossi I, Power C, Slaney J.A comparison of liquid latex and tape for removing surface debris to improve fingerprint quality. Identification Canada. In press.
7. Yamashita B, French M. Latent print development. In: McRoberts A, editor. The fingerprint sourcebook. Washington, DC: U.S. Dept. of Justice, Office of Justice Programs, National Institute of Justice, 2011;7–13.
8. International Fingerprint Research Group. Guidelines for the assessment of fingermark detection techniques. J Forensic Identif 2014;64(2):174–200.
9. Brown RM, Hillman AR. Electronic enhancement of latent fingerprints by poly(3,4-ethylenedioxythiophene): supporting information. Phys Chem Chem Phys 2012;14(24):8653–61. https://doi.org/10.1039/c2cp40733g
10. Deepthi NH, Darshan GP, Basavarji RB, Daruka Prasad B, Nagabhushana H. Large-scale controlled bio-inspired fabrication of SD CeO_2: Eu^{3+} hierarchical structures for evaluation of highly sensitive visualization of latent fingerprints. Sensor Actuat B-Chem 2018;255:3127–47. https://doi.org/10.1016/j.snb.2017.09.138
11. Qin C, Lee C, Ho S, Koh J, Athiviraham A. Complication rates following hip arthroscopy in the ambulatory surgical center. J Orthop 2020;20:28–31. https://doi.org/10.1016/j.jor.2019.12.009
12. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PMM. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 2003;56(11):1129–35. https://doi.org/10.1016/S0895-4356(03)00177-X
13. Campbell I. Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations. Stat Med 2007;26(19):3661–75. https://doi.org/10.1002/sim.2832

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14542
Available online at: onlinelibrary.wiley.com

PAPER

GENERAL

Mohammed A. Almazrouei,[1,2,3] *M.Sc.; Itiel E. Dror*,[1,2] *Ph.D.; and Ruth M. Morgan*,[1,2] *D.Phil.*

Organizational and Human Factors Affecting Forensic Decision-Making: Workplace Stress and Feedback

ABSTRACT: Although forensic examiners operate in a stressful environment, there is a lack of understanding about workplace stress and feedback. These organizational and human factors can potentially impact forensic science judgments. In this study, 150 practicing forensic examiners from one laboratory were surveyed about their experiences of workplace stress, and the explicit and implicit feedback they receive. Forensic examiners reported that their high stress levels originated more from workplace-related factors (management and/or supervision, backlogs, and the pressure to do many cases) than from personal related factors (family, medical, and/or financial). The findings showed that a few (8%) of the forensic examiners sometimes felt strong implicit feedback about what conclusions were expected from them and that some (14%) also strongly felt that they were more appreciated when they helped to solve a case (e.g., by reaching a "match" as opposed to an "inconclusive" conclusion). Differences were found when comparing workplace stress and feedback levels across three core forensic science fields (forensic biology, chemistry, and latent prints) and across career stages (early, mid, and late). Gaining insights into the stress factors within a workplace and explicit and implicit feedback has implications for developing policies to improve the well-being, motivation, and performance of forensic examiners.

KEYWORDS: forensic science, workplace stress, feedback, implicit feedback, forensic decision-making, well-being, human factors

Workplace stress has been shown to have an impact on the quality of decisions made by professionals in a variety of domains, such as medicine (e.g., [1]), policing (e.g., [2]), the military (e.g., [3]), management (e.g., [4]), and psychology (e.g., [5-7]). In the medical domain, for instance, a review of 22 empirical studies indicated that high levels of stress factors (such as bleeding, time pressure, and procedural complexity) can affect the performance of surgeons (1). However, research is still lacking with regard to the impact workplace stress may have on the well-being of forensic examiners (8) as well as the quality of their decisions (9).

Forensic examiners operate in a stressful environment (8,10,11). Some organizational stress factors are common in many workplace environments, such as workload volume and number of working hours (11). There are also stress factors that are specific and unique to the forensic science discipline (9), which further contributes to the "high stress occupation" of forensic examiners ([8], p. 34). One of these unique stress factors is the intensified scrutiny of forensic techniques and criticisms of their validity, as well as working within an adversarial legal system (e.g., [12]). Moreover, there are often unreasonable expectations placed on the forensic examiners not to ever make any mistakes (13,14). In addition, forensic examiners can be directly exposed to emotionally distressing elements from crime scenes or disturbing case details (9,11).

Stress can have positive or negative impacts on human performance and decision-making (15-17). The Yerkes–Dodson law empirically shows an inverted U-shape relationship between stress and performance (17). Performance is lower at low stress; then with increased stress, performance is higher, but this elevation in performance continues only until the level of stress is moderate. As stress becomes high, performance and quality of decisions start to drop (17). In forensic science, quality of judgments includes accuracy, but also other issues, such as confidence levels, documentation of the decision-making process, reporting of the conclusions, ability to justify the decisions, and their presentation in court (18; see also [19] for Hierarchy of Expert Performance).

High levels of stress, or repeated exposure to stress, have been shown to impair the cognitive ability of individuals (20) and the well-being of forensic examiners (21). Workplace stress can result in negative workplace experiences. These occupational experiences can cause physical (e.g., stomach distress and heart disease), psychological (e.g., anger and job dissatisfaction), and behavioral reactions (e.g., substance use and absenteeism) (22,23). For example, Holt and Blevins (21) surveyed 56 digital forensic examiners and found that around 68% were working under a lot of pressure at work. Participates in this study reported a number coping mechanisms, such as drinking alcohol and smoking (21). It was also reported that in some law

[1]UCL Department of Security and Crime Science, University College London, 35 Tavistock Square, London, WC1H 9EZ, U.K.

[2]UCL Centre for the Forensic Sciences, University College London, 35 Tavistock Square, London, WC1H 9EZ, U.K.

[3]Forensic Evidence Department, Abu Dhabi Police General Headquarters, Abu Dhabi, 253, U.A.E.

Corresponding author: Mohammed A. Almazrouei, M.Sc. E-mail: mohammed.almazrouei@ucl.ac.uk

Received 7 May 2020; and in revised form 1 July 2020; accepted 24 July 2020.

enforcement agencies the attrition rates were about 50% within 3 years for staff members responding to critical crime scenes (about 20% of them reported long-term psychological problems) (11).

Feedback is a critical factor in its own right that can impact well-being and performance (24), as it can have implications for the motivation, expectations, and the decision-making of forensic examiners (e.g., questions 8 and 9 in [25]). Therefore, understanding the ways feedback given to forensic examiners and how it may affect the decision-making of forensic examiners is important for understanding the context in which decisions are made (18,26). This has the potential to impact the entire crime reconstruction process (27).

During casework, forensic examiners communicate and receive feedback from a variety of sources, which can be categorized into five domains: forensic services, investigative, legal, public (26), and regulatory (18) domains (see Fig. 1). For example, forensic examiners communicate with top management and/or immediate supervisors (14,28), with police investigators (29-31), and they can be in contact with legal advocates during the preparation of evidence for presentation in court (32,33).

Human factors are not independent and often affect one another. For example, stress and emotions are closely related, as stress can generate negative emotions (34). Similarly, stress and feedback are related (e.g., pressures from feedback can cause stress). Importantly, such pressures can impact conclusions (35):

> “Errors and disagreements among examiners may be due to in part . . . [to] systemic pressures encouraging some decisions more than others. These pressures will vary by agency or among cases, and examiners’ responses to these pressures will vary.” (p. 66)

The study reported here deals with these organizational and human factors of stress and feedback that can affect decision-making. A questionnaire was designed to contain questions about stress and feedback (see Appendix 1). For clarity in presenting the findings, this paper was divided into two parts. The first part focuses on stress experienced at the workplace, examining the existence of and sources of stress in forensic science laboratories. The second part addresses the feedback provided, examining how it is perceived by practicing forensic examiners.

Part One: Workplace Stress

Research addressing the decision-making in forensic science has mainly focused on some key human factors, such as contextual information (e.g., [36]) and emotional factors (e.g., [13,37]). However, other human and organizational factors, such as workplace stress have generally been neglected in the published literature. Only a few studies have paid attention to forensic examiners’ stress at the workplace, such as the stress experienced by forensic digital examiners exposed to internet crimes against children (e.g., [21,38]); crime scene examiners exposed to horrific crimes (e.g., [11]); and forensic odontologists exposed to mass casualties (e.g., [39]).

There is a lack of research addressing workplace stress of examiners working in forensic science in general, and specifically across core forensic science fields (such latent prints and forensic chemistry) and across different stages of their career. It

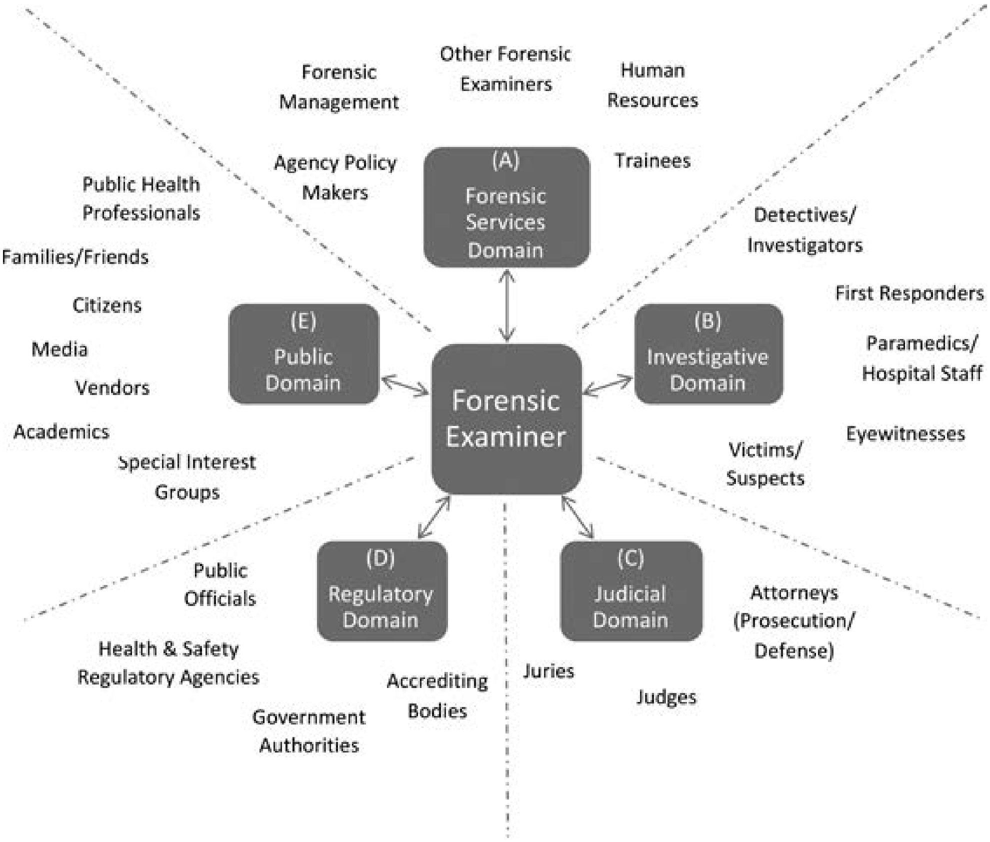


FIG. 1—*Interactions and communications of forensic examiners with five stakeholders (taken from Dror and Pierce [18]). [Color figure can be viewed at wileyonlinelibrary.com]*

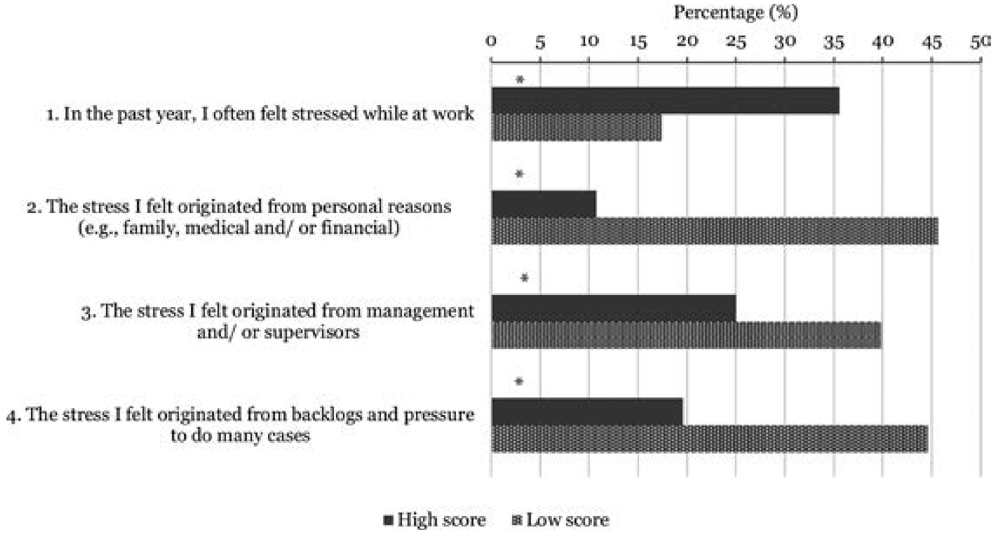


FIG. 2—*Scores of stress levels (*p < .05 for χ^2 of low vs. high scores). [Color figure can be viewed at wileyonlinelibrary.com]*

is argued that research on the psychological consequences of stress experienced by forensic science professionals is lacking because of the general belief that professionals involved in emergency situations are expected to deal with stress and demands as part of their job (39).

Research on stress experienced by forensic examiners can help in understanding the factors that moderate stress, and how different factors play a role in creating, reducing, and managing stress (3,10). This may have implications for developing relevant evidence-based approaches to improve the well-being of experts as well as their decision-making performance. Therefore, this study explores the factors that may cause forensic science examiners to feel stress. It was of interest to examine the contribution of stresses attributed to the workplace as opposed to personal factors; whether there were differences in the stresses felt by examiners working in different forensic science fields; and whether the years of experience moderated the level of stress experienced.

Method

Questionnaire

Following established approaches in decision-making studies within the forensic science discipline (25,40,41), and studies addressing perceptions of workplace stress factors (e.g., [10,38]), a questionnaire was designed to examine workplace stress (Part One) and feedback (Part Two).

Part One contained questions to ascertain whether forensic examiners had felt stressed at work, and how much of the stress they attributed to personal reasons (e.g., family, medical, and/or financial matters) as opposed to relating the stress to the workplace (see Fig. 2). The participants were required to rank their responses on a seven-point Likert-type scale. The participants were also asked to provide demographic information on their primary forensic field and years of experience.

Participants

A total of 150 forensic examiners from a major forensic laboratory in the United States took part in the study (71% response rate; $N = 212$). All the participants were practicing forensic examiners, and they were from the same forensic laboratory, so that it was possible to examine and compare variables (e.g., fields of expertise and years of experience) without introducing interlaboratory variations.

Forensic examiners identified their primary fields as: biology/DNA ($n = 42$), latent prints ($n = 40$), controlled substances ($n = 24$), forensic alcohol ($n = 7$), toxicology ($n = 4$), firearms ($n = 9$), and trace evidence ($n = 5$). Nineteen (13%) did not report their primary field, and three latent print examiners stated that they also work as crime scene examiners as a secondary field. The fields were grouped together on the basis of the type of expertise deployed, giving three field categories: forensic biology ($n = 42$; DNA and biology), latent prints ($n = 40$), and forensic chemistry ($n = 35$; controlled substances, toxicology, and forensic alcohol). The remaining fields (trace evidence, firearms, and crime scene investigation as a secondary field) were excluded from the analysis by field of expertise, because they contained low participant numbers and did not fit within any of the three main field categories.

The mean years of experience was 12 (*SD* = 9.7 years, with a range from 1 to 47 years; did not respond: $n = 12$). Four examiners provided a qualitative written response to the question about their years of experience (e.g., “many” or “lots”) or the number written was illegible and thus not included in the years of experience analysis (i.e., 16 participants (11%) were excluded from the analysis by experience, leaving 134 participants). Following the accepted approach in the published literature to categorize data, such as the years of experience (e.g., [10,42]), we grouped the years of experience into categories of comparable sample sizes: early career (0–5, $n = 36$); mid-career (6–10, $n = 28$) and (11–20, $n = 40$); and late career (>20, with $n = 30$).

Statistical Analysis

Both descriptive and inferential statistics were applied, using SPSS (version 25), to measure the reported stress levels in general and to examine stress by field and years of experience. Following previous research (10), the seven-point Likert-type scale responses were converted to an ordinal, categorical scale of low, moderate, and high scores: scores 1–2 as low (i.e., low feelings of stress), scores 3–5 as medium, and scores 6–7 as high (i.e., strong feelings of stress). Equal categories of low and high scores were made as per previously published research (10). However, it should be emphasized that some of the neighboring scores (e.g., scores 2 and 3) are grouped in different categories (i.e., low and medium) and this is reflected in the interpretation. Likert scales can be categorized (e.g., [25]) and can be statistically treated at an ordinal level (43). This categorization helps to examine the variability of stress experienced by the examiners.

A chi-square test (goodness of fit) was used to determine whether the categorical responses for each question differed significantly (i.e., low *vs.* high stress scores; see Figure 2). An alpha significance level of 0.05 was used for all the statistical tests. In addition to the significance testing, the means and standard deviations are reported.

One-way ANOVA and post hoc (Bonferroni) were used to compare the mean workplace stress levels across the categories of forensic fields and years of experience. In case that the homogeneity of variance assumption was not met, as assessed by Levene's test, then a one-way Welch ANOVA and post hoc (Games–Howell) were used instead. In addition to comparing the means, a chi-square test was used to test whether the responses of the high scores for the three categories of forensic fields differed significantly from one another. The stress scores were particularly important at the high levels where the influence of stress on the well-being and performance of forensic examiners can be most critical (17,20,23).

Results

Workplace Stress

One in three forensic examiners (36%, $n = 53$) strongly felt that they often experience stress while at the workplace (low *vs.* high stress scores, χ^2 (1, $N = 79$) = 9.23, $p = 0.002$; $M = 4.61$, $SD = 1.90$; see Fig. 2). For the high stress levels felt by the examiners, stress was attributed more from the workplace (i.e., 25%, $n = 37$, from management and/or supervisors (χ^2 (1, $N = 96$) = 5.04, $p = 0.025$; $M = 3.62$, $SD = 2.16$), and 20%, $n = 29$, from backlog pressure (χ^2 (1, $N = 95$) = 14.41, $p < 0.001$; $M = 3.30$, $SD = 2.05$)) than from the personal life (11%, $n = 16$; χ^2 (1, $N = 84$) = 32.19, $p < 0.001$; $M = 3.14$, $SD = 1.85$).

Stress by Field and Experience

On average, moderate workplace stress (question 1) was felt by all forensic field categories: biologists ($M = 5.02$, $SD = 1.94$), latent print examiners ($M = 4.75$, $SD = 1.77$), and forensic chemists ($M = 4.09$, $SD = 1.92$). While the mean stress levels did not vary across the three field categories (questions 1–4, $p > 0.05$), the level of high stress differed from backlog pressure only, χ^2 (2, $N = 24$) = 7.75, $p = 0.021$. The percentage of forensic biologists (34%, $n = 14$) who strongly felt that their stress originated from backlog pressure was higher than the other fields, that is, latent print examiners (18%, $n = 7$) and forensic chemists (9%, $n = 3$).

The mean stress levels varied across experience groups, but only due to stress from management and/or supervisors (question 3, *Welch's F* (3, 67.7) = 6.01, $p = 0.001$) and backlog stress (question 4, *Welch's F* (3, 67.7) = 8.15, $p < 0.001$; see Table 1). There were no interactions between the forensic field and years of experience on the reported stress levels (univariate ANOVA for questions 1–4, $p > 0.05$).

Discussion

Workplace Stress

On average, forensic examiners in this study reported a moderate frequency in feeling stressed at the workplace (question 1, $M = 4.61$, $SD = 1.90$). However, there was variability in the data as reflected by the standard deviations and by the low and high stress scores (see Fig. 2). Variability is expected given individual differences in responding to stress factors (44,45). Also worth noting is that although question 1 asked examiners on the *frequency* of their stress at work (i.e., "often"), the responses to this question can also reflect their *level* of stress. It is generally reasonable to assume that people who feel stressed more frequently also feel higher levels of stress (e.g., see transdisciplinary model of stress that describes "stress" as a set of integrated processes, including the history of stressors in the life of an individual [44]).

In this study, 36% of the forensic examiners strongly felt that they are often stressed at work. Published research from other domains has shown that repeated exposure to stress or when stress levels are high, the well-being (23), and decision-making performance drops (17,20). For example, LeBlanc et al. (46) asked 30 paramedics to calculate drug dosage after working in a highly stressful scenario and found that intense stress increased medical errors.

The data from the study reported here concerns the feelings experienced by forensic examiners. It does not include objective measures of the performance and quality of decisions of the participants. Hence, the data reported do not show the nature of the causational relationship, if any, between high stress and performance. Higher levels of stress can impact performance in a number of ways. These data cannot ascertain the impact, but clearly show that stress is felt by forensic examiners, and hence warrant further research.

Future research needs to experimentally examine the impact of stress on the decision-making performance in the forensic science context, as has been studied in other specialized domains (see, for example, Arora et al. (1) for a review of studies that

TABLE 1—*Mean responses (and standard deviations in brackets) for questions 3, 4, and 7 where significant findings were found among the experience groups.*

	Experience Group (Years)			
Question	0–5	6–10	11–20	>20
3. Management/ supervision stress	2.53 (1.63)^{a,b,c,d,e,f}	4.21 (2.83)^{a,b,c,d,e,f}	3.70 (2.12)^{a,b,c,d,e,f}	4.20 (2.28)^{a,b,c,d,e,f}
4. Backlog stress	2.06 (1.51)^{a,b,c,d,e,f}	3.37 (1.98)^{a,b,c,d,e,f}	3.98 (2.19)^{a,b,c,d,e,f}	3.50 (1.94)^{a,b,c,d,e,f}
7. Feedback on expected conclusions	2.18 (1.62)^{g}	2.89 (1.85)	2.76 (1.48)	3.33 (1.94)^{g}

$^{a,b,c,d,e,f}p < 0.05$, post hoc (Games–Howell)

$^{g}p < 0.05$, post hoc (Bonferroni).

investigated the impact of stress in the medical domain). Such experimental research is important given the critical nature of forensic science decisions within the criminal justice system (27,29).

In the current study, 17% of forensic examiners reported feelings stressed at work relatively infrequently (if they felt stressed at all). It has been observed in some contexts that low levels of stress can lead to underload, boredom, and lower performance (47). Conversely, moderate stress can improve performance (17), as it can, among other things, push individuals to meet deadlines (9). Hence, the published literature addressing stress suggests that there could be benefits in maintaining moderate stress levels at the workplace of forensic examiners (by, for example, providing new, interesting tasks to motivate underloaded, low stressed individuals [47]).

The findings of this study suggest that the forensic laboratory management and/or supervision contribute to the stress levels felt by the forensic examiners (the way the question was framed in the survey does not allow us to determine if it was the laboratory management or the supervisor that created the stress, or both —it is only possible to identify that there was stress felt and it was attributed to either or both of these factors). Published research addressing stress suggests that relationships in the workplace are a common organizational-level stress factor, and that they can be one of the primary causes of stress among criminal justice employees in general (10,48). Hence, it would appear that forensic management and/or supervisors may play a key role in optimizing the stress levels and well-being of forensic examiners.

Similarly, the findings of the current study reveal that backlogs and pressure to complete many cases can contribute to the stress felt by the forensic examiners (see Fig. 2). It has been suggested in the published literature that pressure from case backlog is intensified by the increase of requests from prosecutors and law enforcement agencies for rapid forensic analysis and reports, in addition to increasing forensic service requests for nonviolent crimes in an under-resourced and overtaxed forensic science environment (9,12). However, it is acknowledged that backlog pressure is a complex measure and can vary from one forensic organization to another (8).

The findings show that more examiners strongly felt that their stress originated from the workplace than arising due to personal reasons. It is, however, important to note that the questions posed in this study did not directly relate personal and workplace causes of stress in one question so as to offer the opportunity for examiners to rate one type of stress factor directly against the other. Further research on personal life stress is needed, as it has been suggested in the published literature that stress from the personal life can affect the work–life balance, increase work–life conflict, reduce job satisfaction, and lower performance in the workplace (49,50).

Stress by Field and Experience

On average, forensic biologists, forensic chemists, and latent print examiners reported moderate frequencies or levels of stress at the workplace (again, it is important to note that there were individual differences even within the same forensic science field). Previous research targeting specific forensic fields yielded inconsistent findings. For instance, forensic odontologists reported low stress levels when attending mass casualty incidents, for reasons such as having sense of achievement and obtaining invaluable professional experience (39), whereas digital forensic examiners reported moderate levels of stress in undertaking their roles (e.g., examining child pornography [21]). These previous studies were conducted across laboratories; hence, it is not possible to attribute the different findings to the forensic fields, because these differences may arise from other confounding factors, such as the general workplace culture and environment in the laboratory.

The results from this study, within a single laboratory, allow for a better comparison across forensic fields. These data indicate that high levels of stress from backlog pressure vary among the three fields; specifically, more forensic biologists strongly felt stress from backlog pressure in comparison with forensic chemists and latent print examiners. However, as previously mentioned, backlog is a complex measure and has been shown to vary across forensic organizations—even within the same field of expertise—and can change with time (8). The dynamic and complex nature of backlog pressure suggests that each forensic organization may be well advised to evaluate the way they communicate their own backlogs among the different forensic fields, and how it can influence the well-being and performance of their forensic examiners.

The findings also reveal that mid- and late career examiners—that is, over 5 years of experience—felt more stress originating from management and/or supervision and from backlogs in comparison with early career examiners— that is, under 5 years of experience (there were no interactions between field of expertise and years-of-experience categories in all the stress questions). A previous study suggested that examiners with more experience have more workload responsibilities, such as having a supervisory role (21), which may go some way toward offering insight to this trend that was observed in this study.

There are differences in the levels of workplace stress across occupations (51). There are insufficient understanding and data about stress in forensic science to enable a meaningful comparison to other occupations. This study is one of the first to address workplace stress from various forensic science fields (with statistical comparisons of examiners working in core fields, such as forensic biology and chemistry). In addition, since data were collected from one laboratory, the data do not necessarily generalize to other forensic laboratories. However, there are good reasons to believe that forensic science is a high stress occupation in comparison with typical working environments (8,9). Working environment and organizational culture are human factors that impact forensic decision-making (see fifth source of bias in [52]).

Part Two: Workplace Feedback

Feedback is a key component of the conceptual model of communication in forensic science presented by Howes (53). Additionally, feedback received by forensic examiners who perform casework analysis and interpretation is an important component of monitoring and improving performance, and motivating and rewarding examiners for hard work (e.g., [24]). Feedback can be explicit (messages that can be directly codified and articulated) (30,54), such as an immediate supervisor saying "well-done" to the examiner. Feedback can also be implicit, meaning that messages are not direct and less codified (30,54). An example of implicit feedback would be the supervisor "smiling" to the examiner, which can cause subjective interpretation and experiences of emotions (55).

Stress and pressure resulting from explicit and/or implicit feedback can influence forensic science judgments. In an earlier study, some fingerprint examiners reported that they were not allowed or were discouraged from making inconclusive decisions when the latent mark and known prints were of value and included a large area for comparison (56). Moreover, Kassin et al. (57) discussed that a contributing factor of the misidentification in the 2004 Madrid train bombings was the increased "need for closure" (i.e., the desire to provide clear-cut judgments [58]), which resulted in a subsequently established erroneous identification of Mayfield. It is salient that an independent investigation report on this case stated that the criteria for reaching an inconclusive result could lead to implicit pressures on an examiner to reach an identification when making a difficult comparison of marks, particularly when the case was very serious (59).

Previously published research has started to look into the possible relationships between perceived feedback and forensic expert decision-making (e.g., questions 8 and 9 in [25]). Yet its impact and scope are still largely unexplored. This current study assessed the explicit and implicit feedback, as felt by the forensic examiners with the following key actors (see Fig. 1): forensic management and/or supervisors (the forensic services domain), police investigators (the investigative domain), and legal advocates (the legal domain). These have been identified as actors that can impact decisions made during crime scene work, laboratory analysis, and/or judicial procedures (12,33,60,61).

Therefore, the second part of this current study sought to identify the level of explicit and implicit feedback as felt by the forensic examiners, and whether the feedback varied by forensic science field of expertise or years of experience.

Method

The same methodology was followed as outlined in Part One, with the only difference being the inclusion of three questions on feedback. Specifically, the feedback questions addressed whether forensic examiners received feedback about their work from stakeholders, such as from management, supervisors, police investigators, and/or legal advocates (i.e., explicit feedback; see question 5 in Fig. 3). In addition, questions 6 and 7 asked whether the forensic examiners felt that the stakeholders appreciated them more when they help to solve a case (such as when finding a "match" rather than "inconclusive") and whether the examiners sometimes felt they know what the stakeholders expect or want their conclusions to be (i.e., implicit feedback; Fig. 3).

Results

Workplace Feedback

About half (49%, $n = 71$; $M = 3.06$, $SD = 1.93$) of forensic examiners reported low scores for feeling that management, supervisors, police investigators, and/or legal advocates appreciated it more when they were helping to solve cases, and that sometimes they felt they knew what these stakeholders wanted or expected their conclusions to be (53%, $n = 77$; $M = 2.75$, $SD = 1.77$). Nevertheless, some examiners, albeit a small minority, reported high scores for feeling such feedback and expectations, 14%, $n = 20$, $\chi^2(1, N = 91) = 28.58$, $p < 0.001$ and 8%, $n = 11$, $\chi^2(1, N = 88) = 49.50$, $p < 0.001$, respectively. Examiners were equally divided (27%, $n = 40$, high scores *vs.* 28%, $n = 42$, low scores; $p > 0.05$) on whether they receive explicit feedback ($M = 3.95$, $SD = 2.00$; see Fig. 3).

Feedback by Field and Experience

On average, most forensic biologists ($M = 4.49$, $SD = 2.06$), forensic chemists ($M = 3.77$, $SD = 2.00$), and latent print examiners ($M = 3.62$, $SD = 1.96$) felt they received moderate explicit feedback from their management, supervisors, police investigators, and/or legal advocates. Both the explicit and implicit mean feedback levels did not significantly differ by field of expertise (questions 5–7, $p > 0.05$). However, for the high scores of the explicit feedback question, more forensic biologists (41%, $n = 17$) reported receiving feedback than latent print examiners (21%, $n = 8$) and forensic chemists (20%, $n = 7$; approaching statistical significance, $\chi^2(2, N = 32) = 5.69$, $p = 0.058$).

Question 7 on expected conclusions was the only feedback question that varied by experience (approaching significance, $F(3, 126) = 2.54$, $p = 0.060$; see Table 1). There were no interactions between the forensic science field and years of experience on the reported feedback levels (univariate ANOVA for questions 5–7, $p > 0.05$).

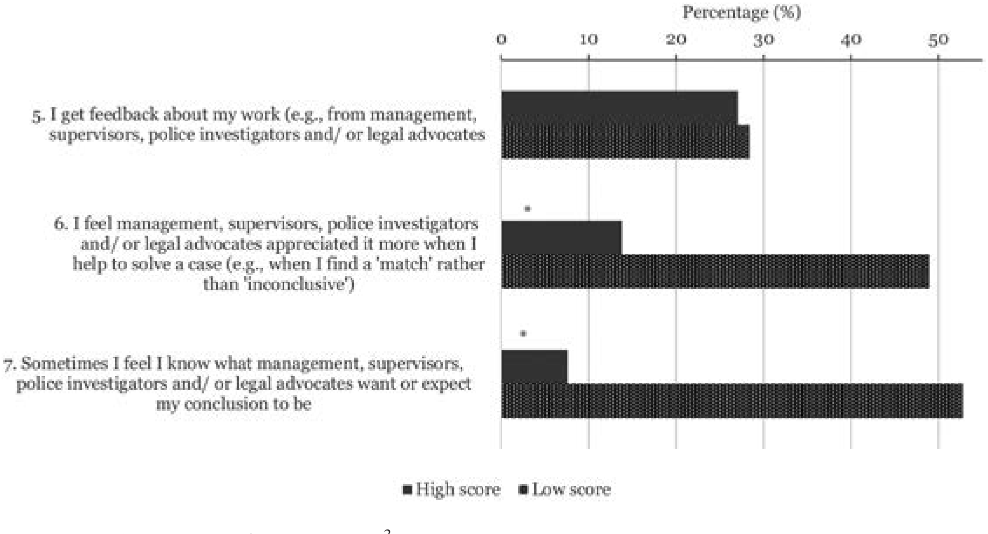


FIG. 3—*Scores of explicit and implicit feedback (*p < 0.05 for χ^2 of low vs. high scores). [Color figure can be viewed at wileyonlinelibrary.com]*

Discussion

Explicit Feedback

Forensic examiners were divided on whether they receive low or high amounts of explicit feedback about their work from the stakeholders they interact with. Additionally, on average, forensic examiners reported receiving similar levels of explicit feedback across the investigated forensic science fields and experience groups. However, more forensic biologists reported receiving high levels of explicit feedback than the latent print examiners and forensic chemists did, while at the same time, more forensic biologists reported experiencing high levels of stress from backlog pressure than the other two fields of expertise (see Part One). The data, however, do not include measures to inform an understanding of how such feedback impacts the well-being and the performance of the forensic examiners. Therefore, in order to consider the explicit feedback within the crime reconstruction process further, it will be important for future research to identify what type and level of feedback is warranted (18,26,62).

Implicit Feedback

A few forensic examiners strongly felt that sometimes they knew what stakeholders wanted their conclusions to be (question 7, 8%; see Fig. 3). Despite being a low proportion, this finding on implicit feedback is concerning because each forensic examiner is involved in casework analysis and interpretation (32). The findings also show that a higher level of implicit feedback was felt by late career (>20 groups) in comparison with early career examiners (0–5 group), in terms of what stakeholders wanted or expected their conclusions to be (see Table 1). This finding is consistent with previous research, which found that 63.6% of forensic examiners agreed (i.e., slightly agreed, agreed, and strongly agreed) that on occasions they know what conclusions they are expected to find (25) and that forensic examiners can be pressured to extend opinions beyond their scientific findings (63).

To be clear, the aforementioned findings do not demonstrate that the examiners are in fact being pressured by the stakeholders to reach expected conclusions. Rather, the data illustrate what the examiners perceive and feel as implicit pressure. It is the perception and feeling of stress that makes a situation stressful rather than there being an actual stress factor (44,45). It is important to consider the context within which decisions are being made to ensure there is transparency in this process to mitigate conditions that exert pressure on examiners to make "expected" decisions.

The findings from this study demonstrate that some (question 6, 14%) forensic examiners strongly felt that stakeholders in the forensic services, investigative, and legal domains appreciated it more when they reported conclusions of high certainty (e.g., a clear-cut, match conclusion as opposed to inconclusive). While this is a low percentage of the sample, this high implicit feedback score is also concerning. It shows that some active casework scientists may feel an implicit pressure to reach certain conclusions. As stated earlier, it is the "cognitive appraisal" of the individual to the situation that makes it pressurizing (44,45), even in the absence of such pressures. It is of course important to note that these data cannot indicate whether conclusions are being influenced by such implicit pressures.

General Discussion

Taking the stress and feedback findings together, many of the forensic examiners in this study perceived that they operate under pressure, and that the level of pressure varies by field and experience, during casework and reporting conclusions. The findings emphasize that one must consider the operating environment that forensic examiners work in, and the importance of managing the levels of workplace stress and feedback.

The insights from the data provide a valuable but limited insight into the possible relationships between feedback, stress, and forensic decision-making. This study clearly cannot identify and characterize the relationships but indicates that this could be a fruitful avenue for future studies. Additionally, as detailed earlier, organizational and human factors (such as stress and feedback) are interrelated and affect one another (34). Hence, it is possible that the questions addressing the feelings of examiners regarding implicit feedback (i.e., questions 6 and 7) can be related to stress and/or other factors.

The current study further contributes to the forensic science literature by synthesizing relevant stress and feedback literature from other domains. It offers a focused theoretical discussion, along with empirical data, on how workplace stress and feedback can affect forensic science judgments (whereas most of the previous research mainly focused on the relationship between stress and well-being of forensic examiners (e.g., [10,38])). In addition, the current paper unpacks the notion of feedback, an under-researched but important organizational factor in forensic science. It is hoped that this study will drive further research directed toward workplace feedback and its potential effects on expert decision-making.

The published literature suggests that there can be individual differences in perceiving and coping with stress (44,45). This means that forensic examiners can perceive and cope with stress and feedback differently, even among those examiners who work in the same laboratory and forensic field, and have the same years of experience. The current data account for interlaboratory variations, as it has been collected from a single laboratory. However, differences in individual stress perceptions and coping styles were not investigated, and so should to be considered in future research and also in practice.

It is important to note that self-reporting from a participant of how they feel about stress or feedback can offer valuable and informative insights. However, individuals cannot accurately describe the rationale of their decision-making and judgments, as this often involves unpacking complex cognitive processes (40,64). It is possible, for example, that the workplace stress felt by the forensic examiners is originating from personal reasons (50) and it could have been difficult for participants to separate the workplace from personal causes of stress. In addition, the responses of forensic examiners may have been affected by social desirability bias (65), in particular for the implicit feedback questions. Although the current study included a large sample size of 150 practicing forensic examiners from the same laboratory, it may not be representative to forensic laboratories worldwide. The reported levels of stress and feedback may vary in other jurisdictions that have different working environments and cultures.

Conclusion

This study surveyed active forensic examiners with different fields of expertise and years of experience working within one laboratory. The examiners reported feeling varying levels of workplace stress and levels of explicit and implicit feedback. More high levels of stress were reported to originate from the workplace (specifically, stress from backlogs and pressure to do many cases, and management and/or supervisors) than from stress derived from personal reasons outside the workplace. More forensic biologists perceived high levels of backlog pressure than latent print examiners and forensic chemists. Mid- and late career examiners (i.e., over 5 years of experience) reported higher stress levels originating from management and/ or supervision, as well as backlog pressure in comparison with early career examiners (i.e., less than 5 years of experience).

It was concerning that a few forensic examiners sometimes felt strongly that they knew what the stakeholders in the forensic services, investigative, and/or legal domains expected or wanted their conclusions to be and that some forensic examiners also strongly felt that the same stakeholders appreciated it more when they helped to solve a case (e.g., by finding a match as opposed to inconclusive).

In a broader context, the creation of working environments that can address the negative impacts of the types of stress examiners are exposed to will be valuable. It is also important to be aware of the impact of both explicit and implicit feedback and to develop practices that ensure the positive assistance and timely explicit feedback. This may include preventive risk management measures (18), such as the evaluation of the how backlogs are measured and communicated to forensic examiners across different fields of expertise. It is also important to consider the context within which decisions are being made to ensure there is transparency in this process to mitigate conditions that exert pressure on examiners to make "expected" decisions.

Acknowledgements

First author PhD at UCL SECReT (Security Science Doctoral Training Centre) is funded by Abu Dhabi Police.

References

1. Arora S, Sevdalis N, Nestel D, Woloshynowych M, Darzi A, Kneebone R. The impact of stress on surgical performance: a systematic review of the literature. Surgery 2010;147(3):318–30. https://doi.org/10.1016/j.surg.2009.10.007.
2. Akinola M, Mendes WB. Stress-induced cortisol facilitates threat-related decision making among police officers. Behav Neurosci 2012;126 (1):167–74. https://doi.org/10.1037/a0026657.
3. Kavanagh J. Stress and performance: a review of the literature and its applicability to the military. Santa Monica, CA: RAND Corporation, 2005. https://www.rand.org/pubs/technical_reports/TR192.html (accessed January 14, 2020).
4. Gok K, Atsan N. Decision-making under stress and its implications for managerial decision-making: a review of literature. Int J Bus Soc Res 2016;6(3):38–47. https://doi.org/10.18533/ijbsr.v6i3.936.
5. Dror IE, Busemeyer JR, Basola B. Decision-making under time pressure: an independent test of sequential sampling models. Mem Cogn 1999;27 (4):713–25. https://doi.org/10.3758/BF03211564.
6. Kerstholt J. The effect of time pressure on decision-making behaviour in a dynamic task environment. Acta Physiol (Oxf) 1994;86(1):89–104. https://doi.org/10.1016/0001-6918(94)90013-2.
7. Yu R, Yang L, Guo X, Zhang Y. Effect of time pressure on dynamic visual search performance. Procedia Manuf 2015;3:4658–64. https://doi.org/10.1016/j.promfg.2015.07.556.
8. National Institute of Justice. Report to congress: needs assessment of forensic laboratories and medical examiner/coroner offices. U.S. Department of Justice, 2019. https://nij.ojp.gov/library/publications/report-congress-needs-assessment-forensic-laboratories-and-medical (accessed January 20, 2020).
9. Jeanguenat AM, Dror IE. Human factors effecting forensic decision-making: workplace stress and well-being. J Forensic Sci 2018;63 (1):258–61. https://doi.org/10.1111/1556-4029.13533.
10. Holt TJ, Blevins KR, Smith RW. Examining the impact of organizational and individual characteristics on forensic scientists' job stress and satisfaction. J Contemp Crim Justice 2017;40(1):34–49. https://doi.org/10.1080/0735648X.2016.1216731.
11. Kelty SF, Gordon H. No burnout at this coal-face: managing occupational stress in forensic personnel and the implications for forensic and criminal justice agencies. Psychiat Psychol Law 2015;22(2):273–90. https://doi.org/10.1080/13218719.2014.941092.
12. National Research Council. Strengthening forensic science in the United States: a path forward. Washington, DC: The National Academies Press, 2009. https://www.ncjrs.gov/pdffiles1/nij/grants/228091.pdf (accessed October 15, 2019).
13. Charlton D, Fraser-Mackenzie PAF, Dror IE. Emotional experiences and motivating factors associated with fingerprint analysis. J Forensic Sci 2010;55(2):385–93. https://doi.org/10.1111/j.1556-4029.2009.01295.x.
14. Mustonen V, Hakkarainen K, Tuunainen J, Pohjola P. Discrepancies in expert decision-making in forensic fingerprint examination. Forensic Sci Int 2015;254:215–26. https://doi.org/10.1016/j.forsciint.2015.07.031.
15. Kowalski-Trakofler KM, Vaught C, Scharf T. Judgment and decision-making under stress: an overview for emergency managers. Int J Emerg Manag 2003;1(3):278. https://doi.org/10.1504/IJEM.2003.003297.
16. Paton D, Flin R. Disaster stress: an emergency management perspective. Disaster Prev Manage 1999;8(4):261–7. https://doi.org/10.1108/09653569910283897.
17. Yerkes RM, Dodson JD. The relation of strength of stimulus to rapidity of habit-formation. J Comp Neurol Psychol 1908;18(5):459–82. https://doi.org/10.1002/cne.920180503.
18. Dror IE, Pierce ML. ISO standards addressing issues of bias and impartiality in forensic work. J Forensic Sci 2020;65(3):800–8. https://doi.org/10.1111/1556-4029.14265.
19. Dror IE. A hierarchy of expert performance. J Applied Res Memory Cognition 2016;5(2):121–7. https://doi.org/10.1016/j.jarmac.2016.03.001.
20. Deligkaris P, Panagopoulou E, Montgomery A, Masoura E. Job burnout and cognitive functioning: a systematic review. Work Stress 2014;28 (2):107–23. https://doi.org/10.1080/02678373.2014.909545.
21. Holt TJ, Blevins KR. Examining job stress and satisfaction among digital forensic examiners. J Contemp Crim Justice 2011;27(2):230–50. https://doi.org/10.1177/1043986211405899.
22. Spector PE. Industrial and organizational psychology: research and practice, 6th rev edn. Singapore: Wiley, 2012;278.
23. Benson H, Casey A, editors. Stress management: approaches for preventing and reducing stress. Boston, MA: Harvard Medical School, 2013. https://www.health.harvard.edu/promotions/harvard-health-publications/stress-management-approaches-for-preventing-and-reducing-stress (accessed October 14, 2019).
24. Choi E, Johnson DA, Moon K, Oah S. Effects of positive and negative feedback sequence on work performance and emotional responses. J Organ Behav Manage 2018;38(2–3):97–115. https://doi.org/10.1080/01608061.2017.1423151.
25. Kukucka J, Kassin SM, Zapf PA, Dror IE. Cognitive bias and blindness: a global survey of forensic science examiners. J Appl Res Mem Cogn 2017;6(4):452–9. https://doi.org/10.1016/j.jarmac.2017.09.001.
26. Almazrouei MA, Dror IE, Morgan RM. The forensic disclosure model: what should be disclosed to, and by, forensic experts. Int J Law Crime Just 2019;59:100330. https://doi.org/10.1016/j.ijlcj.2019.05.003.
27. Morgan RM, Meakin GE, French JC, Nakhaeizadeh S. Crime reconstruction and the role of trace materials from crime scene to court. WIREs Forensic Sci 2020;2(1):1–18. https://doi.org/10.1002/wfs2.1364.
28. Dror IE. Cognitive neuroscience in forensic science: understanding and utilizing the human element. Philos Trans R Soc B Biol Sci 2015;370 (1674):20140255. https://doi.org/10.1098/rstb.2014.0255.
29. Morgan RM. Conceptualising forensic science and forensic reconstruction. Part I: a conceptual model. Sci Justice 2017;57(6):455–9. https://doi.org/10.1016/j.scijus.2017.06.002.
30. Morgan RM. Conceptualising forensic science and forensic reconstruction. Part II: the critical interaction between research, policy/law and practice. Sci Justice 2017;57(6):460–7. https://doi.org/10.1016/j.scijus.2017.06.003.
31. Raymond T, Julian R. Forensic intelligence in policing: organisational and cultural change. Aust J Forensic Sci 2015;47(4):371–85. https://doi.org/10.1080/00450618.2015.1052759.
32. Dror IE. Biases in forensic experts. Science 2018;360(6386):243. https://doi.org/10.1126/science.aat8443.
33. Murrie DC, Boccaccini MT, Guarnera LA, Rufino KA. Are forensic experts biased by the side that retained them? Psychol Sci 2013;24:1889–97. https://doi.org/10.1177/0956797613481812.

34. Du J, Huang J, An Y, Xu W. The relationship between stress and negative emotion: the mediating role of rumination. Clin Res Trials 2018;4(1):1–5. https://doi.org/10.15761/CRT.1000208.
35. Ulery BT, Hicklin RA, Roberts MA, Buscaglia J. Factors associated with latent fingerprint exclusion determinations. Forensic Sci Int 2017;275:65–75. https://doi.org/10.1016/j.forsciint.2017.02.011.
36. van den Eeden CAJ, de Poot CJ, van Koppen PJ. The forensic confirmation bias: a comparison between experts and novices. J Forensic Sci 2019;64(1):120–6. https://doi.org/10.1111/1556-4029.13817.
37. Dror IE, Péron AE, Hind S-L, Charlton D. When emotions get the better of us: the effect of contextual top-down processing on matching fingerprints. Appl Cognitive Psych 2005;19(6):799–809. https://doi.org/10.1002/acp.1130.
38. Burruss GW, Holt TJ, Wall-Parker A. The hazards of investigating internet crimes against children: digital evidence handlers' experiences with vicarious trauma and coping behaviors. Am J Crim Justice 2018;43(3):433–47. https://doi.org/10.1007/s12103-017-9417-3.
39. Webb DA, Sweet D, Pretty IA. The emotional and psychological impact of mass casualty incidents on forensic odontologists. J Forensic Sci 2002;47(3):539–41. https://doi.org/10.1520/JFS2001330.
40. Gardner BO, Kelley S, Murrie DC, Dror IE. What do forensic analysts consider relevant to their decision making? Sci Justice 2019;59(5):516–23. https://doi.org/10.1016/j.scijus.2019.04.005.
41. Hamnett HJ, Jack RE. The use of contextual information in forensic toxicology: an international survey of toxicologists' experiences. Sci Justice 2019;59(4):380–9. https://doi.org/10.1016/j.scijus.2019.02.004.
42. Yoo Y-S, Cho O-H, Cha K-S, Boo Y-J. Factors influencing post-traumatic stress in Korean forensic science investigators. Asian Nurs Res 2013;7(3):136–41. https://doi.org/10.1016/j.anr.2013.07.002.
43. Jamieson S. Likert scales: how to (ab)use them. Med Ed 2004;38(12):1217–8. https://doi.org/10.1111/j.1365-2929.2004.02012.x.
44. Epel ES, Crosswell AD, Mayer SE, Prather AA, Slavich GM, Puterman E, et al. More than a feeling: a unified view of stress measurement for population science. Front Neuroendocrin 2018;49:146–69. https://doi.org/10.1016/j.yfrne.2018.03.001.
45. Lazarus RS, Folkman S. Stress, appraisal, and coping. New York, NY: Springer, 1984;22–5.
46. LeBlanc VR, MacDonald RD, McArthur B, King K, Lepine T. Paramedic performance in calculating drug dosages following stressful scenarios in a human patient simulator. Prehosp Emerg Care 2005;9(4):439–44. https://doi.org/10.1080/10903120500255255.
47. Driskell T, Driskell JE, Salas E. Stress, performance and decision making in organizations. In: Highhouse S, Dalal RS, Salas E, editors. Judgment and decision making at work. New York, NY: Routledge, 2014;251–76.
48. Cullen FT, Lemming T, Link BG, Wozniak JF. The impact of social supports on police stress. Criminology 1985;23(3):503–22. https://doi.org/10.1111/j.1745-9125.1985.tb00351.x.
49. Burke RJ. Stressful events, work-family conflict, coping, psychological burnout, and well-being among police officers. Psychol Rep 1994;75(2):787–800. https://doi.org/10.2466/pr0.1994.75.2.787.
50. Hall GB, Dollard MF, Tuckey MR, Winefield AH, Thompson BM. Job demands, work-family conflict, and emotional exhaustion in police officers: a longitudinal test of competing theories. J Occup Organ Psychol 2010;83(1):237–50. https://doi.org/10.1348/096317908X401723.
51. Johnson S, Cooper C, Cartwright S, Donald I, Taylor P, Millet C. The experience of work-related stress across occupations. J Manage Psychol 2005;20(2):178–87. https://doi.org/10.1108/02683940510579803.
52. Dror IE. Cognitive and human factors in expert decision making: six fallacies and the eight sources of bias. Anal Chem 2020;92(12):7998–8004. https://doi.org/10.1021/acs.analchem.0c00704.
53. Howes LM. The communication of forensic science in the criminal justice system: a review of theory and proposed directions for research. Sci Justice 2015;55(2):145–54. https://doi.org/10.1016/j.scijus.2014.11.002.
54. Ellis R, Loewen S, Erlam R. Implicit and explicit corrective feedback and the acquisition of L2 grammar. Stud Sec Lang Acq 2006;28(02):339–68. https://doi.org/10.1017/S0272263106060141.
55. Söderkvist S, Ohlén K, Dimberg U. How the experience of emotion is modulated by facial feedback. J Nonverbal Behav 2018;42(1):129–51. https://doi.org/10.1007/s10919-017-0264-1.
56. Ulery BT, Hicklin RA, Buscaglia J, Roberts MA. Accuracy and reliability of forensic latent fingerprint decisions. Proc Natl Acad Sci 2011;108(19):7733–8. https://doi.org/10.1073/pnas.1018707108.
57. Kassin SM, Dror IE, Kukucka J. The forensic confirmation bias: problems, perspectives, and proposed solutions. J Appl Res Mem Cogn 2013;2(1):42–52. https://doi.org/10.1016/j.jarmac.2013.01.001.
58. Ask K, Granhag PA. Motivational sources of confirmation bias in criminal investigations: the need for cognitive closure. J Invest Psychol Off 2005;2(1):43–63. https://doi.org/10.1002/jip.19.
59. Office of the Inspector General. A review of the FBI's handling of the Brandon Mayfield case. Washington, D.C.: U.S. Department of Justice, 2006. https://oig.justice.gov/sites/default/files/legacy/special/s0601/PDF_list.htm (accessed October 18, 2019).
60. Julian R, Kelty SF. Forensic science as "risky business": identifying key risk factors in the forensic process from crime scene to court. J Criminol Res Pol Pract 2015;1(4):195–206. https://doi.org/10.1108/JCRPP-09-2015-0044.
61. Kelty SF, Julian R, Bruenisholz E, Wilson-Wilde L. Dismantling the justice silos: flowcharting the role and expertise of forensic science, forensic medicine and allied health in adult sexual assault investigations. Forensic Sci Int 2018;285:21–8. https://doi.org/10.1016/j.forsciint.2018.01.015.
62. Morgan RM, Earwaker H, Nakhaeizadeh S, Harris AJL, Rando C, Dror IE. Interpretation of forensic evidence at every step of the forensic science process: decision-making under uncertainty. In: Wortley R, Sidebottom A, Tilley N, Laycock G, editors. Routledge handbook of crime science. Abingdon, U.K.: Routledge, 2018;408–20.
63. Becker WS, Dale WM, Lambert A, Mangus D. Forensic lab directors' perceptions of staffing issues. J Forensic Sci 2005;50(5):1255–7. https://doi.org/10.1520/JFS2005201.
64. Nisbett RE, Wilson TD. Telling more than we can know: verbal reports on mental process. Psychol Rev 1977;84(3):231–59. https://doi.org/10.1037/0033-295X.84.3.231.
65. Chung J, Monroe GS. Exploring social desirability bias. J Bus Ethics 2003;44(4):291–302. https://doi.org/10.1023/A:1023648703356.

Appendix 1 Anonymous Questionnaire

Please rate the following statements. It is totally anonymous, so please be honest.

	1 (Low)	2	3	4	5	6	7 (High)
1. In the past year, I often felt stressed while at work.							
2. The stress I felt originated from personal reasons (e.g., family, medical and/ or financial).							
3. The stress I felt originated from management and/or supervisors.							
4. The stress I felt originated from backlogs and pressure to do many cases.							
5. I get feedback about my work (e.g., from management, supervisors, police investigators and/or legal advocates).							
6. I feel management, supervisors, police investigators and/ or legal advocates appreciated it more when I help to solve a case (e.g., when I find a “match” rather than “inconclusive”).							
7. Sometimes I feel I know what management, supervisors, police investigators and/ or legal advocates want or expect my conclusion to be.							

Which section do you work at (e.g., DNA, firearms, latent prints, etc)?___________Years of experience:__________

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14546
Available online at: onlinelibrary.wiley.com

PAPER

GENERAL

Jeff Kukucka,[1] *Ph.D.; Alexa Hiley*,[2] *M.A.; and Saul M. Kassin*,[2] *Ph.D*

Forensic Confirmation Bias: Do Jurors Discount Examiners Who Were Exposed to Task-Irrelevant Information?*,†

ABSTRACT: Knowledge of task-irrelevant information influences judgments of forensic science evidence and thereby undermines their probative value (i.e., *forensic confirmation bias*). The current studies tested whether laypeople discount the opinion of a forensic examiner who had *a priori* knowledge of biasing information (i.e., a defendant's confession) that could have influenced his opinion. In three experiments, laypeople (N = 765) read and evaluated a trial summary which, for some, included testimony from a forensic examiner who was either unaware or aware of the defendant's confession, and either denied or admitted that it could have impacted his opinion. When the examiner admitted that the confession could have influenced his opinion, laypeople generally discounted his testimony, as evidenced by their verdicts and other ratings. However, when the examiner denied being vulnerable to bias, laypeople tended to believe him—and they weighted his testimony as strongly as that of the confession-unaware examiner. In short, laypeople generally failed to recognize the superiority of forensic science judgments made by context-blind examiners, and they instead trusted examiners who claimed to be impervious to bias. As such, our findings highlight the value of implementing context management procedures in forensic laboratories so as not to mislead fact-finders.

KEYWORDS: confirmation bias, context management, contextual bias, cross-examination, expert testimony, forensic science, juror decision-making, task-relevant

Misleading forensic science has contributed to over 600 known wrongful convictions (1). In recent years, *forensic confirmation bias*—that is, the phenomenon whereby task-irrelevant information influences judgments of forensic science evidence—has been identified as one source of these costly errors (2). In an early demonstration of this phenomenon, Dror and Charlton (3) found that fingerprint examiners unknowingly changed 17% of their own prior judgments of the same pair of fingerprints after being told that the suspect had either confessed (implying that the prints should match) or provided a verified alibi (implying that the prints should not match). This effect has since been replicated (4,5), and other studies have found similar effects in myriad forensic science disciplines (for reviews, see (6,7))—such as handwriting identification (8,9), arson investigation (10), forensic anthropology (11), bloodstain pattern analysis (12), bitemark analysis (13), crime scene investigation (14), and complex DNA analysis (15).

In 2015, the National Commission on Forensic Science explained that *task-irrelevant* information includes any "contextual information [that] supports inferences about a proposition only through a chain of logic that does not involve assessment of the physical evidence"—such as a suspect's criminal history, confession, or alibi (16). Stated otherwise, forensic examiners "should draw conclusions solely from the physical evidence that they are asked to evaluate… and not from any other evidence in the case" (16). Reliance on task-irrelevant information undermines the probative value of a forensic examiner's opinion because it creates a problematic double-counting of evidence (e.g., [17–20]). For example, if an examiner judges physical evidence as inculpatory only because they had *a priori* knowledge of a suspect's confession, their opinion appears to independently corroborate the confession, but is actually a product of it (20). (For a discussion of this problem from a Bayesian perspective, see (21).)

As such, forensic laboratories have been urged to adopt procedures that blind examiners to task-irrelevant information and thereby minimize the risk of cognitive bias (e.g., [18,22–23]). However, many forensic examiners remain uninformed or unconcerned about how bias can affect their work. In a recent survey of 403 forensic examiners from 21 countries, Kukucka et al (24) found that only 49% felt that they should be shielded from task-irrelevant information, and 71% believed that they can mitigate bias simply by ignoring their expectations. Moreover, the examiners exhibited a "bias blind spot" (25): Most (71%) saw bias as a problem in the forensic sciences as a whole, but fewer (52%) saw bias as a problem in their own domain, and still fewer (26%) believed that bias affects them personally. In other words, many examiners felt that cognitive bias can affect their peers but not themselves.

[1]Department of Psychology, Towson University, 8000 York Road, Towson, MD,.

[2]Department of Psychology, John Jay College of Criminal Justice, New York, NY,.

Corresponding author: Jeff Kukucka, Ph.D. E-mail: jkukucka@towson.edu

*Presented at the 2016 (Studies 1 and 2) and 2017 (Study 3) Annual Meetings of the American Psychology-Law Society, March 9-12, 2016, and March 15-18, 2017, respectively.

†Funding provided by the John Jay College of Criminal Justice of the City University of New York to Dr. Saul Kassin.

Received 29 May 2020; and in revised form 30 June 2020, 22 July 2020; accepted 24 July 2020.

If examiners do not take steps to combat bias in the laboratory, the onus of identifying and discounting biased forensic opinions will fall on judges and jurors. Two independent lines of research suggest pessimism over their ability to do this. First, basic social-cognitive psychology indicates that lay observers tend to accept other people's self-reports at face value, resulting in a truth bias that contributes to poor performance at detecting deception (26). In the related domain of attribution, it is well-established that observers routinely commit the *fundamental attribution error* (or *correspondence bias*), making dispositional attributions for other people's words and deeds, while underestimating the impact of situational factors (27–30). This literature would thus suggest that even observers who are aware of an examiner's prior exposure to potentially biasing information will underappreciate the impact of that situational information on their conclusions.

Second, research has indicated that judges (31,32) and jurors (33,34) are generally not adept at identifying flaws in scientific evidence. Of particular relevance to forensic confirmation bias are two studies in which jurors failed to recognize *experimenter bias* (i.e., a form of self-fulfilling prophecy in which a researcher's expectations unconsciously impact the outcome of their study; (35)) as a threat to the validity of research findings (36,37). Compounding this problem, laypeople tend to believe that forensic science errors are exceptionally rare (38). Hence, jurors tend to trust forensic evidence (39,40), even if it was analyzed using an unvalidated technique (41) and even after being explicitly informed of its limitations (42).

Although jurors left to their own devices may fail to recognize unreliable scientific evidence, some studies suggest that cross-examination that attacks scientific validity rather than credibility can sensitize jurors to the quality of scientific evidence presented by an expert (43,44). Others, however, have found little or no effect of cross-examination (33,45). With respect to the forensic sciences, findings have also been mixed: Lieberman et al (39) found that an "evidence-focused" cross-examination (i.e., noting the potential for contamination and subjectivity involved in the analysis) weakened the credibility of a DNA expert more so than an "expert-focused" cross-examination (i.e., attacking the expert's experience and academic record), whereas Koehler (46) found that a cross-examination highlighting the potential for error had little effect on jurors' appraisals of testimony from a shoeprint expert.

To date, two studies have specifically examined whether cross-examination can sensitize jurors to cognitive bias as a factor that undermines evidence quality. Chorn and Kovera (47) had mock jurors read expert testimony from a clinical psychologist who administered an intelligence test to the plaintiff in a civil case. In some conditions, the clinician was blind to the plaintiff's supposed IQ score when administering the test; in others, the clinician knew *a priori* that the plaintiff was thought to have a low IQ. This variation had no significant effect. Jurors rated the clinician as similarly competent and the test result as similarly compelling regardless of whether its administration was blind or nonblind. This held true even among jurors who heard a cross-examination that explained the importance of blind testing for scientific validity.

Of particular relevance to the current studies, Thompson and Scurich (48) had venirepersons read a summary of an assault case that included testimony from a forensic odontologist who compared a bitemark on the victim against a model of the suspect's teeth. In some conditions, the odontologist admitted that he was exposed to task-irrelevant information (i.e., the suspect's criminal history and gang affiliation) before performing this comparison and either admitted or denied that this information influenced his judgment. In another condition, the odontologist explained that his laboratory's standard protocol kept him blind to this information until after he completed his analysis. Venirepersons rated his testimony as more credible when he was purposefully blinded to task-irrelevant information than when he was exposed to it, regardless of whether he claimed to have ignored or used that information. However, they did not rate the nonblind examiner as less credible than a control examiner who was not exposed to task-irrelevant information. This pattern suggests that laypeople understand that blinding procedures increase the probative value of forensic science evidence, but they fail to understand that nonblind procedures correspondingly decrease its value.

Participants in Thompson and Scurich (48) evaluated testimony on bitemark comparison, which is among the most criticized of forensic science disciplines. In its 2016 report, the President's Council of Advisors on Science and Technology (23) concluded that "bitemark analysis does not meet the scientific standards for foundational validity [and] the prospects of developing bitemark analysis into a scientifically valid method [are] low" (p. 87). Indeed, such comparisons may be highly subjective, as research has shown that the same teeth can leave bitemarks that appear quite different (e.g., [49,50]), while different people can leave bitemarks that are indistinguishable from each other (e.g., [51,52]). As such, it is perhaps unsurprising that laypeople considered a blinded bitemark examiner to be more credible than a nonblind bitemark examiner.

But do jurors understand that bias can affect other forensic disciplines—even those with relatively strong scientific foundations? For example, fingerprint examiners typically follow a standardized procedure (i.e., the ACE-V method; [23]), use sophisticated technologies to identify potential matches to a latent print (53), and exhibit low error rates (e.g., [54,55]). Yet, they are likewise vulnerable to cognitive bias (e.g., [3,56,57]). Moreover, laypeople seem to view bitemark and fingerprint evidence as equally compelling: Koehler (38) found that laypeople believed both fingerprint and bitemark identification errors to be exceptionally rare (1 in 5.5 million and 1 in 1 million, respectively), and Ribeiro et al (58) found that laypeople perceived fingerprint and bitemark identification as equally accurate (88% vs. 89%, respectively) and as entailing a comparable degree of subjective judgment.

The Present Studies

The present studies investigated whether mock jurors would devalue the testimony of a forensic examiner who had *a priori* knowledge of task-irrelevant information that could have influenced his opinion. To be exact, we tested whether jurors would discount the opinion of an examiner who had *a priori* knowledge of the defendant's confession—a form of evidence that creates a uniquely strong expectation of guilt (20) and has repeatedly been shown to impact people's perceptions, memories, and judgments (e.g., [59–61]), including judgments of forensic science evidence (e.g., [8]). Outside of the laboratory, forensic science errors have been found in 67% of DNA exonerations that involved false confessions—and in these cases, the false confession almost always preceded the forensic science error, such that the former may have spawned the latter in some cases (62).

In three studies, we varied whether the examiner was exposed to the defendant's confession prior to his analysis, and if so, whether he admitted or denied that it could have influenced his opinion. Study 1 also tested whether the effect of exposure depended on the scientific validity of the examiner's discipline (i.e., bitemarks vs. fingerprints). Study 2 then aimed to partially replicate the findings of Study 1. Finally, Study 3 tested whether the effect of exposure depended on the credentials of the individual examiner, as opposed to the general validity of their discipline.

Study 1

Study 1 aimed to partially replicate and extend the findings of Thompson and Scurich (48) by comparing across forensic science disciplines with a more powerful form of task-irrelevant information (i.e., confession evidence) and novel stimulus materials (e.g., a different crime). Participants read a trial summary in which either a bitemark or fingerprint examiner testified that the defendant's bitemark or fingerprint matched evidence from the crime scene. For some participants, the examiner testified that he was aware of the defendant's confession before he analyzed the forensic evidence, and under cross-examination, he either admitted or denied that it could have influenced his analysis; for other participants, the examiner testified that he was unaware of the defendant's confession. We also included a control condition with no forensic evidence, allowing us to test whether the confession-unaware examiner strengthened the prosecution's case and/or the confession-aware examiner weakened their case.

Method

Participants and Design

Participants ($N = 377$) were recruited online via Amazon Mechanical Turk (mTurk) and completed the study on an external survey website. Overall, 52.3% of participants were female, 74.5% were White, and 53.0% were college-educated. The sample included at least one individual from 43 of the 50 U.S. states.

Each participant was randomly assigned to one of seven cells in a 2 (Discipline: Bitemark vs. Fingerprint) × 3 (Exposure to Confession: Unexposed, Deny, or Admit) + 1 (Control: No Examiner) between-subjects design, with cell sizes ranging from $n = 49$ to 59. See Table 1 for an explanation of the components of each Exposure condition.

Procedure

After providing consent, participants read a summary of a murder trial. The summary first recounted opening statements by the prosecution and defense, which provided background information on the case, including the fact that the defendant had confessed but recanted his confession prior to trial. Then, the summary recounted the testimony and cross-examination of the defendant's coworker, the defendant's neighbor, and by random assignment, either a bitemark examiner (*Bitemark* condition), a fingerprint examiner (*Fingerprint* condition), or neither (*No Examiner* condition). In the Bitemark and Fingerprint conditions, the defense attorney asked the examiner under cross-examination if he had been aware of the defendant's confession prior to analyzing the forensic evidence and, if so, whether it could have influenced his judgment. By random assignment, the examiner replied that he was either unaware of the confession (*Unexposed* condition), aware of it but denied that it influenced his judgment (*Deny* condition), or aware of it and admitted that it could have influenced his judgment (*Admit* condition). Finally, the summary recounted closing statements by the prosecution and defense. After reading the summary, participants rendered a verdict, rated their confidence in that verdict, and provided judgments of the prosecutor, the defense attorney, and each witness.

Trial Summary

Each participant read one of seven versions of a trial summary (943–1286 words) that were newly created for the current studies. The summary recounted the first-degree murder trial of Michael Thompson. In all seven versions, the prosecutor's opening statement argued that Thompson sexually assaulted and murdered his former boss, Jane Anderson, out of anger over having recently been fired. The prosecutor noted that Thompson had previously been accused of sexual assault while in college and that he could not provide an alibi for his whereabouts on the night of the murder. Finally, the prosecutor argued that Thompson had confessed and that his confession contained details that would only be known to the true perpetrator. The defense attorney's opening statement argued that Thompson's confession was the only evidence against him, that it was false, and that it was coerced by detectives who threatened him and fed him details of the crime during a lengthy interrogation.

All participants then read a summary of testimony given by Thompson's coworker, who testified for the prosecution. The coworker testified that he heard Thompson throw a lamp and threaten Mrs. Anderson upon learning that he had been fired. On cross-examination, the coworker testified that Thompson had a bad temper but never acted on it, explaining that Thompson always calmed down quickly, seemed embarrassed by his behavior, and apologized for it.

Next, participants read the testimony of either a bitemark examiner, a fingerprint examiner, or neither (*Discipline* manipulation). Participants in the *Bitemark* condition read the testimony

TABLE 1—*Components of exposure conditions in studies 1–3.*

Condition	Studies	Defendant confessed	Examiner testified	Examiner exposed to confession	Examiner denied influence of confession	Examiner admitted influence of confession
No confession or examiner	2, 3	–	–	–	–	–
No examiner	1, 2, 3	X	–	–	–	–
Unexposed	1, 2, 3	X	X	–	–	–
Exposed	2, 3	X	X	X	–	–
Deny	1, 2, 3	X	X	X	X	–
Admit	1, 2, 3	X	X	X	–	X

An X indicates that this component was present in the corresponding condition.

of a bitemark examiner who testified for the prosecution. The bitemark examiner held a doctoral degree in dental surgery and was a member of the American Society of Forensic Odontology. He explained the methodology he had used to compare photographs of a bitemark on the victim's neck against a model of Thompson's teeth, and he concluded that the bitemark matched the model. Participants in the *Fingerprint* condition read the analogous testimony of a fingerprint examiner who testified for the prosecution. The fingerprint examiner held a Bachelor's degree in forensic science, was trained at an FBI laboratory, and was certified by the International Association for Identification. He too explained the methodology he had used to compare photographs of fingerprints found in the blood surrounding the victim's body against Thompson's fingerprints, and he concluded that the two sets of prints matched. Participants in the *No Examiner* condition did not read either examiner's testimony.

Participants in the Bitemark and Fingerprint conditions also read a summary of the examiner's cross-examination, in which the defense attorney asked the examiner if he was aware of Thompson's confession prior to his analysis (*Exposure* manipulation). In the *Unexposed* condition, the examiner testified that he had not been aware of the confession. In the Deny and Admit conditions, the examiner testified that he was aware of the confession, and the defense attorney asked the examiner if that knowledge could have influenced his conclusion. In the *Deny* condition, the examiner stated that the confession had no bearing on his conclusion and then reiterated this point. In the *Admit* condition, the examiner conceded that the confession could have influenced his conclusion, and the defense attorney then prompted him to reiterate this point.

In all conditions, the final witness was Thompson's neighbor, who testified for the defense. The neighbor testified that she saw Thompson in their apartment building on the night of the murder and that he said he was going to watch a movie at home that night. On cross-examination, the neighbor clarified that she had seen Thompson around 6 p.m. and could not confirm that he was in his apartment for the entire evening. Finally, the prosecution and defense made their closing arguments, which reiterated the main points of their opening statements.

Dependent Measures

See Table 2 for a summary of our dependent measures. After reading the trial summary, participants first rendered a verdict (guilty or not guilty) and rated their confidence in this verdict on a scale from 1 (very unsure) to 10 (very confident). Then, on a new screen, participants separately rated the convincingness of the prosecutor and defense attorney's arguments, each on a scale from 1 (very unconvincing) to 9 (very convincing).

Next, for each witness whose testimony they read (coworker, neighbor, and/or examiner), participants rated the convincingness of their testimony, the importance of their testimony in rendering a verdict, and the degree to which they were biased toward the side for which they testified—each on a 9-point scale. We did not analyze ratings of the two nonexaminer witnesses; these items were included only to conceal the purpose of the study, and we had no reason to believe that our manipulations would affect ratings of these witnesses.

Manipulation Check

Lastly, participants in the six experimental conditions answered two questions to verify that they had noticed and remembered the details of our manipulations (i.e., which type of evidence the examiner had analyzed, and whether the examiner had known about Thompson's confession prior to his analysis). Unfortunately, these data were lost due to a malfunction of the survey website; hence, we did not exclude data from any participants prior to analysis. We return to this point later.

TABLE 2—*Dependent measures in studies 1–3.*

Measure	Study 1	Study 2	Study 3
Verdict (guilty/not guilty)	X	X	X
Likelihood of commission (0–100)		X	X
Prosecutor convincingness (1–9)	X		X
Defense convincingness (1–9)	X		X
Examiner convincingness (1–9)	X		X
Examiner importance (1–9)	X		X
Examiner pro-prosecution bias (1–9)	X		X
Examiner influence on verdict (1–10)		X	
Examiner influenced by confession (1–10)		X	X

An X indicates that this measure was collected in the corresponding study. The five Examiner ratings were not collected in the two control conditions. Ratings of 'Examiner Influenced by Confession' were not collected in the Unexposed condition.

Results

Verdicts

Across all conditions, 68.7% of participants returned a guilty verdict. A full factorial logistic regression revealed that neither Discipline, Wald $\chi^2(1) = 0.17$, $p = 0.679$, $OR = 1.19$ [95% CI: 0.53, 2.67], nor Exposure, Wald $\chi^2(2) = 1.67$, $p = 0.435$, nor their interaction, Wald $\chi^2(2) = 2.01$, $p = 0.366$, predicted verdict. As shown in Table 3, pairwise comparisons indicated that participants in the Fingerprint/Unexposed (80.4%), $\chi^2(1) = 6.87$, $p = 0.009$, $OR = 3.11$ [95% CI: 1.31, 7.38], and Fingerprint/Deny (76.3%), $\chi^2(1) = 4.94$, $p = 0.026$, $OR = 2.44$ [95% CI: 1.10, 5.39], conditions voted guilty more often than those in the No Examiner control condition (56.9%)—but none of the other experimental conditions differed from the No Examiner condition, *p*s > 0.07.

Prosecutor and Defense Convincingness

Ratings of the prosecutor and defense attorney were negatively correlated, $r = -0.62$, $p < 0.001$; overall, participants rated the prosecutor's argument as more convincing ($M = 6.80$, $SD = 2.03$) than the defense attorney's argument ($M = 4.11$, $SD = 2.29$), $t(376) = 13.41$, $p < 0.001$, $d = 0.69$ [95% CI: 0.58, 0.81].

A 2 (Discipline) X 3 (Exposure) MANOVA on ratings of the prosecution and defense revealed a multivariate effect of Exposure, $F(4,624) = 2.68$, $p = 0.031$, $\eta^2_p = 0.02$, with univariate effects on both prosecutor convincingness, $F(2,313) = 3.38$, $p = 0.035$, $f = 0.15$, and defense convincingness, $F(2,313) = 5.03$, $p = 0.007$, $f = 0.18$. As shown in Table 4, Tukey HSD comparisons showed that participants in the Admit condition rated the prosecutor's argument as less convincing than those in the Deny condition—neither of which differed from the Unexposed condition. Conversely, participants in the Admit condition rated the defense's argument as more convincing than those in the Deny or Unexposed conditions, which did not differ from each other.

TABLE 3—*Effects of exposure, discipline (bitemark vs. fingerprint), and credentials (weak vs. strong) on guilty verdicts (%) in studies 1–3.*

	No confession or examiner	No examiner	Unexposed	Exposed	Deny	Admit
Study 1		56.9	70.8		75.0	66.7
Bitemark			61.8		73.5	64.7
Fingerprint			*80.4*		*76.3*	68.5
Study 2	28.1_{a}	44.8_{b}	77.8_{c}	52.2_{abc}	62.5_{bc}	55.6_{bc}
Study 3	10.3	53.1	74.5_{a}	62.8_{a}	65.7_{a}	40.8_{b}
Weak			62.5	60.0	61.1	34.8
Strong			*87.0*	65.2	70.6	46.2

The suspect had confessed in all conditions unless otherwise noted (see Table 1). Marginal values in the same row that do not share a common subscript differ at $p < 0.05$. Cell values that are shown in italics differ from the corresponding No Examiner control group at $p < 0.05$.

Neither the multivariate effect of Discipline, $F(2,312) = 1.57$, $p = 0.210$, $\eta^2_p = 0.01$, nor the multivariate interaction, $F(4,624) = 0.83$, $p = 0.508$, $\eta^2_p = 0.01$, was significant.

As shown in Table 4, a one-way ANOVA with planned contrasts indicated that participants in the Fingerprint/Unexposed, $p = 0.015$, and Fingerprint/Deny, $p = 0.015$, conditions rated the prosecutor's argument as more convincing than those in the No Examiner control condition—but no other conditions differed from the No Examiner condition, ps > 0.35. Ratings of the defense attorney's argument did not differ between the No Examiner control condition and any of the six experimental conditions, ps > 0.12.

Judgments of the Examiner

A 2 X 3 MANOVA on three judgments of the forensic examiner's testimony (i.e., convincingness, importance, and pro-prosecution bias) revealed a multivariate effect of Exposure, $F(6,624) = 5.39$, $p < 0.001$, $\eta^2_p = 0.05$, with univariate effects on convincingness, $F(2,313) = 13.68$, $p < 0.001$, $f = 0.30$, and pro-prosecution bias, $F(2,313) = 3.22$, $p = 0.041$, $f = 0.14$, but not importance, $F(2,313) = 1.42$, $p = 0.245$, $f = 0.10$. As shown in Table 5, Tukey HSD comparisons indicated that participants in the Unexposed condition rated the examiner's testimony as more convincing than those in the Deny condition, who in turn rated it as more convincing than those in the Admit condition. Participants in the Admit condition rated the examiner's testimony as less biased toward the prosecution than those in the Unexposed condition—neither of which differed from the Deny condition.

Neither the multivariate effect of Discipline, $F(3,311) = 1.03$, $p = 0.378$, $\eta^2_p = .01$, nor the multivariate interaction, $F(6,624) = 0.80$, $p = 0.571$, $\eta^2_p = 0.01$, was significant.

Discussion

Study 1 participants rated a forensic examiner's testimony as less convincing if he had *a priori* knowledge of the defendant's confession than if he did not. This finding implies that laypeople possess at least some understanding that cognitive bias can undermine the value of forensic science evidence, which is consistent with Thompson and Scurich's (48) finding that venirepersons rated a blinded forensic expert as more credible than one who was exposed to potentially biasing information. However, in contrast to Thompson and Scurich, Study 1 participants were no less likely to convict the defendant—the only legally relevant outcome measure—if the examiner had been aware of the defendant's confession.

A closer look at our data reveals an even more concerning departure from Thompson and Scurich's findings. In their study, participants rated the nonblind examiner as less credible regardless of whether he claimed to have used or ignored task-irrelevant information. In Study 1, however, several findings suggest

TABLE 4—*Effects of exposure, discipline (bitemark vs. fingerprint), and credentials (weak vs. strong) on verdict-confidence ratings (study 1), likelihood of commission estimates (studies 2 and 3), and convincingness of the prosecution and defense (studies 1 and 3).*

	No confession or examiner	No examiner	Unexposed	Exposed	Deny	Admit
Likelihood of commission (0-100)						
Study 2	$51.56\ (3.97)_{a}$	$61.03\ (4.28)_{ab}$	$78.33\ (4.30)_{b}$	$67.39\ (4.92)_{ab}$	$62.92\ (6.00)_{ab}$	$68.89\ (5.02)_{ab}$
Study 3	46.21 (4.95)	60.63 (4.21)	$75.32\ (2.96)_{a}$	$71.63\ (3.79)_{ab}$	$75.14\ (3.49)_{a}$	$61.02\ (3.77)_{b}$
Weak			69.58 (4.72)	69.00 (5.47)	72.78 (4.03)	53.04 (5.95)
Strong			*81.30 (3.16)*	*73.91 (5.33)*	*77.65 (5.85)*	68.08 (4.44)
Prosecutor convincingness (1–9)						
Study 1		6.41 (0.25)	$7.05\ (0.17)_{ab}$		$7.10\ (0.20)_{a}$	$6.44\ (0.22)_{b}$
Bitemark			6.76 (0.26)		6.84 (0.31)	6.57 (0.32)
Fingerprint			*7.35 (0.21)*		*7.32 (0.25)*	6.31 (0.30)
Study 3	4.17 (0.41)	5.48 (0.42)	$7.19\ (0.23)_{a}$	$6.81\ (0.29)_{ab}$	$7.20\ (0.24)_{a}$	$5.90\ (0.31)_{b}$
Weak			*6.96 (0.37)*	*6.75 (0.41)*	*6.89 (0.38)*	5.70 (0.46)
Strong			*7.43 (0.25)*	*6.87 (0.42)*	*7.53 (0.27)*	6.08 (0.43)
Defense convincingness (1–9)						
Study 1		4.26 (0.31)	$3.85\ (0.20)_{a}$		$3.76\ (0.23)_{a}$	$4.66\ (0.23)_{b}$
Bitemark			4.02 (0.31)		3.73 (0.32)	4.37 (0.36)
Fingerprint			3.67 (0.25)		3.78 (0.32)	4.93 (0.28)
Study 3	6.62 (0.43)	5.23 (0.36)	$3.96\ (0.29)_{a}$	$4.19\ (0.40)_{a}$	$4.60\ (0.41)_{ab}$	$5.84\ (0.33)_{b}$
Weak			4.04 (0.43)	4.70 (0.56)	4.78 (0.56)	5.87 (0.48)
Strong			*3.87 (0.39)*	*3.74 (0.56)*	4.41 (0.61)	5.81 (0.46)

All values are presented as *M* (*SE*). Marginal means in the same row that do not share a common subscript differ at $p < 0.05$. Cell means that are shown in italics differ from the corresponding No Examiner control group at $p < 0.05$.

TABLE 5—*Effects of exposure, discipline (bitemark vs. fingerprint), and credentials (weak vs. strong) on judgments of the examiner (M [SE]) in studies 1–3.*

	Unexposed	Exposed	Deny	Admit
Convincingness				
Study 1	7.42 (0.14)$_{a}$		6.75 (0.21)$_{b}$	6.02 (0.22)$_{c}$
Bitemark	7.27 (0.20)		6.55 (0.30)	6.20 (0.31)
Fingerprint	7.59 (0.18)		6.92 (0.28)	5.85 (0.30)
Study 3	7.20 (0.31)$_{a}$	6.70 (0.33)$_{a}$	6.69 (0.44)$_{a}$	4.88 (0.37)$_{b}$
Weak	6.54 (0.50)	6.20 (0.48)	6.72 (0.53)	4.26 (0.51)
Strong	7.91 (0.31)	7.13 (0.44)	6.65 (0.74)	5.42 (0.52)
Importance				
Study 1	7.33 (0.16)		7.07 (0.19)	6.91 (0.19)
Bitemark	7.15 (0.25)		6.86 (0.30)	6.78 (0.31)
Fingerprint	7.53 (0.21)		7.25 (0.25)	7.04 (0.22)
Study 3	7.59 (0.27)	7.26 (0.27)	7.40 (0.39)	7.57 (0.22)
Weak	7.29 (0.45)	7.00 (0.38)	7.67 (0.44)	7.65 (0.35)
Strong	7.91 (0.29)	7.48 (0.38)	7.12 (0.65)	7.50 (0.29)
Pro-prosecution bias				
Study 1	4.67 (0.10)$_{a}$		4.32 (0.15)$_{ab}$	4.14 (0.18)$_{b}$
Bitemark	4.76 (0.16)		4.33 (0.23)	4.00 (0.26)
Fingerprint	4.57 (0.13)		4.32 (0.20)	4.28 (0.25)
Study 3	4.48 (0.23)	4.07 (0.23)	4.03 (0.29)	3.61 (0.30)
Weak	4.29 (0.30)	3.70 (0.36)	4.39 (0.40)	3.48 (0.43)
Strong	4.68 (0.34)	4.39 (0.29)	3.65 (0.41)	3.73 (0.42)
Influence on verdict				
Study 2	8.17 (0.69)	7.70 (0.48)	7.83 (0.49)	6.78 (0.52)
Influenced by confession				
Study 2		5.26 (0.70)	4.33 (0.59)	5.44 (0.46)
Study 3		4.58 (0.40)$_{a}$	4.46 (0.50)$_{a}$	5.98 (0.35)$_{b}$
Weak		5.70 (0.51)	4.11 (0.68)	6.61 (0.49)
Strong		3.61 (0.53)	4.82 (0.75)	5.42 (0.47)

Marginal means in the same row that do not share a common subscript differ at $p < 0.05$.

that our participants believed the examiner when he claimed that the confession did not influence him. Overall, participants rated the examiner's testimony as more convincing when he denied influence than when he admitted to the possibility. Moreover, when the fingerprint examiner denied influence, participants voted to convict more often and rated the prosecutor's argument as more convincing compared to the no examiner control condition. For both measures, an examiner who knew of the confession but denied its influence was perceived no differently than an examiner who was unaware of the confession.

This pattern is problematic insofar as many forensic examiners resist claims of cognitive bias. As noted above, survey data from Kukucka et al. (24) revealed a "bias blind spot" among forensic examiners, such that examiners generally recognized that bias can impact their peers but simultaneously denied that bias can impact their own judgments (see also [63]). Hence, even if an examiner were to be cross-examined on this issue, they will presumably deny that task-irrelevant information could have impacted their judgment—and the results of Study 1 suggest that jurors will tend to believe them.

Though latent fingerprint examination is widely considered to have a stronger scientific foundation than bitemark identification (23), participants' ratings generally did not reflect this, as bitemark and fingerprint examiners were rated as equally convincing and equally influential, and their testimony was equally likely to elicit a guilty verdict. However, the unexposed fingerprint examiner—but not the unexposed bitemark examiner—increased the conviction rate and perceived strength of the prosecutor's argument relative to the no examiner condition, thus providing some indication that participants considered fingerprint evidence more probative than bitemark evidence.

A glaring limitation of Study 1 was the unfortunate loss of manipulation check data, which prevented us from verifying that all participants had noticed critical aspects of the examiner's testimony. In theory, including inattentive participants in our final sample would increase statistical noise and make it more difficult to detect significant effects. Still, this problem may have caused us to underestimate the magnitude of the observed effects and/or fail to detect other important effects. For that reason, we sought to replicate the findings of Study 1.

Study 2

The primary aim of Study 2 was to replicate the effects observed in the Fingerprint condition of Study 1. Each participant read one of six versions of a trial summary, four of which were the Unexposed, Deny, Admit, and No Examiner summaries used in Study 1. A fifth version included a fingerprint examiner who admitted to knowing about the defendant's confession but neither admitted nor denied that it could have influenced his judgment (i.e., Exposed condition). A sixth version provided a second control group in which the defendant did not confess and no forensic examiner testified.

Method

Participants and Design

Participants ($N = 186$) were recruited via mTurk and completed the study on an external survey website. We later excluded data from 33 participants (17.7%) who failed a manipulation check, leaving a final sample of $N = 153$. No demographic information was collected.

Each participant was randomly assigned to one of six groups in a one-way design (i.e., Unexposed, Exposed, Deny, Admit, No Examiner, or No Confession or Examiner; see Table 1 for an explanation of the components of each condition).

Procedure

Each participant read one of six versions of a trial summary, which were adapted from the summaries used in Study 1. After reading the summary, participants rendered a verdict, rated their confidence in that verdict, estimated the likelihood that the defendant had committed the murder, and rated the degree to which each witness' testimony (i.e., the coworker, neighbor, and/or fingerprint examiner) had influenced their verdict. Finally, participants completed a manipulation check.

Trial Summary

Each participant read one of six versions of a trial summary (550–1044 words) in which the defendant was charged with first-degree murder. The summaries were identical to those used in Study 1, apart from the exceptions described below. As in Study 1, each summary included opening statements by the prosecution and defense, testimony and cross-examination of the defendant's coworker and neighbor, and closing statements.

Four of the six summaries also included the testimony and cross-examination of a fingerprint examiner who testified that the defendant's fingerprints matched fingerprints found in the blood around the victim's body (i.e., experimental conditions; see Table 1). In each, the defense attorney asked the examiner if he had been aware of the defendant's confession prior to analyzing the fingerprints; the four summaries differed in how the examiner responded to this question. In the *Unexposed* condition, the examiner testified that he was unaware of the confession when he analyzed the fingerprints. In the *Exposed* condition, the examiner testified that he was aware of the confession; however, the examiner was not asked if it could have influenced his judgment. In the *Deny* condition, the examiner testified that he was aware of the confession but insisted that it did not influence his judgment, which he described as “purely objective.” In the *Admit* condition, the examiner testified that he was aware of the confession and admitted that it could have influenced his judgment, which he described as “partly subjective.”

Two of the six summaries did not include testimony from a fingerprint examiner (i.e., control conditions; see Table 1). In the *No Examiner* condition, participants read only opening statements (which mentioned the defendant’s recanted confession), testimony from the defendant’s coworker and neighbor, and closing statements. In the *No Confession or Examiner* condition, they read this same summary, but with all references to the defendant’s confession and fingerprint examiner removed.

Dependent Measures

See Table 2 for a summary of our dependent measures. As in Study 1, participants first rendered a verdict (guilty or not guilty) and rated their confidence in that verdict on a 10-point scale. To obtain a more sensitive measure of belief in the defendant’s guilt, we also asked Study 2 participants to estimate the likelihood that the defendant had committed the murder, using an 11-point scale with options ranging from 0% to 100%.

Whereas Study 1 participants rated the convincingness, importance, and biasedness of each witness’ testimony, Study 2 participants provided only one rating for each witness—namely, the extent to which that witness’ testimony influenced their verdict, on a scale from 1 (not at all) to 10 (very). As in Study 1, we did not analyze participants’ ratings of the nonexaminer witnesses.

Lastly, we asked participants in the Exposed, Deny, and Admit conditions to rate the extent to which they believed that the fingerprint examiner’s analysis was influenced by his knowledge of the defendant’s confession, on a scale from 1 (not at all) to 10 (very). This item was presented after the manipulation check so as not to prime participants’ responses to those items.

Manipulation Check

Participants in the four experimental conditions answered one or two items to verify that they noticed and remembered key aspects of the fingerprint examiner’s testimony. The first item (included in all four experimental conditions) asked if the examiner knew about the defendant’s confession prior to analyzing the fingerprints. The second item (Deny and Admit conditions only) asked if the examiner said it was or was not possible that the confession influenced his analysis. We later excluded data from participants who responded incorrectly to either or both of these items (n = 33; 17.7%), leaving a final sample of N = 153.

Results

Verdicts

Overall, 52.2% of participants returned a guilty verdict—including 28.1% in the No Confession or Examiner control condition and a combined 57.0% in the five conditions where the defendant confessed, $\chi^2(1) = 8.46$, $p = 0.004$, $OR = 3.39$ [95% CI: 1.45, 7.94]. A chi-square test revealed an overall effect of our manipulation on verdicts, $\chi^2(5) = 13.81$, $p = 0.017$, $V = 0.30$ (see Table 3). Compared to the No Confession or Examiner control condition, the conviction rate significantly increased in the Unexposed (77.8%), Deny (62.5%), and Admit (55.6%) conditions. In contrast, only the Unexposed condition increased the conviction rate above the No Examiner control condition (44.8%). However, the conviction rate also did not significantly differ between the four experimental conditions (i.e., Unexposed, Exposed, Deny, and Admit).

Likelihood of Commission

On average, participants estimated a 63.73% likelihood that the defendant had committed the murder (SD = 25.10%). A one-way ANOVA revealed a significant effect of our manipulation on these estimates, $F(5,147) = 3.36$, $p = 0.007$, $f = 0.34$. As shown in Table 4, Tukey HSD comparisons revealed that participants in the Unexposed condition thought it more likely that the defendant had committed the crime compared to those in the No Confession or Examiner condition, and no other conditions differed from each other.

Judgments of the Examiner

Participants in the four experimental conditions reported that the examiner’s testimony strongly influenced their verdict (overall $M = 7.55$, $SD = 2.59$). A one-way ANOVA indicated that these ratings did not differ between conditions, $F(3,88) = 1.27$, $p = 0.289, f = 0.21$ (see Table 5).

Participants in the Exposed, Deny, and Admit conditions believed that the examiner’s analysis was somewhat influenced by his knowledge of the defendant’s confession (overall $M = 5.03$, $SD = 2.89$). A one-way ANOVA indicated that these ratings did not differ between conditions, $F(2,71) = 1.05$, $p = 0.356, f = 0.17$ (see Table 5).

Discussion

Study 2 participants recognized that a fingerprint examiner’s knowledge of a defendant’s confession can influence his judgment, and they believed that the confession would have the same impact regardless of whether the examiner admitted or denied that such influence was possible. This recognition, however, was not reflected in their judgments and verdicts: Participants who read testimony from an unexposed or exposed fingerprint examiner felt equally influenced by his testimony, believed it equally likely that the defendant had committed the crime, and were equally likely to return a guilty verdict. As such, we replicated the finding from Study 1 that the fingerprint examiner’s testimony vis-à-vis bias did not affect judgments of guilt.

As in Study 1, the unexposed examiner increased the conviction rate above the no examiner control condition—but unlike in Study 1, the examiner who denied being influenced by the confession did not increase the conviction rate in Study 2. That is to say, the examiner’s testimony increased guilty verdicts relative to the control condition in Study 2 only when the examiner was unexposed. This pattern may suggest optimism regarding jurors’ ability to detect and discount forensic opinions that have been compromised by cognitive bias. However, the rate of guilty verdicts also increased—albeit nonsignificantly—in each of the three exposed examiner conditions relative to the control condition, and participants in those conditions still reported that the exposed examiner's testimony strongly influenced their verdicts. Thus, this pattern may instead indicate that Study 2 was underpowered to detect differences in conviction rate that are relatively small in magnitude but perhaps still practically important.

In contrast, likelihood of commission ratings was equivalent across the four experimental conditions and the no examiner control condition, which may suggest that participants universally gave little weight to the fingerprint examiner's testimony. As discussed earlier, prior work suggests that laypeople's trust in fingerprint evidence is highly variable (e.g., [38,58]). Accordingly, the weak differences between conditions could be partly attributable to a subset of participants who are generally skeptical of fingerprint evidence and felt that it should carry little weight regardless of the conditions under which it was analyzed.

Perhaps jurors' judgments are influenced more by characteristics of the individual examiner than by that examiner's discipline as a whole. Consistent with dual-process models of persuasion (64), jurors who are unable or unmotivated to fully process a message may be more heavily influenced by peripheral cues—such as the examiner's credentials or confidence—and less influenced by the degree to which the examiner's conclusions are scientifically valid. Supporting this idea, Koehler et al. (41) found that an examiner's level of experience—but not the degree to which their method had been scientifically validated—reliably affected laypeople's judgments of the strength of the evidence. Relatedly, although many forensic examiners agree that experience does not immunize examiners from cognitive bias (24), jurors might believe that a forensic examiner who has more training and experience is necessarily less susceptible to bias. Study 3 tested this possibility.

Study 3

Study 3 tested whether the effect of an examiner admitting exposure to incriminating but task-irrelevant information depends on the examiner's credentials. Similar to Study 2, most participants read a trial summary in which a fingerprint examiner—with either strong or weak credentials—testified that he either was or was not aware of the defendant's confession prior to his analysis, and if he was aware, either admitted or denied that the confession could have influenced his judgment (or neither). As in Study 2, we also included two control conditions that did not include testimony from a fingerprint examiner and/or a confession.

Method

Participants and Design

Participants ($N = 315$) were recruited via mTurk and completed the study on an external survey website. We later excluded data from 80 participants (25.4%) who failed a manipulation check, leaving a final sample of $N = 235$. Our final sample included at least one participant from 35 of the 50 U.S. states, and a slight minority of participants (48.1%) held a Bachelor's degree or higher. No other demographic information was collected.

Each participant was randomly assigned to one of ten cells in a 2 (Credentials: Weak vs. Strong) × 4 (Exposure: Unexposed, Exposed, Deny, or Admit) + 2 (Controls: No Examiner; No Confession or Examiner) between-subjects design.

Procedure

As in Studies 1 and 2, participants read one of ten versions of a trial summary (550–1065 words), which recounted opening and closing statements from the prosecution and defense and testimony from the defendant's coworker and neighbor. Eight of the ten summaries also recounted the testimony of a fingerprint examiner, which varied in terms of the examiner's credentials and his explanation of whether he was aware of and/or influenced by the defendant's confession. Participants then completed an amalgamation of the dependent measures from Studies 1 and 2 (see Table 2) as well as a manipulation check.

Trial Summary

Each participant read one of ten trial summaries, which were identical to those used in Study 2, apart from the *Credentials* manipulation. In the *Strong Credentials* condition, the examiner held a Bachelor's degree in biophysical chemistry from a prestigious college, held a Master's degree in forensic science, and had been working at the crime laboratory for 10 years. In the *Weak Credentials* condition, the examiner held a Bachelor's degree in communications from a relatively unknown college, had completed an online certificate in fingerprint analysis, and had been working at the crime laboratory for 18 months.

Dependent Measures

See Table 2 for a summary of our dependent measures. Study 3 included all of the dependent measures from Study 1, plus two items that were added in Study 2 (i.e., likelihood of commission and—for the Exposed, Deny, and Admit conditions only—a rating of the degree to which the examiner's knowledge of the confession influences his analysis).

Manipulation Check

Participants in the eight experimental conditions answered three or four items to verify that they noticed and remembered key aspects of the fingerprint examiner's testimony. The first two items asked participants to recall how long the fingerprint examiner had been working at the crime lab (i.e., 6 months, 18 months, 4 years, 10 years, or 15 years) and rate the strength of the examiner's credentials on a scale from 1 (not strong) to 10 (very strong). The third and fourth items were identical to the manipulation check items from Study 2.

We later excluded data from participants who responded incorrectly to any of the three nonscalar items ($n = 80$; 25.4%), leaving a final sample of $N = 235$. Confirming the effectiveness of our manipulation, participants in the Strong Credentials condition rated the examiner's credentials ($M = 7.55$, $SD = 1.43$) as much stronger than those in the Weak Credentials condition ($M = 4.49$, $SD = 2.10$), $t(170) = 11.21$, $p < 0.001$, $d = 1.76$ [95% CI: 1.41, 2.11].

Results

Verdicts

Overall, 53.2% of participants returned a guilty verdict—including 10.3% in the No Confession or Examiner control group and a combined 59.2% in the other conditions, $\chi^2(1) = 24.39$,

$p < 0.001$, *OR* = 12.59 [95% CI: 3.69, 42.93]. Compared to the No Confession or Examiner condition, the conviction rate significantly increased in the No Examiner condition (53.1%), $p < 0.001$, and in all eight experimental conditions, *p*s < 0.04. Additional pairwise comparisons indicated that only the Strong/Unexposed condition (87.0%) produced a higher conviction rate than the No Examiner control condition, $p = 0.008$; none of the other experimental conditions differed from the No Examiner condition, *p*s > 0.23 (see Table 3).

A full factorial logistic regression revealed that Exposure significantly predicted verdicts, Wald $\chi^2(3) = 8.33$, $p = 0.040$, such that the conviction rate was lower in the Admit condition (40.8%) compared to the Unexposed (74.5%), Deny (65.7%), or Exposed (62.8%) conditions, which did not differ (see Table 3). Neither Credentials, Wald $\chi^2(1) = 0.35$, $p = 0.556$, *OR* = 0.66 [95% CI: 0.16, 2.68], nor the interaction, Wald $\chi^2(3) = 1.56$, $p = 0.669$, predicted verdicts.

Likelihood of Commission

A 2 X 4 ANOVA on likelihood of commission estimates revealed a main effect of Exposure, $F(3,166) = 4.26$, $p = 0.006$, $f = 0.28$. As shown in Table 4, Tukey HSD comparisons indicated that participants in the Unexposed and Deny conditions thought it more likely that the defendant had committed the crime compared to the Admit condition—none of which differed from the Exposed condition.

A main effect of Credentials also emerged, $F(1,166) = 6.71$, $p = 0.010$, $d = 0.39$ [95% CI: 0.09, 0.69], such that participants who read testimony from a Strong examiner thought it more likely that the defendant committed the crime ($M = 74.83$, $SD = 22.32$) compared to those who read the same testimony from a Weak examiner ($M = 65.65$, $SD = 24.81$). No Credentials X Exposure interaction was found, $F(3,166) = 0.53$, $p = 0.662$, $f = 0.10$.

A one-way ANOVA with planned contrasts indicated that the Strong/Unexposed, $p = 0.001$, Strong/Deny, $p = 0.015$, and Strong/Exposed, $p = 0.037$, conditions produced higher likelihood of commission estimates than the No Examiner control condition (see Table 4). None of the other experimental conditions differed from the No Examiner condition, all *p*s > 0.075.

Prosecution and Defense Arguments

As in Study 1, ratings of the prosecution and defense were negatively correlated, $r = -0.57$, $p < 0.001$, and participants rated the prosecutor's argument as more convincing ($M = 6.25$, $SD = 2.17$) than the defense attorney's argument ($M = 4.99$, $SD = 2.45$), $t(233) = 4.72$, $p < 0.001$, $d = 0.31$ [95% CI: 0.16, 0.45].

A 2 X 4 MANOVA on ratings of the prosecution and defense revealed a multivariate effect of Exposure, $F(6,330) = 3.94$, $p = 0.001$, $\eta^2_p = 0.07$, with univariate effects on both prosecutor convincingness, $F(3,166) = 5.31$, $p = 0.002$, $f = 0.31$, and defense convincingness, $F(3,166) = 6.04$, $p = 0.001$, $f = 0.33$. As shown in Table 4, Tukey HSD comparisons indicated that participants in the Unexposed and Deny conditions rated the prosecutor's argument as more convincing than those in the Admit condition—none of which differed from the Exposed condition. Conversely, participants in the Admit condition rated the defense attorney's argument as more convincing than those in the Unexposed or Exposed conditions—none of which differed from the Deny condition.

Neither the multivariate effect of Credentials, $F(2,165) = 1.12$, $p = 0.327$, $\eta^2_p = .01$, nor the multivariate interaction, $F(6,330) = 0.41$, $p = 0.875$, $\eta^2_p = 0.01$, was significant.

A one-way ANOVA with planned contrasts indicated that six of the eight experimental conditions—that is, Strong/Unexposed, $p < 0.001$, Strong/Deny, $p < 0.001$, Strong/Exposed, $p = 0.009$, Weak/Unexposed, $p = 0.005$, Weak/Deny, $p = 0.014$, and Weak/Exposed, $p = 0.022$—rated the prosecutor's argument as more convincing than the No Examiner control condition, all other *p*s > 0.24 (see Table 4). Conversely, only two conditions—Strong/Unexposed, $p = 0.033$, and Strong/Exposed, $p = 0.020$—rated the defense attorney's argument as less convincing than the No Examiner condition, all other *p*s > 0.059.

Judgments of the Examiner

A 2 X 4 MANOVA on three judgments of the examiner's testimony (i.e., convincingness, importance, and pro-prosecution bias) revealed a multivariate effect of Exposure, $F(9,495) = 3.94$, $p < 0.001$, $\eta^2_p = 0.07$, with a univariate effect on convincingness, $F(3,165) = 9.35$, $p < 0.001$, $f = 0.41$, but no effect on importance, $F(3,165) = 0.37$, $p = 0.777$, $f = 0.08$, or bias, $F(3,165) = 2.02$, $p = 0.113$, $f = 0.19$. As shown in Table 5, Tukey HSD comparisons indicated that participants in the Admit condition rated the examiner's testimony as less convincing than those in the Unexposed, Exposed, and Deny conditions, which did not differ from each other. Neither the multivariate effect of Credentials, $F(3,163) = 2.08$, $p = 0.105$, $\eta^2_p = 0.04$, nor the multivariate interaction, $F(9,495) = 0.72$, $p = 0.692$, $\eta^2_p = 0.01$, was significant.

Overall, participants in the Exposed, Deny, and Admit conditions believed that the examiner's analysis was somewhat influenced by his knowledge of the defendant's confession ($M = 5.09$, $SD = 2.72$). A 2 X 3 ANOVA on these ratings revealed a main effect of Exposure, $F(2,121) = 4.78$, $p = .010$, $f = 0.28$ (see Table 5), such that participants in the Admit condition felt that the examiner was more influenced by the confession than those in the Deny or Exposed conditions, which did not differ from each other. Neither a main effect of Credentials, $F(1,121) = 3.43$, $p = .067$, $f = 0.17$, nor an interaction, $F(2,121) = 2.93$, $p = .057$, $f = 0.22$, was found.

Discussion

As in Studies 1 and 2, participants in Study 3 showed only a limited understanding that exposure to a confession could impact a fingerprint examiner's judgment. As such, exposed and unexposed examiners generally had the same impact on their decision-making. Moreover, Study 3 yielded additional evidence that people tend to believe fingerprint examiners who claim to be impervious to bias. We return to these points in the General Discussion.

The examiner's credentials had little effect on perceptions or judgments. Testimony from the stronger-credentialed examiner increased likelihood of commission ratings relative to the weaker-credentialed examiner, but it did not affect verdicts or the examiner's perceived credibility, nor did it moderate the effect of exposure to the defendant's confession. This pattern was surprising given that jurors consider experience and education to be the most important qualifications of a forensic science expert (65) and that prior studies found that jurors were more heavily influenced by forensic examiners with greater experience (41). Additionally, though many examiners believe that their vulnerability to bias declines with experience (24), this pattern suggests that jurors view examiners as equally affected (or unaffected, as the case may be) by cognitive bias regardless of their credentials. To date, only one study has compared the biasability of forensic experts and novices: van den Eeden, de Poot, and van Koppen (66) found that biasing information had the same impact on forensic science students' and experienced crime scene investigators' judgments of a mock crime scene. Future work should continue to explore how experience impacts professional forensic examiners' credibility and biasability.

General Discussion

Forensic science evidence can be an extremely powerful tool in solving crimes. However, misleading forensic science has also contributed to an alarming number of wrongful convictions (1). In recent years, research has overwhelmingly indicated that when forensic examiners are exposed to information that is not germane to their analysis, that knowledge can color their opinions of forensic evidence, thereby undermining the independent probative value of those opinions (2,17,21). The current studies tested whether laypeople discount the testimony of a forensic examiner who was exposed to incriminating but task-irrelevant information (i.e., a confession) that could have influenced his opinion.

Across three studies, we found little evidence that laypeople gave more weight to the forensic examiner who was not exposed to task-irrelevant information. Although they rated the unexposed examiner's testimony as most convincing (Study 1) and as least influenced by the defendant's confession (Study 2), they also reported that the examiner's testimony had the same impact on their verdict whether he was exposed to the confession or not (Studies 1, 2, and 3). Accordingly, exposed and unexposed examiners elicited equivalent conviction rates (Studies 1, 2, and 3) and likelihood of commission estimates (Studies 2 and 3). These findings suggest that jurors should not be trusted to recognize and discount forensic science opinions that have been impacted by cognitive bias. This conclusion echoes previous research showing that jurors are poor at recognizing flaws in scientific evidence—including experimenter expectancy effects (36,37) and confirmation bias (47).

Although participants were largely unable to recognize bias on their own, they generally recognized bias when it was made explicit. When the forensic examiner openly admitted that his knowledge of the defendant's confession could have influenced his opinion, participants rated the prosecution's argument as weakest, the defense's argument as strongest, and the examiner's testimony as least convincing (Studies 1 and 3), and produced the fewest guilty verdicts and lowest likelihood of commission estimates (Study 3). It is unlikely, however, that real world jurors would hear an examiner make such a concession. A recent study found that defense attorneys failed to recognize when confirmation bias had tainted evidence against their client, such that only 46% of attorneys who read a patently biased autopsy report asked the medical examiner about potential bias during cross-examination (67). Moreover, while many examiners recognize that cognitive bias can affect their peers, few believe that they are personally vulnerable to bias (i.e., a "bias blind spot;" [24–25]). To the contrary, many examiners continue to insist that task-irrelevant information does not affect, and may even benefit, the accuracy of their judgments (e.g., [68–72]).

Thus, even if asked, many forensic examiners would deny being susceptible to bias—and our results suggest that jurors believe these denials. When our examiner denied that the defendant's confession impacted his judgment, participants believed that he was in fact less influenced by it (Study 3) and rated his testimony as more convincing (Study 1) compared to when the examiner admitted to possible influence. Moreover, participants weighted the testimony of the exposed examiner who denied influence just as heavily as that of the unexposed examiner, as evidenced by equivalent conviction rates (Studies 1, 2, and 3), likelihood of commission estimates (Studies 2 and 3), and ratings of the convincingness of the prosecutor's argument (Studies 1 and 3) and the examiner's testimony (Study 3). Jurors can be discerning in their reliance on expert testimony (e.g., (73,74]), but our participants clearly trusted the examiner who claimed to be impervious to bias—even though the extant literature indicates that this trust is misplaced (2,6).

It is important to note that while further studies should seek to replicate our findings, the basic pattern of results is highly consistent with research in nonforensic domains. Specifically, research has shown that observers tend to focus on persons and underestimate the role of situational factors. This pattern was first reported by Jones and Harris (75) who presented participants with an essay purportedly written by a fellow student. As one would expect, participants inferred the student's true opinion from his essay when he freely chose the position he espoused. But they also inferred that student's opinion in a no-choice condition in which he was assigned to the position he took. Across topics and modalities, this finding that observers lack sensitivity to the situational determinants of behavior is robust (28). In one experiment, participants were themselves assigned to take a position, whereupon they swapped essays and rated each other. Remarkably, they still jumped to conclusions about each other's attitudes (76). In another experiment, participants inferred attitudes from a speech even when they were the ones who had assigned the writer to the position to be taken (77).

The tendency of social perceivers to focus on persons and overlook the impact of situations is so pervasive that it has been called the *fundamental attribution error* ([29]; for a retrospective overview, see [30]). In *The Psychology of Interpersonal Relations,* Fritz Heider (78), the architect of attribution theory, noted that behavior "has such salient properties that it tends to engulf the field rather than be confined to its proper position as a local stimulus whose interpretation requires the additional data of a surrounding field—the situation in social perception" (p. 54). Gilbert and Malone (27) offered additional explanations, the most compelling being that people draw quick, reflex-like dispositional inferences from behavior and then fail to adjust or correct for the presence of situational constraints, an effortful process.

The fundamental attribution error can also be seen in studies of how mock juries evaluate confession evidence. Kassin and Sukel (79) presented mock jurors with one of three versions of a murder trial transcript—a low-pressure version in which the defendant had confessed to police immediately upon questioning; a high-pressure version in which the defendant was interrogated aggressively and at length; and a control version that contained no confession. In the high-pressure condition, participants perceived the statement to be coerced. Yet this confession significantly increased their conviction rate—regardless of coercion. This same pattern was also obtained in a study of judges (80). Similarly, jurors tend to trust secondary confession evidence, even when provided by an informant with an obvious incentive to lie (81). In short, just as factfinders do not sufficiently account for the pressures that elicit a confession, our participants evaluated forensic science examiners in a manner that was largely oblivious to their exposure to biasing information.

In Thompson and Scurich (48), venirepersons likewise rated the testimony of a nonblind examiner who denied bias as no less credible than that of an examiner who was not exposed to task-irrelevant information. The current studies therefore replicated this finding with novel stimulus materials that described a different discipline (fingerprints vs. bitemarks), crime (murder vs. assault), and form of task-irrelevant information (confession vs. criminal history). Notably, however, Thompson and Scurich's participants did devalue the nonblind examiner relative to an examiner who was deliberately kept unaware of task-irrelevant information as per his laboratory's standard procedures, whereas the current studies did not include an analogous condition. Perhaps jurors are not affected by the examiner's unawareness *per se*, but rather by his description of his laboratory's purposeful blinding procedures. Future research should explore this possibility.

Along these same lines, the fingerprint examiner in Studies 2 and 3 described his analysis as "partly subjective" when he admitted that bias was possible but described it as "purely objective" when he denied bias, such that his self-described biasability was confounded with his characterization of the analysis. It is therefore possible that our participants did not actually believe that the examiner was impervious to bias but rather that bias was a nonissue because the analysis was objective. In Thompson and Scurich, venirepersons devalued the bias-denying examiner's testimony only if he was also cross-examined about the subjectivity of his analysis, which suggests that the perceived risk of bias may depend on the perceived objectivity of the analysis. Future research should more carefully examine how jurors' general understanding of the analytic procedure interacts with case-specific details to shape decision-making.

Future research should also directly investigate *why* jurors fail to detect and/or discount biased forensic expert testimony. One possibility is that jurors are *unable* to identify biased testimony as such because they are uninformed about the phenomenon of cognitive bias and/or unaware of the subjectivity inherent to many forensic science judgments. If that is the case, perhaps opposing expert testimony on cognitive bias would better equip jurors to detect bias testimony. Alternatively, and consistent with a cognitive coherence framework (82), jurors may be *unwilling* to discount biased testimony if that testimony accords with their belief in the defendant's guilt. In our study, for example, if participants did not believe the defendant's recantation of his confession, they may have been less motivated to scrutinize the forensic examiner's testimony due to a prevailing belief in the defendant's guilt. Future studies on this topic should incorporate additional measures (e.g., open-ended items) with an eye toward clarifying why exactly jurors find exposed and unexposed examiners equally persuasive.

In any case, our findings provide further indication that jurors cannot be trusted to differentiate between valid and biased forensic science testimony. In turn, jurors may weight forensic and other evidence inappropriately. As Thompson (18) explained, task-irrelevant information that shapes a forensic examiner's opinion is "double-counted" in a way that undermines the independent probative value of the examiner's opinion. Similarly, Kassin (20) explained that, once obtained, confession evidence can corrupt the interpretation of other evidence, thereby creating an illusion of independent corroboration. For example, if a forensic examiner's knowledge of a confession leads that examiner to judge the forensic evidence as also incriminating, then the confession is effectively being counted twice. At trial, however, jurors will view the confession and fingerprints as two independent but congruent pieces of evidence and give them more weight than they deserve. Notably, Gardner et al. (70) recently found that 34% of forensic examiners said they would review a suspect's confession if it was available, so this example is hardly farfetched.

Policy Implications

Rather than relying on jurors to serve as the safety net in cases involving dubious forensic science evidence, this problem would be better addressed further upstream—by reforming laboratory procedures and keeping dubious forensic opinions out of the courtroom in the first place. For example, Thompson (18) proposed a case manager model to protect examiners against the biasing effects of task-irrelevant information. Under this model, a qualified case manager acts as an intermediary between police investigators and the forensic examiner working on the case; the case manager would receive both task-relevant and task-irrelevant information from investigators, however, convey only the task-relevant information to the examiner. Examiners thus base their analysis solely on task-relevant information and provide the results of their analysis to the case manager, who relays this information back to investigators.

A second approach, called *linear sequential unmasking* (LSU; 22), may be less onerous than the case manager model insofar as it does not require a second examiner to act as case manager. Under the LSU approach, the examiner first analyzes the crime scene evidence in isolation and carefully documents their analysis, including their confidence in their initial judgment. Only after this is done may the examiner compare the crime scene evidence against the suspect's sample and consider other task-relevant information. The examiner is then permitted to revisit and revise their initial judgment, but all revisions must be documented. LSU therefore does not eliminate the possibility of bias, but it does make it more transparent. Moreover, by requiring examiners to first analyze trace evidence in isolation of any reference material, LSU also mitigates attentional biases associated with comparing two forensic samples (22,83).

Several forensic laboratories have described positive experiences of implementing either or both of these context management procedures. Found and Ganas (84) reported on a forensic handwriting laboratory whose examiners implemented a case manager approach and revised their evidence submission forms to eliminate information that was unnecessary and potentially biasing (see also [85]). Several months later, these examiners reported "no negative outcomes," adding that these procedures have boosted confidence in their opinions and were not "complex, overly time-consuming or expensive" to implement (p. 158; see also [86]). Mattijssen, Kerkhoff, Berger, Dror, and Stoel (87) recounted the implementation of a similar context management procedure in a firearms examination laboratory and noted that its examiners have been "very supportive and receptive" of the changes (p. 121). Finally, Archer and Wallman (88) described the adoption of case manager and sequential unmasking procedures in a forensic entomology laboratory; they concluded that these changes have served to "increase the value of the [examiners'] opinion to the court by endeavoring to remove contextual influences" (p. 1276). In short, examiners who use context management procedures have almost unanimously reported that their implementation was not difficult or disruptive, and furthermore, that these procedures have increased the value of their opinions (for an alternate perspective, see [89]).

Like all humans, forensic science examiners are vulnerable to unconscious cognitive biases, and research has overwhelmingly indicated that these biases can influence the opinions they present in court. The current findings suggest that jurors cannot be relied on to distinguish between the reliable testimony of a context-blind forensic examiner and the dubious testimony of one who was exposed to task-irrelevant information. Rather than seeking to instruct or educate jurors about these effects, perhaps a more expedient solution would be to minimize the bias before it gets to court—by encouraging forensic laboratories to adopt context management procedures that will ensure the independent probative value of that evidence.

References

1. National Registry of Exonerations. https://www.law.umich.edu/special/exoneration/Pages/about.aspx (Accessed June 29, 2020).
2. Kassin SM, Dror IE, Kukucka J. The forensic confirmation bias: problems, perspectives, and proposed solutions. J Appl Res Mem Cogn 2013;2:42–52. https://doi.org/10.1016/j.jarmac.2013.01.001
3. Dror IE, Charlton D. Why experts make errors. J Forensic Identif 2006;56:600–16.
4. Smalarz L, Madon S, Yang Y, Guyll M, Buck S. The perfect match: do criminal stereotypes bias forensic evidence analysis? Law Hum Behav 2016;2016(40):420–9. 0.1037/lhb0000190
5. Stevenage SV, Bennett A. A biased opinion: demonstration of cognitive bias on a fingerprint matching task through knowledge of DNA test results. Forensic Sci Int 2017;276:93–106. https://doi.org/10.1016/j.forsciint.2017.04.009
6. Cooper GS, Meterko V. Cognitive bias research in forensic science: a systematic review. Forensic Sci Int 2019;297:35–46. https://doi.org/10.1016/j.forsciint.2019.01.016
7. Kukucka J. Confirmation bias in the forensic sciences: causes, consequences, and countermeasures. In: Koen WJ, Bowers CM, editors. The psychology and sociology of wrongful convictions: forensic science reform. New York, NY: Elsevier, 2018;223–45.
8. Kukucka J, Kassin SM. Do confessions taint perceptions of handwriting evidence? An empirical test of the forensic confirmation bias. Law Hum Behav 2014;38:256–70. https://doi.org/10.1037/lhb0000066
9. Miller LS. Bias among forensic document examiners: a need for procedural changes. J Police Sci Admin 1984;12:407–11.
10. Bieber P. Measuring the impact of cognitive bias in fire investigation. In Proceedings of the 5th International Symposium on Fire Investigation, Science and Technology; 2012 Oct 15-17. Adelphia, MD. Bradenton, FL: National Association of Fire Investigators, 2012.
11. Nakhaeizadeh S, Dror IE, Morgan R. Cognitive bias in forensic anthropology: visual assessment of skeletal remains is susceptible to confirmation bias. Sci Justice 2014;54:208–14. https://doi.org/10.1016/j.scijus.2013.11.003
12. Osborne NK, Taylor MC, Healey M, Zajac R. Bloodstain pattern classification: accuracy, effect of contextual information and the role of analyst characteristics. Sci Justice 2016;56:123–8. https://doi.org/10.1016/j.scijus.2015.12.005
13. Osborne NK, Woods S, Kieser J, Zajac R. Does contextual information bias bitemark comparisons? Sci Justice 2014;54:267–73. https://doi.org/10.1016/j.scijus.2013.12.005
14. van den Eeden CAJ, de Poot CJ, van Koppen PJ. Forensic expectations: investigating a crime scene with prior information. Sci Justice 2016;56:475–81. https://doi.org/10.1016/j.scijus.2016.08.003
15. Dror IE, Hampikian G. Subjectivity and bias in forensic DNA mixture interpretation. Sci Justice 2011;51:204–8. https://doi.org/10.1016/j.scijus.2011.08.004
16. National Commission on Forensic Science. Ensuring that forensic analysis is based upon task-relevant information. 2015 Dec. https://www.justice.gov/archives/ncfs/file/818196/download (Accessed July 24, 2020).
17. Kukucka J. People who live in ivory towers shouldn't throw stones: a refutation of Curley. Forensic Sci Int Synergy 2020;2:110–3. https://doi.org/10.1016/j.fsisyn.2020.03.001
18. Thompson WC. What role should investigative facts play in the evaluation of scientific evidence? Aust J Forensic Sci 2011;43:123–34. https://doi.org/10.1080/00450618.2010.541499
19. Dror IE. Biases in forensic experts. Science 2018;360:243. https://doi.org/10.1126/science.aat8443
20. Kassin SM. Why confessions trump innocence. Am Psychol 2012;67:431–45. https://doi.org/10.1037/a0028212
21. Thompson WC. Determining the proper evidentiary basis for an expert opinion: what do experts need to know and when do they know too much? In: Robertson C, Kesselheim A, editors. Blinding as a solution to bias: strengthening biomedical science, forensic science, and law. New York, NY: Elsevier, 2015;133–50.
22. Dror IE, Thompson WC, Meissner CA, Kornfield I, Krane D, Saks M, et al. Context management toolbox: a linear sequential unmasking (LSU) approach for minimizing cognitive bias in forensic decision making. J Forensic Sci 2015;60:1111–2. https://doi.org/10.1111/1556-4029.12805
23. President's Council of Advisors on Science and Technology. Forensic science in criminal courts: ensuring scientific validity of feature-comparison methods. 2016 Sept. https://www.justice.gov/archives/ncfs/page/file/933476/download (Accessed July 24, 2020).
24. Kukucka J, Kassin SM, Zapf PA, Dror IE. Cognitive bias and blindness: a global survey of forensic science examiners. J Appl Res Mem Cogn 2017;6:452–9. https://doi.org/10.1016/j.jarmac.2017.09.001
25. Pronin E, Lin DY, Ross L. The bias blind spot: perceptions of bias in self versus others. Per Soc Psychol Bull 2002;28:369–81. https://doi.org/10.1177/0146167202286008
26. Bond CF, DePaulo BM. Accuracy of deception judgments. Pers Soc Psychol Rev 2006;10:214–34. https://doi.org/10.1207/s15327957pspr1003_2
27. Gilbert DT, Malone PS. The correspondence bias. Psychol Bull 1995;117:21–38. https://doi.org/10.1037/0033-2909.117.1.21
28. Jones EE. Interpersonal perception. New York, NY: Freeman, 1990.
29. Ross L. The intuitive psychologist and his shortcomings: distortions in the attribution process. Adv Exp Soc Psychol 1977;10:174–221. https://doi.org/10.1016/S0065-2601(08)60357-3
30. Ross L. From the fundamental attribution error to the truly fundamental attribution error and beyond: my research journey. Perspect Psychol Sci 2018;13:750–69. https://doi.org/10.1177/1745691618769855
31. Gatowski SI, Dobbin SA, Richardson JT, Ginsburg GP, Merlino ML, Dahir V. Asking the gatekeepers: a national survey of judges on judging expert evidence in a post-Daubert world. Law Hum Behav 2001;25:433–58. https://doi.org/10.1023/A:1012899030937
32. Kovera MB, McAuliff BD. The effects of peer review and evidence quality on judge evaluations of psychological science: are judges effective gatekeepers? J Appl Psychol 2000;85:574–86. https://doi.org/10.1037/0021-9010.85.4.574
33. Kovera MB, McAuliff BD, Hebert KS. Reasoning about scientific evidence: effects of juror gender and evidence quality on juror decisions in a hostile work environment case. J Appl Psychol 1999;84:362–75. https://doi.org/10.1037/0021-9010.84.3.362
34. McAuliff BD, Kovera MB. Juror need for cognition and sensitivity to methodological flaws in expert evidence. J Appl Soc Psychol 2008;38:385–408. https://doi.org/10.1111/j.1559-1816.2007.00310.x
35. Rosenthal R. Experimenter effects in behavioral research. East Norwalk, CT: Appleton-Century-Crofts, 1966.
36. McAuliff BD, Duckworth TD. I spy with my little eye: jurors' detection of internal validity threats in expert evidence. Law Hum Behav 2010;34:489–500. https://doi.org/10.1007/s10979-010-9219-3
37. McAuliff BD, Kovera MB, Nuñez G. Can jurors recognize missing control groups, confounds, and experimenter bias in psychological science? Law Hum Behav 2009;33:247–57. https://doi.org/10.1007/s10979-008-9133-0
38. Koehler JJ. Intuitive error rate estimates for the forensic sciences. Jurimetrics 2017;57:153–68.
39. Lieberman JD, Carrell CA, Miethe TD, Krauss DA. Gold versus platinum: do jurors recognize the superiority and limitations of DNA evidence compared to other types of forensic evidence? Psychol Public Pol L 2008;14:27–62. https://doi.org/10.1037/1076-8971.14.1.27
40. Mitchell G, Garrett BL. The impact of proficiency testing information and error aversions on the weight given to fingerprint evidence. Behav Sci Law 2019;37:195–210. https://doi.org/10.1002/bsl.2402
41. Koehler JJ, Schweitzer NJ, Saks MJ, McQuiston DE. Science, technology, or the expert witness: what influences jurors' judgments about forensic science testimony? Psychol Public Pol L 2016;22:401–13. https://doi.org/10.1037/law0000103
42. McQuiston-Surrett D, Saks MJ. The testimony of forensic identification science: what expert witnesses say and what factfinders hear. Law Hum Behav 2009;33:436–53. https://doi.org/10.1007/s10979-008-9169-1
43. Austin JL, Kovera MB. Cross-examination educates jurors about missing control groups in scientific evidence. Psychol Public Pol L 2015;21:252–64. https://doi.org/10.1037/law0000049
44. Salerno JM, McCauley MR. Mock jurors' judgments about opposing scientific experts: do cross-examination, deliberation and need for cognition matter? Am J Forensic Psychol 2009;27:37–60.

45. Kovera MB, Levy RJ, Borgida E, Penrod SD. Expert testimony in child sexual abuse cases. Law Hum Behav 1994;18:653–74. https://doi.org/10.1007/BF01499330
46. Koehler JJ. If the shoe fits they might acquit: the value of forensic science testimony. J Empir Legal Stud 2011;8:21–48. https://doi.org/10.1111/j.1740-1461.2011.01225.x
47. Chorn JA, Kovera MB. Variations in reliability and validity do not influence judge, attorney, and mock juror decisions about psychological expert evidence. Law Hum Behav 2019;43:542–57. https://doi.org/10.1037/lhb0000345
48. Thompson WC, Scurich N. How cross-examination on subjectivity and bias affects jurors' evaluations of forensic science evidence. J Forensic Sci 2019;64:1379–88. https://doi.org/10.1111/1556-4029.14031
49. Bush MA, Bush PJ, Sheets HD. A study of multiple bitemarks inflicted in human skin by a single dentition using geometric morphometric analysis. Forensic Sci Int 2011;211:1–8. https://doi.org/10.1016/j.forsciint.2011.03.028
50. Bush MA, Miller RG, Bush PJ, Dorion RB. Biomechanical factors in human dermal bitemarks in a cadaver model. J Forensic Sci 2009;54:167–76. https://doi.org/10.1111/j.1556-4029.2008.00908.x
51. Holtkötter H, Sheets HD, Bush PJ, Bush MA. Effect of systematic dental shape modification in bitemarks. Forensic Sci Int 2013;228:61–9. https://doi.org/10.1016/j.forsciint.2013.02.024
52. Miller RG, Bush PJ, Dorion RB, Bush MA. Uniqueness of the dentition as impressed in human skin: a cadaver model. J Forensic Sci 2009;54:909–14. https://doi.org/10.1111/j.1556-4029.2009.01076.x
53. Dror IE, Mnookin J. The use of technology in human expert domains: challenges and risks arising from the use of automated fingerprint identification systems in forensics. Law Probab Risk 2010;9:47–67.
54. Tangen JM, Thompson MB, McCarthy DJ. Identifying fingerprint expertise. Psychol Sci 2011;22:995–7. https://doi.org/10.1177/0956797611414729
55. Ulery BT, Hicklin RA, Buscaglia J, Roberts MA. Accuracy and reliability of forensic latent fingerprint decisions. Proc Natl Acad Sci 2011;108:7733–8. https://doi.org/10.1073/pnas.1018707108
56. Dror IE, Charlton D, Perón A. Contextual information renders experts vulnerable to making erroneous identifications. Forensic Sci Int 2006;156:174–8. https://doi.org/10.1016/j.forsciint.2005.10.017
57. Fraser-Mackenzie PA, Dror IE, Wertheim K. Cognitive and contextual influences in determination of latent fingerprint suitability for identification judgments. Sci Justice 2013;53:144–53. https://doi.org/10.1016/j.scijus.2012.12.002
58. Ribeiro G, Tangen JM, McKimmie BM. Beliefs about error rates and human judgment in forensic science. Forensic Sci Int 2019;297:138–47. https://doi.org/10.1016/j.forsciint.2019.01.034
59. Elaad E, Ginton A, Ben-Shakhar G. The effects of prior expectations and outcome knowledge on polygraph examiners' decisions. J Behav Decis Making 1994;7:279–92. https://doi.org/10.1002/bdm.3960070405
60. Hasel LE, Kassin SM. On the presumption of evidentiary independence: can confessions corrupt eyewitness identifications? Psychol Sci 2009;20:122–6. https://doi.org/10.1111/j.1467-9280.2008.02262.x
61. Marion S, Kukucka J, Collins C, Kassin SM, Burke TM. Lost proof of innocence: the impact of confessions on alibi witnesses. Law Hum Behav 2016;40:65–71. https://doi.org/10.1037/lhb0000156
62. Kassin SM, Bogart D, Kerner J. Confessions that corrupt: evidence from the DNA exoneration case files. Psychol Sci 2012;23:41–5. https://doi.org/10.1177/0956797611422918
63. Zapf PA, Kukucka J, Kassin SM, Dror IE. Cognitive bias in forensic mental health assessment: evaluator beliefs about its nature and scope. Psychol Public Pol L 2018;24:1–10. https://doi.org/10.1037/law0000153
64. Petty RE, Cacioppo JT. Communication and persuasion: central and peripheral routes to attitude change. New York, NY: Springer, 1986;1–24.
65. McCarthy-Wilcox A, NicDaeid N. Jurors' perceptions of forensic science expert witnesses: experience, qualifications, testimony style and credibility. Forensic Sci Int 2018;291:100–8. https://doi.org/10.1016/j.forsciint.2018.07.030
66. van den Eeden CAJ, de Poot CJ, van Koppen PJ. The forensic confirmation bias: a comparison between experts and novices. J Forensic Sci 2019;64:120–6. https://doi.org/10.1111/1556-4029.13817
67. Despodova NM, Kukucka J, Hiley A. Can defense attorneys detect forensic confirmation bias? Effects on evidentiary judgments and trial strategies. Z Psychol 2020;228:216–20. https://doi.org/10.1027/2151-2604/a000414
68. Butt L. The forensic confirmation bias: problems, perspectives, and proposed solutions: commentary by a forensic examiner. J Appl Res Mem Cogn 2013;2:59–60. j.jarmac.2013.01.012
69. Elaad E. Psychological contamination in forensic decisions. J Appl Res Mem Cogn 2013;2:76–7. https://doi.org/10.1016/j.jarmac.2013.01.006
70. Gardner BO, Kelley S, Murrie DC, Dror IE. What do forensic analysts consider relevant to their decision making? Sci Justice 2019;59:516–23. https://doi.org/10.1016/j.scijus.2019.04.005
71. Leadbetter M. Letter to the editor. Fingerprint World 2007;33:231.
72. Oliver WR. Comment on Dror, Kukucka, Kassin, and Zapf (2018): When expert decision making goes wrong. J Appl Res Mem Cogn 2018;7:314–5. https://doi.org/10.1016/j.jarmac.2018.01.010
73. Cutler BL, Kovera MB. Expert psychological testimony. Curr Dir Psychol Sci 2011;20:53–7. https://doi.org/10.1177/0963721410388802
74. Vidmar N, Diamond SS. Juries and expert evidence. Brooklyn L Rev 2001;66:1121–80.
75. Jones EE, Harris VA. The attribution of attitudes. J Exp Soc Psychol 1967;3:1–24. https://doi.org/10.1016/0022-1031(67)90034-0
76. Miller AG, Jones EE, Hinkle S. A robust attribution error in the personality domain. J Exp Soc Psychol 1981;17:587–600. https://doi.org/10.1016/0022-1031(81)90041-X
77. Gilbert DT, Jones EE. Perceiver-induced constraint: Interpretations of self-generated reality. J Pers Soc Psychol 1986;50:269–80. https://doi.org/10.1037/0022-3514.50.2.269
78. Heider F. The psychology of interpersonal relations. Hillsdale, NJ: Lawrence Erlbaum Associates, 1958.
79. Kassin SM, Sukel H. Coerced confessions and the jury: an experimental test of the "harmless error" rule. Law Hum Behav 1997;21:27–46. https://doi.org/10.1023/A:1024814009769
80. Wallace DB, Kassin SM. Harmless error analysis: how do judges respond to confession errors? Law Hum Behav 2012;36:151–7. https://doi.org/10.1007/s10979-010-9262-0
81. Neuschatz JS, Lawson DS, Swanner JK, Meissner CA, Neuschatz JS. The effects of accomplice witnesses and jailhouse informants on jury decision making. Law Hum Behav 2008;32:137–49. https://doi.org/10.1007/s10979-007-9100-1
82. Simon D. A third view of the black box: cognitive coherence in legal decision making. U Chicago L Rev 2004;71:511–86.
83. Dror IE, Champod C, Langenburg G, Charlton D, Hunt H, Rosenthal R. Cognitive issues in fingerprint analysis: inter-and intra-expert consistency and the effect of a 'target' comparison. Forensic Sci Int 2011;208:10–17. https://doi.org/10.1016/j.forsciint.2010.10.013
84. Found B, Ganas J. The management of domain irrelevant context information in forensic handwriting examination casework. Sci Justice 2013;53:154–8. https://doi.org/10.1016/j.scijus.2012.10.004
85. Gardner BO, Kelley S, Murrie DC, Blaisdell KN. Do evidence submission forms expose latent print examiners to task-irrelevant information? Forensic Sci Int 2019;297:236–42. https://doi.org/10.1016/j.forsciint.2019.01.048
86. Stoel RD, Dror IE, Miller LS. Bias among forensic document examiners: still a need for procedural changes. Aust J Forensic Sci 2014;46:91–7. https://doi.org/10.1080/00450618.2013.797026
87. Mattijssen EJAT, Kerkhoff W, Berger CEH, Dror IE, Stoel RD. Implementing context information management in forensic casework: minimizing contextual bias in firearms examination. Sci Justice 2016;56:113–22. https://doi.org/10.1016/j.scijus.2015.11.004
88. Archer MS, Wallman JF. Context effects in forensic entomology and use of sequential unmasking in casework. J Forensic Sci 2016;61:1270–7. https://doi.org/10.1111/1556-4029.13139
89. Langenburg G. Addressing potential observer effects in forensic science: a perspective from a forensic scientist who uses linear sequential unmasking techniques. Aust J Forensic Sci 2017;49:548–63. https://doi.org/10.1080/00450618.2016.1259433

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14527
Available online at: onlinelibrary.wiley.com

PAPER

GENERAL; CRIMINALISTICS

Yang Yu,[1] Ph.D.; Yaping Luo,[2] Ph.D.; Wenyi Xie,[3] Ph.D.; Sheng Lin,[1] M.S.; and Luoxi Liu,[1] Ph.D.

The Impact of Fatigue on Decision-Making in the Footwear Examination: Evidence from Questionnaires and Eye-Tracking Test

ABSTRACT: To assess the influence of fatigue on decision-making and performance in footwear examination, questionnaires and eye-tracking techniques are employed. We ask 23 volunteers to wear shoes of four different outsole patterns and obtained 50 image pairs of questioned and known impressions under controlled conditions. Among which, 10 image sets were "repeated" as benchmarks by being horizontally flipped. To evaluate the accuracy of comparison, 12 qualified footwear examiners' response and eye metrics in both morning and afternoon session were recorded and analyzed by descriptive statistics, correlation analysis and gaze plots. The results revealed that as a group there is no significant statistic differences in either subjects' behavioral performance or their eye gaze data after fatigue. When examined via the Earth Mover Metric method, the consistency of examiners as a group did not change after fatigue either. Statistically, the pupil size got smaller during the observation in the afternoon. Study also demonstrated that the visualization process of forensic footwear evidence through eye-tracking method could be a promising technique for training and education.

KEYWORDS: footwear examination, fatigue, decision-making, questionnaires, eye-tracking, eye metrics

Footwear marks and impressions are considered to be one of the most frequently encountered evidence at crime scenes (1). Practitioners may make a basic judgment about the suspect for investigation process, including gender, height or shoes characteristics, etc. The footwear evidence can be used by the police to determine the nature of the case, delimit and narrow the scope of investigation, and it is also important for court trials. However, the comparison conclusions drawn by forensic footwear examiners (FFEs) do not always remain the same (2). This refers to the consistency in judgment across examiners, whereby different examiners reach different conclusions on the same footmarks.

Forensic decision-making is influenced by human factors, such as workload volume, tight deadlines, and exposure to case details and fatigue (3). Thereinto, fatigue not only impairs physical performance, but also interferes with cognitive activity (4), such as attention (5), cognitive processing strategy (6), etc. FFEs might be on duty for extended periods, be on call for crime scene investigation and work extended hours doing multiple comparisons. They are constantly working under the environment of time pressure, difficult tasks, and irregular life that could affect their performance. However, there are few articles about the influence of fatigue factors on the performance in footwear examination such as visual searching pattern.

Footwear examination relies heavily on traditional side-by-side comparison methods. OSAC Footwear & Tire Subcommittee's suggest eye tracker should be used to document footwear comparisons (7). Eye movement can directly reflect the process of cognitive processing, and eye-tracking systems can better document the process of footwear examination compared with previous traditional technologies. Eye-tracking technology has been applied in many fields, such as medicine and health, linguistics and reading, education and training, and forensic science (8–10). Through within-subject comparison, the variability and consistency of FFE's decisions can be examined before and after fatigue. In our approach, we ask FFEs to compare the same pair of shoeprints before and after fatigue.

Van den Linden et al. (11) investigated the impact of fatigue on exploration behavior through a complex computer task. The subjects were asked to participate in two tasks: involving the spreadsheet program (Excel) and the graphical program (ClarisDraw), respectively. Given the results of subjects' exploration behavior and the number of subtasks being solved in certain time, the authors suggested that not only did the fatigued participants use significantly less systematic exploration cognitive function, but they also made more errors than nonfatigued participants. However, there was no difference in the number of subtasks being solved.

Fatigue can also interfere with an individual's cognitive process and consume cognitive resources. Wang et al. (12) hold that when fatigue individuals are faced with decision-making tasks, they may be more inclined to choose conservative options that require less cognitive processing rather than consume more cognitive resources to choose risky options.

[1] School of Forensic Science, People's Public Security University of China, Beijing, 100038, China.

[2] Graduate School, People's Public Security University of China, Beijing, 100038, China.

[3] Earth Science and Engineering College, Nanjing University, No. 163, Xianlin Road, Qixia District, Nanjing, 210023, China.

Corresponding author: Yaping Luo, Ph.D. E-mail: yaping_luo@126.com

[Correction added 11 August 2020. The Highlights section was removed.]

Received 14 Mar. 2020; and in revised form 6 July 2020; accepted 9 July 2020.

The impact of fatigue on the evaluation of fingerprint examination process was measured by Busey et al. (13) with both eye-tracking and behavioral methods to study the performance of fingerprint examiners. The researchers found that participants' searching accuracy and working memory capacity declined.

From the research results above, fatigue appear to affect the comparison conclusions, attention, and cognitive strategies. However, up to now, there is no relevant research in footwear examinations. This study aims to explore the feasibility of using eye tracker to assess the performance of footwear examination. As a preliminary study, we try to gain a greater understanding whether FFEs make different conclusions when they are tired. To evaluate the effects of fatigue in actual work, fatigue was induced through the analysis and discussion of complex footmarks. Moreover, we documented both comparison conclusions and eye-tracking results of all subjects in two experiments (before and after fatigue).

More specifically, eye-tracking technology is employed to evaluate how FFEs make their decisions and to record what are the features of footwear images they focus on during the examination. After a preliminary understanding of the relationship between fatigue factors and footwear examination, this can provide a basic reference for further relevant research and expanding the number of participants and samples.

Materials and Methods

In this study, subjects were told to make decisions about known and questioned footprints that appeared side by side on the screen of an Apple® computer. The process of the experiments was a within-subjects design, where all subjects were required to participate in two sessions in the morning and in the afternoon, during which eye movements and questionnaire accuracy of the subjects were recorded quantitatively.

Subjects

The experimental group consisted of 12 subjects from four forensic science laboratories in the police department and forensic research institutions. The age of subjects ranged from 31 to 51 with 5 to 25 years of footwear examination experience. All the 12 subjects were qualified examiners who could issue expert reports regarding shoeprints. Subjects were informed about the nature of study and signed informed consents prior to participation.

Shoeprints

To ensure the scientific validity of experiment, all samples used in this study were under controlled conditions. The knowns were formed by normal walking, i.e., an individual with inked footwear walked across the floor covered with paper. All questioned impressions and knowns were photographed with Nikon® D850 or Nikon® D700 camera mounted on either a tripod or a copy stand.

We asked 23 volunteers to wear shoes of four different outsole patterns. The known shoeprints were collected up to 10 days after the collection of the questioned shoeprints. To prepare challenging nonmatch shoeprints, we asked volunteers to wear shoes of same shoe size and sole pattern. Such design is due to the actual cases where criminal gang suspects wear the same shoes.

We selected 50 pairs of shoeprints from the image sets to represent typical shoeprints that a FFEs may encounter in casework. The resolution of each image pair (composed of a questioned shoeprint and a known shoeprint) is 300 ppi. When the shoeprints are presented to an FFE, the image on the left was questioned shoeprint, and the known shoeprint was shown on the right side of the screen. Two sets of images are shown in Fig. 1. In each set, the physical size and sole pattern of the shoeprint pairs are the same so that the image sets can be directly compared by FFEs without concern for the difference between size and shape. In the image sets, FFEs might encounter a shoeprint of the same source with complex background, blur or incomplete display, or a challenging close nonmatch shoeprint.

The study was divided into two separate sessions, one in the morning and the other in the afternoon. So, the 50 shoeprints pairs were divided evenly into two parts in the morning and afternoon. In the 25 pairs of images in the morning, 13 images were known matches (from the same source) and 12 images were known nonmatches (from various sources). In the 25 pairs of images in the afternoon, 15 images were known matches, and 10 images were known nonmatches.

It should be noted that 10 image sets were "repeated" in both sessions as benchmarks, where these pictures were flipped horizontally. The flip was to eliminate the influence of memory as much as possible while maintaining the same difficulty of stimuli. In the afternoon, we would also remind the subjects not to be influenced by the residual memory of work in the morning, however, the decision-making of the subjects would to some extent be affected in the afternoon.

Experimental Procedure

Chronologically, the experiment was divided into two separate phases. Two separate experimental phases were completed according to the workflow illustrated in Fig. 2.

Phase 1 in the morning: FFEs were informed of the significance of experiment, signed the consent form, and completed a simple background survey. Then, the subjects were assigned to complete 25 image comparisons (A1–A25). They can complete the comparison tasks at their own pace without any time limit. Before starting image comparison, information including deposition objects, substrate conditions, and methods of recovery was provided so that the FFEs could be as informed as they are in real cases. After each image set was accomplished, the subjects were asked to answer two behavioral questionnaires (Fig. 3) regarding their conclusion and the assessment of difficulty.

Phase 2 in the afternoon: To cause the fatigue of subjects, FFEs are engaged into a discussion on the detection, recovery, and examination of footwear impression evidence in complex cases for an hour right after a big lunch. Before the session in the afternoon, the subjects were asked to fill out a fatigue scale and record their subjective feelings. The self-designed fatigue scale included six fatigue ratings ranging from 0 (none) to 5 (the strongest). In addition, the subjects all claimed to have varying degrees of fatigue. Then FFEs were asked to complete another 25 image comparisons (B1–B25).

Data Recording

Eye gaze data and human behavior were recorded via Tobii® X3-120 eye tracker. The eye tracker has a sample rate of

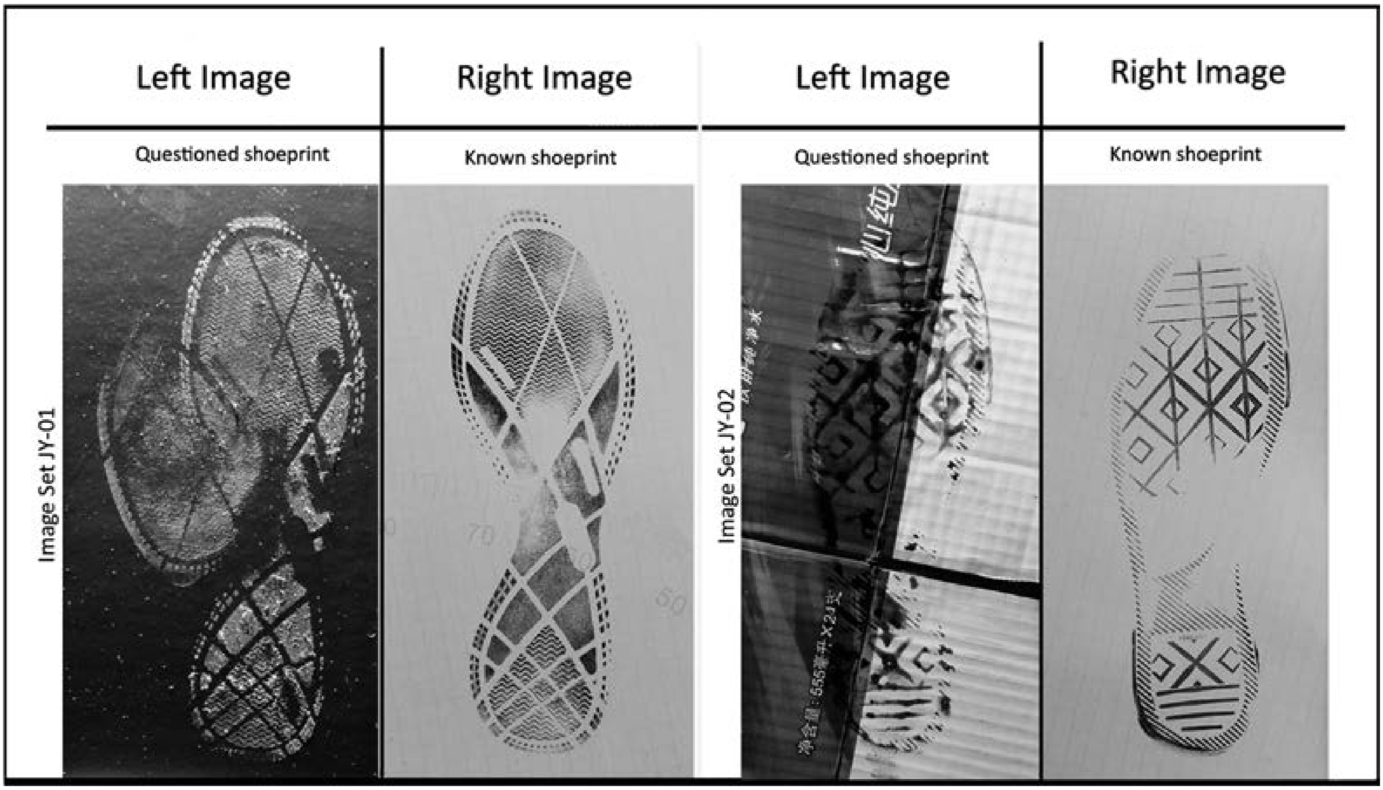


FIG. 1—*Two sets of images used in the experiment. The questioned shoeprint may be with complex background, blur, incomplete display, or a challenging close nonmatch shoeprint. [Color figure can be viewed at wileyonlinelibrary.com]*

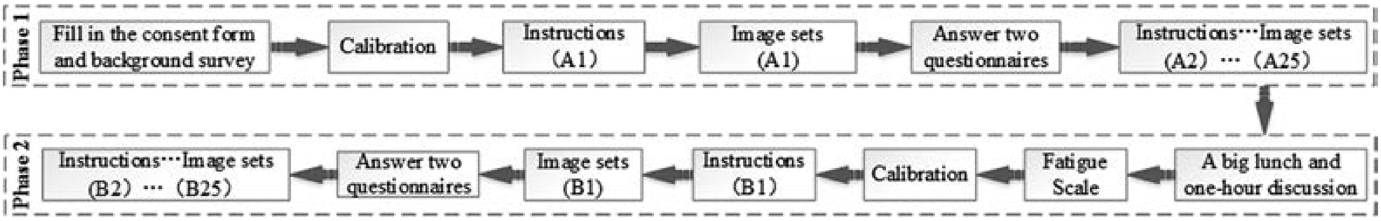
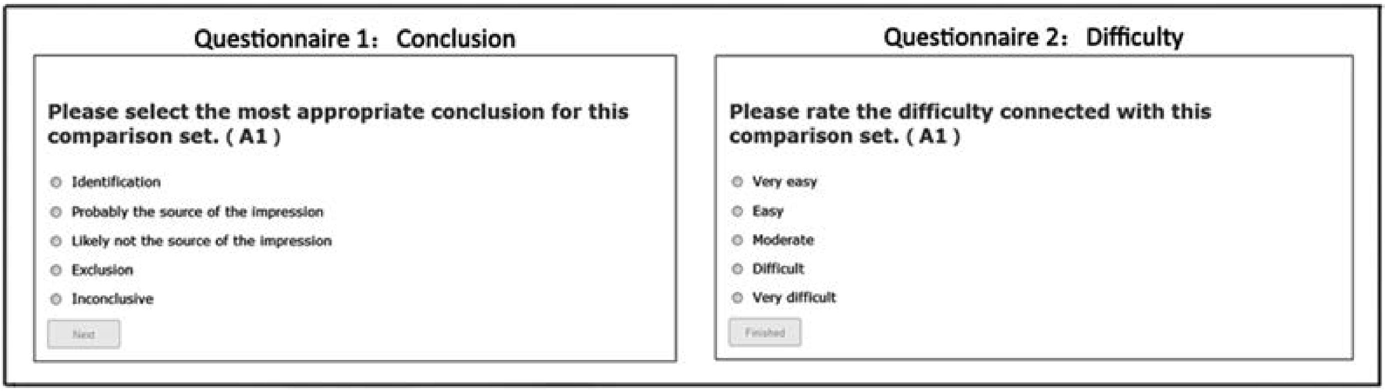

Phase 1
Fill in the consent form and background survey
Calibration
Instructions (A1)
Image sets (A1)
Answer two questionnaires
Instructions···Image sets (A2) ··· (A25)
Phase 2
A big lunch and one-hour discussion
Fatigue Scale
Calibration
Instructions (B1)
Image sets (B1)
Answer two questionnaires
Instructions···Image sets (B2) ··· (B25)

FIG. 2—*Workflow of the experiment. [Color figure can be viewed at wileyonlinelibrary.com]*



Questionnaire 1：Conclusion

Please select the most appropriate conclusion for this comparison set. (A1)

○ Identification
○ Probably the source of the impression
○ Likely not the source of the impression
○ Exclusion
○ Inconclusive

Questionnaire 2：Difficulty

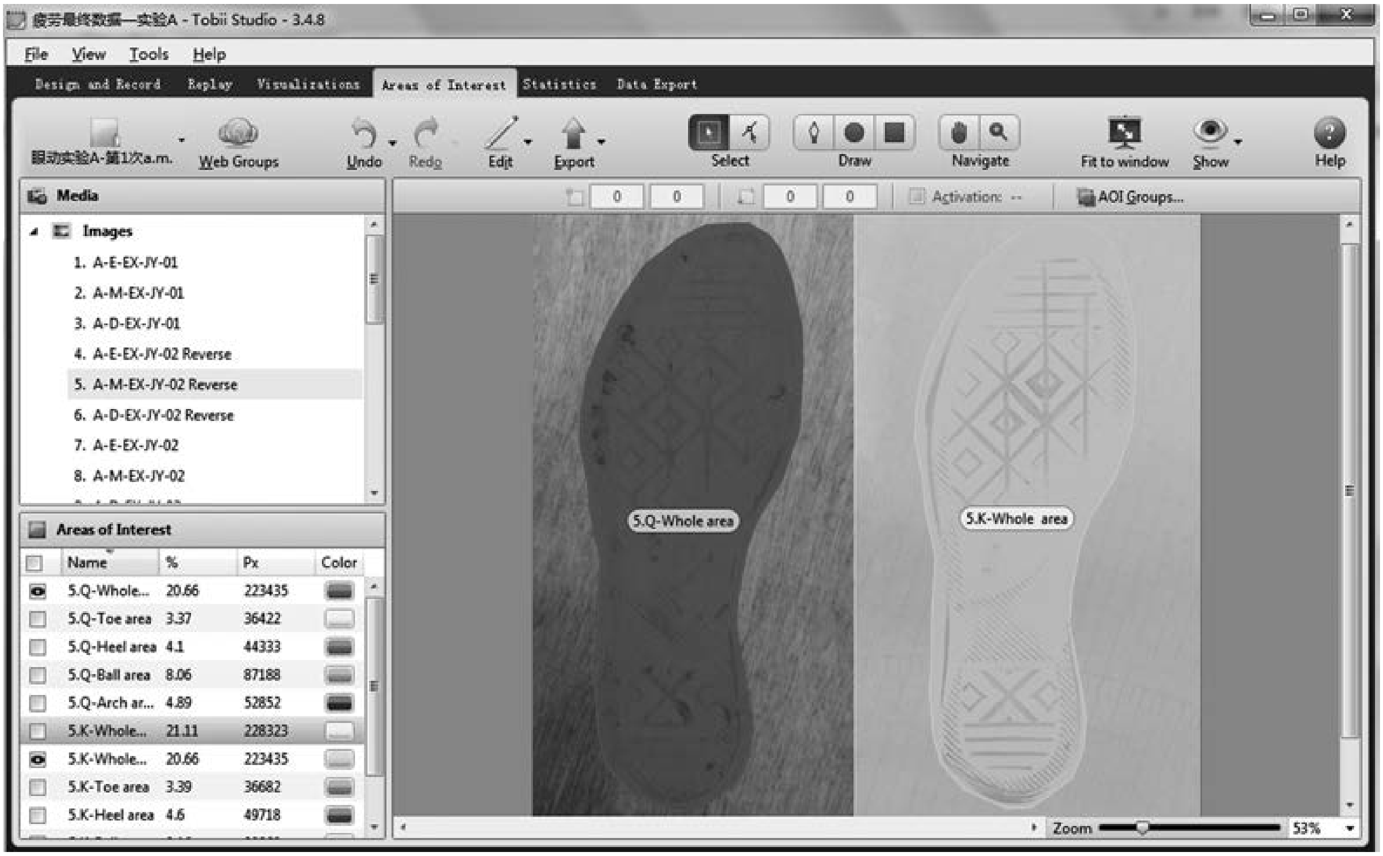
Please rate the difficulty connected with this comparison set. (A1)

○ Very easy
○ Easy
○ Moderate
○ Difficult
○ Very difficult

FIG. 3—*Two behavioral questionnaires FFEs needed to fill in after each comparison.*

approximately 120 Hz, and it takes about 8.333 ms to collect an eye movement sample with measurement precision of gaze point is 0.24°. To record data, the subject only needs to sit comfortably in front of the 16″ MacBook Pro laptop.

Before each data collection, calibration was performed, and the distance between the subject’s eyes and the eye tracker was kept at 65 ± 5 cm so that eye movements within 31 visual degrees could all be recorded. Behavioral data, eye-tracking



Media

Images

1. A-E-EX-JY-01
2. A-M-EX-JY-01
3. A-D-EX-JY-01
4. A-E-EX-JY-02 Reverse
5. A-M-EX-JY-02 Reverse
6. A-D-EX-JY-02 Reverse
7. A-E-EX-JY-02
8. A-M-EX-JY-02

Areas of Interest

□	Name	%	Px	Color
●	5.Q-Whole...	20.66	223435	■
□	5.Q-Toe area	3.37	36422	□
□	5.Q-Heel area	4.1	44333	■
□	5.Q-Ball area	8.06	87188	■
□	5.Q-Arch ar...	4.89	52852	■
□	5.K-Whole...	21.11	228323	□
●	5.K-Whole...	20.66	223435	□
□	5.K-Toe area	3.39	36682	■
□	5.K-Heel area	4.6	49718	■

FIG. 4—*Area of Interest of questioned and known shoeprints. [Color figure can be viewed at wileyonlinelibrary.com]*

metrics, and division of areas of interest (AOIs) were processed through Tobii® Studio 3.4.8 software. The data preprocessing of eye metrics was carried out with two-tailed *t*-test, Wilcoxon test, Cohen's *d*, and Earth Mover's Distance (EMD) via SPSS statistics 24 software, R software, MATLAB software, and Microsoft® Excel 2016 software.

As can be seen in Fig. 4, areas of interest (AOIs) of image set were defined and divided into two sub-areas based on analysis needs. The primary eye-tracking metrics recorded contained total fixation duration, fixation count, average saccadic amplitude, and average pupil diameter. Total fixation duration is the entire duration of all fixations within the AOI. Fixation count is the number of times that the subjects fixates on the AOI. Fixation duration is the duration of each individual fixation within the AOI. Average saccadic amplitude is the distance in visual degrees between the earlier fixation location and the current fixation location, and average pupil diameter is size of the left/right eye pupil.

Data Analysis

Before comparing the eye-tracking metrics data in two experiments, if normality assumptions were satisfied, the data will be applied with the parametric tests (two-tailed *t*-test); otherwise, it will be processed with equivalent nonparametric test (Wilcoxon test).

Results

We focused on the results of 10 image pairs that were repeated in reverse patterns in the morning and afternoon. That is, a total of 120 data sets were recorded and completed by 12 FFEs.

Questionnaire Results

The response frequencies for the two sessions are shown in Table 1. Note that 12 subjects need to make 120 decisions from 48 mated pairs and 72 nonmated pairs in each session. The response rate was computed based on the mated total or nonmated total. In the morning session, 46% of the known match images were correctly identified and 19% of them were decided as "probably the source of the impressions". In addition, there were eight mistaken exclusions (17%), eight erroneous decisions (17%) of "likely not the source of the impressions" and one inconclusive decision (2%) being drawn. With regard to the known nonmatch samples, there were 36 exclusions (50%) and nine decisions of "likely not the source of the impressions" (13%) and no inconclusive decision being made. They made 17 mistaken identifications (24%), 10 mistaken decisions of "probably the source of the impressions" (14%).

In the afternoon session, the subjects provided more inconclusive conclusions. For mated pairs, the subjects showed 4 (8%) inconclusive conclusions, while only 1 (2%) in the morning session. For nonmated pairs, the subjects draw 3 (4%) inconclusive conclusions, while zero in the morning session. Compared with the morning data, number of erroneous exclusion (errors) increased by two and that of "no conclusion" increased by three. Furthermore, number of erroneous identification (errors) increased by three and that of no conclusion increased by three.

TABLE 1—*Response frequencies for two sessions.*

	Mated						Nonmated					
	ID	Prob ID	EX	Prob EX	INC	Mated total	ID	Prob ID	EX	Prob EX	INC	Nonmated total
am	0.46	0.19	0.17	0.17	0.02	1	0.24	0.14	0.50	0.13	0.00	1
pm	0.46	0.23	0.21	0.02	0.08	1	0.28	0.10	0.44	0.14	0.04	1

ID: Identification; EX: Exclusion; INC: Inconclusive; Prob ID: Probably the source of the impression; Prob EX: Likely not the source of the impression.

TABLE 2—*Area under the curve (AUC) value of 12 subjects for morning and afternoon sessions.*

	1	2	3	4	5	6	7	8	9	10	11	12
AUC (am)	0.750	0.542	0.750	0.792	0.625	0.542	0.917	0.542	0.667	0.542	0.542	0.583
AUC (pm)	0.708	0.500	0.583	0.750	0.500	0.667	0.633	0.792	0.625	0.667	0.917	0.792

Null hypothesis: true area = 0.5.

TABLE 3—*Response frequencies to shoeprints of various difficulties in two sessions.*

	Very Easy	Easy	Moderate	Difficult	Very Difficult	Total
am	3	25	58	28	6	120
pm	1	19	54	33	13	120

In Table 2, The area under the ROC curve was used to evaluate each subject performance of Questionnaire 1 tests against the ground truth in the morning and afternoon sessions (14). Note that both "ID" and "probably ID" were consider as the correct answer. We compared of the AUC levels of two sessions, through Delong's test (15). There was no significant difference ($p = 0.584$), suggesting subjects' efficiency of telling mated pairs from nonmated one was not affected by fatigue.

As shown in Table 3, the number of "difficult" and "very difficult" shoeprints being reported was on the rise in the afternoon. Correspondingly, we listed both "difficult" and "very difficult" rates of each subject in Table 4. However, there was no significant increase in either the "difficult" rate ($t(11) = -0.64$, $p = 0.54$) or "very difficult" rate ($t(11) = -1.63$, $p = 0.13$) due to fatigue, which suggested that footwear examiners as a group would not alter their judgment on the difficulty ratings under fatigue. The detailed response of the two questionnaires was shown in the Appendix S1.

Eye-Tracking Results

In the two sessions, the eye-tracking data changed with behavioral frequencies. The summary statistics indicating the effects of fatigue are as follows.

AOI and Saccadic Metrics

All subjects were regarded as a whole in the systematic analysis between testing results of am and pm. We compared two sessions and set alpha = 0.05 in two-tailed test. However, for total fixation duration, no statistic difference was found between morning with afternoon ($t(11) = 1.08$, $p = 0.30$). In addition, there was no significant difference in the fixation count of two sessions ($t(11) = 0.71$, $p = 0.49$). Furthermore, the saccadic amplitude did not change significantly before and after fatigue ($t(11) = 0.74$, $p = 0.47$). In a word, the statistics of the three eye-tracking metrics above were not affected by fatigue (Table 5).

TABLE 4—*Difficult and very difficult rates of 12 subjects in morning and afternoon sessions.*

Subject Number	Morning Difficult Rate	Afternoon Difficult Rate	Morning Very Difficult Rate	Afternoon Very Difficult Rate
01	0.7	0.4	0.2	0.6
02	0.2	0.3	0	0.2
03	0.3	0.2	0	0
04	0.4	0.4	0	0
05	0.2	0.3	0	0
06	0.1	0.3	0	0
07	0.3	0	0	0
08	0.2	0.2	0	0
09	0	0	0	0
10	0.2	0.7	0	0
11	0.2	0.2	0.4	0.4
12	0	0.3	0	0.1
Mean	0.23	0.28	0.05	0.12

TABLE 5—*Paired* t*-test result of the AOIs and saccadic amplitude, systematic analysis between morning and afternoon testing.*

Eye-Tracking Metrics	t	df	p-Value	Cohen's d
Total fixation duration	1.084	11	0.301	0.443
Fixation count	0.712	11	0.491	0.291
Saccadic amplitude	0.738	11	0.476	0.301

Pupil Metrics

Table 6 and Fig. 5 illustrate descriptive statistics of subjects' pupil diameter prior and post fatigue. We collected the fatigue feedback of the subjects, quantifying from level 0 (none) to level 5 (the strongest). After fatigue, the pupil diameter of most subjects decreased to a varying extent. In the afternoon, FFEs tended to have a smaller left pupil size between the two shoeprints (2.942 vs. 2.711; $t(11) = 2.81$, $p = 0.017$). However, there was no significant changes in the size of right pupil (2.964 vs. 2.803; $t(11) = 1.826$, $p = 0.095$). This finding is consistent with the result previously reported (16), indicating that as fatigue deepens, pupil diameter decreases, and fluctuations increase.

Pearson Correlation Coefficient

We found that total fixation duration was positively correlated with age, while saccade amplitude was negatively correlated

TABLE 6—*Descriptive statistics of subjects' pupil diameter between morning and afternoon, fatigue scale, mean (SD).*

		Pupil diameter (L)		Pupil diameter (R)	
Subject number	Fatigue scale (0–5)	Morning	Afternoon	Morning	Afternoon
01	2	3.557 (0.730)	2.941 (0.840)	3.976 (0.912)	3.363 (0.862)
02	0	3.163 (0.712)	3.084 (0.743)	3.643 (0.822)	3.398 (0.884)
03	2	3.357 (0.872)	3.484 (0.712)	3.291 (0.877)	3.444 (0.736)
04	1	2.921 (0.673)	2.933 (0.737)	2.811 (0.854)	2.875 (0.838)
05	3	2.399 (0.648)	2.124 (0.573)	2.377 (0.594)	2.041 (0.377)
06	3	2.482 (0.641)	2.102 (0.394)	2.145 (0.587)	2.318 (0.616)
07	4	2.911 (0.740)	2.719 (0.689)	3.073 (0.797)	2.795 (0.714)
08	5	3.163 (0.799)	2.712 (0.816)	3.132 (0.918)	2.958 (0.718)
09	2	2.608 (0.913)	2.137 (0.701)	2.617 (0.944)	2.375 (0.846)
10	2	3.748 (1.046)	3.135 (1.195)	3.725 (1.099)	3.141 (1.199)
11	2	2.124 (0.507)	2.029 (0.445)	1.944 (0.612)	1.695 (0.379)
12	2	2.869 (1.136)	3.127 (0.989)	2.832 (1.099)	3.232 (0.987)
Mean	2.333	2.942	2.711	2.964	2.803

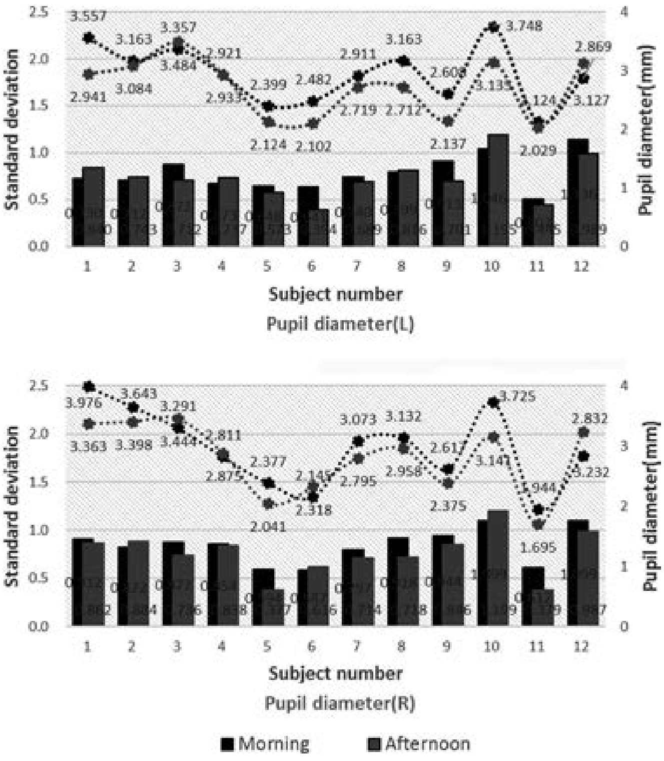


FIG. 5—*Subjects' pupil diameter and its standard deviation in morning and afternoon. [Color figure can be viewed at wileyonlinelibrary.com]*

TABLE 7—*Correlations between the stability of eye-tracking metrics and age/experience/fatigue level.*

Standard Deviation	Age	Experience	Fatigue
Fixation count	0.56	0.51	0.19
Total fixation duration	0.65*	0.49	0.44
Saccadic amplitude	−0.36	−0.20	−0.71†

*Correlation is significant at the 0.05 level.
†Correlation is significant at the 0.01 level.

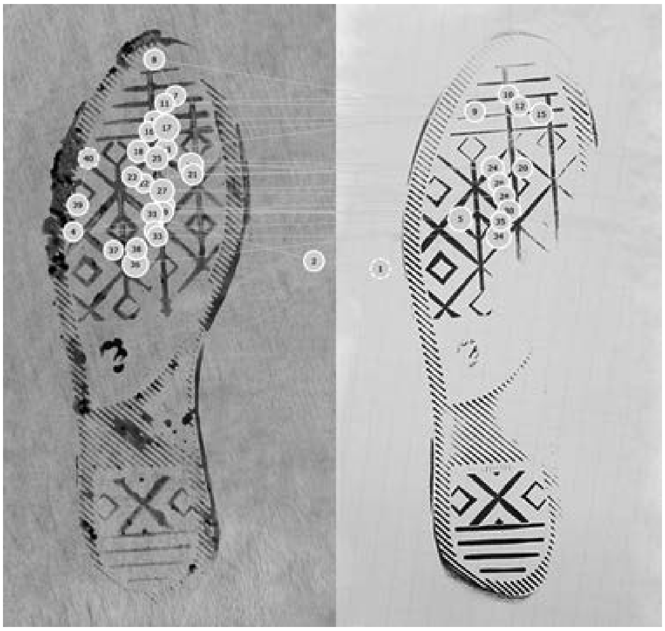


FIG. 6—*Gaze plot of subject A in the morning session. The yellow line represents the gaze trace. The yellow dots are the fixation points, and the size of dot stands for the fixation duration, where larger dot suggest longer fixation duration. [Color figure can be viewed at wileyonlinelibrary.com]*

with fatigue level (Table 7). However, there is no significant correlation between eye-tracking metrics and experience.

Gaze Plot

In order to visualize the processing of subjects, four gaze examples from two subjects were presented. Figure 6 illustrates 40 eye movements by an FFE (subject A) in the morning recorded. The gaze plot visually displays which feature did the examiner observe during examination, and the sequence of each eye movement. It can be seen that the examiner employed a side-by-side search pattern and was consistent with his logical and systematic processing strategies repeatedly from unknown shoeprints (on the left side) to the known samples (on the right side). Correspondingly, no obvious change in the examiner's processing strategy was observed when examiners were presented with the horizontally flipped shoeprint of the same image pair in the afternoon. Besides, the examiner's searching pattern in the afternoon was generally consistent with that in the morning (Fig. 7).

Similarly, when the footwear impression was interfered with complex background, based on the gaze plot in the morning

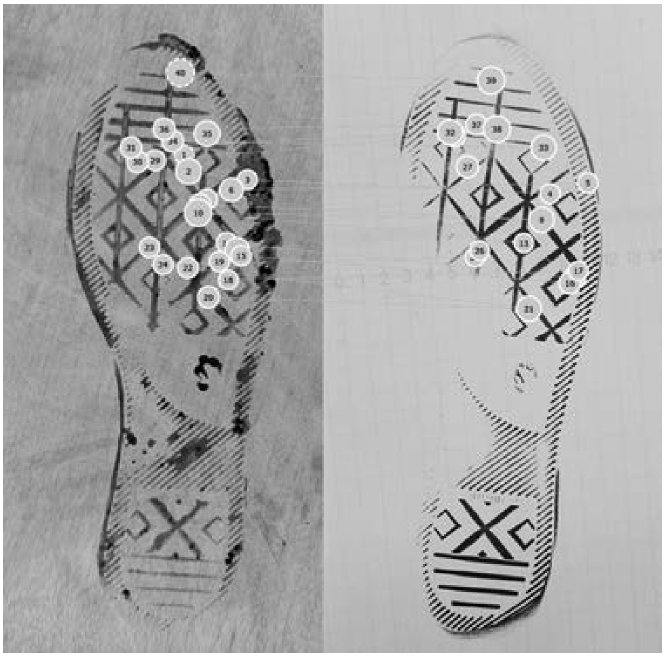


FIG. 7—*Gaze plot of subject A in the afternoon session. When examining the horizontally reversed image of footwear impression in Fig. 6, where the searching pattern did not change. [Color figure can be viewed at wileyonlinelibrary.com]*

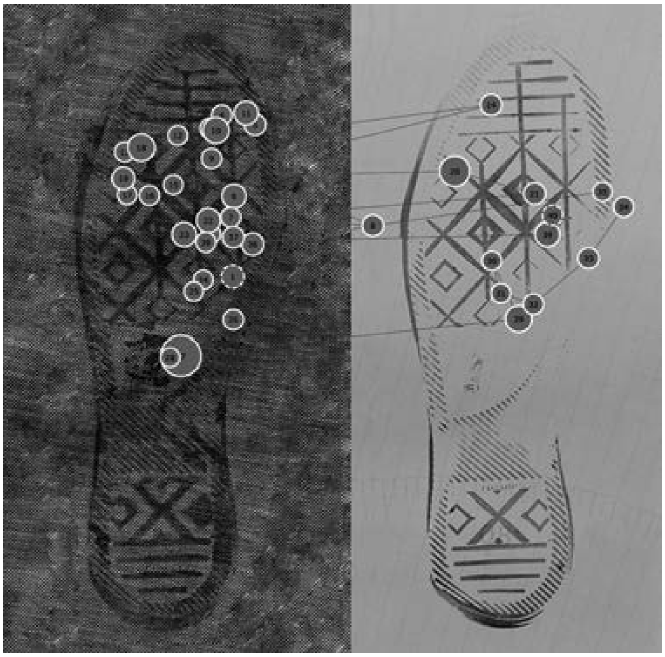


FIG. 9—*Searching pattern of subject B in the afternoon session when examining the horizontally reversed image of footwear impression in Fig. 8, where the searching area and distribution are essentially the same. [Color figure can be viewed at wileyonlinelibrary.com]*

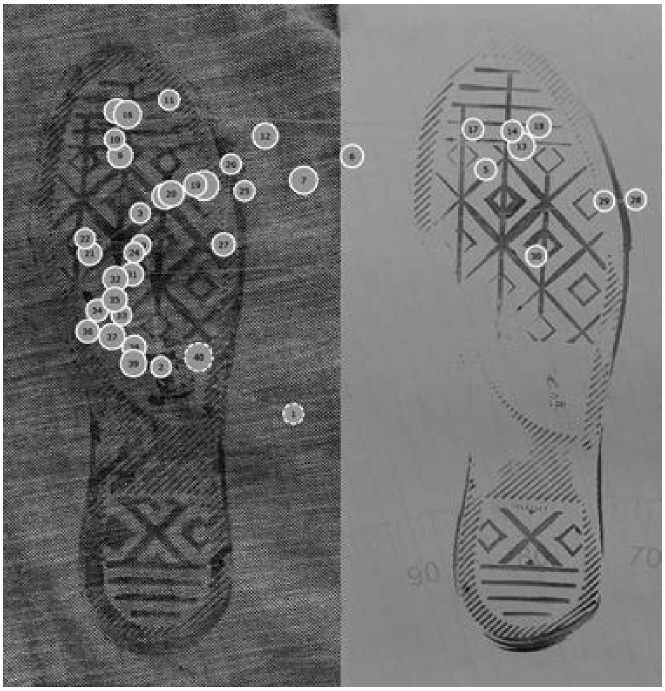


FIG. 8—*Searching pattern of subject B in the morning session. The gaze points were mainly distributed in the ball area. [Color figure can be viewed at wileyonlinelibrary.com]*

(Fig. 8) and afternoon (Fig. 9) similar searching strategy and searching range of examiner (subject B) was observed. In other words, the searching strategies of subjects did not change significantly when they are working under fatigue.

However, when the fatigued subjects are regarded as a group, will the inter-examiner consistency be reduced, and will they observe different regions? Earth Mover's Distance (EDM) is a distance metric for evaluating the similarity of histograms proposed by Rubner et al. (17). EMD is actually the optimal solution to the transportation problem in Linear Programming, that is, the minimum cost of changing from one distribution to another. EMD is a statistical method used to solve the problem of how similar two gaze sets (18). In the process of footwear comparison, the examiner's fixation can be used as a feature point in EMD, and the fixation duration is the weight of this feature point. The combination of all fixation positions (X, Y) and fixation duration in a footprint image is defined as a distribution. By calculating the EMD between the two examiners, we can judge the similarity between examiners in the comparison process.

We compute the distance between each subject and the remaining subjects in the morning and afternoon. In the morning, we compare subject 1 to the combined data of all of the other subjects so that we could know how different is subject 1 from everyone else. By calculating the subject 1 versus subjects 2–12, we obtain one number that tells us the distance between subject 1 and everybody else. Then we repeat that by calculating the distance between subject 2 and the combined data of subject1 and subjects 3–12. In this way, we obtained 12 data for each session (Table 8). Therefore, a paired-samples *t*-test could be employed to compare the EMD difference between the morning session and the afternoon session. The result suggested no EMD difference between two sessions ($t(11) = -1.47$, $p = 0.17$), which proved the consistency of examiners with each other when they are fatigued.

Discussion

This paper proposes the eye-tracking method and questionnaires to investigate the impact of fatigue on footwear examination. All 12 subjects were reported to have varying levels of

TABLE 8—*Earth Mover Metric (EMD) value of 12 subjects for morning and afternoon sessions.*

Subject Number	Morning EMD	Afternoon EMD
1	122.83	104.19
2	125.23	174.4
3	81.67	122.65
4	234.16	201.85
5	115.26	195.63
6	125.39	96.87
7	76.12	62.13
8	134.92	129.7
9	71.38	142.44
10	73.67	68.63
11	66.06	85.08
12	124.96	167.2

fatigue before the second session. The manipulation of fatigue was not strong; however, it should have reached the high level of fatigue that might be encountered in daily work.

The results of this study not only validate the applicability of eye tracker technology to the assessment of footwear comparisons skill, but also provide the results of conclusion, difficulty level evaluation and searching strategy produced by tired subjects. On the one hand, there was no significant change in the eye-tracking metrics, indicating that fatigue did not decrease the examiners' processing ability of stimulus materials. On the other hand, the accuracy did not drop significantly, suggesting that the decision-making of the subjects as a group was not affected by fatigue.

During the two sessions, left pupil diameter was found to be wider in the morning, which suggested that the subjects were more concentrated with a higher cognitive demand at that time. With fatigue, left pupil diameter decreased significantly, and the fluctuation of pupil diameter increased.

Overall, the statistic results showed that the subjects as a group were not significantly affected by fatigue. Nonetheless, several subjects might be affected by fatigue according to the preliminary analysis of individual eye gaze data. We think this is an interesting and important phenomenon that might call for further investigation in the future. Accordingly, even if the examiners' performance as a group were not significantly affected by fatigue, we still advocate not carrying out difficult tasks when FFEs are tired.

Limitations

Nevertheless, our study still has its limitations. First, the subjects were required to operate in front of a computer without access to physical photos, which was slightly different from their usual working habits. Second, owning to the restraints of eye-tracking devices, the shoeprint pairs had to be presented side by side on the laptop screen. Accordingly, instead of multiple images of known samples being presented to the experts, there is only one image of known sample, which might increase the difficulty in comparison, especially the close nonmatch shoeprint pairs. Third, even though we tried to simulate questioned shoeprints samples in the real case as much as possible, there could only be one questioned sample being presented. In other words, not every detail of the footwear impression could be reflected. Fourthly, for individual differences, the subject's biological clock was not taken into account. Besides, subjects were not asked to fill out the fatigue survey in the morning. Therefore, to a certain extent, some of the limitations above might affect the objectivity of decision-making and may also explain why there is a higher error rate in the sessions.

Conclusion

The results of this study demonstrate that fatigue did not produce significant differences in questionnaires and eye-tracking tests. In the two sessions, the FFEs' decision-making in footwear examination was not affected significantly either.

Eye-tracking proves to be useful for revealing the searching strategies and technical skills of an examiner objectively and, thus, could be a promising means for relevant training and education. In this study, we discussed the differences in pupil diameter resulted by the fatigue. That is to say, it is possible to monitor fatigue through pupil diameter in the future.

Moreover, prospect research could focus on the differences in the searching patterns between experts and novices through eye metrics. Furthermore, in view of the complex and interactive decision-making process, in addition to external factors such as fatigue, more in-depth research should also be conducted on the internal factors affecting the examiner.

Acknowledgments

The authors are thankful for enlightening communications with Thomas Busey and John Vanderkolk. Thanks to the FFEs who sacrificed their personal time to participate in the sessions. Their responsible attitude and constructive suggestions are important foundations for the completion of the experiment. Special thanks to Yongzhi Quan for his programming support. And thanks to the support by Shiquan Liu, Jing Li, and Hongxia Xu. Acknowledgments are also due to the Nanjing University of Science and Technology (NJUST) for assistance with apparatuses.

References

1. Bodziak WJ. Forensic footwear evidence, 2nd edn. Boca Raton, FL: CRC Press, 2017;24–5.
2. Majamaa H, Ytti A. Survey of the conclusions drawn of similar footwear cases in various crime laboratories. Forensic Sci Int 1996;182(1):109–20. https://doi.org/10.1016/0379-0738(96)01972-X
3. Jeanguenat AM, Dror IE. Human factors effecting forensic decision making: workplace stress and well-being. J Forensic Sci 2018;63 (1):258–61. https://doi.org/10.1111/1556-4029.13533
4. Marcora SM, Staiano W, Manning V. Mental fatigue impairs physical performance in humans. J Appl Physiol 2009;106(3):857–64. https://doi.org/10.1152/japplphysiol.91324.2008
5. Boksem MAS, Meijman TF, Lorist MM. Effects of mental fatigue on attention: an ERP study. Brain Res Cogn Brain Res 2005;25(1):107–16. https://doi.org/10.1016/j.cogbrainres.2005.04.011
6. Van den Linden D, Eling P. Mental fatigue disturbs local processing more than global processing. Psychol Res 2006;70(5):395–402. https://doi.org/10.1007/s00426-005-0228-7
7. OSAC Footwear & Tire Subcommittee's response to the President's Council of Advisors on Science and Technology's (PCAST) request for information. Submitted December 2015. https://www.nist.gov/system/files/documents/2016/12/16/osac_footwear_tire_subcommittees_response_to_the_presidents_council_of_advsiors_on_science_and_technologys_pcast_request_for_information_-_submitted_december_2015.pdf (accessed July 6, 2020).
8. Breen CJ, Bond R, Finlay D. An evaluation of eye tracking technology in the assessment of 12 lead electrocardiography interpretation. J Electrocardiol 2014;46(6):922–9. https://doi.org/10.1016/j.jelectrocard.2014.08.008
9. Ozcelik E, Karakus T, Kursun E, Cagiltay K. An eye-tracking study of how color coding affects multimedia learning. Comput Educ 2009;53 (2):445–53. https://doi.org/10.1016/j.compedu.2009.03.002

10. Henderson JM. Gaze control as prediction. Trends Cogn Sci 2017;21(1):15–23. https://doi.org/10.1016/j.tics.2016.11.003
11. Van den Linden D, Frese M, Sonnentag S. The impact of mental fatigue on exploration in a complex computer task: rigidity and loss of systematic strategies. Hum Factors 2003;45(3):483–93. https://doi.org/10.1518/hfes.45.3.483.27256
12. Wang LL, Li YJ. 心理疲劳与任务框架对风险决策的影响 [The effect of mental fatigue and framing on risk decision-making]. Adv Psychol Sci 2012;20(11):1546–50. https://doi.org/10.3724/SP.J.1042.2012.01546
13. Busey T, Swofford HJ, Vanderkolk J, Emerick B. The impact of fatigue on latent print examinations as revealed by behavioral and eye gaze testing. Forensic Sci Int 2015;251:202–8. https://doi.org/10.1016/j.forsciint.2015.03.028
14. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett 2006;27(8):861–74. https://doi.org/10.1016/j.patrec.2005.10.010
15. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45. https://doi.org/10.2307/2531595
16. Jin HB, Yu GH, Liu HB. 瞳孔直径检测管制疲劳的有效性分析 [Effectiveness analysis of pupil diameter detection for airtraffic controller's fatigue]. J Beijing Univ Aero Astro 2018;44(7):1402–7. https://doi.org/10.13700/j.bh.1001-5965.2017.0553
17. Rubner Y, Tomasi C, Guibas LJ. The Earth mover's distance as a metric for image retrieval. Int J Computer Vis 2000;40(2):99–121. https://doi.org/10.1023/A:1026543900054
18. Busey T, Yu C, Wyatte D, Vanderkolk JR, Parada FJ, Akavipat R. Consistency and variability among latent print examiners as revealed by eye tracking methodologies. J Forensic Identif 2011;61(1):60–91.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Each examiner's conclusion of the 10 shoeprints in the morning session.

Table S2. Each examiner's conclusion of the 10 shoeprints in the afternoon session.

Table S3. Subjects' difficulty rating of shoeprints in the morning session.

Table S4. Subjects' difficulty rating of shoeprints in the afternoon session.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14507
Available online at: onlinelibrary.wiley.com

PAPER

ODONTOLOGY

Jun Ai Chong,[1] B.D.S., M.F.D.S., R.C.S.Ed.; Alizae Marny Fadzlin Syed Mohamed,[1] B.D.S., M.Sc., M.Orth.R.C.S.; Murshida Marizan Nor,[1] D.D.S., M.Sc.D., M.Orth.R.C.S.; and Allan Pau,[2] B.D.S., M.Sc., Ph.D., F.D.S., R.C.S.Ed.

The Heritability of Palatal Rugae Morphology Among Siblings*'†

ABSTRACT: Although there is clinical applicability of the palatal rugae as an identification tool in forensic odontology, controversy exists whether the palatal rugae patterns are stable or variable. The greater the genetic component, the higher the probability that palatal rugae patterns are stable. The aim of this study was to compare the palatal rugae morphology between full siblings and the proportion of variability due to genetic component. This cross-sectional study was conducted on digital models of 162 siblings aged 15–30 years old. The palatal rugae patterns were assessed with Thomas and Kotze (1983) classification using Geomagic Studio software (3D Systems, Rock Hill, SC). The palatal rugae morphology between siblings showed significantly similar characteristics for total number of left rugae (p = 0.001), left primary rugae (p = 0.017), secondary rugae for right (p = 0.024) and left sides (p = 0.001), right straight rugae (p = 0.010), and right convergent rugae (p = 0.005) accounting for at least 6.25%-12.8% of the variability due to heredity. Despite the similarities found, the palatal rugae patterns showed significant differences between siblings of at least 46.9% (p = 0.001). Zero heritability was found in 9 of the 14 rugae patterns. Meanwhile, total number of rugae, primary, backward, and convergent rugae showed moderate heritability (h^2 > 0.3) and total number of secondary rugae showed high heritability (h^2 > 0.6). In conclusion, despite the individuality characteristics, an appreciable hereditary component is observed with significant similarities found between sibling pairs and the palatal rugae patterns were both environmentally and genetically influenced.

KEYWORDS: identification tool, forensic odontology, heredity, siblings, palate, stable, rugae

Palatal rugae are asymmetric ridges of connective tissue located behind the incisive papilla at the anterior part of the hard palate (1). There are three to five rugae in each palatal half which can vary at both sides with no bilateral symmetry in rugae patterns. The palatal rugae patterns have been described by various researchers according to its number, position, length, shape, direction, and unification. A more detailed system is said to have a greater discrimination power, and various attempts have been made toward that approach throughout the nineteenth century. However, a complex system to transform the subjective observations and interpretations of the palatal rugae patterns into objective data could make it harder to apply and a faster classification may provide better results. Additionally, the lack of uniformity among the range of criteria used across the different classifications has hindered meaningful comparisons among studies (2). Comparisons of the latest studies of the palatal rugae patterns worldwide have shown that the Thomas and Kotze (1983) classification is the most widely used (3).

Previous studies suggest a hereditary link with palatal rugae patterns. Patel et al. (2015) found significant resemblance of palatal rugae patterns between parents and offspring. Besides that, rugae patterns of monozygote twins were shown to be related to near identical measurements of rugae lengths (4). Meanwhile, Mala and colleagues studied the palatal rugae of 30 individuals from three consecutive generations of ten different families. They discovered self-repetition of the palatal rugae patterns among the generations with 10% showing repetition in all the three generations, another 10% showed repetition in alternate generations, and 20% showed repetition in two consecutive generations. However, 60% showed no repetition in any generation (5).

Controversy remains whether palatal rugae patterns are stable or variable because studies have shown that the position of the rugae can change by stretching to accommodate the dentoalveolar growth with the most affected rugae being the first rugae which is closest to the maxillary incisor (6). Additionally, the lengths, numbers, and positions of the rugae are affected by orthodontic treatment depending on the magnitude and type of tooth movement. Thus, a better understanding of the relative effects of the environment and genes to the palatal rugae is essential as to whether we can adopt this anatomical structure as a stable landmark. The greater the genetic component, the higher the probability that palatal rugae patterns are stable.

Heritability estimate (h^2) expresses the proportion of phenotypic variance that is due to variation in breeding value of the

[1] Discipline of Orthodontics, Department of Family Oral Health, Universiti Kebangsaan Malaysia, Jalan Raja Muda Abdul Aziz, Kuala Lumpur, Federal Territory of Kuala Lumpur, 50300, Malaysia.

[2] Dental Public Health, School of Dentistry, International Medical University, 126, Jln Jalil Perkasa 19, Bukit Jalil, Kuala Lumpur, Federal Territory of Kuala Lumpur, 57000, Malaysia.

Corresponding author: Alizae Marny Fadzlin Syed Mohamed, B.D.S., M.Sc., M.Orth.R.C.S. E-mail: alizaemarny@ukm.edu.my

*Supported by the Universiti Kebangsaan Malaysia Research Grant (GUP-2019-029).

†Presented in part at the 2nd Postgraduate Dental Symposium at Universiti Kebangsaan Malaysia, March 11, 2020, in Kuala Lumpur, Malaysia.

Received 27 May 2020; and in revised form 18 June 2020; accepted 22 June 2020.

sampled population at that point in time, indicating the relative importance of nature versus nurture on a given trait (7). It is a ratio ranging from 0 for no heritability to 1 for completely inherited trait. Thus, there will be higher resemblances between siblings when the h^2 value is closer toward 1.00 and vice versa (8). Intraclass correlation coefficient (ri) is commonly used to quantify the degree to which individuals with a fixed degree of relatedness (e.g., full siblings) resemble each other in terms of a quantitative trait. Therefore, calculation of h^2 will be considered as twice the intraclass correlation coefficient (h^2 = 2ri) in a full sibling model. The application of heritability estimate in family studies is a preliminary step to analyze the factors that affect the phenotype with the goal of identifying causative agents (7). Thus, if the palatal rugae patterns have trivial heritability, research should be directed toward factors that can influence the rugae patterns. The aim of this study is to investigate the variation and similarities in palatal rugae morphology between siblings and the relative contributions of heredity to the palatal rugae patterns.

Methods

This was a cross-sectional study conducted on digital casts of 81 pairs of Malaysian siblings aged 15–30 years old. Ethical approval was obtained from the Universiti Kebangsaan Malaysia Research Ethics Committee prior to its commencement (UKM PPI/111/8/JEP-2018-507).

The subjects were recruited from patients who sought dental treatment at the Faculty of Dentistry, Universiti Kebangsaan Malaysia (UKM), through convenience sampling. All subjects were healthy, aged 15–30 years with complete permanent dentition (excluding third molars) that were sound or with minor restorations and no obvious attritions. The sibling pairs were first-degree siblings as avowed by the parents. No serological test was used to confirm consanguinity. In addition, the subjects had no history of or were currently undergoing orthodontic treatment, craniofacial surgery, and mucogingival surgery. Twins and those with presence of any active periodontal and dental diseases were excluded.

The sample size was estimated with a power of 80%, at a margin error of 5%, confidence interval of 95%, mean difference of 1.6 (9), and standard deviation as 2.79. Hence, 24 pairs of siblings were required per group. However, there was an expected attrition rate of 10%. Therefore, total subjects per group were 27 pairs of siblings or 54 subjects per group. The three sibling groups were female–female (F-F), male–male (M-M), and female–male (F-M).

Dental Casts Preparation

The subjects had dental impression of their maxillary arch taken with alginate of which the dental casts were constructed upon. Then, all dental casts were digitized using a 3-dimensional laser scanner, Rexcan CS + scanner (Solutionix Corp., Seoul, Korea), which has a built-in image software (EZ-scan). The scanner was calibrated following the manufacturer's instructions prior to the digital scans. The data obtained were exported in a stereolithographic (STL) file format into the Geomagic Studio 2014 software (3D Systems, Rock Hill, SC).

Prior to the palatal rugae assessment, the digital casts were randomized, and the examiner (JAC) blinded to the subjects of the digital casts. The palatal rugae were digitally drawn using the Geomagic Studio 2014 software (3D Systems).

The palatal rugae was assessed based on the Thomas and Kotze (10) classification which includes the position, number, length, shape, direction, and unification of the rugae.

Position and Number of Rugae

The position of the rugae is designated based on the side that they are located and to which zone they belonged. The mid-palatine raphe divides the palatal region to the right and left side. The dental cast was further divided into the following five zones based on six horizontal lines (Fig. 1):

- Line I: Transverse line passing through the incisal third at the palatal of the central incisors.
- Line II: Transverse line from the mesial side of right lateral incisor to mesial side of left lateral incisor.
- Line III: Transverse line through the mesial side of the right canine to the mesial side of left canine.
- Line IV: Transverse line through the mesial side of the right first premolar to the mesial side of the left first premolar.
- Line V: Transverse line through the mesial side of the right second premolar to the mesial side of the left second premolar.
- Line VI: Transverse line through the distal side of the second premolar to the right side of the distal of left second premolar.

Therefore, zone 1 is between line 1 and 2 (central incisor region), zone 2 is between line 2 and 3 (lateral incisor region), zone 3 is between line 3 and 4 (canine region), zone 4 is between line 4 and 5 (first premolar region), and zone 5 is between line 5 and 6 (second premolar region).

Length of the Rugae

The length of each rugae was measured from the most lateral point to the most medial point with the digital caliper in Geomagic Studio 2014 software (3D Systems). The software allows rotation of the digital model and magnification of the images for better identification of the anatomic landmarks. The origin of the axis of rotation was placed at the medial or lateral point of the rugae accordingly

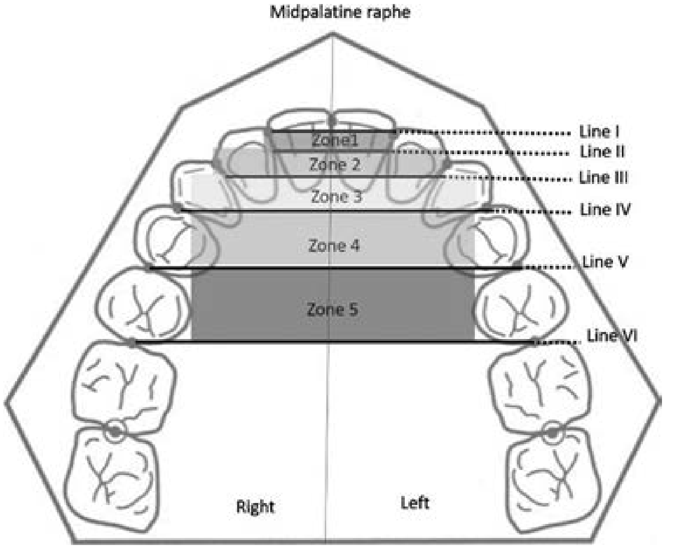


FIG. 1—*Position of the rugae. [Color figure can be viewed at wileyonlinelibrary.com]*

with the axes of rotation set parallel to the mid-palatine raphe for the *Y*-axis, lateral movement from the mid-palatine raphe for the *X*-axis, and vertically to the occlusal plane for the *Z*-axis. The rugae were classified based on their length as follows:

- If the length of the rugae was 5 mm or more, it was classified as a primary rugae.
- If the length of the rugae was 3–5 mm, it was classified as a secondary rugae.
- If the length of the rugae was 2–3 mm, it was classified as a fragmentary rugae.
- Rugae measuring <2 mm was discarded.

Shape of the Rugae

The rugae shape was classified into four major types:

- Straight type: runs in a straight line directly from their origin to termination.
- Curved type: simple crescent shape with a slightest bend in the middle.
- Wavy type: basic serpentine shape or presence of slight curves at the origin or termination.
- Circular type: definite continuous ring formation.

Direction of the Rugae

The rugae direction was determined by measuring the angle between the line joining its origin and termination and a line perpendicular to the mid-palatine raphe (MPR) and divided into three types:

- Forward directed rugae: positive angle formed with MPR perpendicular.
- Backward directed rugae: negative angle formed with MPR perpendicular.
- Perpendicular rugae: angle of zero degrees.

Type of Rugae Unification

The unification pattern was categorized into two types:

- Diverging: immediate branching of the rugae from a common origin at the midline.
- Converging: rugae with different origins from midline but are joined on their lateral portions.

Reliability Testing

Ten percent of the sample size (17 dental casts) were randomly selected for the intra-examiner reliability testing. The assessment of the palatal rugae morphology of these dental casts was conducted twice with a two-week interval by the same examiner. The intra-examiner reliability was deemed good to excellent as the Cohen's Kappa (κ) scores and intraclass correlation coefficient for categorical and continuous variables, respectively, ranged from 0.838 to 1.000. Another independent examiner (AMFSM) measured the same 17 study models for the inter-examiner reliability. The inter-examiner agreement was also deemed good to excellent as it was above 0.80 for each variable.

Statistical Analysis

Data obtained from this study were analyzed using IBM Statistical Package for the Social Sciences 25.0 Software (SPSS, Chicago, IL). Pearson's chi-square test of contingencies was used to compare the palatal rugae patterns between the sibling groups. Meanwhile, Pearson's correlation coefficient and Spearman's rho analysis were used to investigate the similarities between the sibling pairs. Discriminant function analysis was applied to produce a discriminant function score, and the percentage of correct prediction for sibling groups was calculated. Heritability estimates for each of the palatal rugae patterns were also obtained.

Results

Baseline Characteristics of the Subjects

The mean age of the sibling groups was 18 ± 3.2 years for F-F, 21 ± 4.4 years for M-M, and 18 ± 4.0 years for F-M as shown in Table 1. Most of the subjects were Malays with 78% for F-F and 52% for both the M-M and F-M siblings while Chinese accounted for 18% of F-F, 44% of M-M, and 41% of F-M sibling groups. The remainder of the subjects was Indian.

Number of Palatal Rugae According to Location

The rugae were predominantly located at the fourth and fifth zone which is at the premolar region. F-F (n = 223) had significantly more rugae at the fourth zone, followed by F-M (n = 197) and M-M (n = 167) (p = 0.004) as shown in Table 2.

Number of Palatal Rugae Based on Length

M-M had significantly longer first right rugae (9.65 ± 1.57 mm) than the F-F (8.62 ± 1.52 mm) and F-M sibling groups (8.67 ± 1.56 mm). Similarly, the average rugae lengths bilaterally present the same pattern of longer rugae in M-M as compared to F-F and F-M sibling groups.

TABLE 1—*Baseline characteristics of the subjects.*

	Female–Female (n = 54)		Male–Male (n = 54)		Female–Male (n = 54)	
Baseline Characteristics	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range
Age	18.30 (3.15)	15-27	21.04 (4.37)	15–30	18.13 (3.96)	15–30
Race	n	%	n	%	n	%
Malay	42	78	28	52	28	52
Chinese	10	18	24	44	22	41
Indian	2	4	2	4	4	7

n, number; SD, standard deviation; %, percentage.

TABLE 2—*Number of palatal rugae according to location among sibling groups.*

	Female–Female (n = 54)		Male–Male (n = 54)		Female–Male (n = 54)		Comparison Between Groups
Zone	n	%	n	%	n	%	p Value
2	4	1	4	1	7	1	0.698
3	128	23	134	26	133	26	0.726
4	223	41	167	33	197	39	0.004**
5	191	35	201	39	169	33	0.227
6	0	0	3	1	3	1	0.232

n, number; p value by chi-square.
**Significance level $p < 0.01$.

Meanwhile, F-F had the most number of rugae (n = 545) compared to the M-M (n = 505) and F-M (n = 509) sibling groups but it was not significantly different between the sibling groups (p = 0.391). The rugae predominantly consist of primary rugae followed by secondary rugae and lastly, fragmentary rugae. Only the number of secondary rugae showed significant difference among the sibling groups of which the F-F had the most right secondary rugae (n = 61, p = 0.002) and total secondary rugae (n = 107, p = 0.020) as compared to the F-M and M-M sibling groups as shown in Table 3.

Number of Palatal Rugae According to Shape

Table 4 illustrates the total number of rugae based on shape of which the rugae was predominantly wavy, followed by straight, then curve, and lastly circle. There were no significant differences between the shape of the palatal rugae among the sibling groups.

Number of Palatal Rugae According to Direction

Backward directed rugae was the most common in this study group, while forward directed rugae was the least common as shown in Table 5. Perpendicular rugae showed significant difference among the sibling groups (p = 0.001) with F-F showing the highest amount (n = 225) followed by M-F (n = 160) and the least in the M-M sibling group (n = 130).

TABLE 3—*Number of rugae based on length among sibling groups.*

Number of Rugae		Female–Female (n = 54)		Male–Male (n = 54)		Female–Male (n = 54)		Comparison Between Groups
		n	%	n	%	n	%	p Value
Primary	Right	201	75	214	86	221	87	0.750
	Left	224	81	225	87	220	86	0.512
	Total	425	78	439	87	441	87	0.222
Secondary	Right	61	23	28	11	30	12	0.002**
	Left	46	17	31	12	31	12	0.126
	Total	107	20	59	12	61	12	0.020*
Fragmentary	Right	6	2	6	3	3	1	0.516
	Left	6	2	1	1	4	2	0.191
	Total	12	2	7	1	7	1	0.608
Total	Right	268	49	248	49	254	50	0.145
	Left	277	51	257	51	255	50	0.413
	Total	545	100	505	100	509	100	0.391

n, number; p value by chi-square.
*Significance level $p < 0.05$.
**Significance level $p < 0.01$.

TABLE 4—*Total number of rugae based on shape among sibling groups.*

Shape		Female–Female (n = 54)		Male–Male (n = 54)		Female–Male (n = 54)		Comparison Between Groups
		n	%	n	%	n	%	p Value
Curve	Right	52	20	47	19	49	19	0.791
	Left	55	20	57	22	55	22	0.739
	Total	107	20	104	21	104	21	0.798
Circle	Right	3	2	6	2	8	3	0.287
	Left	9	3	5	2	3	1	0.159
	Total	12	2	11	2	11	2	0.642
Straight	Right	53	18	31	13	38	15	0.206
	Left	38	14	35	14	29	11	0.466
	Total	91	17	66	13	67	13	0.302
Wavy	Right	160	60	164	66	159	63	0.340
	Left	175	63	160	62	169	66	0.246
	Total	335	61	324	64	328	64	0.490

n, number; p value by chi-square.

TABLE 5— *among sibling groups.*

	Female–Female (n = 54)		Male–Male (n = 54)		Female-Male (n = 54)		Comparison Between Groups
Direction	n	%	n	%	n	%	p Value
Backward	211	39	204	40	244	48	0.121
Forward	110	20	140	28	134	26	0.383
Perpendicular	225	41	160	32	130	26	0.001**

n, number; p value by chi-square.
**Significance level $p < 0.01$.

Number of Palatal Rugae According to Unification

This study group had more divergent type of rugae than convergent rugae. However, the unification pattern was not statistically different between the sibling groups as presented in Table 6.

Similaraties of Rugae Morphology among Sibling Pairs

There were only statistically significant association between the sibling pairs for total number of left rugae (p = 0.001), left primary rugae (p = 0.017), secondary rugae for right (p = 0.024) and left sides (p = 0.001), right straight rugae (p = 0.010), and right convergent rugae (p = 0.005) accounting for at least 6.25%-12.8% of the variability due to heredity.

TABLE 6—*Number of rugae based on unification among sibling groups.*

Unification		Female–Female (n = 54)		Male–Male (n = 54)		Female–Male (n = 54)		Comparison Between Groups
		n	%	n	%	n	%	p Value
Divergent	Right	54	50	72	52	50	52	0.081
	Left	54	50	66	48	46	48	0.370
	Total	108	100	138	100	96	100	0.297
Convergent	Right	22	48	22	50	12	38	0.342
	Left	24	52	22	50	20	62	0.906
	Total	46	100	44	100	32	100	0.801

n, number; p value by chi-square.

Differences of Rugae Morphology Among Sibling Pairs

The palatal rugae patterns showed significant differences between siblings of at least 46.9% (p = 0.001). The direction of the left rugae showed the most significant difference between siblings (91.4%). Meanwhile, the length of the third rugae had the highest mean difference of 4.25 mm while the average length of the right rugae showed the least difference of 1.53 mm. All the rugae were significantly different in length between siblings (p = 0.001).

Figure 2 shows the palatal rugae pattern of a pair of siblings from the F-F sibling group. The number of rugae between the siblings differs at both sides with four right rugae and seven left rugae in sibling 1, whereas there were seven right rugae and five left rugae in sibling 2. Although there were some similarities among the siblings such as the first right rugae was straight in shape in both siblings, generally, the palatal rugae patterns were uniquely different between siblings.

Sibling Group Identification with Discriminant Analysis

Discriminant analysis was used to determine the ability of palatal rugae patterns in distinguishing sibling pairs into their respective groupings. The results are presented in Table 7 which shows the relative contribution of each of the rugae pattern to the discriminant function. Predictor variables were differences of average rugae length, total number of primary rugae, secondary rugae, shape, direction, and unification between siblings. All of the predictors did not show significant mean differences. However, the predictor with the greatest discriminating ability was the difference in average rugae length as it had the highest coefficient value of 0.74 whereas difference in total number of primary rugae was a poor predictor with a low coefficient value of −0.141. Meanwhile, differences of average length, shape, direction, and unification had positive correlation with the discriminant function. The mean discriminant score for the F-F was −0.315, M-M was 0.532, and F-M sibling group was −0.217.

The first function has a low discriminatory ability due to the low eigenvalue of 0.149, low canonical correlation of 0.360, high wilk's lambda score of 0.792, and p value showing non-significance (p = 0.127). Furthermore, the function explained only 59.8% of the total variance. The following discriminant function was obtained as a predictive equation to classify sibling pairs into their respective groups based on the differences of each palatal rugae patterns:

Function 1: −3.27 + 2.18 (direction) +2.72 (unification) −0.91 (shape) −0.23 (total number of primary rugae) −0.55 (total number of secondary rugae) +0.69 (average length).

TABLE 7—*Discriminant function coefficients for differences in rugae characteristics between siblings.*

			Group Centroids		
Variables	Structure Matrix	Unstandardized Coefficients	Female–Female	Male–Male	Female–Male
Average length	0.681	0.694	−0.315	0.532	−0.217
TN primary rugae	−0.121	−0.234			
TN secondary rugae	−0.258	−0.553			
Shape	0.055	−0.909			
Direction	0.431	2.18			
Unification	0.462	2.72			

TN, total number.

TABLE 8—*Accuracy of the discriminant function for sibling group identification.*

		Predicted Group Membership							
		Male–Male		Female–Female		Female–Male		Total	
		n	%	*n*	%	*n*	%	*n*	%
Original	Male–Male	18	66.7	4	14.8	5	18.5	27	100.0
	Female–Female	6	22.2	15	55.6	6	22.2	27	100.0
	Female–Male	6	22.2	4	14.8	17	63.0	27	100.0
Cross-validated	Male–Male	15	55.6	6	22.2	6	22.2	27	100.0
	Female–Female	8	29.6	12	44.4	7	25.9	27	100.0
	Female–Male	7	25.9	6	22.2	14	51.9	27	100.0

n, number of subjects.

The original model correctly classified 66.7% in the M-M, 55.6% in the F-F, and 63% in the F-M sibling group as shown in Table 8. Meanwhile, the precision of the cross validation function in discriminating the sibling groups was lower (51%) as compared to the original results (62%).

Heritability Estimates of the Palatal Rugae Patterns

Table 9 shows the h^2 of the palatal rugae patterns among sibling groups. Total number of primary rugae (M-M), total number of secondary rugae (M-M and F-F), total number of wavy rugae (F-M), and total number of convergent rugae (F-M) had

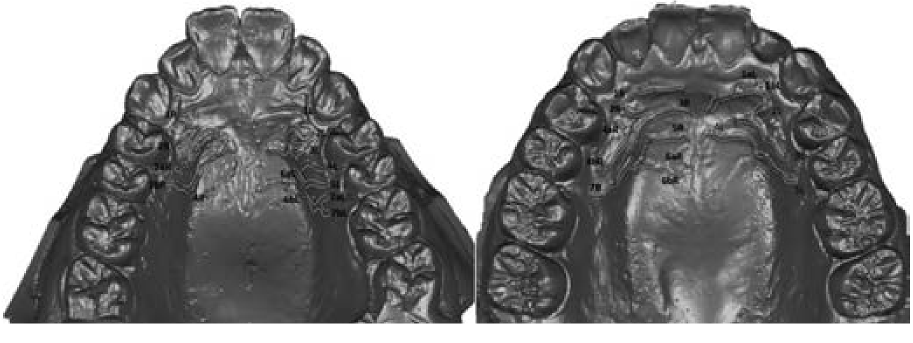


FIG. 2—*Palatal rugae patterns of a pair of siblings (Sibling 1: Right, Sibling 2: Left). [Color figure can be viewed at wileyonlinelibrary.com]*

TABLE 9—*Heritability estimates of the palatal rugae among sibling groups.*

	Heritability Estimates (h^2)		
Palatal Rugae Patterns	Male–Male	Female–Female	Female–Male
TN	0.854	0.54	0.31
Average length	0	0.508	0.29
TN primary	1.042	0.904	0.39
TN secondary	1.102	1.014	0.61
TN fragmentary	0	0.488	0.548
TN curve	0.286	0	0.74
TN wavy	0.478	0	1.152
TN straight	0.886	0	0.574
TN circle	0	0.444	0.568
TN forward	0.198	0	0
TN perpendicular	0	0	0.518
TN backward	0.328	0.506	0.316
TN divergent	0	0.506	0
TN convergent	0.414	0.912	1.314

TN, total number.

invariably inflated h^2 values ($h^2 > 1$). On the contrary, zero heritability was found in 9 of the 14 variables. Overall, total number of rugae, total number of primary rugae, total number of backward rugae, and total number of convergent rugae showed moderate heritability ($h^2 > 0.3$) and total number of secondary rugae showed high heritability ($h^2 > 0.6$).

Discussion

This study is pertaining to the palatal rugae patterns of 81 pairs of young Malaysian siblings. The findings demonstrated that every individual had unique rugae patterns with significant differences between siblings of at least 47%. However, the variability due to heredity did exist. The h^2 showed that the palatal rugae patterns were both environmentally and genetically influenced. Nonetheless, the findings of this study may lack generalizability because convenience sampling methodology was used. Constraining the sampling frame to reduce sociodemographic heterogeneity was done to limit the amount of bias (11) such as limiting the age group to 15–30 years old. Additionally, the age group was chosen because maxillary length increases from ages 6–12 years but stabilizes after 14 years (12).

The rugae in this study were predominantly located at the fourth and fifth zone which is at the premolar region. These findings were similar with a study by Hermosilla and co-workers who examined the rugae location of 120 subjects aged 15–50 years old in Chile, South America. They observed that 40% of the rugae were commonly observed at the second premolar region and 30% at the first premolar region (13). Additionally, the palatal rugae of 100 subjects aged 17–25 years at Davangere, South of India, were commonly found at the first premolar area (45% in males and 42% in females) followed by the second premolar area (29% in males and 32% in females). However, the distribution of the rugae did not show any sexual dimorphism in their study ($p = 0.440$) (14). This is in contrast with our study of which the palatal rugae at the first premolar region showed significant difference ($p = 0.004$) with the F-F having more rugae at the first premolar region (41%) as compared to the M-M (33%) and M-F sibling group (39%).

In our study, the M-M had significantly longer first right rugae as compared to the F-F and F-M sibling groups. Studies have reported that males had longer rugae lengths than females due to larger dimensions of the head in males in comparison with females (15–17). Similarly, it has been postulated that males had more primary rugae and lesser number of secondary rugae as compared to females because males have wider palates. Hence, the rugae length could be a predictor to differentiate between gender in identification. F-F siblings had the least number of primary rugae ($n = 425$) but the greatest number of secondary rugae ($n = 107$). This concurs with the observation that there is an indirect relationship between the proportion of primary and secondary rugae of which the development of one type is at the expense of the other (18). Only the secondary rugae showed significant difference among the sibling groups. Correspondingly, the secondary rugae has been shown to have stronger discriminatory ability between different populations than primary rugae (19).

The rugae were predominantly wavy which is commonly documented in most populations such as in Serbians, Turkish, Australians, Iraqi, Bosnians, Indians and Africans (20–26) and Chileans with sinuous patterns (13). In contrast, Egyptians and Indonesians had predominantly line rugae (3,27) alike Iranians and Koreans who have straight rugae patterns (16,19). Although the terminology is different due to the various classifications, wavy and sinuous patterns are similar in shape whereas line and straight patterns are analogous.

The shape of the rugae was not significantly different among the sibling groups. In contrast, Sudanese Nubians (50 males and 50 females) showed positive sexual dimorphism using shape with prediction accuracy of 62–68% (27). Besides, shape has been shown to discriminate between population with prediction accuracy of 70% as reported by Nayak and colleagues. They reported that southern Indians had higher number of straight rugae than Western Indians and those with curved rugae were more likely to be Western Indians (26). Usage of rugae shape is preferred over rugae length to facilitate population identification because it is a discrete variable and the rugae shape is stable throughout life (10). It is believed that the rugae shape is mostly genetically controlled because the genes determine the orientation of collagen fibers within the connective tissue of the rugae, thereby governing the rugae pattern of the diverse racial groups (19).

This study group showed more backward directed rugae compared to forward and perpendicular rugae which is similar to the Bengali population (28). In contrast, forward directed rugae was predominant in Sudanese, Egyptians, Turkish, Gujarati, and Indians (22,27,29–33). Generally, the population worldwide would show more forward directed rugae because there is a decrease in backward directed rugae with age. This is because there is forward movement of the lateral terminal points of the rugae with forward growth of the dental arch (25).

Besides, this study group had more divergent type of rugae than convergent rugae. This finding was similar to the unification pattern observed in Egyptians and Indians but in contrast with Sudanese, Libyans, Serbians, and Bengalis which were predominantly convergent (23,27–29,32,34). The unification pattern was not statistically different between the sibling groups in this study. In contrast, Sudanese and Serbian females showed significantly more convergent rugae than males (23,27). Meanwhile, Iranian and southeast, North, and West Indian males showed significantly more divergent rugae than their counterparts (35–37). Besides, the variation between findings of differrent studies is due to different classification used such as some studies classify unification as additional patterns (29).

In this study, there were statistically significant similarities between sibling pairs accounting for at least 6.25–12.8% of the variability due to heredity. Despite the similarities found, the palatal rugae patterns showed significant differences between siblings of at least 46.9%. These findings are consistent with the findings of a study on 30 subjects that consisted of five families (father, mother, and two siblings) and five pairs of dizygotic twins which showed individualistic rugae patterns among the subjects. However, some similar forms within a family were observed (38). This finding was confirmed by another study of 30 families consisting of parents and their offspring which showed nonidentical rugae configuration but with significant resemblance of the rugae patterns between child and parents indicating the role of heredity (39).

The predictor with the greatest discriminating ability in this study was the difference in average rugae length. However, there was still low accuracy in predicting sibling groups which may be due to individual variations and the complex interplay of genetic and environmental factors in the palatal rugae morphology. Additionally, different population groups may show similar patterns rendering population identification challenging (40). In contrast, discriminatory ability between gender of 130 Iranians using rugae length, shape, and unification yielded prediction accuracy of 70% with discriminant function analysis (35). Therefore, the palatal rugae patterns can still be a useful identification tool due to its diverse configuration with distinguishing features (41).

Heritability is an essential parameter that determines the role of environmental and genetic influence in the palatal rugae patterns. However, it is population-specific with large between-population differences. A strong correlation between phenotype and genotype is considered with a high heritability estimate whereas lower heritability would have a predominant environmental exposure (42). Correlations between full siblings are a possible source of bias with inflated heritability estimates as shown by our results. This is because there would be acquired similarities from “co-habitational effect” such as dietary habits besides having half of the same genes (43). Nonetheless, full siblings of larger difference in age may be raised in different social environments reducing the shared environment effect.

There is an increase in palatal rugae lengths with overall craniofacial growth (16). Additionally, studies have shown that the rugae patterns such as the unification, number of fragmentary rugae, shape, and direction are influenced by orthodontic treatment which is in concordance with zero heritability values in our study (1). However, the changes did not affect the individuality characteristic of the palatal rugae patterns indicating the influence of genetic factors. There are genes within the connective tissue of the rugae that govern the orientation of the collagen fibers into its unique rugae pattern. The total number of secondary rugae showed high heritability ($h^2 > 0.6$) which parallel findings of the discriminatory ability of secondary rugae (19).

This study found that the rugae patterns were environmentally and genetically influenced. This provides the justification for further genetic studies regarding the palatal rugae patterns. Moreover, future studies should focus on the total number of secondary rugae as a predictor for population differentiation because it had high heritability estimates ($h^2 > 0.6$). Additionally, future studies should increase the sample size to account for racial differences and there should be more constraints in the sampling frame such as excluding certain malocclusions (posterior crossbites and anterior open bite) to limit confounding factors. Although measurements on digitized study models have been found to be a reliable and valid method, future studies should obtain digital impressions with intraoral scanners to reduce the cost and time required to digitize the conventional casts (44).

Conclusion

No two individuals exhibited identical rugae patterns which demonstrated the uniqueness of the rugae patterns. Despite the individuality characteristics, an appreciable hereditary component is observed with the significant similarities found between sibling pairs. The heritability estimates also showed a combination of rugae patterns that were environmentally and genetically influenced.

References

1. Mustafa AG, Allouh MZ, Alshehab RM. Morphological changes in palatal rugae patterns following orthodontic treatment. J Forensic Leg Med 2015;31:19–22. https://doi.org/10.1016/j.jflm.2015.01.002
2. Chowdhry A. A simple working type Integrated Rugoscopy Chart proposed for analysis and recording rugae pattern. J Forensic Dent Sci 2016;8(3):171–2. https://doi.org/10.4103/0975-1475.195106
3. Suhartono AW, Syafitri K, Puspita AD, Soedarsono N, Gultom FP, Widodo PT, et al. Palatal rugae patterning in a modern Indonesian population. Int J Legal Med 2016;130(3):881–7. https://doi.org/10.1007/s00414-015-1272-5
4. Taneva E, Evans C, Viana G. 3D evaluation of palatal rugae in identical twins. Case Rep Dent 2017;2017:2648312. https://doi.org/10.1155/2017/2648312
5. Mala S, Rathod V, Pundir S, Dixit S. Pattern self-repetition of fingerprints, lip prints, and palatal rugae among three generations of family: a forensic approach to identify family hierarchy. J Forensic Dent Sci 2017;9(1):15–9. https://doi.org/10.4103/jfo.jfds_115_15
6. Christou P, Kiliaridis S. Vertical growth-related changes in the positions of palatal rugae and maxillary incisors. Am J Orthod Dentofac Orthop 2008;133(1):81–6. https://doi.org/10.1016/j.ajodo.2007.07.009
7. Harris EF. Interpreting heritability estimates in the orthodontic literature. Semin Orthod 2008;14(2):125–34. https://doi.org/10.1053/j.sodo.2008.02.003
8. Johannsdottir B, Thorarinsson F, Thordarson A, Magnusson TE. Heritability of craniofacial characteristics between parents and offspring estimated from lateral cephalograms. Am J Orthod Dentofac Orthop 2005;127(2):200–7. https://doi.org/10.1016/j.ajodo.2004.07.033
9. Cassidy KM, Harris EF, Tolley EA, Keim RG. Genetic influence on dental arch form in orthodontic patients. Angle Orthod 1998;68(5):445–54. https://doi.org/10.1043/0003-3219(1998)068<0445:GIODAF>2.3.CO;2
10. Thomas CJ, Kotze TJ. The palatal ruga pattern: a new classification. J Dent Assoc South Africa 1983;38(3):153–7.
11. Jager J, Putnik DL, Bornstein MH. More than just convenient: the scientific merits of homogeneous convenience samples. Monogr Soc Res Child Dev 2017;82(2):13–30. https://doi.org/10.1111/mono.12296
12. Ochoa B, Nanda R. Comparison of maxillary and mandibular growth. Am J Orthod Dentofac Orthop 2004;125(2):148–59. https://doi.org/10.1016/j.ajodo.2003.03.008
13. Hermosilla V, Valenzuela SP, López M, Galdames S. Palatal rugae: systematic analysis of its shape and dimensions for use in human identification. Int J Morphol 2009;27(3):819–25.
14. Selvamani M, Hosallimath S, Madhushankari, Basandi P, Yamunadevi A. Dimensional and morphological analysis of various rugae patterns in Kerala (South India) sample population: a cross-sectional study. J Nat Sci Biol Med 2015;6(2):306–9. https://doi.org/10.4103/0976-9668.159985
15. Gautam N, Patil SG, Krishna RG, Agastya H, Mushtaq L, Kumar KV. Association of palatal rugae pattern in gender identification: an exploratory study. J Contemp Dent Pract 2017;18(6):470–3. https://doi.org/10.5005/jp-journals-10024-2067
16. Kim N, Im Y, Kim J, Kim B. Palatal rugae pattern in Korean children and adolescents. J Oral Med Pain 2019;44(4):169–73.
17. Saadeh M, Ghafari JG, Haddad RV, Ayoub F. Sex prediction from morphometric palatal rugae measures. J Forensic Odontostomatol 2017;35(1):9–20.
18. Dawasaz AA, Dinkar AD. Rugoscopy: predominant pattern, uniqueness, and stability assessment in the Indian Goan population. J Forensic Sci 2013;58(6):1621–7. https://doi.org/10.1111/1556-4029.12190

19. Sheikhi M, Zandi M, Ghazizadeh M. Assessment of palatal rugae pattern for sex and ethnicity identification in an Iranian population. Dent Res J 2018;15(1):50–6. https://doi.org/10.4103/1735-3327.223611
20. Sandeep G, Sonia G. Study of palatal rugae pattern of Rwandan patients attending the dental department at King Faisal Hospital, Kigali, Rwanda: a preliminary study. Rwanda Med J 2013;70(1):19–25.
21. Mohamed TJ. A comparison of rugae pattern in males and females as a samples of Iraqi population. Tikrit J Dent Sci 2016;4:1–5.
22. Gezer R, Deniz M, Uslu AI. Morphological characteristics and individual differences of palatal rugae. J Craniofac Surg 2019;30(6):1906–10. https://doi.org/10.1097/SCS.0000000000005599
23. Filipovic G, Janosevic M, Janosevic P, Radojicic J, Ajdukovic Z, Janjic OT. Palatal rugae patterns in the Serbian population. Arch Biol Sci 2014;66(3):1131–4. https://doi.org/10.2298/ABS1403131F
24. Muhasilovic S, Hadziabdic N, Galic I, Vodanovic M. Analysis of palatal rugae in males and females of average age of 35 in the Bosnia and Herzegovina population sample (Sarajevo Canton). J Forensic Leg Med 2016;39:147–50. https://doi.org/10.1016/j.jflm.2016.01.029
25. Kapali S, Townsend G, Richards L, Parish T. Palatal rugae patterns in Australian aborigines and Caucasians. Aust Dent J 1997;42(2):129–33. https://doi.org/10.1111/j.1834-7819.1997.tb00110.x
26. Nayak P, Acharya AB, Padmini AT, Kaveri K. Differences in the palatal rugae shape in two populations of India. Arch Oral Biol 2007;52 (10):977–82. https://doi.org/10.1016/j.archoralbio.2007.04.006
27. Ahmed AA, Hamid A. Morphological study of rugae palatinae in Sudanese Nubians. Folia Morphol 2015;74(3):303–10. https://doi.org/10.5603/FM.2015.0046
28. Pramanik A, Madhumita D, Moulik D. A comparative study of gender difference in palatal rugae patterns among Bengali subjects in Murshidabad. Int J Anatomy Radiol Surg 2019;8(1):6–10. https://doi.org/10.7860/IJARS/2019/36281:2449
29. Sherif AF, Hashim AA, Ashmawy MH, Soliman M. A pilot cross sectional study of palatal rugae shape and direction among Egyptians and Malaysians. Egypt J Forensic Sci 2018;8:17. https://doi.org/10.1186/s41935-018-0050-1
30. Gondivkar SM, Patel S, Amol RG, Gaikwad RN, Chole R, Parikh RV. Morphological study of the palatal rugae in Western Indian population. J Forensic Leg Med 2011;18(7):310–2. https://doi.org/10.1016/j.jflm.2011.06.007
31. Jibi PM, Gautam KK, Basappa N, Raju OS. Morphological pattern of palatal rugae in children of Davangere. J Forensic Sci 2011;56(5):1192–7. https://doi.org/10.1111/j.1556-4029.2011.01831.x
32. Surekha R, Anila K, Vikram SR, Hunasgi S, Ravikumar S, Ramesh N. Assessment of palatal rugae patterns in Manipuri and Kerala population. J Forensic Dent Sci 2012;4(2):93–6. https://doi.org/10.4103/0975-1475.109896
33. Dwivedi N, Nagarajappa AK. Morphological analysis of palatal rugae pattern in central Indian population. J Int Soc Prev Community Dent 2016;6(5):417–22. https://doi.org/10.4103/2231-0762.192947
34. Mahesh S, Salim A, Jothikumar K, Aruna DS, Akansha K. Cross-sectional survey on palatal rugae patterns in Libyan population – a rugoscopic study. Baba Farid Univ Dent J 2016;6:9–14.
35. Malekzadeh AR, Pakshir R, Ajami S, Pakshir F. The application of palatal rugae for sex discrimination in forensic medicine in a selected Iranian population. Iran J Med Sci 2018;43(6):612–22.
36. Harchandani N, Marathe S, Rochani R, Nisa S. Palatal rugoscopy: a new era for forensic identification. J Indian Acad Oral Med Radiol 2015;27(3):393–8.
37. Thabitha RS, Reddy RE, Manjula M, Sreelakshimi N, Rajesh A, Vinay LK. Evaluation of palatal rugae pattern in establishing identification and sex determination in Nalgonda children. J Forensic Dent Sci 2015;7 (3):232–7. https://doi.org/10.4103/0975-1475.172447
38. Mohamadzowharsajid, Prachi S, Kumar S, Kaminikiran, Agrawal S, Rameshwar S, et al. Study of palatal rugae pattern for establishing individuality. J Dent Med Sci 2016;15(8):37–43.
39. Patel RN, Umesh K, Patel R, Patel H, Patel N, Patel CK, et al. Assessing the inheritance of palatal rugae patterns. Int J Adv Res 2015;3 (6):297–301.
40. Pillai J, Banker A, Bhattacharya A, Gandhi R, Patel N, Parikh S. Quantitative and qualitative analysis of palatal rugae patterns in Gujarati population: a retrospective, cross-sectional study. J Forensic Dent Sci 2016;8(3):126–34. https://doi.org/10.4103/0975-1475.195110
41. Kolude B, Akinyele A, Joshua OT, Ahmed L. Ethnic and gender comparison of rugae patterns among clinical dental trainees in Ibadan. Nigeria. Pan Afr Med J 2016;23:204. https://doi.org/10.11604/pamj.2016.23.204.8584
42. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era – concepts and misconceptions. Nat Rev Genet 2008;9:255–66. https://doi.org/10.1038/nrg2322
43. Harris EF, Johnson MG. Heritability of craniometric and occlusal variables: a longitudinal sib analysis. Am J Orthod Dentofac Orthop 1991;99(3):258–68. https://doi.org/10.1016/0889-5406(91)70007-J
44. Reuschl RP, Heuer W, Stiesch M, Wenzel D, Dittmer MP. Reliability and validity of measurements on digital study models and plaster models. Eur J Orthod 2016;38(1):22–6. https://doi.org/10.1093/ejo/cjv001

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14524
Available online at: onlinelibrary.wiley.com

PAPER

PATHOLOGY/BIOLOGY

Joyce L. deJong,[1] D.O.; Jenelle Lee,[1] B.S.; Abigail Grande,[1] M.P.H.; Cuyler Huffman,[1] M.S.; Chloe Bielby,[1] M.P.H.; and Theodore Brown,[1] M.D.

Positional Asphyxia in Opioid-Related Deaths: Is It Being Overlooked?

ABSTRACT: The contribution of positional asphyxia in opioid-related deaths is currently unknown. Diagnostic criteria for positional asphyxia include finding the decedent in a position that does not allow for adequate respiration and an inability to extricate themselves from the position due to various conditions. Our primary objective was to assess whether positional asphyxia and the resulting airway compromise were a contributing factor to death due to the toxic effects of opioids. We evaluated 225 deaths where the death scene investigation contained adequate information to evaluate for positional asphyxia and performed a Pearson chi-square test to determine if the proportion of deaths found in an airway compromising position was higher when opioid(s) caused the death. The proportion of decedents found in a potential airway compromising position was greater when the death was related to opioid use (p < 0.0001). Further, narrowing the dataset to decedents who were definitely in an airway compromising position [Yes (24.49%) vs. No (11.02%)] showed a statistically significant association between positional asphyxia and deaths related to opioid use (p = 0.0021). Carefully documenting the position in which the decedent was initially found may be a significant factor in accurate reporting and in harm reduction efforts to decrease the opioid mortality rate.

KEYWORDS: forensic pathology, opioids, positional asphyxia, death scene investigation, drug-related fatalities, harm reduction, autopsy

Opioid use is a leading cause of death in the United States, resulting in more than 47,000 deaths per year (1). For medical examiners, investigation of deaths possibly related to opioids and the determination of whether opioid(s) contributed to a death require careful consideration of many factors as guided by a recent position paper by the American College of Medical Toxicology and the National Association of Medical Examiners. The investigation of a possible death related to opioids or other drugs requires a scene investigation, full medical history, complete autopsy, comprehensive toxicology testing, and interpretation of all information gathered from these various sources by a board-certified forensic pathologist (2). The scene investigation of any unexpected death routinely includes noting the presence of illicit drugs and drug paraphernalia at the scene, counting prescription drugs and comparing the remaining number to what should remain based on the prescription instructions and fill dates, obtaining the medical and social history, and searching for any other scene conditions which may have caused or contributed to the death, unrelated to drug use.

General observations regarding the location and position of the body are routine in any death investigation. Some types of death investigations, such as sudden and unexpected deaths of infants, have a much higher focus on describing the placed position and the found position of the deceased infant, with careful attention to any potential obstruction of the airways. When the death is of an adult, unless the decedent appears wedged or is in an especially awkward position with potential concerns about whether asphyxia caused or contributed to the death, thorough descriptions of whether the airways were covered or compressed are not routinely documented in the narrative report from medicolegal death investigators at the death scene.

In this project, we reviewed death scene information of adults to understand if the position in which the decedent was found could cause or contribute to death through asphyxia. Criteria for the diagnosis of fatal positional asphyxia include finding the decedent in a position that compromises breathing; the investigation indicates the position was inadvertent; there was a reason the individual would not be able to extricate from the fatal position; and various other conditions that may have caused death are ruled out (3). Conditions such as acute alcohol intoxication and dementia may predispose individuals to assume a position that inadvertently results in partial or complete airway obstruction and prevents the individual from self-extrication from a fatal position (4). Positional asphyxia may also result from any entity which impairs consciousness (head injury or substances), restraints (whether by clothing or other items), or decreased mobility (immaturity or disease).

The mechanism of death for individuals whose death was opioid-related is largely respiratory depression (5). The impact of opioids is not limited to respiratory depression, however, as there is also a generalized decrease in mental responsiveness and a clear potential that individuals may not be able to extricate themselves from a position that results in obstructed airways. While much research has been conducted to investigate the mechanism of death for opioid-related deaths, currently it is unclear if positional asphyxia is a significant contributing factor to many opioid-related fatalities.

[1] WMU Homer Stryker M.D. School of Medicine, 300 Portage Street, Kalamazoo, MI, 49007.

Corresponding author: Joyce L. deJong, D.O. E-mail: joyce.dejong@med.wmich.edu

Received 21 April 2020; and in revised form 9 July 2020; accepted 9 July 2020.

Materials and Methods

Our medical examiner's office serves multiple counties with a total population of slightly more than 1.2 million based on 2017 census estimates. All deaths reported to the office are examined by board-certified forensic pathologists who determine the cause and manner of death. For this study, our medical examiner database was searched for deaths occurring at the decedent's residence, another's residence, or at a hotel/motel from April 2016 to December 2019. The focus was to identify individuals who were found dead, and the manner of death was classified as natural, accident, or undetermined; suicides and homicides were excluded. The identified death records were reviewed to determine whether the narrative provided by the medicolegal death investigator in combination with scene photographs contained adequate information to determine whether the position in which the decedent was initially found may have resulted in airway obstruction. Of the 361 deaths reviewed, 136 were excluded as they did not have adequate scene photographs or information to determine the position in which the body was originally found.

A board-certified forensic pathologist evaluated the narrative description and scene photographs of the deaths with adequate information to make a determination as to whether the position in which the decedent was found would cause airway compromise, categorized as "Yes," "Maybe," or "No." Figures 1 and 2 provide examples of deaths in which the position likely contributed to the death and were classified as "Yes." Figure 3 provides an example in which the position may have contributed to the death and was classified as "Maybe."

Evaluators also determined whether the facial features around the mouth and nose appeared to be compressed or flattened, using an evaluation of livor mortis and soft-tissue compression with the option to answer either "Yes" or "No."

Pearson chi-square tests were used to assess each individual objective. Statistical Analysis Software (SAS) 9.4 was used to perform all analyses. Significance was determined at an $\alpha = 0.05$ level.

Results

Of 361 records reviewed, we identified 225 deaths with adequate information contained within the investigative report, along with scene photographs, to determine whether the position may have contributed to the death: 115 of the deaths were classified as natural and 110 of the deaths were classified as accident or undetermined manners.

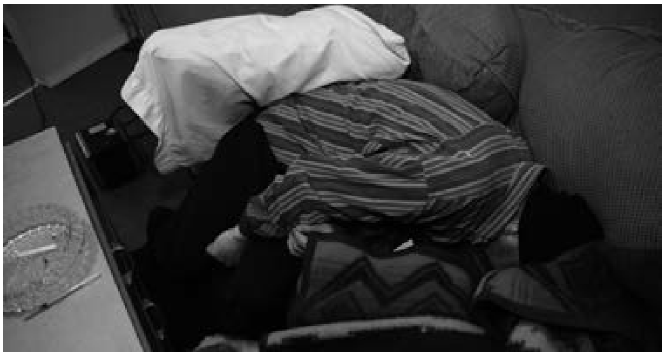


FIG. 1—*A 54-year-old man was found seated on his couch and slumped over to his left with his face directly on the seat cushion and blanket. His oral and nasal airways were obstructed and the forensic pathologist reviewing this case for this study believed the position would have contributed to his death. His accidental death was due to the toxic effects of fentanyl and morphine. [Color figure can be viewed at wileyonlinelibrary.com]*

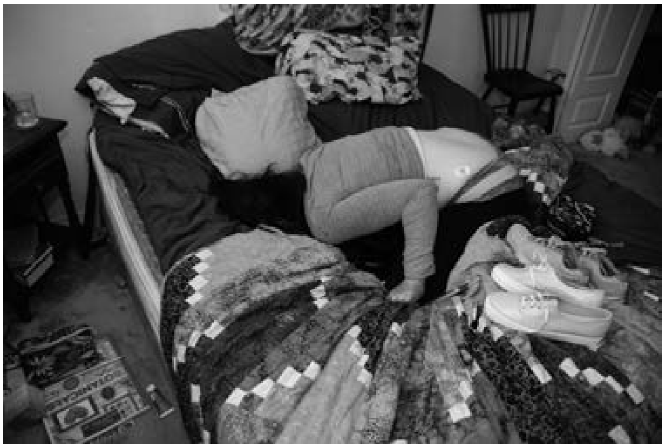


FIG. 2—*A 28-year-old woman was found seated on her bed in a frog-leg position and slumped forward with her face directly on the mattress and bedding. Her oral and nasal airways were obstructed, and the forensic pathologist reviewing this case for this study believed the position would have contributed to her death. Her accidental death was due to the toxic effects of fentanyl and acetyl fentanyl. [Color figure can be viewed at wileyonlinelibrary.com]*

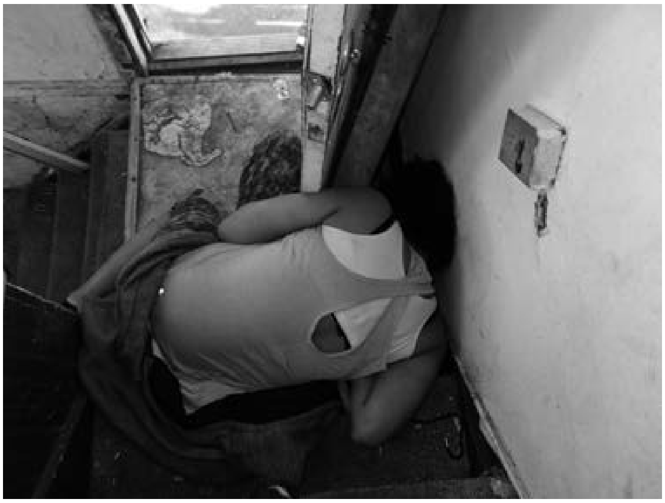


FIG. 3—*A 29-year-old woman was found seated in a stairwell, slumped forward with her neck hyperflexed. Her oral and nasal airways were not directly obstructed, and the forensic pathologist reviewing this case for this study believed the position may have contributed to her death. Her accidental death was due to the toxic effects of fentanyl, morphine, and acetyl fentanyl. [Color figure can be viewed at wileyonlinelibrary.com]*

Of the 110 deaths classified as accident or undetermined, seven were not related to drugs. Of the seven accidental and undetermined deaths not related to drugs, one death was caused by positional asphyxia associated with dementia, and all others were classified as accident due to nondrug-related conditions, such as injuries secondary to a fall or hypothermia. Complete autopsies and comprehensive toxicology were performed in all of the deaths believed to be drug-related. In 98 of the 103 deaths considered to be drug-related, one or more opioids were a contributing factor. Two of the five deaths which were nonopioid drug-related were caused by cocaine, one by methamphetamine, one by duloxetine, and one by the combined toxic effects of amlodipine, lurasidone, ethanol, and fluoxetine (Table 1).

Of the opioid-related deaths, 24 (24.49%) were determined to be found in an airway compromising position, and 12 (12.24%) were found in a position that was potentially airway compromising (Table 2). Of the deaths not related to opioid use, there were 14 (11.02%) determined to be in an airway compromising position, and three (2.36%) were found in a position that was potentially airway compromising (Table 3).

Of decedents found in an airway compromising position or a potential airway compromising position, 36 (36.73%) in opioid-related deaths versus 17 (13.39%) for nonopioid-related deaths, there was sufficient evidence to conclude that the proportion of bodies found in either an airway compromising position or a potentially airway compromising position was higher when the death was due to opioid-related drug use (p-value of <0.0001; Fig. 4). Further analysis of the decedents found in a definite airway compromising position showed similar results: 24 (24.49%) for the opioid-related deaths and 14 (11.02%) for the nonopioid-related deaths (Table 4; Fig. 4). The p-value of 0.0021 is evidence that chance alone is unlikely to account for the disproportionately high number of decedents that were positive for opioids who were found in an airway compromising position.

As expected, a significantly higher proportion of decedents were found in a position in which the facial features appeared to be compressed or flattened, for the opioid-related deaths (n = 27, 28.13%) than the for the nonopioid-related deaths (n = 16, 12.60%), (p-value of 0.0057).

Discussion

Overall, our data suggest that positional asphyxia may have been a contributing factor to as many as 37% of the 98 deaths in which opioids caused or contributed to death compared to 13% of the 127 nonopioid-related deaths. This suggests that the respiratory depression and decreased mental responsiveness caused by opioid use frequently result in decedents succumbing to positional asphyxia which in turn likely contributes to their ultimate death.

For many opioid-related deaths, the role of positional asphyxia was not initially considered based upon the original evidence and investigation; however, in our retrospective analysis of data it is likely that positional asphyxia contributed to some of these deaths. An example of a death where positional asphyxia may have been a contributing factor includes decedents found prone in bed with their face pressed into a pillow. In these cases, the investigation, autopsy, and toxicology findings provided information typical of opioid-related deaths and there was not a search for other reasons why the individual may have died. Similarly, in individuals with lethal natural disease who were found in an airway compromising position, positional asphyxia was not typically included as a cause or contributing factor to the death. In positional asphyxia deaths related to acute alcohol intoxication, the position in which the decedent is found is typically quite noteworthy. Furthermore, the alcohol level alone may be inadequate to explain the death but would explain why the individual did not extricate from the airway compromising position. Thus for opioid-related deaths, positional asphyxia is a contributing factor that is often overlooked.

TABLE 1—*Total deaths by manner of death.*

Manner of death	n
Natural	115
Accident or undetermined	110
Opioid drug-related	98
Nonopioid drug-related	5
Nondrug-related	7
Total	225

TABLE 2—*Drugs causing death in opioid-related deaths in which the forensic pathologist answered "yes" or "maybe" to whether the position could cause airway compromise.*

Age	Drugs Causing Death in Opioid-Related Fatalities With Possible Positional Asphyxia	Could Position Cause Airway Compromise?
24	Heroin, fentanyl, methadone, alprazolam	Maybe
25	Heroin, fentanyl, acetyl fentanyl	Yes
26	Fentanyl, acetyl fentanyl, heroin	Yes
26	Fentanyl, acetyl fentanyl, heroin	Yes
26	Morphine, clonazepam, fentanyl	Yes
27	Fentanyl	Yes
28	Fentanyl, acetyl fentanyl	Yes
28	Clonazepam, fentanyl, methadone, morphine, olanzapine	Yes
28	Morphine, methoxyacetyl fentanyl, fentanyl	Yes
29	Fentanyl, acetyl fentanyl, heroin, cocaine, alcohol	Maybe
29	Diazepam, demoxepam, fentanyl, heroin	Maybe
30	Alprazolam, methadone	Yes
30	Loperamide	Maybe
32	Fentanyl, acetyl fentanyl, loperamide	Maybe
34	Fentanyl, methoxyacetyl fentanyl, U47700	Maybe
35	Fentanyl, acetyl fentanyl, heroin, alprazolam, hydrocodone, diphenhydramine	Maybe
35	Methamphetamine, fentanyl, acetyl fentanyl, heroin, hydromorphone	Yes
35	Cocaine, diphenhydramine, fentanyl, acetyl fentanyl	Yes
36	Heroin, clonazepam, methamphetamine	Yes
36	Methamphetamine, fentanyl	Yes
36	Methamphetamine, fentanyl	Yes
37	Heroin, cocaine	Yes
39	Methadone	Maybe
39	Fentanyl, alprazolam, buprenorphine, amphetamine	Yes
40	Methoxyacetyl fentanyl	Maybe
40	Methamphetamine, fentanyl	Yes
43	Morphine, diphenhydramine, fentanyl, acetyl fentanyl	Maybe
45	Alprazolam, fentanyl	Maybe
46	Fentanyl, acetyl fentanyl	Yes
48	Fentanyl, oxycodone	Yes
54	Fentanyl, cyclopropyl fentanyl, methamphetamine, amphetamine	Yes
55	Fentanyl, acetyl fentanyl, methamphetamine	Yes
57	Fentanyl, acetyl fentanyl, cocaine	Yes
62	Diazepam, methadone, diphenhydramine	Yes
63	Morphine, fentanyl, acetyl fentanyl	Yes
69	Morphine, cyclobenzaprine, fentanyl, zolpidem	Maybe

In evaluating death records, a very high number of cases needed to be excluded because the medicolegal death investigator did not provide adequate information regarding the original position in which the decedent was found unresponsive or dead. Investigators seldom asked this information of the individual who found the decedent. In some cases, it was unclear whether the position in which the body was found reflected the position

TABLE 3—*Cause and manner of death in nonopioid-related deaths in which the forensic pathologist answered "yes" or "maybe" to whether the position could cause airway compromise.*

Manner of Death	Immediate Cause of Death	Other Conditions Contributing to Death	Could Position Cause Airway Compromise?
Natural	Severe aortic valve stenosis	Chronic obstructive pulmonary disease	Yes
Natural	Atherosclerotic cardiovascular disease		Yes
Natural	Chronic obstructive pulmonary disease		Yes
Natural	Bilateral pulmonary thromboemboli	Obesity	Yes
Natural	Lung adenocarcinoma	Chronic small vessel ischemic disease of the brain	Yes
Natural	Hypertensive and atherosclerotic cardiovascular disease		Yes
Natural	Cerebral infarct	Hypertension	Yes
Natural	Hypertensive and atherosclerotic cardiovascular disease	Probable metastatic lung cancer	Yes
Natural	Massive aneurysm of the left main coronary artery with thrombosis		Yes
Natural	Alcohol use disorder		Yes
Natural	Diabetes mellitus, Type 2	Hypertension, mixed hyperlipidemia	Yes
Accident	Blunt force injuries of the torso due to fall	Chronic obstructive pulmonary disease, chronic ethanol use, hypertensive and atherosclerotic cardiovascular disease	Yes
Accident	Positional asphyxia	Dementia	Yes
Accident	Toxic effects of methamphetamine	Chronic obstructive pulmonary disease, atherosclerotic cardiovascular disease, diabetes mellitus	Yes
Natural	Hypertensive cardiovascular disease		Maybe
Natural	Coronary artery disease		Maybe
Natural	Hypertensive and atherosclerotic cardiovascular disease	Pulmonary emphysema	Maybe

the body was initially found or if the position being described was the position of the decedent upon arrival of the investigator. Sometimes, photographs aided in clarifying these questions. Not having this information documented in the investigator's narrative report is not unexpected as this has not been a routine process followed by investigators.

To adequately assess whether positional asphyxia contributed to a death, photographs with documentation of the position of the airways should be obtained before the body is moved, when possible. Individuals and first responders who find someone unresponsive or dead often move the body to perform an assessment and initiate resuscitation, if appropriate. This occurs before arrival of the medicolegal death investigator and obtaining photographs in those situations is not possible. In all cases, the medicolegal death investigator should identify the individual who found the decedent and obtain detailed information of the

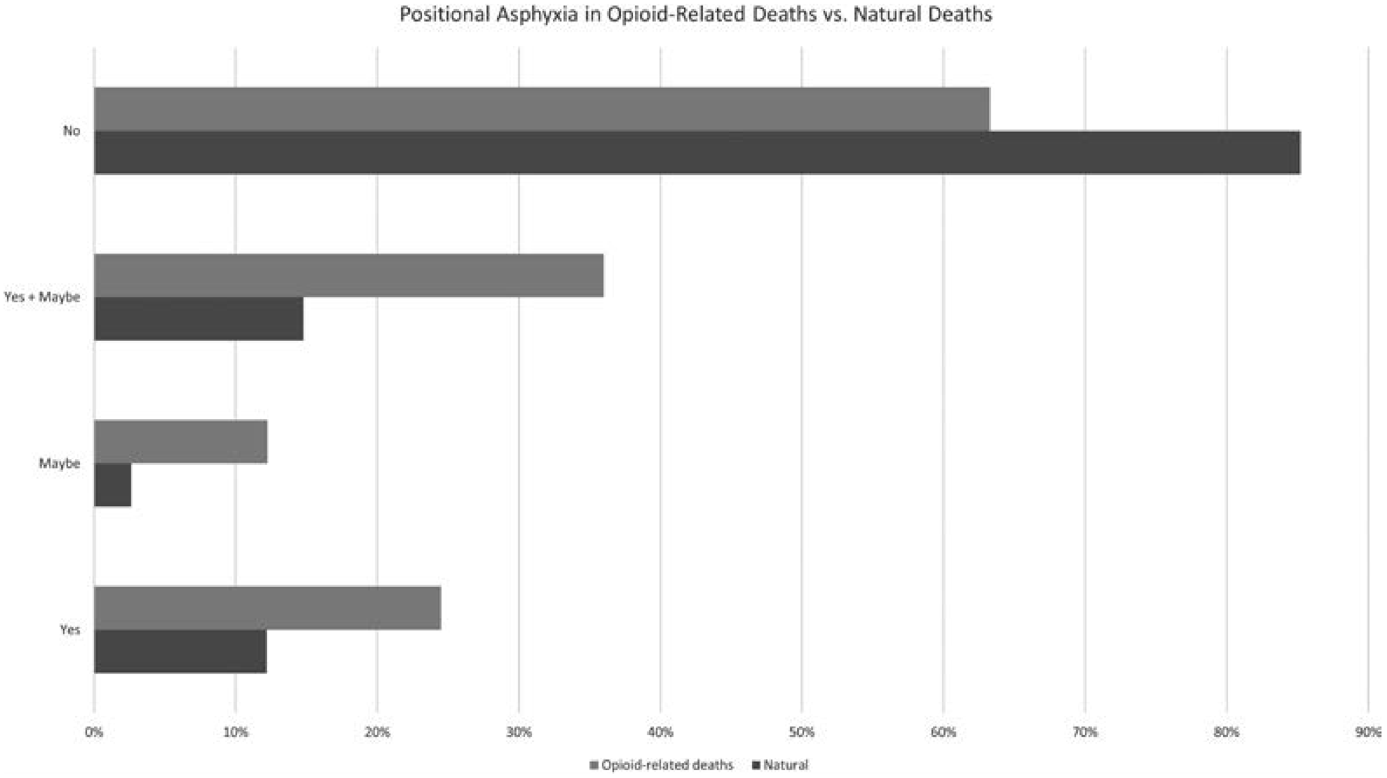


FIG. 4—*This bar graph demonstrates the percentage of cases in which the position would have likely contributed to the death, could have contributed, or did not contribute for 115 natural deaths and 98 opioid-related deaths. [Color figure can be viewed at wileyonlinelibrary.com]*

position of the body when found with special attention to the position found, including whether the nasal and oral airways were obstructed or whether the neck may have been in a position to compromise breathing. For these reasons, we propose that documentation of the decedent's position is an important component of the death scene investigation and has recommended some key information that should be gathered in Table 5.

The potential that positional asphyxia is a significant contributing factor to many opioid-related deaths may have public health implications. Various harm reduction strategies for people who use drugs have been proposed, and there is evidence to support the implementation of harm reduction behaviors (6). Examples of harm reduction behaviors include instructions to use test doses of opioids to evaluate the strength of the drug and not to use alone so that another person is available to administer naloxone or provide other lifesaving interventions if needed. People still do use alone, and similar to the recognition of sleep position being a significant contributing factor for infant deaths, a change in the position during opioid use as a harm reduction behavior for people who use drugs may result in a decrease in mortality (7). Additional research is needed on this topic as well as an evaluation of implementation procedures for any proposed harm reduction behavior.

TABLE 4—*Numbers and percentages of all deaths evaluated with regard to whether the position could have contributed to the death and numbers and percentages of facial compression.*

	Opioid Cause of Death		Nonopioid Cause of Death		
	n	%	*n*	%	*p*-Value
Decedent found in airway compromising position	98	100	127	100	
Yes*	24	24.49	14	11.02	0.0021
Maybe	12	12.24	3	2.36	
Total†	36	36.73	17	13.38	<0.0001
Decedent had compressed or flattened facial features	98	100	127	100	
Yes*	27	27.55	16	12.60	0.0057
No	71	72.45	111	87.40	

Sample size = 225.
*Significant at the α = 0.010 level.
†Significant at the α = 0.001 level.

TABLE 5—*Recommendations of information and images to obtain in all death investigations.*

Recommendations for Investigations
Photograph decedent from multiple angles with attention to airways and position before the decedent is moved
Identify the individual who found the decedent unresponsive or dead and gather detailed information regarding the position of the head and neck when found
Interview medical first responders when applicable to gather additional detailed information about potential airway compromise
Document livor mortis and compression of the nose and mouth with photographs and in the narrative report

In summary, our findings are strongly suggestive that positional asphyxia contributes to greater than one-third of all opioid-related deaths. However, due to the limited research available, outside of this study, further confirmational research is required before recommending that positional asphyxia be reported as a contributing factor in these deaths. However, this study aims to raise awareness of the potential role of positional asphyxia in opioid-related deaths to improve the accuracy of the mechanisms and causes of death, but more importantly to reduce the number of deaths by helping to inform behavioral interventions. Medicolegal death investigators are encouraged to include additional information about the decedent's original position and condition to better clarify the role of positional asphyxia in opioid-related deaths.

Acknowledgments

The authors thank Laura Bauler, Ph.D., for editing the manuscript for improved language and readability.

References

1. Scholl L, Seth P, Kariisa M, Wilson N, Baldwin G. Drug and opioid-involved overdose deaths—United States, 2013–2017. Morb Mortal Wkly Rep 2018;67(5152):1419–27. https://doi.org/10.15585/mmwr.mm675152e1
2. Davis GG, National Association of Medical Examiners and American College of Medical Toxicology Expert Panel on Evaluating and Reporting Opiod Deaths. Complete republication: National Association of Medical Examiners position paper: recommendations for the investigation, diagnosis, and certification of deaths related to opioid drugs. J Med Toxicol 2014;10(1):100–6. https://doi.org/10.1007/s13181-013-0323-x
3. Bell MD, Rao VJ, Wetli CV, Rodriguez RN. Positional asphyxiation in adults. A series of 30 cases from the Dade and Broward County Florida Medical Examiner Offices from 1982 to 1990. Am J Forensic Med Pathol 1992;13(2):101–7.
4. Byard RW, Wick R, Gilbert JD. Conditions and circumstances predisposing to death from positional asphyxia in adults. J Forensic Leg Med 2008;15(7):415–9. https://doi.org/10.1016/j.jflm.2008.01.001
5. Boom M, Niesters M, Sarton E, Aarts L, Smith TW, Dahan A. Non-analgesic effects of opioids: opioid-induced respiratory depression. Curr Pharm Des 2012;18(37):5994–6004. https://doi.org/10.2174/138161212803582469
6. Rouhani S, Park JN, Morales KB, Green TC, Sherman SG. Harm reduction measures employed by people using opioids with suspected fentanyl exposure in Boston, Baltimore, and Providence. Harm Reduct J 2019;16(1):39. https://doi.org/10.1186/s12954-019-0311-9
7. Ponsonby AL, Dwyer T, Gibbons LE, Cochrane JA, Wang Y-G. Factors potentiating the risk of sudden infant death syndrome associated with the prone position. N Engl J Med 1993;329(6):377–82. https://doi.org/10.1056/NEJM199308053290601

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14518
Available online at: onlinelibrary.wiley.com

PAPER

PATHOLOGY/BIOLOGY

Gregory M. Dickinson,[1] M.D.; Gene X. Maya,[2] M.D.; Yungtai Lo,[3] Ph.D.; and Hannah C. Jarvis,[4] M.B.B.S., A.I.C.S.M., B.Sc. (Hons) M.R.C.S.

Hypothermia-related Deaths: A 10-year Retrospective Study of Two Major Metropolitan Cities in the United States*

ABSTRACT: Hypothermia-related deaths affect vulnerable populations and are preventable. They account for the vast majority of weather-related deaths in the United States. The postmortem diagnosis of hypothermia can be challenging, as there are no pathognomonic signs. The electronic databases of the New York City Office of Chief Medical Examiner and Harris County Institute of Forensic Sciences were searched for all fatalities where the primary cause of death included hypothermia, between January 2009 and July 2019. There were 139 hypothermia deaths in New York City (NYC) with an average annualized rate of 1.7 per million. During this same time, there were 50 hypothermia deaths in Houston with an average annualized rate of 2.4 per million. Males were more likely to die of hypothermia compared to females in both cities. The rate ratio (RR) in NYC was 3.55 (95% CI 2.40, 5.25), while the RR in Houston was 2.83 (95% CI 1.50, 5.32). Age- and sex-specific standardized hypothermia mortality rates were 18.2 (95% CI 15.1, 21.2) per million in NYC and 30.1 (95% CI 21.7, 38.6) per million in Houston. The comparative hypothermia death ratio was 1.66 (95% CI 1.19, 2.30), indicating hypothermia mortality in Houston was 66% higher than in NYC. There was no correlation between zip code poverty rates and hypothermia-related deaths. The most consistent autopsy finding was Wischnewski spots (56.6%), and ethanol was the most common toxicological finding (36.5%). Local agencies can use this data to target these higher-risk populations and offer appropriate interventions to try to prevent these deaths.

KEYWORDS: hypothermia, weather-related fatalities, environmental cold exposure, mortality, autopsy, forensic pathology

There are approximately 1300 deaths due to hypothermia in the United States every year. Most of these deaths affect vulnerable populations, including the homeless, elderly, infants, and people with substance abuse and mental illness. The majority of hypothermia-related deaths occur in the Midwest or West; however, Southern states may exhibit rapid temperature drops at night, in stark contrast to the daytime heat, leaving people unprepared (1–3). These deaths are the principal cause of weather-related mortality, representing twice the number of heat-related deaths. The manner of death in the majority of hypothermia deaths is accident (1).

Autonomic thermoregulatory mechanisms exist to maintain a controlled core body temperature, allowing optimal enzyme function, and survival in a wide range of environmental temperatures. Intrinsic heat production is generated at a cellular level by metabolism, which can be adjusted to meet demands. The body can also raise its internal temperature by involuntary motor responses in skeletal muscle (shivering). Hypothermia, defined as a core body temperature of less than 95°F (35°C) (4,5), occurs when the body is unable to generate enough heat to overcome heat loss. There are four mechanisms of heat loss: evaporation, convection, conduction, and radiation. Radiation accounts for up to 60% of heat loss from the body. Heat loss can be accelerated by water, where the conductive transfer of heat is 100 times that of the air.

The environmental temperature at which hypothermia can occur may be as high as 75°F (23.9°C) in susceptible populations, such as the elderly whose autonomic regulatory response is less effective, or in substance or alcohol abuse where physiologic responses are depressed (5). Morbidity and mortality related to cold exposure can be prevented with targeted interventions and public health campaigns, including insulation of buildings, access to homeless shelters, wearing appropriate clothing, and anticipation of cold weather.

The postmortem diagnosis of hypothermia can be challenging, as there are no pathognomonic signs, and the diagnosis may rely on circumstantial evidence (6). Wischnewski spots, nonulcerative hemorrhagic lesions of the gastric mucosa, were first described in 1895, and despite much research, they remain the most consistent autopsy finding (7). Other findings at autopsy include hemorrhagic pancreatitis, intramuscular hemorrhage involving

[1] Pathology Department, Montefiore Medical Center, 111 E 210th Street, Bronx, NY, 10467-2490.

[2] Office of the Chief Medical Examiner - Northern District, 10850 Pyramid Place, Suite 121, Manassas, VA, 20110.

[3] Albert Enstein College of Medicine, Jack and Pearl Resnick Campus, 1300 Morris Park Avenue, Block, Room 311, Bronx, NY, 10461.

[4] Harris County Institute of Forensic Sciences, 1861 Old Spanish Trail, Houston, TX, 77030.

Corresponding author: Gregory M. Dickinson, M.D. E-mail: gdickinsonmd@gmail.com

*Presented at the 72nd Annual Scientific Meeting of the American Academy of Forensic Sciences, February 17-22, 2020, in Anaheim, CA.

[Correction added 13 August 2020. The Highlights section was removed.]

Received 20 Jan. 2020; and in revised form 10 May 2020; 24 June 2020; accepted 29 June 2020.

large muscles (such as iliopsoas), pulmonary edema, and pink lividity, particularly over the extensor surfaces.

This study examined the medical examiner case files on all deaths due to hypothermia over a 10-year period in two large populous cities in distinct geographic regions of the United States (New York City, New York and Houston, Texas) and describes the similarities and differences in scene investigation, autopsy findings, toxicological results, and the epidemiological patterns of these fatalities. The information can be used to help forensic pathologists evaluate hypothermic deaths, and most importantly, for public health campaigns to prevent these deaths.

Materials and Methods

The electronic databases of the New York City Office of Chief Medical Examiner and Harris County Institute of Forensic Sciences were screened for all fatalities where the cause of death included hypothermia between January 1, 2009, and July 31, 2019. Cases in which hypothermia was listed on the death certificate and identified as a primary cause of death or a contributing factor to death were included in the study. All cases were anonymized and de-identified prior to analysis. All documents in the medical examiner case files for all 189 fatalities were reviewed, and the following information was manually extracted: age, gender, ethnicity, admission date (if hospitalized), place of injury, place of death, cause of death, manner of death, toxicology results, autopsy findings, body temperature, and pertinent scene investigation findings. Hypothermia deaths in New York City (NYC) and Houston were stratified by age, gender, ethnicity, and borough (for NYC only). Incidence rate ratios and crude hypothermia mortality rates were calculated using 2010 U.S. Census data for NYC and Houston populations. The standard error of the logarithm of rate ratio was estimated according to Altman (8); confidence intervals were calculated using a normal distribution. Differences in characteristics between NYC and Houston residents were compared using chi-square or Fisher's exact tests. Age- and sex-specific standardized mortality rates for NYC and Houston were calculated by using the direct standardization method with the U.S. national population from 2010 Census as the reference population. The standard error of the logarithm of directly standardized mortality rate was estimated by using the method of statistical differentials (9). Confidence intervals for standardized mortality rates were calculated based on normal distributions. All statistical analyses were performed using SAS version 9.4 (SAS Institute Inc, Cary, NC). Population demographic and poverty data were gathered from the U.S. Census Bureau (https://factfinder.census.gov/faces/nav/jsf/pages/community_facts.xhtml). This included demographic data for both cities from the 2010 U.S. Census and poverty data from 2013 to 2017 American Community Survey 5-year Estimates. We obtained weather data from the National Climatic Data Center and chose the NY City Central Park weather station for New York City data and the Houston William P Hobby Airport weather station for Houston data. Both sites were chosen for their complete data sets and central locations.

Results

Between January 1, 2009, and July 31, 2019, there were 189 hypothermia deaths in NYC and Houston combined. There were 143 males and 46 females, and the average age was 60.1 years in NYC and 64 years in Houston; ranging from a neonate to 97 years old.

There were 139 hypothermia deaths in New York City (NYC) with an average annualized rate of 1.7 per million (Table 1).

TABLE 1—*Demographic characteristics of hypothermia-related deaths in NYC, January 1, 2009–July 31, 2019.*

		Hypothermia-related Deaths			
Decedent Characteristics	*n*	2010 Census Population	Rate Per Million	Rate Ratio	95% CI
Total	139	8,175,133	17.0	–	–
Gender					
Male	106	3,882,544	27.3	3.55	2.40, 5.25
Female	33	4,292,589	7.7	reference	–
Female age-group (years)					
0–19	1	978,610	1.0	0.11	0.01, 0.82
20–39	2	1,355,217	1.5	0.15	0.03, 0.68
40–64	13	1,360,444	9.6	reference	–
65–84	10	500,121	20.0	2.09	0.92, 4.77
85+	7	98,197	71.3	7.46	2.98, 18.70
Male age-group (years)					
0–19	2	1,016,260	2.0	0.04	0.01, 0.16
20–39	11	1,267,220	8.7	0.18	0.09, 0.34
40–64	59	1,204,224	49.0	reference	–
65–84	30	351,631	85.3	1.74	1.12, 2.70
85+	3	43,209	69.4	1.42	0.44, 4.52
Unknown	1	–	–	–	–
Ethnicity stratification*					
Black or African American	59	1,861,295	31.7	2.01	1.36, 2.97
White	43	2,722,904	15.8	reference	-
Hispanic	26	2,336,076	11.1	0.71	0.43, 1.15
Asian	11	1,028,119	10.7	0.68	0.35, 1.31
Borough stratification					
Bronx	19	1,385,108	13.7	0.68	0.39, 1.20
Queens	49	2,230,722	22.0	1.09	0.70, 1.70
Manhattan	32	1,585,873	20.2	reference	-
Brooklyn	35	2,504,700	14.0	0.69	0.43, 1.12
Staten Island	4	468,730	8.5	0.42	0.15, 1.20

*As recorded by death certificate

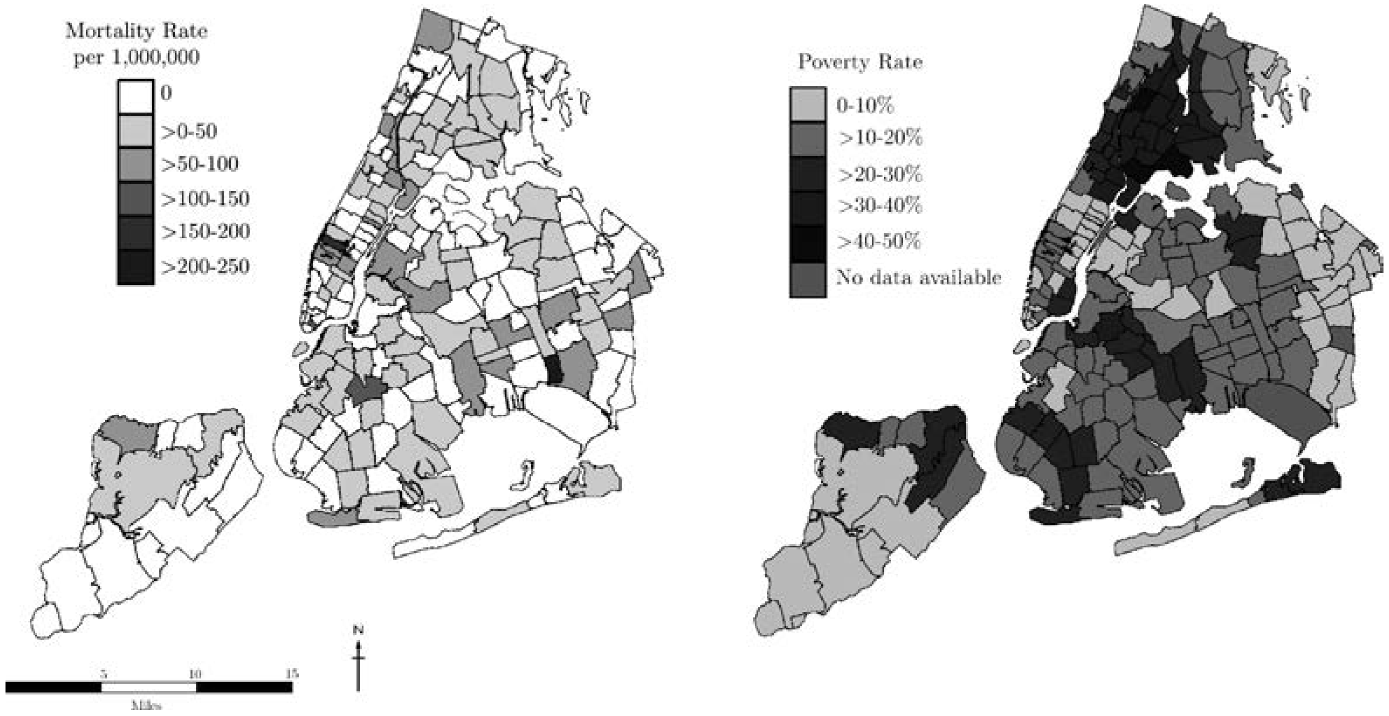


FIG. 1—*Geographic distribution of hypothermia-related mortality rates and poverty rates by zip code for New York City, January 1, 2009–July 31, 2019. [Color figure can be viewed at wileyonlinelibrary.com]*

TABLE 2—*Demographic characteristics of hypothermia-related deaths in the city of Houston, January 1, 2009–July 31, 2019.*

		Hypothermia-related Deaths			
Decedent Characteristics	*n*	2010 Census Population	Rate per Million	Rate Ratio	95% CI
Total	50	2,099,451	23.8	–	–
Gender					
Male	37	1,053,517	35.1	2.83	1.50, 5.32
Female	13	1,045,934	12.4	reference	–
Female age-group (years)					
0–19	0	294,270	–	–	–
20–39	0	336,647	–	–	–
40–64	6	305,472	19.6	reference	–
65–84	2	93,819	21.3	1.09	0.22, 5.38
85+	4	15,726	254.4	12.95	3.65, 45.89
Male age-group (years)					
0–19	0	307,648	–	–	–
20–39	2	362,086	5.5	0.08	0.02, 0.32
40–64	22	303,386	72.5	reference	–
65–84	11	72,748	151.2	2.09	1.01, 4.30
85+	3	7,649	392.2	5.41	1.62, 18.07
Ethnicity stratification*					
Black or African American	18	498,466	36.1	1.08	0.56, 2.07
White	18	537,901	33.5	reference	–
Hispanic	13	919,668	14.1	0.42	0.21, 0.86
Asian	1	126,378	8.0	0.24	0.03, 1.79

*As recorded by death certificate

Males were more likely to die of hypothermia compared to females (rate ratio (RR) = 3.55, 95% CI 2.40, 5.25). Hypothermia mortality rates increased with age with the highest rates among men aged 65–84 (85.3 per million) and women aged 85 or older (71.3 per million). Of 139 decedents, 59 (42.4%) were Black or African American, 43 (30.9%) were White, 26 (18.7%) were Hispanic, and 11 (7.9%) were Asian. Black or African Americans were more likely to die of hypothermia compared to Whites (RR = 2.01, 95% CI 1.36, 2.97). Forty-nine (35.3%) hypothermia deaths took place in the borough of Queens, 35 (25.2%) in Brooklyn, 32 (23.0%) in Manhattan, 19 (13.7%) in the Bronx, and 4 (2.9%) in Staten Island. Zip code 11225 (Prospect-Leffert), in Brooklyn, had the most deaths (*n* = 7), and zip code 11436 (South Jamaica), in Queens, had the highest mortality rate at 223 per million (Fig. 1). During this same time, there were 50 hypothermia deaths in Houston with an average annualized rate of 2.4 per million. Males were more likely to die of hypothermia compared to females (RR = 2.83, 95% CI 1.50, 5.32; Table 2). Hypothermia mortality rates increased with age with the highest rates among men aged 85 or older (392.2 per

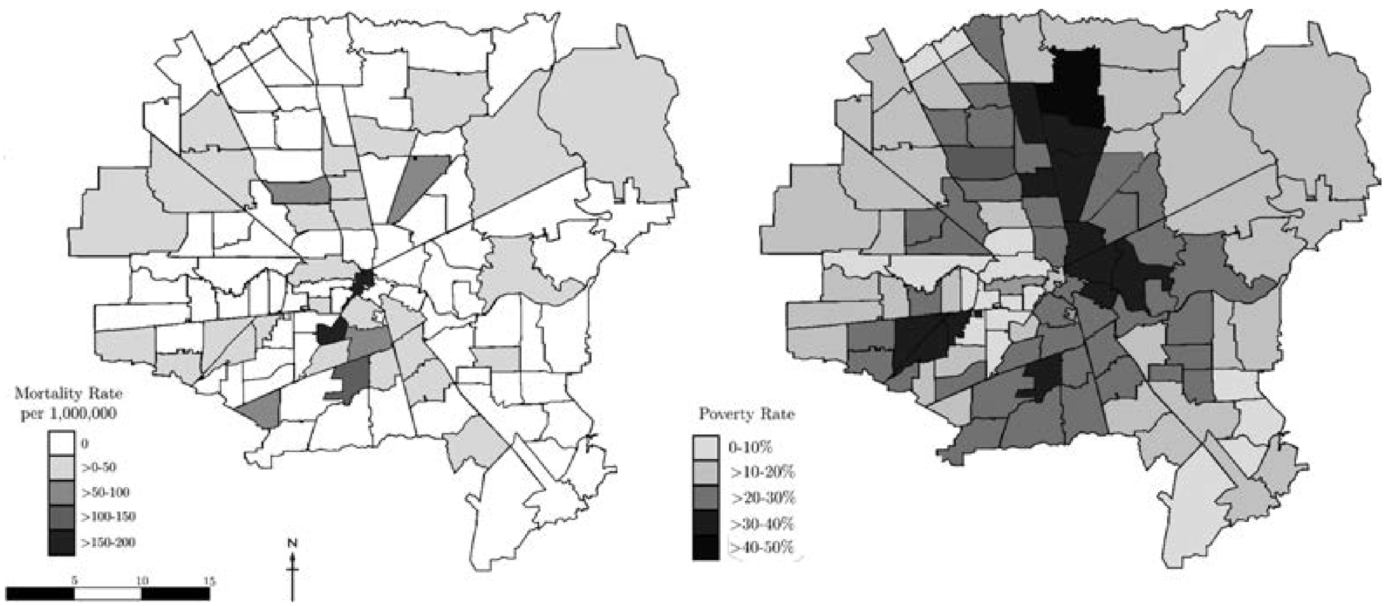


FIG. 2—*Geographic distribution of hypothermia-related mortality rates and poverty rates by zip code for central Houston, January 1, 2009–July 31, 2019. [Color figure can be viewed at wileyonlinelibrary.com]*

million) and women aged 85 or older (254.4 per million). Among 50 hypothermia deaths, 18 (36%) were Black or African American, 18 (36%) were White, 13 (26%) were Hispanic, and 1 (2%) was Asian. Hispanics were less likely to die of hypothermia compared to Whites (RR = 0.42, 95% CI 0.21, 0.86). In Houston, zip code 77002 (Downtown Houston) had the most deaths (*n* = 3) and zip code 77030 (University Place) had the highest mortality rate at 195 per million (Fig. 2). There was no correlation between zip code poverty rate and hypothermia-related mortality rate in either city (Figs 1 and 2).

Both Houston and NYC experienced peak hypothermic deaths in 2018 (Fig. 3) with a total of 23 deaths (2.8 per million) in NYC and 11 (5.2 per million) in Houston in that year alone. The majority of hypothermic deaths in NYC occurred in January (*n* = 54, 38.8%), February (*n* = 30, 21.6%), and December (*n* = 20, 14.3%), while the majority of deaths in Houston occurred during the months of January (*n* = 18, 36%), December (*n* = 16, 32%), and November (*n* = 9, 18%; Fig. 4). Age- and sex-specific standardized hypothermia death rates were 18.2 (95% CI 15.1, 21.2) per million in NYC and 30.1 (95% CI 21.7, 38.6) per million in Houston for the time period between January 1, 2009, and July 31 2019. The comparative hypothermia death ratio was 1.66 (95% CI 1.19, 2.30), indicating hypothermia mortality in Houston was 66% higher than in NYC (Tables S1 and S2).

Of the 189 combined cases, 127 (67.2%) were found outside, exposed to the environment, while there was evidence that about half (*n* = 93, 49.2%) had evidence of living in or having access to a fixed residence. The manner of death was reported as an accident in 181 cases (95.8%), followed by undetermined (*n* = 5, 2.6%), and suicide (*n* = 3, 1.6%). A survival interval occurred in 91 cases, ranging from the time required to complete the rewarming protocol (hours) up to 33 days; of whom, the coldest recorded body temperature was 68.3°F (20.2°C), and the highest was 92°F (33.3°C). On the date of injury, the average daytime high temperature in NYC was 42°F (5.5°C) and in Houston was 54°F (12.2°C); and the average nighttime low in NYC was 27°F (−2.7°C) and in Houston was 36°F (2.2°C). In both cities, precipitation (rain and snow) was noted in the weather reports for 56 (30%) cases (Table 3).

Wischnewski spots were found in 107 (56.6%) cases, while pancreatitis was found in 29 (15.3%) cases and pink lividity was

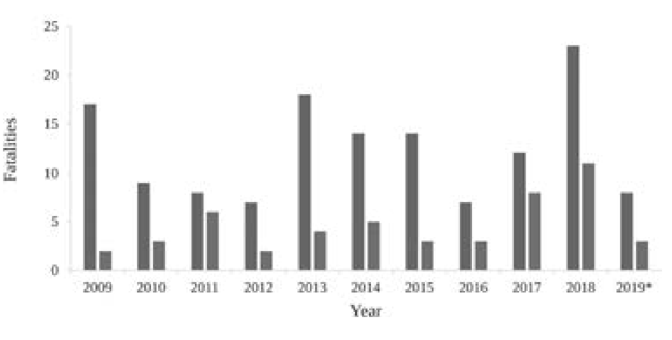


FIG. 3—*The absolute number of hypothermia-related deaths stratified by year in Houston and NYC for January 1, 2009–July 31, 2019*. *2019 data limited to July 31, 2019. [Color figure can be viewed at wileyonlinelibrary.com]*

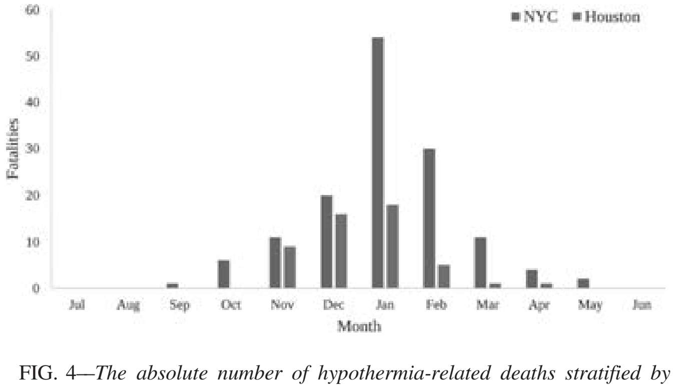


FIG. 4—*The absolute number of hypothermia-related deaths stratified by month for NYC and Houston for January 1, 2009–July 31, 2019. [Color figure can be viewed at wileyonlinelibrary.com]*

TABLE 3—*Characteristics of hypothermia-related deaths in New York City (NYC) and Houston, January 1, 2009–July 31, 2019.*

	NYC		Houston		
Decedent Characteristics	*n*	%	*n*	%	*P* value¶
Total	139	100%	50	100%	
Residence*					
Homeless	60	43.2	14	28.0	0.042
Nonhomeless†	67	48.2	26	52.0	
Unknown	12	8.6	10	20.0	
Place of injury*					
Indoors	37	26.6%	13	26.0%	0.852
Outdoors‡	90	64.7%	37	74.0%	
Unknown	2	1.4%	0	–	
Manner of death					
Accident	131	94%	50	100%	0.350
Undetermined	5	4%	0	–	
Suicide	3	2%	0	–	
Findings at autopsy					
Wischnewski spots	74	53%	33	66%	0.021
Pancreatitis	19	14%	10	20%	
Pink lividity	14	10%	0		
Ethanol testing, blood					
Positive	57	41	12	24	0.061
Negative	77	50	34	68	
Not tested or objected	5	4	0	–	
Average (of highest blood site)	0.20 gm%		0.27 gm %		
Cocaine testing, all fluids					
Positive for cocaine	8	–	1	–	0.449
Past medical history§					
Cardiovascular disease	44	32	37	74	0.016
Pulmonary disease	5	4	9	18	
Dementia	6	4	2	4	
Schizophrenia	7	5	0	–	
Bipolar disorder	3	2	0	–	

*Residence and place of injury determined by evaluating case files.

†All those with a fixed address/clear living circumstances.

‡Outdoors to include subway station, on the street, encampments, and other situations with minimal coverage.

§Past medical history determined by findings at autopsy or scene investigation.

¶*P* values were obtained from chi-square or Fisher's exact tests.

found in 14 (7.4%) cases. Toxicological analysis detected ethanol in 69 cases (36.5%) with an average blood concentration of 0.20 gm% in NYC and 0.27 gm% in Houston. Other intoxicants included cocaine (n = 9, 4.8%), opioids, benzodiazepines, stimulants, and antipsychotics. Underlying medical conditions/risk factors include cardiovascular disease (n = 81, 42.9%) and pulmonary disease (n = 14, 7.4%). Pre-existing diagnoses of dementia (n = 8, 4.2%) or a mental illness such as schizophrenia or bipolar disorder (n = 10, 5.3%) were also noted (Table 3).

Discussion

This study combines forensic data from two of the largest cities in the United States, New York City and Houston, Texas. As of the 2010 Census, New York City, NY (NYC) is the largest city in the country with a population of 8,175,133 and covers approximately 302.6 sq mi (783 km^2). The city is located in the Northeast, along the Atlantic Ocean at an approximate latitude of 41° N. Houston is the fourth largest city in the country with a population of 2,099,451 and has twice as much land at 637.4 sq mi (1623 km^2). It is located in the South, along the Gulf of Mexico at an approximate latitude of 30° N. These two cities are drastically different in geography, climate, and population. Houston is a considerably warmer city with winter temperatures on average 10–20°C higher than NYC and has a much smaller population that is distributed over a much larger area. Between January 1, 2009, and July 31, 2019, there were 139 hypothermia-related deaths in NYC with an annualized mortality rate of 1.7 per million. This is consistent with recently published studies (10) and considerably lower than the 2003–2013 national average with unadjusted annual rates ranging from 0.3 to 0.5 per 100,000 persons (or 3–5 per million) (11). Houston experienced fewer deaths (n = 50) but had a higher annualized mortality rate at 2.4 per million, again lower than the national average. After directly standardizing the data to age and sex, the mortality rate from hypothermia was 66% higher in Houston compared to NYC. Our data demonstrate the risk of hypothermia in all areas of the country, especially in the temperate and warmer regions. Similar results are seen in other published studies (12,13).

Hypothermia-related death is associated with advanced age, the male sex, comorbidities such as cardiovascular disease and pulmonary disease, homelessness, and intoxication (5,10–15). The most common comorbidities among decedents in our review were cardiovascular disease, pulmonary disease, and mental illness. Cardiovascular disease was appreciably higher in Houston where it was seen in 74% of the cases, compared with NYC, where only 32% of cases demonstrated disease. Both pulmonary disease and mental illness were less frequently seen in both Houston and NYC. Approximately one-third of all cases that underwent toxicological analysis were positive for ethanol (69 of 189). Detectable ethanol levels were slightly more common in NYC. The blood alcohol levels in both cities were similar and ranged from 0.13 gm% to 0.50 gm%. A wide variety of other medication and recreational substances such as opioids, benzodiazepines, antipsychotics, and other stimulants and their metabolites were also detected.

Determining the cause of death, without a perimortem core body temperature, is difficult and requires a careful and thorough evaluation of the case file including scene investigation, autopsy results, and other reports such as toxicology and neuropathology. The manner of death in the vast majority of our 189 cases was listed as accident; however, there were five undetermined cases and three cases were certified as suicides. The cases that were listed as undetermined involved complicated histories which likely provided insufficient certainty to make a clear distinction between other manners of death (16). In all three cases of suicide, the remains were found in a body of water. Suicide by hypothermia is rare and may be more frequently associated with complex suicides, involving more than one modality (17).

Autopsy findings associated with hypothermia are subtle and often nonspecific (6). A literature review by Tsokos et al revealed that Wischnewski spots were identified in 40–91% of all autopsies associated with hypothermia (7). Our results were consistent with this finding. There is some controversy in the literature regarding the association between hypothermia and pancreatitis. Only 14–20% of our cases had gross signs of acute pancreatitis, ranging from focal hemorrhage to diffuse involvement. Pink lividity was only identified in 10% of the NYC cases but not identified in Houston cases. Paradoxical undressing was infrequently identified in the case files.

There is a growing body of evidence that environmental temperature plays only a limited role in winter excess mortality rate (18–20) and speculation that other factors, such as respiratory infection, may play a larger role. The relationship between comorbidities and hypothermia-related death should be further investigated.

There are several limitations to our study. It is a retrospective analysis, and the data were limited to case reports. Also, determining the cause of death, without a core body temperature, is difficult and requires a careful and thorough evaluation of the case file, autopsy results, and other reports such as toxicology and neuropathology. It is difficult to assess a decedent's homeless status even with a thorough review of the case file, medical examiner record, and death certificate. Using these sources alone likely underestimates the number of homeless individuals in our study.

Local agencies can use these data to target higher-risk populations and offer appropriate intervention to try to prevent these deaths. The lack of pathognomonic autopsy findings emphasizes the importance of a thorough scene investigation; otherwise, these deaths may be underreported. The role of the medical examiner is crucial in the accurate classification of these deaths, to assist in improving public education, targeted interventions, and emergency response planning.

Acknowledgments

We are very grateful to the New York City Office of Chief Medical Examiner and the Harris County Institute of Forensic Sciences for their assistance with this project.

References

1. Rolf CM, Gallagher KE. Hypothermic death in the Arctic state. Acad Forensic Pathol 2018;8(1):64–82. https://doi.org/10.23907/2018.005.
2. Berko J, Ingram DD, Saha S, Parker JD. Deaths attributed to heat, cold, and other weather events in the United States, 2006–2010. Natl Health Stat Report 2014;76:1–15.
3. Xu J. CDC QuickStats: number of hypothermia-related deaths, by sex – National Vital Statistics System, United States, 1999–2011. United States, 2011. MMWR Morb Mortal Wkly Rep 2013;61(51):1050. https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6151a6.htm (accessed July 15, 2019).
4. DiMaio D, DiMaio V. Forensic pathology (practical aspects of criminal and forensic investigations), 2nd edn. Boca Raton, FL: CRC Press, 2001;592.
5. Nixodorf-Miller A, Hunsaker DM, Husaker JC. Hypothermia and hyperthermia medicolegal investigation of morbidity and mortality from exposure to environmental temperature extremes. Arch Pathol Lab Med 2006;130(9):1297–304. https://doi.org/10.1043/1543-2165(2006)130[1297:HAHMIO]2.0.CO;2.
6. Palmiere C, Teresinski G, Hejna P. Postmortem diagnosis of hypothermia. Int J Legal Med 2014;128(4):607–14. https://doi.org/10.1007/s00414-014-0977-1.
7. Tsokos M, Rothschild MA, Madea B, Rie M, Sperhake JP. Histological and immunohistochemical study of Wischnewsky spots in fatal hypothermia. Am J Forensic Med Pathol 2006;27(1):70–4. https://doi.org/10.1097/01.paf.0000202716.06378.91.
8. Altman DG. Practical statistics for medical research. London, U.K.: Chapman and Hall Publishers, 1991;267.
9. Elandt-Johnson RC, Johnson NL. Survival models and data analysis. New York, NY: John Wiley & Sons, 1980;70–1.
10. Lane K, Ito K, Johnson S, Gibson EA, Tang A, Matte T. Burden and risk factors for cold-related illness and death in New York City. Int J Environ Res Public Health 2018;15(4):632. https://doi.org/10.3390/ijerph15040632.
11. Meiman J, Anderson H, Tomasallo C; Centers for Disease Control and Prevention (CDC). Hypothermia-related deaths–Wisconsin, 2014, and United States, 2003–2013. MMWR Morb Mortal Wkly Rep 2015;64(6):141–3.
12. Taylor AJ, McGwin G Jr, Davis GG, Brissie RM, Holley TD, Rue LW 3rd. Hypothermia deaths in Jefferson County, Alabama. Inj Prev 2001;7(2):141–5. https://doi.org/10.1136/ip.7.2.141.
13. Bright FM, Winskog C, Walker M, Byard RW. A comparison of hypothermic deaths in South Australia and Sweden. J Forensic Sci 2014;59(4):983–5. https://doi.org/10.1111/1556-4029.12451.
14. Zhang P, Wiens K, Wang R, Luong L, Ansara D, Gower S, et al. Cold weather conditions and risk of hypothermia among people experiencing homelessness: implications for prevention strategies. Int J Environ Res Public Health 2019;16(18):3259. https://doi.org/10.3390/ijerph16183259.
15. Brändström H, Eriksson A, Giesbrecht G, Angquist KA, Haney M. Fatal hypothermia: an analysis from a sub-arctic region. Int J Circumpolar Health 2012;71:1–7. https://doi.org/10.3402/ijch.v71i0.18502.
16. Hanzlick R, Hunsaker JC, Davis GJ. National Association of Medical Examiners: a guide for manner of death classification. Marceline, MO: National Association of Medical Examiners, 2002;2–5.
17. Petković S, Maletin M, Durendić-Brenesel M. Complex suicide: an unusual case with six methods applied. J Forensic Sci 2011;56(5):1368–72. https://doi.org/10.1111/j.1556-4029.2011.01821.x.
18. Kinney PL, Schwartz J, Pascal M, Petkova E, Tertre AL, Medina S, et al. Winter season mortality: will climate warming bring benefits? Environ Res Lett 2015;10(6):064016. https://doi.org/10.1088/1748-9326/10/6/064016.
19. Kl Ebi, Mills D. Winter mortality in a warming climate: a reassessment. Wiley Interdiscip Rev Clim Change 2013;4(3):203–12. https://doi.org/10.1002/wcc.211.
20. Staddon PL, Mongomery HE, Depledge MH. Climate warming will not decrease winter mortality. Nat Clim Chang 2014;4:190–4. https://doi.org/10.1038/nclimate2121.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Age-sex-specific standardized hypothermia-related mortality rate in NYC, January 1, 2009–July 31, 2019 using the U.S. national population from the 2010 census as the reference population.

Table S2. Age-sex-specific standardized hypothermia-related mortality rate in Houston, January 1, 2009–July 31, 2019 using the U.S. national population from the 2010 census as the reference population.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14502
Available online at: onlinelibrary.wiley.com

PAPER

PATHOLOGY/BIOLOGY

Jack Garland,[1] *B.Med.; Benjamin Ondruschka,*[2] *M.D.; Simon Stables,*[3] *M.B.Ch.B.; Paul Morrow,*[3] *M.D.; Kilak Kesha,*[3] *M.B.B.S.; Charley Glenn,*[3] *M.D.; and Rexson Tse,*[3,4] *M.D.*

Identifying Fatal Head Injuries on Postmortem Computed Tomography Using Convolutional Neural Network/Deep Learning: A Feasibility Study

ABSTRACT: Postmortem computed tomography (PMCT) is a relatively recent advancement in forensic pathology practice that has been increasingly used as an ancillary investigation and screening tool. One area of clinical CT imaging that has garnered a lot of research interest recently is the area of "artificial intelligence" (AI), such as in screening and computer-assisted diagnostics. This feasibility study investigated the application of convolutional neural network, a form of deep learning AI, to PMCT head imaging in differentiating fatal head injury from controls. PMCT images of a transverse section of the head at the level of the frontal sinus from 25 cases of fatal head injury were combined with 25 nonhead-injury controls and divided into training and testing datasets. A convolutional neural network was constructed using Keras and was trained against the training data before being assessed against the testing dataset. The results of this study demonstrated an accuracy of between 70% and 92.5%, with difficulties in recognizing subarachnoid hemorrhage and in distinguishing congested vessels and prominent falx from head injury. These results are promising for potential applications as a screening tool or in computer-assisted diagnostics in the future.

KEYWORDS: head, injuries, traumatic brain injury, postmortem computed tomography, convoluted neural network, deep learning, subarachnoid hemorrhage, SAH, autopsy, forensic radiology

Postmortem computed tomography (PMCT) is one of the most recent advancements and ancillary investigations in the field of forensic pathology. Differing from clinical applications, PMCT commonly scans the head and torso, acts as a permanent record for court purposes, and is used as a screening tool and for identification purposes (1–4). In New Zealand, all PMCT scans are interpreted by a forensic pathologist in the first instance and subsequently by a forensically trained radiologist. Although the applications of PMCT can differ from clinical computed tomography (CT) imaging, advances in clinical CT imaging are often the "intellectual father" of PMCT and help further the development of postmortem imaging (5).

One area of clinical CT imaging and clinical radiology in general that has garnered a lot of research interest recently is the area of "artificial intelligence" (AI) such as screening and computer-assisted diagnostics (6,7). Not intended to replace radiologists, there are several proposed reasons for why AI should be able to be used in this field, in particular, that computers are able to quickly process large amounts of information and detect patterns that traditional statistical analysis cannot (7,8).

The applications of AI in clinical radiology, while emerging, are sufficiently established and not entirely theoretical. Recent research has demonstrated the utility and accuracy of AI radiological image analysis as a triage screening tool in head trauma, as well as an adjunct test in the form of computer-assisted diagnostics (6,8–10). While AI analysis of simple data such as variables of age, sex, or given quantities is comparatively straightforward, analysis of complex data such as applies to radiological images is more difficult because the information is not a discrete data-set. Various forms of AI image analysis have been proposed and the computational power required to investigate them has only become available in recent years (6,11).

One of the most promising approaches to AI image analysis is deep learning, a category of machine learning that uses artificial neural networks, loosely based on human cognition (6,11). Convolutional neural network (CNN), the most common subset of deep learning used in image classification and thus in radiology, works by layered analysis of input images by AI programs against a training data-set of prediagnosed images, which can be later tested against a testing dataset of new images (6,11). The layered analysis allows specific features to be recognized and aggregated with respect to the diagnosis of interest.

The use of and research into AI in postmortem imaging is currently very limited in the scientific literature, with only one publication on recognizing hemopericardium (12). The published

[1] Forensic and Analytical Science Service, 480 Weeroona Rd, Lidcombe, NSW 2141, Australia.

[2] Institute of Legal Medicine, University Medical Center Hamburg-Eppendorf, Martinistraße 52 20251, Hamburg, Germany.

[3] Department of Forensic Pathology, LabPLUS, Auckland City Hospital, 2 Park Road, Grafton, Auckland, 1023, New Zealand.

[4] University of Auckland Faculty of Medical and Health Sciences, 85 Park Road, Grafton, Auckland, 1023, New Zealand.

Corresponding author: Rexson Tse, M.D. E-mail: rexsont@adhb.govt.nz

Received 27 Mar. 2020; and in revised form 7 June 2020; accepted 15 June 2020.

literature on AI in postmortem imaging is largely composed of proposed uses of this technology, as well as concepts to one day integrate it with other AI research areas such as histology, with scant research on actual cases of AI postmortem image analysis (13–15). One particular area of postmortem imaging outside of AI that has shown use as both a screening tool and even an alternative to autopsy is in cases of head injury (16,17). This feasibility study was performed to determine whether CNN, as used in deep learning, is able to differentiate fatal head injuries on PMCT scans from normal ("uninjured") cases.

Materials and Methods

Case and Control Selection

This study was conducted at the Department of Forensic Pathology, LabPLUS, Auckland City Hospital, Auckland, New Zealand. Ethical approval was waived as all cases selected for this work underwent coronial postmortem examination authorized by the Coroner and PMCT was a routinely performed part of the postmortem examination in all cases selected. All the PMCT scans performed in the department are initially reviewed by forensic pathologists with radiology experience and subsequently by radiologists with postmortem experience in weekly multidisciplinary meetings. The pathologist review was done before autopsy and therefore blinded to the autopsy findings, and the radiologist reviews were later performed blinded to the autopsy findings.

Twenty-five consecutive cases in which PMCT showed non-survivable head injuries with the final autopsy diagnosis "fatal traumatic brain injury" were selected between 2018 and 2020. Further, 25 consecutive suicide hanging deaths were selected during the same time period and used as controls (all showing no head injuries on PMCT). All cases underwent full three-cavity autopsies. The age and sex for each group were recorded. A transverse image of a section of the head at the level of the frontal sinus in the soft tissue view was used for each case. The PMCT scan of the head was performed in a supine position using a helix 32-slice CT (Siemens Somatom Scope CT scanner) before postmortem examination. The images were reconstructed in the soft tissue window and exported out in Jpeg format with RGB color space. The gold standard for the cranial pathology diagnosis in this study was the finding at autopsy. In all included cases, there was no difference in diagnosis between the prospective review of the PMCT by the forensic pathologist (or the later blinded radiologist review) and the autopsy findings.

Exclusion Criteria

1. All suspicious, homicidal, and pediatric deaths (age < 10) due to potential legal issues.
2. All cases with signs of decomposition.
3. All cases with neurosurgical procedures.

Model Construction

The analysis was performed in R (version 3.4.1, The R Foundation for Statistical Computing). The CNN was constructed using Keras with default TensorFlow backend (18). The exported images were divided into two sets, being the training and the testing sets. The training set had 40 images (20 images with fatal head injury and 20 controls). The testing set had 10 images (5 images with fatal head injury and 5 without). A commercially available 64-bit laptop capable of running R and Keras applications was used, with training time taking approximately 15 min and each test diagnosis taking seconds.

The computer speed was capable of being enhanced via an external graphics processor unit (eGPU) but was not needed in this study due to the small sample size.

The original images had a size of 512 × 512 pixels and were resized to 150 × 150 pixels for ease of computing. The CNN was constructed with four convolutional layers followed by two dense/closed-loop output layers. The convolutional layers had sequential filters of 32, 64, 128, and 128 and a kernel size of 3 × 3. Activation function was set as rectified linear unit ("*relu*"). Corresponding max-pooling and drop out layers were inserted in each convolutional layer. For the two output layers, the units were set as 256 and 2, with the initial activation function set as *relu* and the final activation as softmax.

For compilation, the loss function was set as binary cross-entropy, optimizer as stochastic gradient descent, and metrics as accuracy. Loss function in machine learning is a measure of how well a model fits the data (similar to the R-squared value in linear regression models), with the goal of the learning process being to generate the minimal loss function so as to have the best model fit. In this study, the specific loss function used was binary cross-entropy, which uses the formula "loss function = $-t\log(s) - (1 - t)\log(1 - s)$", where t is the target and s is the prediction by the CNN as a probability. For model fitting, the batch size was 50, epoch was 50 and validation set was 20%. This type of CNN construction is commonly used in image classification.

Results

In the cases with fatal head injuries, all were transport related and accidental (drivers, passengers or pedestrians). The mean age was 37.8 years (range: 17–88), with a male-to-female ratio of 19:6. The head injuries identified on PMCT scan included skull fractures, epidural hemorrhage, subdural hemorrhage, subarachnoid hemorrhage, brain contusion, and/or parenchymal hemorrhage. In the cases without head injuries (all hanging deaths), the mean age was 48.84 years (range: 18–81), with a male-to-female ratio of 22:3. There were no statistical differences between age (t test) and sex (Fisher exact test) between the two groups ($p > 0.05$). Examples of the PMCT images are shown in Fig. 1.

For the training set, the loss function (the measurement between model prediction and the target in deep learning) and accuracy were 0.33 and 92.5%, respectively. The model was able to classify all the cases with head injuries correctly (20 cases) and misclassified three out of 20 cases without head injuries as having one. The misclassified cases all showed features of prominent congestion and cerebral edema (Fig. 2).

For the testing set, the loss function and accuracy were 0.61 and 70%, respectively. The model misclassified one out of five cases with head injuries as having no injuries and two out of five cases without head injuries as having one. The misclassified case with fatal head trauma had only subarachnoid hemorrhage resulting from torn vertebral arteries (Fig. 3), and the misclassified cases without head trauma had congested vessels or prominent falx on PMCT, similar to Fig. 2.

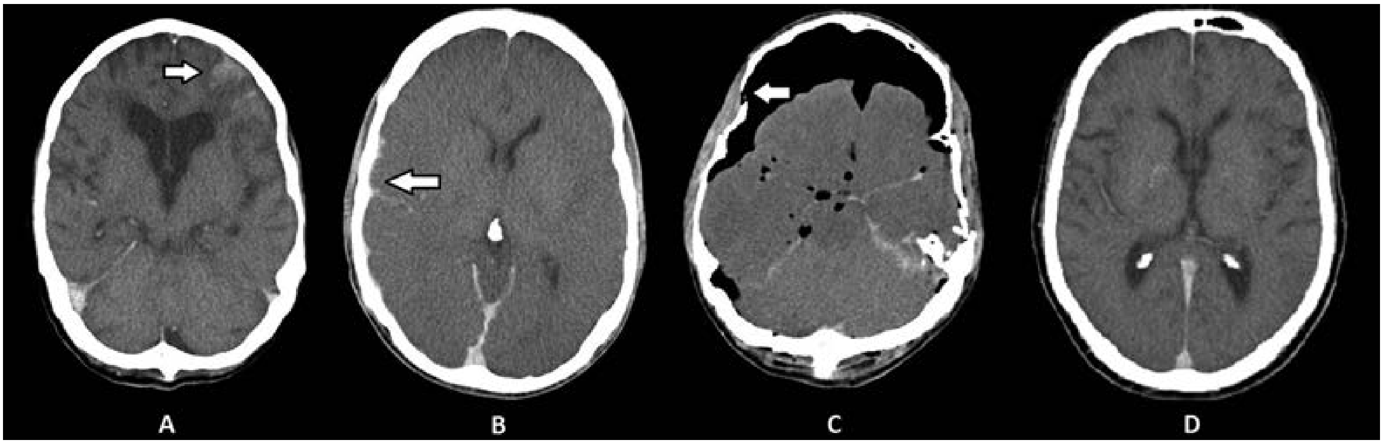


FIG 1—*Examples of postmortem computed tomography (PMCT) transverse section images of the head, including brain contusion (A, marked with arrow), subdural hemorrhage (B, marked with arrow), skull fracture with pneumocranium (C, marked with arrow), and normal (D). All images in Fig. 1 are from the training dataset.*

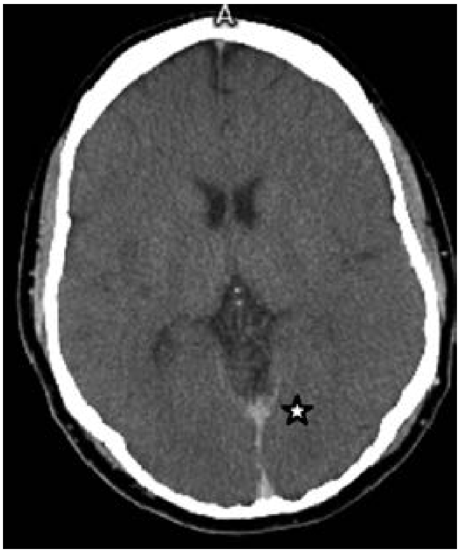


FIG 2—*Postmortem computed tomography (PMCT) image showing congested blood vessels (star) and brain swelling but no head injury, misclassified as having one in the training set (Anterior marked "A").*

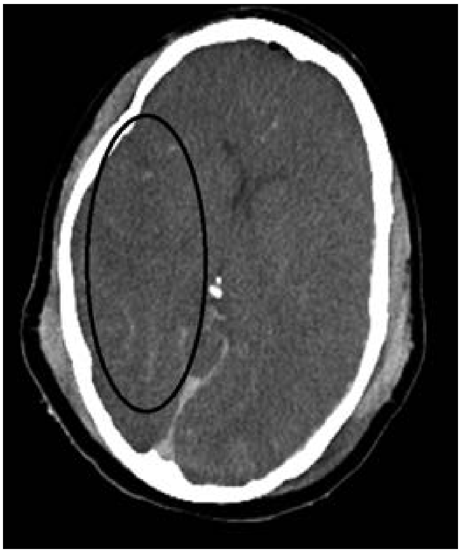


FIG 3—*Postmortem computed tomography (PMCT) image with subarachnoid hemorrhage (circled) misclassified as being normal in the testing set.*

Discussion

This study investigated the novel use of CNN in the postmortem diagnosis of fatal head injury via PMCT. The results of this study were promising and were useful as a proof of concept for AI work in PMCT head injury screening, with training dataset accuracy of 92.5% and testing dataset accuracy of 70%. These results are not dissimilar to previous studies on deep learning AI in clinical/ ante-mortem CT head diagnostics. A 2017 study by Prevedello et al. used CNN to analyze a variety of head and brain pathologies on CT and demonstrated an area under the curve (AUC) of 0.91 in hemorrhage, mass effect, and hydrocephalus (19). A 2018 study by Chilamkurthy et al. used deep learning algorithms to detect intracranial hemorrhage, fractures and mass-effect in CT head images and showed an AUC of between 0.86 and 0.97 depending on the pathology (9). Accuracy in binary problems (such as in the present study) is essentially calculated as true positive + true negative/ total number of tests. AUC is another assessment of test accuracy in binary classification problems when receiver operating characteristic curves are plotted.

For the present study, the single head injury case misclassified as having no head injury showed a subarachnoid hemorrhage, which may be less apparent than other injuries included in this study such as skull fractures, subdural hemorrhage, and parenchymal hemorrhage. Subarachnoid hemorrhage can be challenging to differentiate from postmortem vascular congestion along the cisterns, even by expert radiologists (20). Of the two nonhead injury cases that were misclassified as head injury, the images used showed congested vessels or a prominent falx (a common PMCT artifact), suggesting that with a wider dataset including more variations in normal head CTs, these findings may not have been misinterpreted as pathological/ traumatic.

Limitations

Study Design

This study, being a feasibility/proof of concept study, had relatively low numbers and simplified the complexity of head injuries. Although (i) having only 50 cases, (ii) using only one transverse image from the head PMCT scan as the input, and (iii) grouping all different head injury diagnoses as one single group, a promising result was able to be demonstrated. A refinement of this study would be to use the entire DICOM dataset image stack reconstructed from different planes (sagittal, transverse, and coronial), including different windows, and stratifying the different types of head injuries. This would greatly increase the potential accuracy of the CNN over the jpeg image approach used in this feasibility study, however it would require more complex CNN analysis and greater computational power.

Applications

This study by no means had a view to replace radiologists and/or pathologists in interpreting PMCT, as the minimal error in all deep learning methods is the human error. This study shows that there is a potential role in interpreting PMCT scans using more advanced computational methods, which may aid the radiologist/pathologist as a screening tool or even assisted-diagnostics tool, by drawing attention to traumatic pathologies.

Recommendation and Conclusion

The findings of this study show particular potential in screening and assisted diagnostics for the forensic pathology context, where unlike clinical CT imaging, clinical history and symptoms may be absent or described wrong by relatives, large numbers of normal cases are scanned for screening purposes, and the images are often interpreted by forensic pathologists without radiologist input or expertise. To the best of the authors' knowledge, this is one of the very few studied applications of AI to PMCT, and the very first application of PMCT to head injury assessment.

In future studies, it would be ideal to use the entire DICOM dataset, possibly with additional alterations such as putting increased diagnostic "weight" on particular slices. Additionally, as many deep learning applications use thousands of samples in their data sets, it is likely that the accuracy of the CNN used in this study could be substantially improved by use of larger data clusters.

References

1. Le Blanc-Louvry I, Thureau S, Duval C, Papin-Lefebvre F, Thiebot J, Dacher JN, et al. Post-mortem computed tomography compared to forensic autopsy findings: a French experience. Eur Radiol 2013;23(7):1829–35.
2. Flach PM, Gascho D, Schweitzer W, Ruder TD, Berger N, Ross SG, et al. Imaging in forensic radiology: an illustrated guide for postmortem computed tomography technique and protocols. Forensic Sci Med Pathol 2014;10(4):583–606.
3. Thomsen AH, Jurik AG, Uhrenholt L, Vesterby A. An alternative approach to computerized tomography (CT) in forensic pathology. Forensic Sci Int 2009;183(1–3):87–90.
4. Moskala A, Wozniak K, Kluza P, Romaszko K, Lopatin O. The importance of post-mortem computed tomography (PMCT) in confrontation with conventional forensic autopsy of victims of motorcycle accidents. Leg Med (Tokyo) 2016;18:25–30.
5. Grabherr S, Egger C, Vilarino R, Campana L, Jotterand M, Dedouit F. Modern post-mortem imaging: an update on recent developments. Forensic Sci Res 2017;2(2):52–64.
6. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. Nat Rev Cancer 2018;18(8):500–10.
7. Miller DD, Brown EW. Artificial intelligence in medical practice: the question to the answer? Am J Med 2018;131(2):129–33.
8. Yasaka K, Abe O. Deep learning and artificial intelligence in radiology: current applications and future directions. PLoS Med 2018;15(11): e1002707.
9. Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet 2018;392 (10162):2388–96.
10. Chan S, Siegel EL. Will machine learning end the viability of radiology as a thriving medical specialty? Br J Radiol 2019;92(1094):20180416.
11. Zhu G, Jiang B, Tong L, Xie Y, Zaharchuk G, Wintermark M. Applications of deep learning to neuro-Imaging techniques. Front Neurol 2019;10:869.
12. Ebert LC, Heimer J, Schweitzer W, Sieberth T, Leipner A, Thali M, et al. Automatic detection of hemorrhagic pericardial effusion on PMCT using deep learning – a feasibility study. Forensic Sci Med Pathol 2017;13(4):426–31.
13. O'Sullivan S, Holzinger A, Zatloukal K, Saldiva P, Sajid MI, Wichmann D. Machine learning enhanced virtual autopsy. Autops Case Rep 2017;7 (4):3–7.
14. O'Sullivan S, Heinsen H, Grinberg LT, Chimelli L, Amaro E, do Nascimento Saldiva PH, et al. The role of artificial intelligence and machine learning in harmonization of high-resolution post-mortem MRI (virtopsy) with respect to brain microstructure. Brain Informatics 2019;6(1):3.
15. Lefevre T. Big data in forensic science and medicine. J Forensic Leg Med 2018;57:1–6.
16. Legrand L, Delabarde T, Souillard-Scemama R, Sec I, Plu I, Laborie JM, et al. Comparison between postmortem computed tomography and autopsy in the detection of traumatic head injuries. J Neuroradiol 2020;47(1):5–12.
17. Schmitt-Sody M, Kurz S, Reiser M, Kanz KG, Kirchhoff C, Peschel O, et al. Analysis of death in major trauma: value of prompt post mortem computed tomography (pmCT) in comparison to office hour autopsy. Scand J Trauma Resusc Emerg Med 2016;24:38. https://doi.org/10.1186/s13049-016-0231-6.
18. Keras CF. GitHub repository, 2015. https://keras.io (accessed May 20, 2020).
19. Prevedello LM, Erdal BS, Ryu JL, Little KJ, Demirer M, Qian S, et al. Automated critical test findings identification and online notification system using artificial intelligence in imaging. Radiology 2017;285(3):923–31.
20. Shirota G, Gonoi W, Ikemura M, Ishida M, Shintani Y, Abe H, et al. The pseudo-SAH sign: an imaging pitfall in postmortem computed tomography. Int J Legal Med 2017;131(6):1647–53.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14521
Available online at: onlinelibrary.wiley.com

PAPER

PATHOLOGY/BIOLOGY

Milad Webb,[1] M.D., Ph.D.; Sarah S. Sherman,[2] Ph.D.; LokMan Sung,[1] M.D.; Carl J. Schmidt,[1] M.D.; and Leigh Hlavaty,[1] M.D.

Abusive Pediatric Thoracolumbar Fracture Due to Forced Hyperextension: Case Report, Biomechanical Considerations, and Review of the Literature

ABSTRACT: Pediatric thoracolumbar fractures are rare due to the physiological differences which afford greater resilience to the immature spine. Most pediatric thoracolumbar fractures occur as the result of high energy trauma, such as motor vehicle accidents, and modes of reasonable accidental injuries are limited by age and developmental capabilities of the child. These fractures can occur as the result of inflicted blunt force trauma and child abuse, and in most cases, the mechanism of injury to the spine is not known. We report the death of a 29-month-old man due to blunt force trauma to the back and forced hyperextension of the thoracolumbar spine causing fracture of the fourth lumbar (L4) vertebral body. A complete forensic examination revealed a previous healing fracture of the anterior aspect of the L4 vertebral body, with acute disruption of the anterior longitudinal ligament overlying the fracture site, complete fracture of the vertebral body, and fatal retroperitoneal hemorrhage. We present a review of the biomechanical considerations of the pediatric spine, a survey of pediatric spinal fractures, and a review of the literature on pediatric abusive thoracolumbar fractures. In this case, there was never a provided explanation for how the injury occurred; however, understanding the biomechanics of the pediatric spine allowed for the determination of the mechanism, force required to produce this specific pattern of abusive spinal injury, and the manner of death.

KEYWORDS: lumbar fracture, biomechanics, child abuse, hyperextension, autopsy, forensic pathology

Thoracolumbar fractures are rare in children and represent less than 3% of the traumatic pediatric injuries reported in the literature (1–3). The low incidence of these fractures is explained by the increased flexibility of the pediatric spine. The spine in young children has increased elasticity of the ligaments, increased water content in the intervertebral disks, and greater capacity for growth and remodeling than the spine of an older child, adolescent, or adult. The spines in young children can thus withstand and dissipate greater forces (4–6).

The most common causes of fractures of the thoracolumbar spine in children involve various forms of high energy trauma. They are most commonly due to motor vehicle collisions at all ages, and falls and sports-related injuries in school-aged children and adolescents (1,5). In toddler-aged children, accidental traumas typically lack the force necessary to result in thoracolumbar fractures.

Due to the rarity of these injuries and lack of supporting literature, abusive spinal trauma is difficult to establish in the medicolegal setting. Defense counsels often allude to rare pediatric medical conditions (e.g., rickets), infections (e.g., Kingella), and congenital abnormalities (e.g., osteogenesis imperfecta) to obfuscate inflicted trauma (5,7,8). Although radiology and clinical medicine may experience significant challenges in differentiating certain pathological fractures from inflicted injury (7,9), postmortem examination and the direct visualization of these injuries with histologic confirmation provides significant insight to confirm abusive injuries (10).

Child abuse is an uncommon cause of thoracolumbar fractures, and in cases of abusive spinal fractures, the mechanism of injury is often unknown (11). Reports of such cases are of significant medicolegal value to establish a known mechanism and pattern of injury and to elucidate the prevalence of inflicted pediatric spinal trauma. We report a case of fatal child abuse due to a lumbar fracture from forced hyperextension of the spine, with discussion of the biomechanics of the pediatric lumbar spine and a review of the current literature.

Case Report

The decedent was a 29-month-old man who was found unresponsive at home. The caregiver reported to emergency medical personnel and police that the child had choked, evidenced by baby wipes in the child's proximity and in his mouth. The caregiver and other family members reported no known significant medical history for the child, and the family stated that the child had been acting normally and in his usual state of health in the weeks preceding. The child arrived at the hospital unresponsive,

[1] Wayne County Medical Examiner's Office, Michigan Medicine, 1300 Warren Avenue, Detroit, Michigan, 48207.

[2] Exponent, Farmington Hills, Michigan, 48331.

Corresponding author: Milad Webb, M.D., Ph.D. E-mail: webbm@hillsboroughcounty.org

Received 23 Mar. 2020; and in revised form 20 May 2020, 30 June 2020; accepted 6 July 2020.

with severe pallor, and a mildly distended abdomen. There were no foreign materials identified in the mouth or airways. After extensive resuscitation efforts, he was pronounced dead approximately two hours after arrival. No ante-mortem or postmortem radiographs were taken at the hospital.

The child was taken into the custody of the medical examiner for postmortem examination. Postmortem radiographs were taken prior to autopsy. The body was that of a normally developed male toddler who was below the lowest percentile for weight for age (<1%) and 49th percentile for length for the stated age. External examination revealed no significant injuries; however, there was obvious loss of lumbar lordosis with flattening of the buttocks (Fig. 1). Reflection of the skin revealed two purple subcutaneous hemorrhages that each measured 1/2 inch in diameter in the left lower back.

Internal examination revealed significant injury to the lumbar spine (Fig. 2). There was a complete fracture of the 4th lumbar (L4) vertebral body and rupture of the anterior longitudinal ligament. The anterior longitudinal ligament was disrupted in the proximity of the fracture site and replaced by a 4 × 2-1/2 × 1/8 inch layer of granulation tissue and soft callus formation. The presence of granulation tissue and soft callus indicated remote injury with superimposed acute re-injury. The initial, healing fracture was presumed to be incomplete; however, the extent of this previous injury was unknown due to the re-injury at the same location. The acute fracture extended completely through the L4 vertebral body and caused a 1 × ½ inch defect in the granulation tissue overlying the spine.

The caudal region of the spinal cord in the proximity of the fracture site was intact but contused, and the sacral spinal canal was distended with both acute and resolving hemorrhage. There was extensive acute bilateral retroperitoneal hemorrhage and acute hemorrhage into the paraspinal, back, and pelvic soft tissues. There were acute contusions on the posterior upper lobe of the left lung, posterior bases of the lungs, bilaterally, and posterior diaphragm, bilaterally (Fig. 3). There were acute serosal hemorrhages on the transverse colon and loops of small bowel, acute hemorrhage into the mesentery, and multifocal acute adventitial hemorrhages on the aorta.

The spinal column and spinal cord were removed en bloc from the level of T12 through the sacrum and fixed in formalin prior to microscopic sampling. Microscopic examination of the L4 fracture showed a fracture and acute hemorrhage with inflammation and induction (hallmarked by osteolytic activity and deposition of granulation tissue, fibroblastic proliferation, and new osteoid deposition), as well as soft callus formation (hallmarked by osteoblastic/chrondoblastic activity, woven bone deposition, and cartilaginous nodules; Fig. 4). Indications of hard callus formation were not identified (i.e., periosteal and endosteal lamellar bone formation). New bone deposition is a key indicator of soft callus formation signifying approximately 14 days of healing postinitial injury. There was no evidence of lamellar bone formation and calcification which would indicate hard callus formation (3–4 weeks of healing) (12,13). Examination of the granulation tissue and soft callus overlaying the fracture site showed proliferation of fibroblasts, angiogenesis, mixed mononuclear, and neutrophilic inflammatory infiltration in a loose fibrotic extracellular matrix and superimposed acute hemorrhage. Examination of the spinal cord at the level and below the level of fracture showed acute hemorrhage as well as clot resorption with hemosiderin deposition and histiocytic activity.

Distal to the fracture, the marrow compartment showed normal, trilineage hematopoiesis, and normal trabecular and cortical bone. There was no evidence of bony dysplasia, infection, or metabolic derangement.

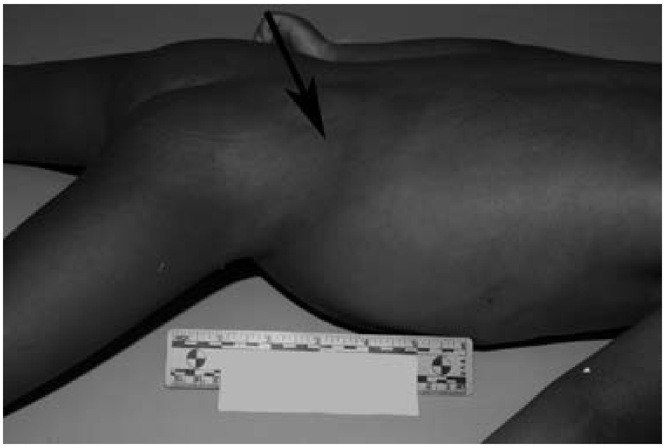


FIG. 1—*Loss of lumbar lordosis. Flattening of the pelvis and buttocks indicating structural compromise of the thoracolumbar spine. The buttocks had congenital dermal melanocytosis. There was a contusion on the right low-back (arrow). [Color figure can be viewed at wileyonlinelibrary.com]*

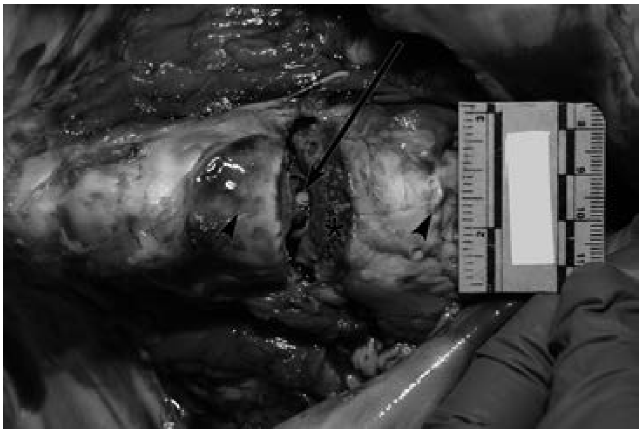


FIG. 2—*Complete fracture of the 4th lumbar (L4) vertebral body. Fracture of the L4 vertebral body with rupture of the anterior longitudinal ligament (arrowhead), granulation tissue and soft callus formation (*), and exposed spinal cord (arrow). [Color figure can be viewed at wileyonlinelibrary.com]*

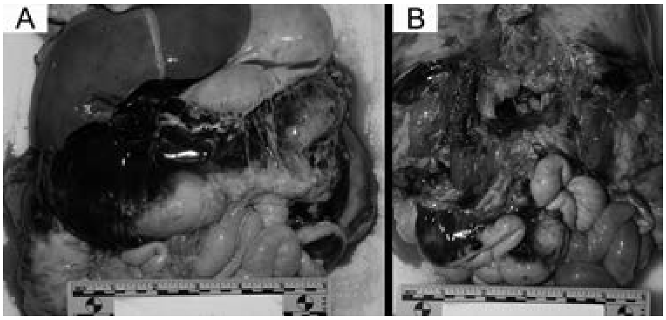


FIG. 3—*Abdominal and retroperitoneal hemorrhage associated with lumbar fracture. (A) Anterior view of the organ block. (B) Posterior view. [Color figure can be viewed at wileyonlinelibrary.com]*

Based on histologic observations, it was estimated that remote (initial) injury to the lumbar spine most likely occurred approximately 14 days prior to the acute re-injury and subsequent death of the child. Acute hemorrhage at the site of the fracture indicated acute re-injury being only hours old.

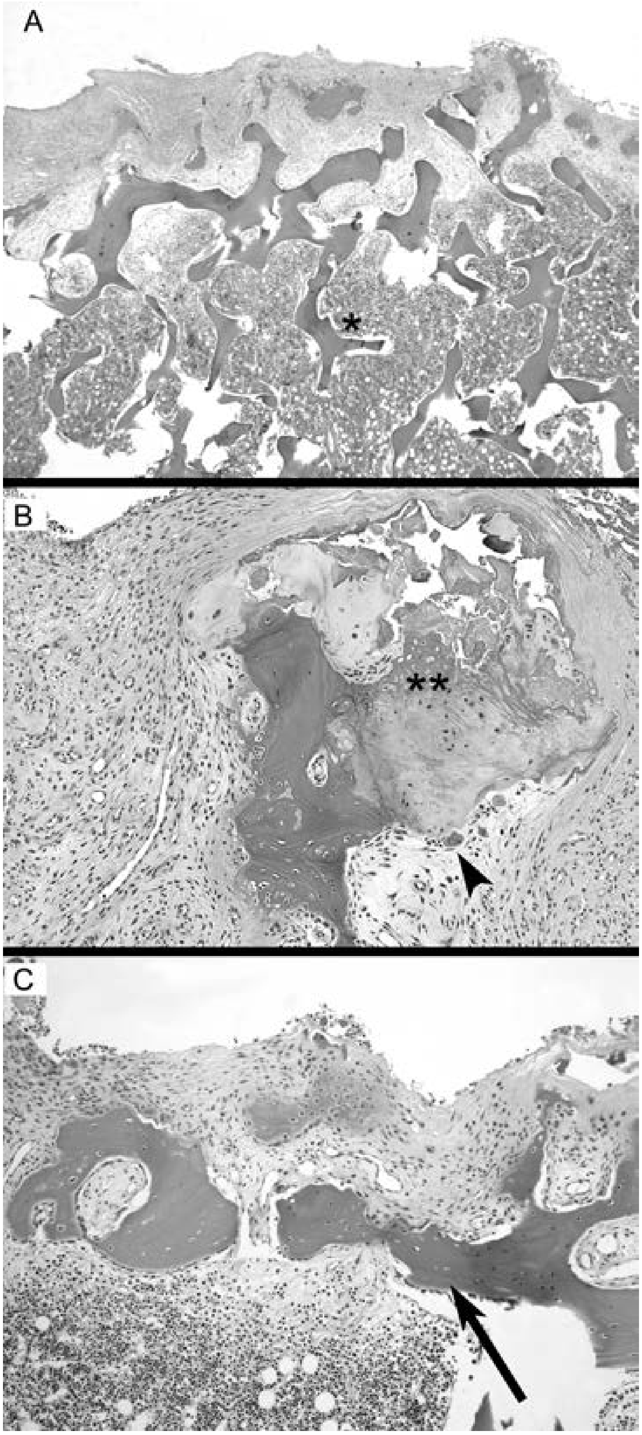


FIG. 4—*Lumbar fracture showing soft callus formation with acute re-injury. (A) Fracture site with multiple fragments of trabecular bone (*) in a loose fibrotic matrix (40× magnification). (B) Inflammation and induction of fracture healing showing mixed inflammation and osteoclasts (arrowhead) removing necrotic bone and soft tissue (100× magnification). There was soft callus formation (**). (C) New bone deposition (arrow, 100× magnification). [Color figure can be viewed at wileyonlinelibrary.com]*

The brain with attached cervical spinal cord and the eyes were retained and fixed in formalin prior to dissection with no additional injuries present. The child exhibited a failure to thrive due to exceptionally low body weight. Failure to thrive in an infant or toddler is defined as the lack of expected normal physical growth or failure to gain weight. Although it can be a result of malabsorptive and existing congenital or chronic medical conditions, it is also a well-documented stigmata of child abuse and neglect (14). The autopsy did not reveal any natural/medical diseases that caused or contributed to death and postmortem toxicology was noncontributory. The cause of death was certified as blunt force trauma to the back and the manner of death was ruled homicide. Additional police investigation did not reveal any information from the primary caregiver or other family members regarding any changes in the child's behavior or any explanation as to how the child sustained the injuries.

Discussion

Biomechanics of the Lumbar Spine

The human spine is comprised of 24 articulating vertebrae and nine fused vertebrae (five in the sacrum and four in the coccyx). The articulating vertebrae are interconnected by fibrous and cartilaginous intervertebral disks, articular facets, ligamentous fibrous tissues, and musculature. The vertebrae consist of the vertebral body (located anteriorly), vertebral arch, laminae, facet joints, spinous process, and transverse process. In this analysis, biomechanics of the spine are best discussed within the framework of the three-column system (Fig. 5) where the anterior

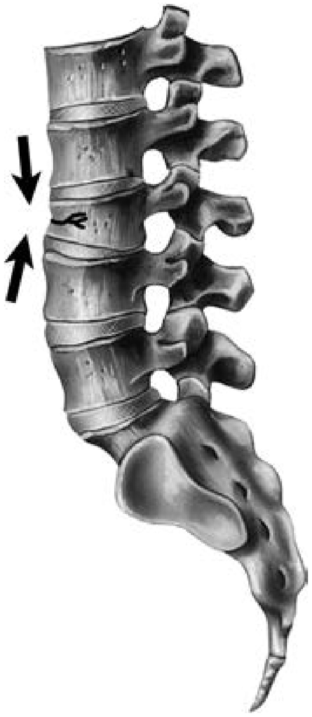


FIG. 5—*Three-column spine concept depicting compression fracture of the thoracolumbar spine. Colors depict the columns of the spine (green, anterior; red, middle; and blue, posterior). Although the uninvolved middle column is pathognomonic, severe wedge fractures can cause middle column injury. [Color figure can be viewed at wileyonlinelibrary.com]*

TABLE 1—*Fracture types and resultant column injuries.*

Type of Fracture	Anterior	Middle	Posterior
Compression (Wedge)	Compression	None	None or distraction (severe)
Burst	Compression	Compression	None
Flexion-distraction (Seat-belt-type)	None or minor compression	Distraction	Distraction
Fracture–dislocation (Shear)	Compression, rotation, shear		
Hyperextension	Distraction	Distraction/Compression	None or compression (severe)

column is comprised of the anterior longitudinal ligament, the anterior annulus fibrosus, and the anterior portion of the vertebral body. The middle column is formed by the posterior longitudinal ligament, the posterior annulus fibrosus, and the posterior wall of the vertebral body; the posterior column is formed by the bony posterior arch and the ligamentous complex (supraspinous ligament, interspinous ligament, capsule, and ligamentum flavum). This model was initially developed for the classification of thoracolumbar spinal fractures by Denis (15). It is widely used in routine clinical practice and medicolegal settings with the expectation that those involved will be familiar with the concept of anterior, middle, and posterior columns. Spinal instability occurs when two contiguous columns are injured. Using this model, fractures of the spine are described by the mechanism of failure of each column (Table 1).

Pediatric spines are significantly different when compared to their adult counterparts (4,16). Children have significant resilience to trauma due to the amount of cartilaginous tissue present within the spinal column, increased water content of the intervertebral disks, and increased laxity and elasticity (2,3,17–22). In addition, the joint facets are smaller and more horizontally oriented resulting in less overall stability (1,23). For these reasons, traumatic spine injuries are rare in children (3), and even among large regional spinal trauma centers, pediatric spinal injuries account for <5% of admission (5,6,20,23,24). Because of the biomechanical differences, pediatric spines have fewer fractures and more incidence of spinal cord injury without radiographic abnormality (SCIWORA) (5,25,26). Of those with pediatric spinal injuries, children under the age of 5 have a higher incidence (5) and up to 75% of injuries are of the thoracolumbar spine. The mechanism of injury is predominantly due to motor vehicle collisions, but also has associations with falls from height, sports, and abuse (1,3,23,27,28). The falls are not simple short distance falls onto a flat surface; they involve either a component of height (>4 m), such as out of a tree, acceleration from playground equipment, or onto an uneven surface such as stairs (27).

The fundamental biomechanical functions of the spine are true for both pediatrics and adults and include flexibility and bending mobility; load bearing and transfer during mobility; and protection of the spinal cord (29,30). The five lumbar vertebrae are the largest and provide the greatest axial stability (31). There are three natural curvatures of the spine, as viewed from the sagittal plane, that give the spinal column increased flexibility and absorption of loading while maintaining adequate stiffness and stability (30). The spine is aligned with a convex curvature anteriorly in the cervical and lumbar regions, clinically known as lordosis. The posterior curvature of the thoracic region is clinically known as kyphosis. Flexibility and bending mobility of the spine result from the articulation of the adjacent vertebrae where soft tissue facet joints guide and limit motion in flexion, extension, axial rotation, and lateral bending (29).

Stability of the spine is maintained by three mechanisms: the musculoskeletal system, the spinal column, and the actions of the nervous system (32). When axial, bending, and/or shear force is applied, a functional unit of the spine (which is composed of a vertebra, intervertebral disk, and its associated musculature and ligaments) will displace from a neutral position to a position where resistance and strain become appreciable (32). In general, the spine has a nonlinear, elastic behavior (32,33). At smaller forces, there is a relatively large displacement where the spine easily deforms with little resistance. At larger forces, there is less displacement with increased resistance (30,32). Additionally, the spine is viscoelastic, meaning the mechanical behavior of the spine varies with loading speed (30). Once the elastic limit of the tissue is met (the physiological limit), any additional strain will cause them to fail and result in permanent deformation (i.e., injury) (31).

Injuries to the lumbar spine are the result of acceleration or applied loads, directly to the region of injury or indirectly, causing the spine to flex (lateral or anterior), rotate, extend, shear, and/or axial displacement (1,29,30,33). There are three primary mechanisms of spinal injury in children—flexion with or without compression, distraction, and shear. These mechanisms cause four distinct types of fractures: compression, burst, flexion-distractions, and fracture–dislocations (5,6,15,29). Although the pediatric injury biomechanics literature related to lumbar spine and the mechanism of pediatric lumbar spine injuries is limited, a review of spinal biomechanics literature was conducted to characterize the mechanism by which spinal fractures occur.

Compression fractures, resulting from axial loading, are the most common fractures of the lumbar spine and occur frequently as a result of falling from height or motor vehicle collision (1,22). Compression fractures are characterized by failure of the anterior column and preservation of the middle column, which acts as a hinge, leaving the radiologic profile of a wedge (Fig. 5). The uninvolved middle column is pathognomonic of this fracture. The posterior column is typically spared, but may suffer partial distraction in severe injuries. There are multiple subtypes of compression fractures dependent on the directionality of flexion (anterior vs lateral). The elastic nature of the pediatric posterior column is vulnerable to compression trauma as well.

When compression force is applied through the center of rotation, it results in a burst fracture of the vertebral body. The key difference between compression fractures and burst fractures is that burst fractures do not result in flexion of the spine, but instead the force is transferred directly onto the vertebral body. Burst fractures are due to uniform compression and cause loss of vertebral height of the anterior and middle columns of the vertebra (Fig. 6). This can cause neurological deficits from bony fragments entering the spinal canal. There are multiple subtypes of burst fractures dependent on the degree of offset of the axial load from the center of rotation (causing anterior or lateral flexion) (15). Burst fractures occur commonly in adults; however, they are relatively rare injuries in children due to the greater

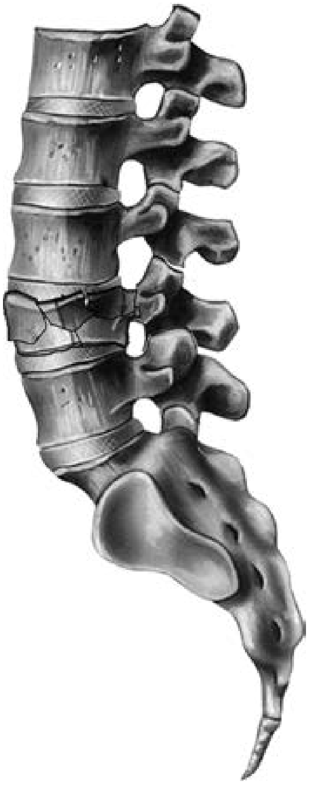


FIG. 6—*Burst fracture of the thoracolumbar spine. Anterior and middle column fractures due to axial force through the center of rotation. [Color figure can be viewed at wileyonlinelibrary.com]*

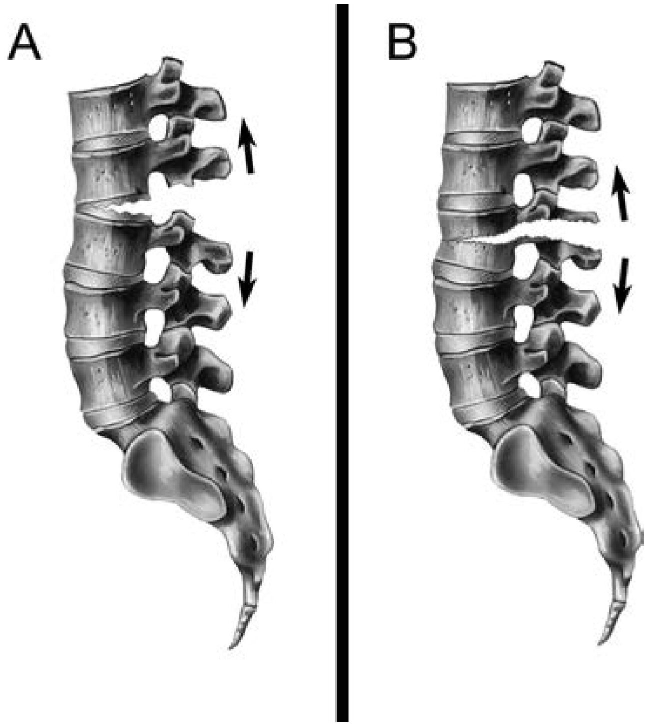


FIG. 7—*Flexion-distraction fracture of the thoracolumbar spine. (A) Distraction of the spine is defined as failure of the posterior and middle columns due to violent anterior flexion. (B) The Chance fracture is one subtype of this injury. [Color figure can be viewed at wileyonlinelibrary.com]*

flexibility, thicker periosteum, and the dense annular fibers of the disk structures allowing greater distribution of forces without disruption of the bony structures (34,35).

Flexion-distraction fractures of the spine have been classically referred to as "seat-belt-type injury" because partially or improperly restrained occupants (including children) are prone to flexion-distraction motion of the spine, especially at the lumbar spine level (Fig. 7) (36). Distraction of the spine is defined by failure of the posterior and middle columns under tension force generated by violent anterior flexion (5,36). The chance fracture describes one subtype of this injury where the fracture is propagated through the bony components of the posterior column and the most anterior portion of the anterior column functions as a hinge and may also suffer anterior column compression injury as a result of hyperflexion (37).

Fracture–dislocation injuries can occur under compression, tension, rotation, or shear force (Fig. 8) (6,15,29). The fundamental difference between fracture–dislocation and essential compression fracture is the rupture of the interspinous ligaments allowing dislocation of the vertebra. Fracture–dislocation refers to failure of all three columns (29). There are multiple subtypes dependent on the directionality of shear and position of locking facets. This injury tends to be associated with high probability of severe neurological complications due to translation of force to the spinal cord.

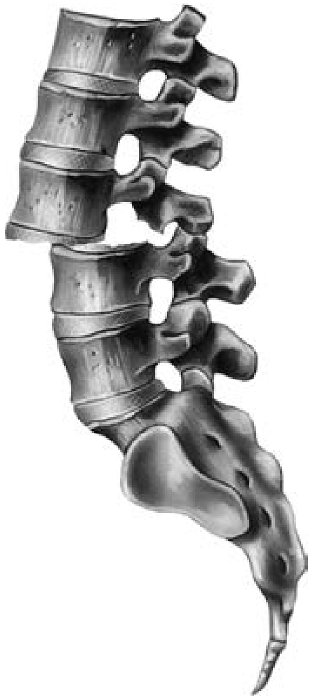


FIG. 8—*Fracture–dislocation of the thoracolumbar spine. Occurs though a myriad of mechanisms, including compression, tension, rotation, or shear (shown). Characterized by failure of all three columns. [Color figure can be viewed at wileyonlinelibrary.com]*

Lumbar Fractures and Child Abuse

Child abuse is an uncommon cause of thoracolumbar fractures and represents a minority of all pediatric thoracolumbar fractures (3,19,38–40). Our review of the literature revealed only 41 cases of lumbar or thoracolumbar fractures in infants and toddlers from child abuse, as reported in 16 papers (11,18,19,21,38,40–48). These fractures are commonly located from T11 to L3.

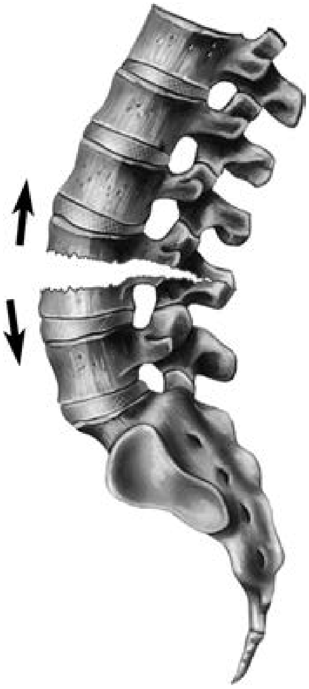


FIG. 9—*Hyperextension fracture of the thoracolumbar spine. Overstretching of the spine causing rupture of the anterior longitudinal ligament and anterior-to-posterior fracture of the vertebral body. [Color figure can be viewed at wileyonlinelibrary.com]*

They are more common in children under two years of age, and the actual number of cases is widely acknowledged to be higher due to underreporting and difficulty in diagnosing such injuries in the very young.

Vertebral body fractures are generally considered to have moderate specificity for child abuse. These fractures are felt to be highly specific for child abuse if the injury is unexplained or inconsistent with the history, or if the injury is not consistent with the child's developmental age and abilities (3,18,20,38, 40,49). It is also considered highly specific for child abuse when the child is under two years of age (50).

In most of the reported cases of abusive thoracolumbar fractures, the mechanism of fracture is unknown primarily due to the lack of causal information and self-preservation of involved parties. When thoracolumbar fractures are caused by child abuse, there is a greater incidence of concomitant skeletal and visceral injuries and death (17,23,27,51). One of the common life-threatening injuries associated with thoracolumbar vertebral body fractures is laceration or transection of the abdominal aorta, as it directly overlies the spine until its distal bifurcation into the common iliac arteries. The aorta can be injured through direct blunt force trauma applied anteriorly from the abdomen or indirect blunt force trauma applied posteriorly from the back (18,44,52).

Hyperextension injuries of the thoracolumbar spine are rare; acute hyperextension fracture–dislocation injuries are exceedingly rare, reported in only a few case reports. It can occur when there is a forceful blow to the face/forehead, forceful blow to the lower back with the abdomen against a fixed object like a wall, or when there is substantial force to the lower back applied over a fulcrum, such as a knee or arm of a sofa (18–20,45). This exaggerated overstretching of the lower spine is characterized by rupture of the anterior longitudinal ligament and fracture of the posterior column (Fig. 9). This injury was previously known as "lumberjack paraplegia" due to its occurrence among timber workers being hit by falling trees (37,53). Because the anterior spinal column is, in part, supported by significant musculature (e.g., muscles of the abdominal wall), these injuries require substantial force typically not seen in low energy trauma. Spinal cord injuries are common with hyperextension fractures, as the increased flexibility of the pediatric vertebral column leads to relative fixation and functional shortening of the spinal cord that renders it more vulnerable to injury (19).

Review of current literature only yielded two cases of abusive lumbar fracture due to hyperextension of the spine in a child, and both were fatal. Lieberman (2010) reported the death of a 3-1/2 year old with complete fracture–dislocation at level of the L2/L3. Histologic evaluation of the fracture site showed acute and chronic changes indicating repetitive injury. Dudley (2014) reported the death of a 30-month old with fracture of the vertebral body at L1/L2. Both cases were associated with disruption of the anterior longitudinal ligament, transection of the abdominal aorta, and massive retroperitoneal hemorrhage.

Two additional cases of lumbar fractures in young children who survived their injuries were reported. In these reports, the authors speculated that hyperextension of the spine may have been the mechanism of injury, but there is no mention of rupture of the anterior longitudinal ligament or other substantive evidence in the reports to confirm their hypotheses. These reports were of a 16 month old with fracture of the L3 vertebral body and transection of the abdominal aorta (44), and a 15 month old with fracture–dislocation of the L1/L2 vertebral body, soft tissue hemorrhage, and spinal cord compression (33).

Conclusion

We report the death of a 29-month-old man due to blunt force trauma to the back-causing hyperextension fracture of the L4 vertebral body. The most likely mechanism was forced hyperextension of the lumbar spine over a fulcrum causing anterior to posterior propagation of the fracture through the L4 vertebral body. Modes of reasonable accidental injuries must be limited not only by the age of the individual but also by their specific set of developmental capabilities. Injuries which appear out of proportion to the described mechanism should be meticulously scrutinized.

Understanding of the biomechanics of the pediatric spine is vital to reconstruction of the necessary mechanism and force required to produce the specific pattern of spinal injury.

References

1. Sayama C, Chen T, Trost G, Jea A. A review of pediatric lumbar spine trauma. Neurosurg Focus 2014;37(1):E6. https://doi.org/10.3171/2014.5.FOCUS1490.
2. Vialle LR, Vialle E. Pediatric spine injuries. Injury 2005;36(2):S104–12. https://doi.org/10.1016/j.injury.2005.06.021.
3. Akbarnia BA. Pediatric spine fractures. Orthop Clin North Am 1999;30 (3):521–36. https://doi.org/10.1016/S0030-5898(05)70103-6.
4. Yoganandan N, Stemper B, Kaufman B, Pintar F. Biomechanics of the pediatric spine. In: Kim DH, Betz RR, Huhn SL, Newton PO, editors. Surgery of the pediatric spine. Stuttgart, Germany: Georg Thieme Verlag, 2008;11–22. https://doi.org/10.1055/b-0034-72563.
5. Clements DH, McCarthy KP. Pediatric thoracolumbar spine trauma. In: Kim DH, Betz RR, Huhn SL, Newton PO, editors. Surgery of the pediatric spine. Stuttgart, Germany: Georg Thieme Verlag, 2008;501–7. https://doi.org/10.1055/b-002-72236.
6. Clark P, Letts M. Trauma to the thoracic and lumbar spine in the adolescent. Can J Surg 2001;44(5):337–45.
7. Paddock M, Sprigg A, Offiah AC. Imaging and reporting considerations for suspected physical abuse (non-accidental injury) in infants and young

children. Part 2: axial skeleton and differential diagnoses. Clin Radiol 2017;72(3):189–201. https://doi.org/10.1016/j.crad.2016.11.015.

8. Klotzbach H, Delling G, Richter E, Sperhake JP, Püschel K. Post-mortem diagnosis and age estimation of infants' fractures. Int J Legal Med 2003;117(2):82–9. https://doi.org/10.1007/s00414-002-0338-3.
9. Paddock M, Sprigg A, Offiah AC. Imaging and reporting considerations for suspected physical abuse (non-accidental injury) in infants and young children. Part 1: initial considerations and appendicular skeleton. Clin Radiol 2017;72(3):179–88. https://doi.org/10.1016/j.crad.2016.11.016.
10. Raynor E, Konala P, Freemont A. The detection of significant fractures in suspected infant abuse. J Forensic Leg Med 2018;60:9–14. https://doi.org/10.1016/j.jflm.2018.09.002.
11. Levin TL, Berdon WE, Cassell I, Blitman NM. Thoracolumbar fracture with listhesis – an uncommon manifestation of child abuse. Pediatr Radiol 2003;33(5):305–10. https://doi.org/10.1007/s00247-002-0857-6.
12. Mindell E, Robbard S, Kwasman B. Chondrogenesis in bone repair. Clin Orthop Relat Res 1971;79:187–96. https://doi.org/10.1097/00003086-197109000-00026.
13. Reith J. Bone and joints. In: Goldblum J, Lamps L, McKenney J, Myers J, editors. Rosai and Ackerman's surgical pathology, 11th edn. London, U.K.: Elsevier, 2018;1740–809.
14. Block RW, Krebs NF, Hibbard RA, Jenny C, Kellogg ND, Spivak BS, et al. Failure to thrive as a manifestation of child neglect. Pediatrics 2005;116(5):1234–7. https://doi.org/10.1542/peds.2005-2032.
15. Denis F. The three column spine and its significance in the classification of acute thoracolumbar spinal injuries. Spine 1983;8(8):817–31. https://doi.org/10.1097/00007632-198311000-00003.
16. Siminoski K, Lee K-C, Jen H, Warshawski R, Matzinger MA, Shenouda N, et al. Anatomical distribution of vertebral fractures: comparison of pediatric and adult spines. Osteoporos Int 2012;23(7):1999–2008. https://doi.org/10.1007/s00198-011-1837-1.
17. Cirak B, Ziegfeld S, Knight VM, Chang D, Avellino AM, Paidas CN. Spinal injuries in children. J Pediatr Surg 2004;39(4):607–12. https://doi.org/10.1016/j.jpedsurg.2003.12.011.
18. Dudley MH, Garg M. Fatal child abuse presenting with multiple vertebral and vascular trauma. J Forensic Sci 2014;59(2):386–9. https://doi.org/10.1111/1556-4029.12326.
19. Kemp AM, Joshi AH, Mann M, Tempest V, Liu A, Holden S, et al. What are the clinical and radiological characteristics of spinal injuries from physical abuse: a systematic review. Arch Dis Child 2010;95 (5):355–60. https://doi.org/10.1136/adc.2009.169110.
20. Muñiz AE, Liner S. Lumbar vertebral fractures in children. Pediatr Emerg Care 2011;27(12):1157–62. https://doi.org/10.1097/PEC.0b013e31823b009c.
21. Sieradzki JP, Sarwark JF. Thoracolumbar fracture-dislocation in child abuse: case report, closed reduction technique and review of the literature. Pediatr Neurosurg 2008;44(3):253–7. https://doi.org/10.1159/000121475.
22. Stemper BD, Pintar FA, Baisden JL. Lumbar spine injury biomechanics. In: Yoganandan N, Nahum AM, Melvin JW, editors. Accidental injury. New York, NY: Springer New York, 2015;451–70.
23. Daniels AH, Sobel AD, Eberson CP. Pediatric thoracolumbar spine trauma. J Am Acad Orthop Surg 2013;21(12):707–16. https://doi.org/10.5435/JAAOS-21-12-707.
24. Slotkin JR, Lu Y, Wood KB. Thoracolumbar spinal trauma in children. Neurosurg Clin N Am 2007;18(4):621–30. https://doi.org/10.1016/j.nec.2007.07.003.
25. Mortazavi MM, Mariwalla NR, Horn EM, Shane Tubbs R, Theodore N. Absence of MRI soft tissue abnormalities in severe spinal cord injury in children: case-based update. Child's Nerv Syst 2011;27(9):1369–73. https://doi.org/10.1007/s00381-011-1472-3.
26. Pang D, Wilberger JE. Spinal cord injury without radiographic abnormalities in children. J Neurosurg 1982;57(1):114–29. https://doi.org/10.3171/jns.1982.57.1.0114.
27. Kim C, Vassilyadi M, Forbes JK, Moroz NWP, Camacho A, Moroz PJ. Traumatic spinal injuries in children at a single level 1 pediatric trauma centre: report of a 23- year experience. Can J Surg 2016;59(3):205–12. https://doi.org/10.1503/cjs.014515.
28. Reddy SP, Junewick JJ, Backstrom JW. Distribution of spinal fractures in children: does age, mechanism of injury, or gender play a significant role? Pediatr Radiol 2003;33(11):776–81. https://doi.org/10.1007/s00247-003-1046-y.
29. King A. Injury to the thoracolumbar spine and pelvis. In: Nahum AM, Melvin JW, editors. Accidental injury: biomechanics and prevention. New York, NY: Springer New York, 2002;454–90.
30. White A, Panjabi M. Clinical biomechanics of the spine, 2nd edn. Philadelphia, PA: Lippincott-Raven, 1990.
31. Miele VJ, Panjabi MM, Benzel EC. Anatomy and biomechanics of the spinal column and cord. In: Verhaagen J, McDonald JW, editors. Handbook of clinical neurology. Amsterdam, the Netherlands: Elsevier BV, 2012;31–43. https://doi.org/10.1016/B978-0-444-52137-8.00002-4
32. Izzo R, Guarnieri G, Guglielmi G, Muto M. Biomechanics of the spine. Part I: spinal stability. Eur J Radiol 2013;82(1):118–26. https://doi.org/10.1016/j.ejrad.2012.07.024.
33. Gabos PG, Tuten HR, Leet A, Stanton RP. Fracture-dislocation of the lumbar spine in an abused child. Pediatrics 1998;101(3):473–6. https://doi.org/10.1542/peds.101.3.473.
34. Lalonde F, Letts M, Yang J, Thomas K. An analysis of burst fractures of the spine in adolescents. Am J Orthop 2001;30(2):115–20.
35. Vander Have KL, Caird MS, Gross S, Farley FA, Graziano GA, Stauff M, et al. Burst fractures of the thoracic and lumbar spine in children and adolescents. J Pediatr Orthop 2009;29(7):713–9. https://doi.org/10.1097/BPO.0b013e3181b76a44.
36. Arkader A, Warner WC, Tolo VT, Sponseller PD, Skaggs DL. Pediatric chance fractures: a multicenter perspective. J Pediatr Orthop 2011;31 (7):741–4. https://doi.org/10.1097/BPO.0b013e31822f1b0b.
37. Erb RE, Glassman SB, Edwards JR, Nance EP. Hyperextension fracture-dislocation of the thoracic spine. Emerg Radiol 1995;2(4):237–40. https://doi.org/10.1007/BF02615825.
38. Barber I, Perez-Rossello JM, Wilson CR, Silvera MV, Kleinman PK. Prevalence and relevance of pediatric spinal fractures in suspected child abuse. Pediatr Radiol 2013;43(11):1507–15. https://doi.org/10.1007/s00247-013-2726-x.
39. Bode KS, Newton PO. Pediatric nonaccidental trauma thoracolumbar fracture-dislocation. Spine 2007;32(14):E388–E393. https://doi.org/10.1097/BRS.0b013e318067dcad.
40. Diamond P, Hansen C, Christoferson M. Child abuse presenting as a thoracolumbar spinal fracture dislocation: a case report. Pediatr Emerg Care 1994;10(2):83–6. https://doi.org/10.1097/00006565-199404000-00005.
41. Carrion WV, Dormans JP, Drummond DS, Christofersen MR. Circumferential growth plate fracture of the thoracolumbar spine from child abuse. J Pediatr Orthop 1996;16(2):210–4. https://doi.org/10.1097/00004694-199603000-00015.
42. Cullen J. Spinal lesions in battered babies. J Bone Joint Surg Br 1975;57(3):364–6.
43. Faure C, Steadman C, Lalanda G, Al Moudares N, Marsault C, Bennet J. The wandering vertebral artery. Ann Radiol 1979;22:96–9.
44. Fox JT, Huang YC, Barcia PJ, Beresky RE, Olsen D. Blunt abdominal aortic transection in a child. J Trauma Inj Infect Crit Care 1996;41 (6):1051–3. https://doi.org/10.1097/00005373-199612000-00020.
45. Lieberman I, Chiasson D, Podichetty VK. Aortic disruption associated with L2–L3 fracture-dislocation in a case of child abuse. J Bone Joint Surg Am 2010;92(7):1670–4. https://doi.org/10.2106/JBJS.I.01404.
46. Renard M, Tridon P, Kuhnast M, Renauld JM, Dollfus P. Three unusual cases of spinal cord injury in childhood. Spinal Cord 1978;16(1):130–4. https://doi.org/10.1038/sc.1978.22.
47. Swischuk LE. Spine and spinal cord trauma in the battered child syndrome. Radiology 1969;92(4):733–8. https://doi.org/10.1148/92.4.733.
48. Tran B, Silvera M, Newton A, Kleinman PK. Inflicted T12 fracture-dislocation: CT/MRI correlation and mechanistic implications. Pediatr Radiol 2007;37(11):1171–3. https://doi.org/10.1007/s00247-007-0594-y.
49. Cramer KE. Orthopedic aspects of child abuse. Pediatr Clin North Am 1996;43(5):1035–51. https://doi.org/10.1016/S0031-3955(05)70449-1.
50. Jauregui JJ, Perfetti DC, Cautela FS, Frumberg DB, Naziri Q, Paulino CB. Spine injuries in child abuse. J Pediatr Orthop 2019;39(2):85–9. https://doi.org/10.1097/BPO.0000000000000877.
51. Leonard M, Sproule J, Mc Cormack D. Paediatric spinal trauma and associated injuries. Injury 2007;38(2):188–93. https://doi.org/10.1016/j.injury.2006.09.019.
52. Inaba K, Kirkpatrick AW, Finkelstein J, Murphy J, Brenneman FD, Boulanger BR, et al. Blunt abdominal aortic trauma in association with thoracolumbar spine fractures. Injury 2001;32(3):201–7. https://doi.org/10.1016/S0020-1383(00)00203-5.
53. Denis F, Burkus J. Shear fracture–dislocations of the thoracic and lumbar spine associated with forceful hyperextension (Lumberjack Paraplegia). Spine 1992;17(2):156–61. https://doi.org/10.1097/00007632-199202000-00007

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14523
Available online at: onlinelibrary.wiley.com

PAPER

PATHOLOGY/BIOLOGY

Asia Sampson,[1,†] B.S.; and Derek S. Sikes,[1,2,†] Ph.D.

A Preliminary Forensic Entomological Study of Beetles (Coleoptera) in Interior Alaska, USA*

ABSTRACT: Forensic entomology uses knowledge of arthropod ecology to help solve crimes. There has been no published forensic entomological research in Alaska. We used one piglet carcass split in half to create two carcass plots in Fairbanks (~64.8°N, subarctic) that were sampled over a period of 59 days in 2019. Four pitfall traps were placed around each carcass, and four similarly arranged pitfall traps were placed 40 m distant as controls. Traps were emptied approximately weekly covering the first four stages of decomposition. We focused on adults of the larger-bodied (>9 mm) families and subfamilies of Coleoptera: Staphylinidae (subfamily Staphylininae), Carabidae, and Silphidae. A total of 621 specimens were collected and processed: 29 staphylinines, 210 carabids, and 382 silphids. A one-way ANOVA showed no significant difference between the mean numbers of staphylinines or carabids caught in carcass versus control traps. Silphids showed significantly greater mean number of beetles caught in carcass traps relative to control traps. Four species of Silphidae were documented, but contrary to similar studies, the vast majority of specimens belonged to two species of *Nicrophorus* (*N. vespilloides* Herbst and *N. investigator* Zetterstedt). Each of the three target taxa showed a peak in the number of specimens collected during the bloat stage of decomposition despite the carabid peak being driven by a phytophagous species.

KEYWORDS: decomposition, subarctic, high latitude, forensic entomology, Coleoptera, Silphidae

Forensic entomology applies knowledge of arthropods to help investigators solve criminal cases (1). Insect evidence can help in identifying key suspects and narrow down estimates of the postmortem Interval. Alaska, the largest and northernmost US state, has a uniquely cold climate, acted in part as a glacial refugium, and has had historical biogeographic connections to Asia. These factors may have resulted in unique aspects of the insects associated with corpse decomposition. Studies on the decomposition of salmon carcasses (2–4), vertebrate scavenging on moose carcasses (5), and basic faunistics of carrion-associated insects such as blow flies and carrion beetles (6,7) have been conducted in Alaska. However, to the best of our knowledge there are no published forensic entomology studies that have been conducted in Alaska. There is therefore a lack of information on the insect species' diversity and timing of events associated with decomposing terrestrial carrion in Alaska. The northernmost forensic entomology study in North America was conducted in Whitehorse (60.7°N), Yukon Territory, Canada (8), approximately 820 km southeast of Fairbanks, Alaska. Weather conditions in interior Alaska can be extreme, winters are long, with temperatures dropping at times lower than −40°C, with summers, although short, consisting of long days with sunlight in midsummer lasting up to 21+ hours. These high-latitude subarctic conditions could potentially cause changes in the duration of decomposition stages and the succession of insects on a corpse.

Typically, five decomposition stages are associated with large corpses although different names for these stages are in use (8,9): fresh, bloat, active decay, advanced decay, and dry remains. Arthropods play a huge role in the rate of corpse decomposition and arrive sometimes only a few hours after the time of death and continue to arrive as the corpse decays until the final stage of decomposition ends (9). Carrion is known to attract many Diptera, primarily Calliphoridae (blow flies) and Sarcophagidae (flesh flies), and many studies are focused on these taxa but various families of Coleoptera are also key participants in corpse decomposition (1,10).

We focused on the larger-bodied families of Coleoptera (>9 mm) that would be most obvious to those investigating a corpse at a crime scene. We wanted to determine the strength of association of each taxon with carrion and during which decomposition stage each taxon was most abundant.

Methods

Study Site Conditions

This study was carried out in Fairbanks, Alaska (approximately 64.8°N latitude, Table 1), in the subarctic, within the boreal forest consisting primarily of a mix of white spruce (*Picea glauca* (Moench) Voss), Alaskan paper birch (*Betula neoalaskana* Sarg.), quaking aspen (*Populus tremuloides* Michx.), and balsam poplar (*Populus balsamifera* L.). Fairbanks is located at 131 m above sea level, and average daily temperatures during

[1] Department of Biology and Wildlife, University of Alaska Fairbanks, 2090 Koyukuk Dr., Fairbanks, Alaska, 99775, USA.

[2] Institute of Arctic Biology, University of Alaska Museum, University of Alaska Fairbanks, 1962 Yukon Dr., Fairbanks, Alaska, 99775, USA.

Corresponding author: Derek S. Sikes, Ph.D. E-mail: dssikes@alaska.edu

*Presented at the 2020 Annual Meeting of the Alaska Entomological Society, February 15, 2020, in Anchorage, AK.

[†] Authors contributed equally.

[Correction added 11 August 2020. The Highlights section was removed.]

Received 28 May 2020; and in revised form 25 June 2020; accepted 6 July 2020.

TABLE 1—*Latitude and longitude coordinates (WGS84 datum) of pig carcass and control plots where specimens were collected in Fairbanks, Alaska.*

CODE	Latitude	Longitude
P1	64.871305°N	147.863849°W
C1	64.871244°N	147.864722°W
P2	64.870865°N	147.865126°W
C2	64.870833°N	147.866389°W

our study ranged between 11.1°C (52°F) and 23.3°C (74°F) with an average of the daily temperature over the study period of 18.08°C (64.56°F) (11). Day length at the start of the study was 21 h and 49 min and by the end of the study had dropped to 16 h and 22 min.

Field Sampling

One previously deceased and frozen *Sus scrofa domesticus* of 8.2 kg donated to the University of Alaska Fairbanks Department of Veterinary Medicine by a local meat market was used to model human decomposition. No IACUC review process was needed since the pig was already dead before the start of the study. This relatively small-bodied pig would model an infant human. Pig carrion was used in this study to model human remains due to its similarities such as lack of heavy fur and similar body composition (12). The pig was thawed and cut in half. Each half was weighed and coded. The front half of the pig was coded P1, which included the head, front legs and upper abdomen, and weighed 4.2 kg. The rear half of the pig was coded P2, which included the lower half of the abdomen, hind legs, and rear, weighing 4.0 kg (Table 1).

The pig halves were transferred to a young experimentally managed *Populus balsamnifera* orchard on the University of Alaska Fairbanks campus "T-Field" surrounded by a >2 m metal fence to exclude large vertebrates. The pig halves were placed 40 m apart in metal, unlined dog cages (19″ × 12″ × 14″) which were secured to the ground with ground stakes to eliminate possible disturbance from smaller vertebrates such as ravens but allowed full access by insects. Ground vegetation was primarily grasses and forbs: *Chamaenerion angustifolium* (L.) Scop. and *Vicia cracca* L.

Collections of trap contents were made for a period of 59 days. The first trap sample was made after 3 days to correspond with the fresh stage of decomposition with weekly samples during the bloat and active decay stages and bi- or tri-weekly samples during the advanced decay stage (Table 2). Four pitfall traps, each a yellow plastic cup 10 cm diameter, containing nontoxic propylene glycol-based Sierra © brand antifreeze, were set into the ground around each side of the cages, within ~16 cm of the cages, in a North, South, East, West arrangement (Fig. 1). Each pitfall trap was covered with a metal pie pan held about 8 cm above the ground using ground stakes to eliminate rain. Control traps lacking pig carcasses were set into the ground in the same pattern about 40 m from each of the pig cages, in identical habitat, and were coded C1 and C2 (Table 1).

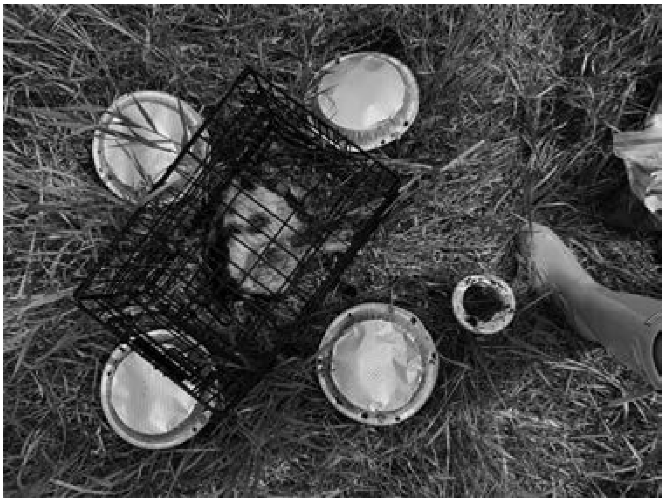


FIG. 1—*Four pitfall traps arranged around a cage holding half a piglet carcass. [Color figure can be viewed at wileyonlinelibrary.com]*

Pitfall trap contents were poured through a ~10 × 10 cm mosquito net filter cloth which drained excess propylene glycol into a new pitfall trap cup and held the specimens on the filter cloth. Samples were each labeled with the corresponding date and location code, and placed into Whirl-Pak © bags and frozen until sorting and preparation of the specimens. Trap periods are listed in Table 2. The four traps at each of the four plots were pooled for each sample period yielding 24 samples.

Unhealthy air conditions from thick smoke as a result of numerous forest fires in the area resulted in some skipped sample weeks during the advanced decay stage, reducing the total sample periods from 9 planned to 6 actual (Table 2). To handle these missed sample weeks for phenology graphing, we corrected the data by dividing the resulting catch by the number of weeks. For example, if 1 week was skipped, the resulting catch would include 2 weeks of accumulated specimens (eg. 28 specimens) which were then divided in half (e.g., 14 specimens for each of the 2 weeks). This procedure allowed a more accurate representation of abundance over time and a more accurate estimate of sample means for statistical analysis.

TABLE 2—*Trap periods in which 24 samples were collected at approximately weekly intervals with corresponding approximate decay stage and day length (hours:minutes) indicated.*

DATES (2019)	Decay Stage	Day Length†	P1	C1	P2	C2
1) June 19–21	Fresh	21:49	1	1	1	1
2) June 21–29	Bloat	21:41	1	1	1	1
3) June 29–July 5	Active decay	21:19	1	1	1	1
4) July 5–19	Advanced decay	20:20	1	1	1	1
5) July 19–August 9	Advanced decay	18:23	1	1	1	1
6) August 9–16	Advanced decay	16:46	1	1	1	1

†Daylength data from https://www.timeanddate.com/sun/usa/fairbanks

Laboratory Methods

During the sorting process, specimens were emptied into a U.S.A. standard test Sieve (0.0041 inch equivalent), rinsed with water, placed in 70% ethanol in a sorting tray and then examined under a Leica dissecting microscope, pinned or pointed and placed in unit trays to be identified and counted. Because of considerable debris in the trap samples sorting, a sample would often require multiple pours into a sorting tray to reduce the density of specimens and debris and increase the likelihood of finding all target taxa. Target taxa included adults of the following coleopteran families: Carabidae, Silphidae, and large-bodied (>9 mm) Staphylinidae. Silphidae were identified to species using the key in (7), Staphylinidae to genus using the keys in (13), and Carabidae to species using the keys in (14) and (15).

A one-way analysis of variance (ANOVA) with the factor being the four plot types (C1, C2, P1, P2) and the response variable being the mean number of beetles trapped was performed using R version 4.0.0 (16) in RStudio version 1.2.5042 (17). A Tukey multiple comparisons of means test was performed in R to determine if the replicates were significantly different (P1 vs. P2, and C1 versus C2) using the full Coleoptera data and they were not ($p < 0.92$), thus indicating each set of traps was sampling the same community in the same way. To determine if temperatures influenced our beetle counts, weekly mean temperature (°C) was used as a predictor variable in a regression in R with Coleoptera counts as the response variable. We used Pearson's chi-squared test with a Bonferroni correction at alpha = 0.05 to test the null hypothesis that each species would be equally abundant in each trap type, limiting our tests to species that had total counts greater than or equal to six.

Google Sheets and Microsoft Excel were used to prepare graphs of changing abundance over time. Voucher specimens of all target taxa have been archived in the University of Alaska Museum Insect Collection. Specimen data can be retrieved from the following URL http://arctos.database.museum/saved/forensic. Data used and R code for statistical analyses are available at https://doi.org/10.6084/m9.figshare.12330551.

Results

Over the 59-day collection period, four of the five decomposition stages were observed: fresh, bloat, active decay, and advanced decay, in which the study ended. Weekly average temperatures recorded in the Fairbanks area ranged from a low of 14.72°C (58.5°F) to a high of 21.46°C (70.64°F) and showed no relationship with beetle abundance (Fig. 2, $R^2 = 0.1524$, $F = 1.259$, $p = 0.2989$). A total of 621 Coleoptera in our focal taxa were collected and processed. The majority of specimens (382, 62%) were silphids. A minority of specimens (29, 4.7%) were large-bodied staphylinines, and 210 specimens (34%) were carabids. A one-way ANOVA showed no statistically significant difference between the mean number of staphylinines or carabids in carcass versus control traps ($p > 0.05$) (Table 3). The mean number of silphids and all Coleoptera combined was significantly greater ($p < 0.016$) in carcass versus control traps (Table 3). Each of the three target taxa showed a peak in the number of specimens collected during the second trap period beginning on 21 June 2019 and ending on 29 June 2019 which corresponds with the bloat stage of decomposition (Fig. 3). Nine species of carabids, four species of silphids, and four species of staphylinines were documented (Table 4). Of the 621 specimens collected, just three species (*Harpalus somnulentus*, *Nicrophorus investigator*, and *N. vespilloides*) represented 78% of the catch (Table 4). Nine species of beetles had total catches larger than or equal to six specimens allowing us to perform a chi-square test of the null hypothesis that they would be equally abundant in both carcass and control traps. Of those nine, six had significant values rejecting the null hypothesis (Table 4).

Discussion

Four of the five decomposition stages were observed, and the majority of specimens were caught during the first three stages (fresh, bloat, and active decay) with a peak in numbers during the bloat stage. Temperature did not covary with counts of captured Coleoptera (Fig. 2). This is despite the well-known relationship between temperature and insect activity. Temperature varied relatively little over the course of the study. Decomposition stage was a stronger predictor of beetle activity than temperature.

Carabid abundance peaked in both carcass and control traps suggesting carabid activity was likely not driven primarily by the presence of the carcasses. The majority of carabids (77%) belonged to the genera *Amara* and *Harpalus*, which are phytophagous as adults—feeding on seeds and pollen—unlike most carabids which are predators (14). Thus, their presence at carcass traps was likely due more to random chance than an attraction to the carrion. Only the most abundant carabid, *H. somnulentus*, was significantly more common in carcass traps than control traps (Table 4)—an observation that is hard to explain given this species is phytophagous. *Calathus ingratus* was the only species

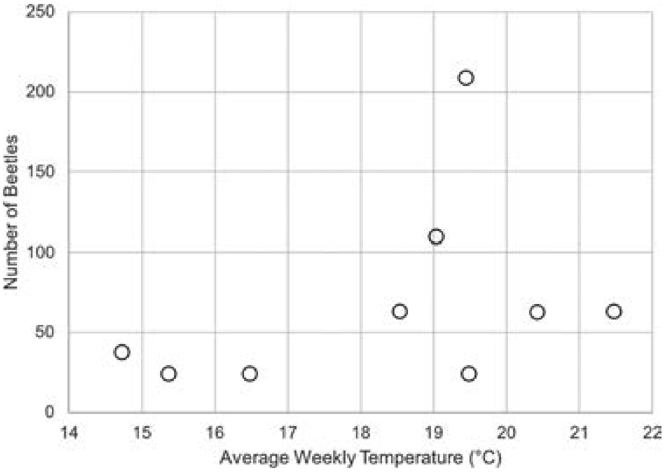


FIG. 2—*Nonsignificant relationship* (p = *0.2989) between Coleoptera counts and weekly mean temperatures in degrees C during the 59-day period from June 19, 2019, through August 16, 2019. Temperature data from The Alaska Climate Research Center (5).*

TABLE 3—*ANOVA results comparing mean number of beetles caught in carcass vs control traps by beetle taxon.*

taxon	*F* Value	*p* Value
Staphylinidae	1.283	0.297
Carabidae	0.326	0.806
Silphidae	6.049	0.0022*
Coleoptera	4.017	0.0156*

* indicates significant *p* values.

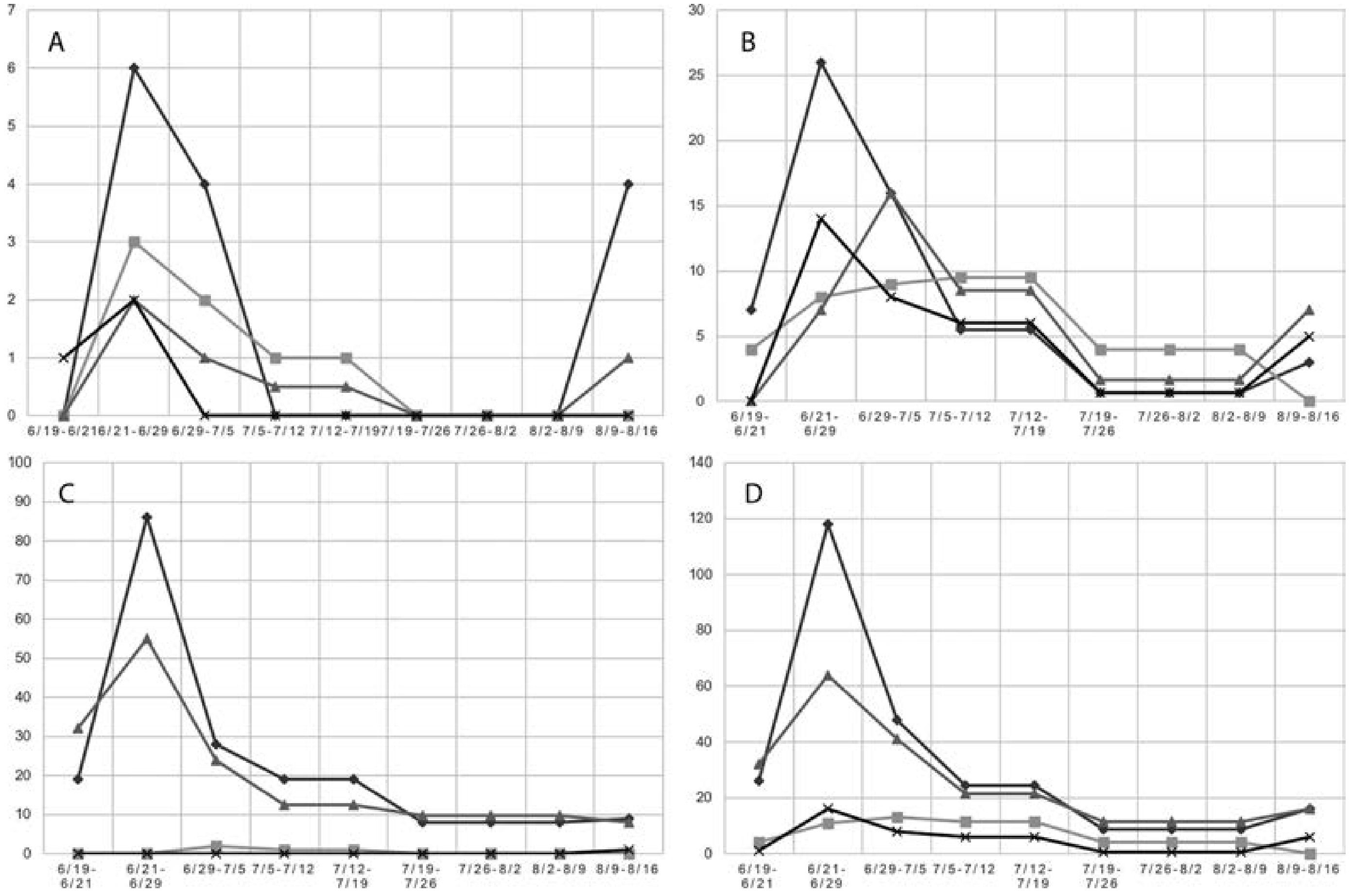


FIG. 3—*The number of large-bodied (A) Staphylininae, (B) Carabidae, (C) Silphidae, and (D) all Coleoptera specimens collected per trap period at two pig carcasses (P1-diamonds, P2-triangles) and two control sites (C1-squares, C2-x) during the 59-day period from June 19, 2019, through August 16, 2019. Beetle specimen count on* y*-axis, date on* x*-axis.*

in the entire study with significantly more specimens caught in control than carcass traps—as if these beetles were actively avoiding the carcasses (Table 4). This species also showed a significant negative association with carcass traps in Yellowstone National Park (18), so perhaps this species does actively avoid carcasses. Bornemissza's work in Australia (19) showed many

TABLE 4—*List of target taxon species trapped with number of specimens caught listed by trap sample and decay stage. See Table 2 for dates of sample periods. Sample period decay stages 1 = fresh, 2 = bloat, 3 = active decay, 4–6 = advanced decay. Specimen counts are presented as # of specimens at carcass traps/# of specimens at control traps; zeroes are indicated by dashes. Chi-square value for species with 6 or more total specimens, with df = 1 and associated* p *value with Bonferroni adjustment with significance at* p *value < 0.005.*

		Sample Period + Specimen Counts							
Family	Species	1	2	3	4	5	6	Chi-square	*p* Value
Carabidae	*Amara sinuosa* (Casey 1918)	–/1	–/–	–/–	–/1	–/2	–/–		
Carabidae	*Amara torrida* (Panzer, 1796)	–/1	–/2	–/2	–/1	–/–	–/–	6	0.0143
Carabidae	*Calathus ingratus* Dejean 1828	–/1	2/6	2/3	1/6	–/3	–/–	8.167	0.0043*
Carabidae	*Carabus chamissonis* Fischer 1820	–/–	–/–	–/–	–/3	1/4	5/1	0.286	0.5930
Carabidae	*Harpalus laticeps* LeConte 1850	–/–	–/–	–/–	1/–	–/–	–/–		
Carabidae	*Harpalus nigritarsis* Sahlberg 1827	1/–	–/–	1/–	1/1	–/–	–/–		
Carabidae	*Harpalus somnulentus* Dejean 1829	6/1	29/13	27/12	25/18	4/5	4/3	12.578	0.0004*
Carabidae	*Pterostichus adstrictus* Esch. 1823	–/–	1/–	2/–	–/–	2/–	1/1	3.571	0.0588
Carabidae	*Pterostichus* (*Cryobius*) sp.	–/–	1/1	–/–	–/1	–/–	–/–		
Silphidae	*Nicrophorus investigator* Zett. 1824	17/–	65/–	22/–	43/–	34/–	7/–	188	0.0000*
Silphidae	*Nicrophorus vespilloides* Herbst 1783	24/–	57/–	21/2	14/2	18/–	10/1	129.67	0.0000*
Silphidae	*Thanatophilus lapponicus* (Herbst)	8/–	19/–	9/–	6/–	1/–	–/–	43	0.0000*
Silphidae	*Thanatophilus sagax* (Mann. 1853)	2/–	–/–	–/–	–/–	–/–	–/–		
Staphylinidae	*Creophilus maxillosus* (Linn. 1758)	–/–	2/–	–/–	–/–	–/–	–/–		
Staphylinidae	*Dinothenarus capitatus* (Bland 1864)	–/–	–/–	–/–	1/–	–/–	–/–		
Staphylinidae	*Philonthus* sp.	–/–	3/–	4/–	–/–	–/–	5/–	12	0.0005*
Staphylinidae	*Quedius labradorensis* Smetana 1965	–/1	3/5	1/2	–/2	–/–	–/–	2.571	0.1088

* indicates significant values.

soil and litter dwelling taxa in the study region disappeared once decomposition of pig carcasses began and returned only long after the carcasses had entered the final stage of decomposition. This deterrent effect of carrion has received far less study than the attractive effect of carrion decay (20).

Carabid numbers peaked during sample periods 2-4 simultaneously with silphid and staphylinid numbers, corresponding to the bloat and active decay stages. Given there was no significant difference in the mean number of carabids in carcass versus control traps and no relation to weekly mean temperatures, perhaps the peak in numbers during the early sample periods resulted from normal seasonal population increases. The only other study to document seasonal activity of carabids in Alaska found no increase in numbers until September and October, near the end of the season (21). However, the dominant species of carabids sampled by Pantoja et al. (21) included no *Harpalus* species and *H. somnulentus* was by far the dominant carabid species driving the early season peak in numbers in our study. Unfortunately, there is no compelling explanation for the peak in carabid numbers coincident with silphid and staphylinid numbers. It is likely the decay of these pig carcasses would provide nutrient boosts to the plants adjacent to them. Numerous studies have shown carrion decomposition provides a flush of nutrients to the soil, creating islands across a landscape in which plant growth is increased (20,22–25). However, this effect on plants would not be immediate so it remains hard to explain the significantly greater catch of the phytophagous *H. somnulentus* in carcass traps so early in the decay process. If fertilization of nearby plants is part of explanation, important questions remain—do these beetles detect the carcass and are drawn toward it to feed on more nutrient-rich vegetation? This is an example with relevance to forensic investigations of how a taxon with no clear association with carcasses can appear to have an association. Only by use of control, noncarcass, sampling can the strength of association be tested. This also stresses the importance of identifying specimens at least to genus because by doing so one can access information on diet preference.

Staphylininae are species-rich predators and have been the focus of prior forensic entomological research (10) which documented habitat specificity and seasonality that could help indicate season of death and relocation of a corpse. Contrary to expectation, we found no significant difference in the mean numbers of Staphylininae captured in carcasses versus control traps. However, our sample size may have been too small with only 29 specimens. Three of the four staphylinine taxa we documented are well-known carrion and dung associates (*Creophilus*, *Dinothenarus*, and some species of *Philonthus* [13), and indeed, none of the specimens of these genera were caught in control traps. *Philonthus* sp. was the only staphylinine caught in large enough numbers to show a significant association with carcass traps (Table 4). The lack of a significant association with carcasses for the mean values of the family was caused by the species *Quedius labradorensis*, which was caught in relatively high numbers in control traps.

It is possible that the dense smoke that prevented us from collecting trap samples also reduced insect activity and accentuated the magnitude of the peak in numbers prior to the smoke, during the early sample periods. All focal taxa showed their lowest counts in late July and early August when the smoke was most intense and subsequently increased in numbers, albeit slightly, thereafter. However, the spike in beetle numbers at the start of the study with a subsequent decline is typical of carrion decomposition studies (e.g., 1,19).

Unsurprisingly, Silphidae were the dominant large-bodied Coleoptera at each of the two pig carcasses. However, surprisingly, nicrophorines far outnumbered silphines. The study in Yukon, Canada (8), found the silphine *Thanatophilus lapponicus* as their dominant silphid but did not provide specimen counts within species making it difficult to compare our results to theirs. They reported the silphid *Nicrophorus hybridus* Hatch & Angell as well, although presumably in lower numbers since they did not describe it to be a dominant species. We expect that they likely misidentified the *Nicrophorus* they collected because *N. hybridus* does not occur as far north as the Yukon (there are no verified records north of 54°N latitude)—it is more likely they had captured one of the species we documented, *N. investigator*, a close relative of *N. hybridus*, and well known to occur at high latitudes throughout the holarctic (7,26).

This dominance by silphines differs from our findings in which two *Nicrophorus* species were dominant with a total of 337 (88.2%) specimens compared with only 45 (11.8%) specimens of silphines. Our results are unusual because *Nicrophorus* are well-known specialists on small carcasses (<300 g, typically <100 g [27) that they can bury for reproduction and our pig carcasses were well above this size (≥4 kg). Although *Nicrophorus* use small carcasses for breeding, they will come to larger carcasses for feeding and mating. This may explain their presence at the pig carcasses in our study. It is also possible the pig carcasses, perhaps due to having been split in half, gave off a reduced odor signal similar to that emitted by small carcasses. Silphines like *T. lapponicus* are well known to be reproductively active at large carcasses (7). Working with bison and elk carcasses in Yellowstone National Park in 1993, Sikes (18) reported over 7,000 specimens of *T. lapponicus* trapped around large carcasses with a mere four specimens of this species collected in control traps. A similar study in Yellowstone National Park (28), based on (18), also reported massive numbers of silphids at large carcass traps relative to control traps (14,861 specimens vs. 1114 specimens), with *T. lapponicus* representing 60.6% of their total catch of 24,209 beetle specimens at all sites. Our finding of dominance by nicrophorines is unusual and requires further study to explain. This and other questions raised by our results might be answered by a larger, multi-year replication of this study.

Acknowledgments

We thank those who helped us obtain and prepare the pig carcass: Eric Zucker, John Blake, and Christine Terzi. We also thank those who helped us find and prepare a suitably safe field site: Milan Shipka, Tracey Martinson, Thaddeus Williamson, Mark Nelson, Dan Uliassi, Matt Olson, and Jan Dawe. Kipling Will is thanked for helpful advice regarding identification of *Harpalus* and Adam Brunke is thanked for his identification of our *Quedius* specimens. This study was undertaken to fulfill the capstone requirement of the first author's undergraduate degree in Biology at the University of Alaska Fairbanks.

References

1. Benecke M. Arthropods and corpses. In: Tsokos M, editor. Forensic pathology reviews. vol. 2. Totowa, NJ: Humana Press Inc, 2005;207–40.
2. Pechal JL, Benbow ME. Microbial ecology of the salmon necrobiome: evidence salmon carrion decomposition influences aquatic and terrestrial insect microbiomes. Environ Microbiol 2016;18(5):1511–22. https://doi.org/10.1111/1462-2920.13187
3. Chaloner DT, Wipfli MS, Caouette JP. Mass loss and macroinvertebrate colonisation of Pacific salmon carcasses in south-eastern Alaskan

streams. Freshw Biol 2002;47(2):263–73. https://doi.org/10.1046/j.1365-2427.2002.00804.x

4. Meehan EP, Seminet-Reneau EE, Quinn TP. Bear predation on Pacific salmon facilitates colonization of carcasses by fly maggots. Am Midl Nat 2005;153(1):142–51. https://doi.org/10.1674/0003-0031(2005)153[0142:BPOPSF]2.0.CO;2
5. Lafferty DJ, Loman ZG, White KS, Morzillo AT, Belant JL. Moose (*Alces alces*) hunters subsidize the scavenger community in Alaska. Polar Biol 2016;39(4):639–47. https://doi.org/10.1007/s00300-015-1819-4
6. Tantawi TI, Whitworth TL, Sinclair BJ. Revision of the Nearctic *Calliphora* Robineau-Desvoidy (Diptera: Calliphoridae). Zootaxa 2017;4226(3):301–47. https://doi.org/10.11646/zootaxa.4226.3.1
7. Anderson RS, Peck SB.The carrion beetles of Canada and Alaska. Coleoptera: Silphidae and Agyrtidae. (Insects and Arachnids of Canada, Part 13). Ottawa, Canada: Canadian Government Pub. Centre, Supply and Services Canada, 1985;1–121.
8. Bygarski K, LeBlanc HN. Decomposition and arthropod succession in Whitehorse, Yukon Territory, Canada. J Forensic Sci 2013;58(2):413–8. https://doi.org/10.1111/1556-4029.12032
9. Goff ML. Early post-mortem changes and stages of decomposition in exposed cadavers. Exp Appl Acarol 2009;49(1–2):21–36. https://doi.org/10.1007/s10493-009-9284-9
10. Mądra A, Konwerski S, Matuszewski S. Necrophilous Staphylininae (Coleoptera: Staphylinidae) as indicators of season of death and corpse relocation. Forensic Sci Int 2014;242:32–7. https://doi.org/10.1016/j.forsciint.2014.06.011
11. Alaska Climate Research Center. http://akclimate.org (accessed July 6, 2020).
12. Matuszewski S, Hall MJ, Moreau G, Schoenly KG, Tarone AM, Villet MH. Pigs vs people: the use of pigs as analogues for humans in forensic entomology and taphonomy research. Int J Legal Med 2020;134(2):793–810. https://doi.org/10.1007/s00414-019-02074-5
13. Newton AF, Thayer MK, Ashe JS, Chandler DS. Staphylinidae. In: Arnett RH Jr, Thomas MC, editors. American beetles: Archostemata, Myxophaga, Adephaga, Polyphaga: Staphyliniformia. Vol. 1. Boca Raton, FL: CRC Press LLC, 2000;272–418.
14. Lindroth CH. The ground beetles (Carabidae, excl. Cicindelinae) of Canada and Alaska, Parts 1–6. Lund, Sweden: Opuscula Entomologica (Suppl), 1961–1969;1–1192.
15. Noonan GR. Classification, cladistics, and natural history of native North American *Harpalus* Latreille (Insecta: Coleoptera: Carabidae: Harpalini), excluding subgenera *Glanodes* and *Pseudophonus*. Thomas Say Foundation Monographs, Vol. 13. Lanham, MD: Entomological Society of America, 1991.
16. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2020. https://www.R-project.org/ (accessed July 6, 2020).
17. RStudio Team. RStudio: integrated development for R. Boston, MA: RStudio, Inc., 2020. http://www.rstudio.com/ (accessed July 6, 2020).
18. Sikes DS. Influences of ungulate carcasses on Coleopteran communities in Yellowstone National Park, USA [MS thesis]. Bozeman, MT: Montana State University, 1994.
19. Bornemisszа GF. An analysis of arthropod succession in carrion and the effects of its decomposition on the soil fauna. Australian J Zool 1957;5(1):1–12. https://doi.org/10.1071/ZO9570001
20. Strickland MS, Wickings K. Carrion effects on belowground communities and consequences for soil processes. In: Benbow ME, Tomberlin JK, Tarone AM, editors. Carrion ecology, evolution, and their applications. Boca Raton, FL: CRC Press, 2015;93–106. https://doi.org/10.1201/b18819-7
21. Pantoja A, Sikes DS, Hagerty AM, Emmert SY, Rondon S. Ground beetle (Coleoptera: Carabidae) assemblages in the Conservation Reserve Program crop rotation systems in interior Alaska. J Entomol Soc B C 2013;110:6–18.
22. Parmenter R, MacMahon J. Carrion decomposition of nutrient cycling in a semiarid shrub-steppe ecosystem. Ecol Monogr 2009;79(4):637–61. https://doi.org/10.1890/08-0972.1
23. Hoback WW, Freeman L, Payton M, Peterson BC. Burying beetle (Coleoptera: Silphidae: *Nicrophorus Fabricius*) brooding improves soil fertility. Coleopt Bull 2020;74(2):427–33. https://doi.org/10.1649/0010-065X-74.2.427
24. Putnam RJ. Flow of energy and organic matter from a carcass during decomposition of small mammal carrion in temperate systems. Oikos 1978;31(1):58–68. https://doi.org/10.2307/3543384
25. Macdonald B, Farrell M, Tuomi S, Barton PS, Cunningham S, Manning AD. Carrion decomposition causes large and lasting effects on the soil amino acid and peptide flux. Soil Biol Biochem 2014;69:132–40. https://doi.org/10.1016/j.soilbio.2013.10.042
26. Sikes DS, Venables C. Molecular phylogeny of the burying beetles (Coleoptera: Silphidae: Nicrophorinae). Mol Phylogenet Evol 2013;69(3):552–65. https://doi.org/10.1016/j.ympev.2013.07.022
27. Sikes DS. Carrion beetles (Coleoptera: Silphidae). In: Capinera JL, editor. Encyclopedia of entomology, 2nd edn. New York, NY: Springer Press, 2008;749–58.
28. Barry JM, Elbroch LM, Aiello-Lammens ME, Sarno RJ, Seelye L, Kusler A, et al. Pumas as ecosystem engineers: ungulate carcasses support beetle assemblages in the Greater Yellowstone Ecosystem. Oecologia 2019;198(3):577–86. https://doi.org/10.1007/s00442-018-4315-z.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14531
Available online at: onlinelibrary.wiley.com

PAPER

PATHOLOGY/BIOLOGY

Lauren M. Weidner,[1,2] Ph.D.; Gregory Nigoghosian,[2] M.S.; Kelie C. Yoho,[2] M.S.; Jonathan J. Parrott,[1] Ph.D.; and Krystal R. Hans,[2] Ph.D.

An Analysis of Forensically Important Insects Associated with Human Remains in Indiana, U.S.A.*

ABSTRACT: Research documenting insect colonization of human remains is limited in North America, and currently nonexistent for the American Midwest. Such research is essential for forensic entomologists to identify species of research interest in a region. In this study, we collected insects from human remains in 24 cases across Indiana from June 2016 through September 2018. We analyzed species composition across scene type and season. Eight species of blow flies were collected as larvae from human remains, with *Phormia regina* and *Lucilia sericata* as the two predominant colonizers. *Phormia regina* was the most numerous species collected from outdoor scenes (73.6% of total collections) while *L. sericata* was the most numerous from the indoor scenes (60.4% of total collections). With scene types pooled, *Calliphora vicina* and *Cochliomya macellaria* were the predominant species in the fall (55.6% and 42.2%, respectively); *P. regina* was the dominant colonizer in the spring (68.6%); and *P. regina* and *L. sericata* were the predominant colonizers (46.5% and 44.4%, respectively) in the summer. In addition to these findings, we confirmed the first record of *Lucilia cuprina* colonizing human remains in Indiana having collected this species from three cases. A single adult *Chrysomya megacephala* was collected from an indoor scene in southern Indiana, which represents its second collection in the state. Beetles belonging to the families Staphylinidae, Silphidae, Histeridae, Cleridae, Trogidae, Dermestdae, and Nitidulidae were also collected from two outdoor scenes. This study provides important baseline data for forensic entomologists in Indiana, as well as surrounding states with similar environments.

KEYWORDS: forensic entomology, Calliphoridae, blow flies, human remains, *Phormia regina*, *Lucilia sericata*, *Lucilia cuprina*, *Calliphora vicina*, *Cochliomyia macellaria*

In legal investigations, insect evidence is often overlooked and underrated. In medicolegal death investigations, forensic entomologists utilize such evidence to provide supplementary information to a case, such as if the body has been relocated, if drugs are present in the decedent, and approximate time of death (1,2). Insect and other arthropod evidence can be used to determine the time of (insect) colonization (TOC), which in turn can be used to estimate the minimum postmortem interval (mPMI), or time between death and discovery of the decedent (3).

In these investigations, one of the primary tasks of forensic entomologists is the identification of the insects and other arthropods that have colonized human remains. Although there are several surveys of forensically relevant insects collected in different geographic regions, many of these surveys are limited in scope, and many regions in North America have not been surveyed at all. For North America, there is currently published information pertaining to the insects collected from human remains for Hawaii (4), British Columbia (5), and Texas (3,6). Sanford (3,6) are the most comprehensive of these surveys, providing an analysis of over 200 forensic entomology cases in Texas, and an examination of multiple factors associated with medicolegal death investigations. Sanford (3) describes the types of scenes, stages of decomposition, temperature, humidity, and manners of death, and provides a comprehensive list of all species collected in each case. Sanford (6) also discusses the differences in the trapping methods of adult insects collected at scenes, giving suggestions for proper methods to collect and rear insects in medicolegal death investigations. No such comprehensive study exists in the American Midwest, however. The goal of this present study is to identify insects associated with human remains in Indiana.

Materials and Methods

Collection and Morphological Identification

Insects were obtained from human remains in 24 investigations handled by the Indiana State Police (ISP), Tippecanoe County Coroner's Office (TCC), or a private Forensic Pathologist office—Central Indiana Forensic Associates, LLC (CIFA)—

[1]School of Mathematical and Natural Sciences, Arizona State University West Campus, 4701 W Thunderbird Road, Glendale, AZ, 85306.

[2]Department of Entomology, Purdue University, 901 West State Street, West Lafayette, IN, 47907.

Corresponding author: Lauren M. Weidner, Ph.D. E-mail: lauren.weidner@asu.edu

*Presented in part at the 2019 Entomological Society of America Annual Meeting, November 17–20, 2019, in St. Louis, MO; the 16th European Association of Forensic Entomology Meeting, June 5–8, 2019, in Bordeaux, France; and the 70th Annual Scientific Meeting of the American Academy of Forensic Sciences, February 19–24, 2018, in Seattle, WA.

Received 29 May 2020; and in revised form 10 July 2020; accepted 13 July 2020.

between June 2016 and September 2018. Crime scene investigators, deputy coroners, a forensic pathologist technician, or a forensic entomologist made the collections either at the scene or during autopsy. All larvae were collected using forceps and/or a plastic spoon and were placed into a 250 mL plastic deli container with a breathable lid (air holes) and a food source, either tuna fish or beef liver, for rearing. In the laboratory, the specimens were transferred to a 946 mL plastic deli container with a breathable lid (mesh top) and placed on pine shavings. The larval specimens were reared out under a fume hood at ~23°C. The number of specimens collected in each case varied depending on the number of life stages present and the collector. In some cases, adult flies and beetles were also collected.

Larvae were reared to adulthood and were identified morphologically (7–9). One case resulted in unsuccessful eclosion of any adults but contained five unhatched pupae. Molecular analyses were used to determine the identification of these specimens. Beetles were identified down to the lowest taxonomic level possible (10,11). For each scene, environmental conditions (temperature, humidity, and precipitation), time of year (season), and scene location (indoor or outdoor) were recorded (Table 1). Relative abundance of larval collections was analyzed by season and scene type.

DNA Extraction and Cytochrome Oxidase I (COI) Sequencing

DNA was extracted from five individual unhatched pupae using the Qiagen DNeasy Blood and Tissue Kit as per the manufacturer's protocols. DNA was re-suspended in 100 μL TE buffer. An approximate 1100 bp region of the cytochome oxidase I (COI) gene was amplified. The PCR reaction mix consisted of 1× PCR buffer, 4 mM $MgCl_2$, 200 μM dNTP's, 25 pm of each primer, C1-J-1751 (5′ - GGATCACCTGATATAGCATTCCC - 3′) and TL2-N-3014 (5′ - CGAGGTATTCCAGCAAGTCC - 3′) and 0.5 U *Taq* polymerase. PCR reactions were run on a SimpliAmp thermocycler (Applied Biosystems, Foster City, CA, USA). Cycling conditions consisted of 35 cycles at 95°C for 1 min, followed by 45°C for 1 min, followed by 72°C for 2 min. There was a final extension period at 72°C for 5 min. Products were visualized on a 1.5% agarose gel using UV transillumination.

Forward and reverse sequencing reactions with internal primers (C1-J-2183 5′ - CAACATTTATTTTGATTTTTTGG - 3′) and C1-N-2329 (5′ - ACTGTAAATATATGATGAGCTCA - 3′) were performed by the Arizona State University Core Research Facility. Sequences were visualized in 4 Peaks for assessment of quality and transferred into MEGAX for sequence trimming. Sequences (accession numbers MT681197–MT691199) were compared against both NCBI BLAST and Barcode of Life Database (BOLD) for species identity.

Results

Insects Collected

The 24 cases analyzed represented 14 counties in Indiana (Fig. 1), including northern, central, and southern Indiana. Of the larval specimens collected, five genera and eight species of blow fly were represented (Table 2), which included the blue bottle, *Calliphora vicina* (Robineau-Desvoidy), hairy maggot blow fly, *Chrysomya rufifacies* (Macquart), secondary screwworm, *Cochliomyia macellaria* (Townsend), bronze bottle fly,

TABLE 1—*Environmental variables collected at each of the 24 scenes.*

Case Number	Season	Scene Type	Avg. Temp (°C) (Day Temp Range)	Relative Humidity (%) Day Range	Total Precipitation (cm)	Precipitation (Y/N)	Larval Species Collected
1	Fa	Indoor	6.11 (3.33–12.78)	74–100	1.75	Y	*C. vicina*
2	Fa	Indoor	20.00 (12.78–27.22)	49–100	0	Y	Sarcophagidae
3	Fa	Outdoor	18.33 (13.89–23.89)	51–100	0	Y	*C. macellaria*, *P. regina*
4	Sp	Indoor	13.89 (7.22–12.78)	77–97	0.61	Y	*C. vicina*, *P. regina*
5	Sp	Indoor	6.42 (1.67–11.67)	59–100	0	Y	*C. vicina*, *L. sericata*, *P. regina*
6	Sp	Indoor	18.61 (12.22–26.11)	17–65	0	Y	*L. sericata*, *P. regina*
7	Sp	Indoor	21.94 (18.33–28.89)	42–100	1.09	Y	*L. cuprina*, *L. sericata*, *P. regina*
8	Sp	Indoor	19.72 (12.22–26.11)	40–97	0	Y	*L. sericata*, *P. regina*
9	Sp	Outdoor	26.11 (20.56–34.44)	29–93	0	N	*C. macellaria*, *P. regina*
10	Su	Indoor	23.58 (19.44–32.33)	59–97	0.66	Y	*L. sericata*, *P. regina*
11	Su	Indoor	21.39 (16.67–30.00)	30–93	0	Y	*C. macellaria*, *L. illustris*, *P. regina*
12	Su	Outdoor	20.56 (15.56–28.33)	46–100	0	N	*C. macellaria*, *P. regina*
13	Su	Indoor	22.50 (18.33–28.33)	58–100	0	N	*L. coeruleiviridis*, *L. sericata*
14	Su	Indoor	18.92 (12.22–26.11_	47–100	0	Y	*L. sericata*
15	Su	Outdoor	19.83 (18.15–26.57)	65–100	0	Y	*C. macellaria*, *P. regina*
16	Su	Outdoor	20.00 (17.22–28.89)	61–100	0.13	Y	*P. regina*
17	Su	Outdoor	23.33 (16.67–29.44)	56–100	0	Y	*P. regina*
18	Su	Outdoor	27.50 (19.44–26.67)	54–91	0.03	Y	*L. sericata*, *P. regina*
19	Su	Outdoor	22.44 (21.67–23.33)	72–97	1.42	Y	*L. coeruleiviridis*
20	Su	Outdoor	17.50 (16.67–23.33)	76–100	0	N	*P. regina*
21	Su	Outdoor	21.78 (19.81–24.63)	47–91	0.03	Y	*C. macellaria*, *L. sericata*, *P. regina*
22	Su	Indoor	24.33 (20.56–31.11)	40–87	1.68	Y	*C. macellaria*, *L. cuprina*, *L. sericata*, *P. regina*
23	Su	Indoor	24.00 (21.11–31.67)	40–84	0	Y	*C. rufifacies*, *C. macellaria*, *L. cuprina*, *L. sericata*, *P. regina*
24	Su	Outdoor	23.89 (20.56–24.44)	76–100	0	Y	*C. macellaria*, *L. cuprina*, *L. sericata*, *P. regina*

For season, Su = summer, Sp = spring, and Fa = fall. Temperature (outside average and range), relative humidity range, and total precipitation are provided for the day the remains were discovered. The last column (Precipitation Y/N) is referencing if precipitation occurred within the week prior to the discovery date.

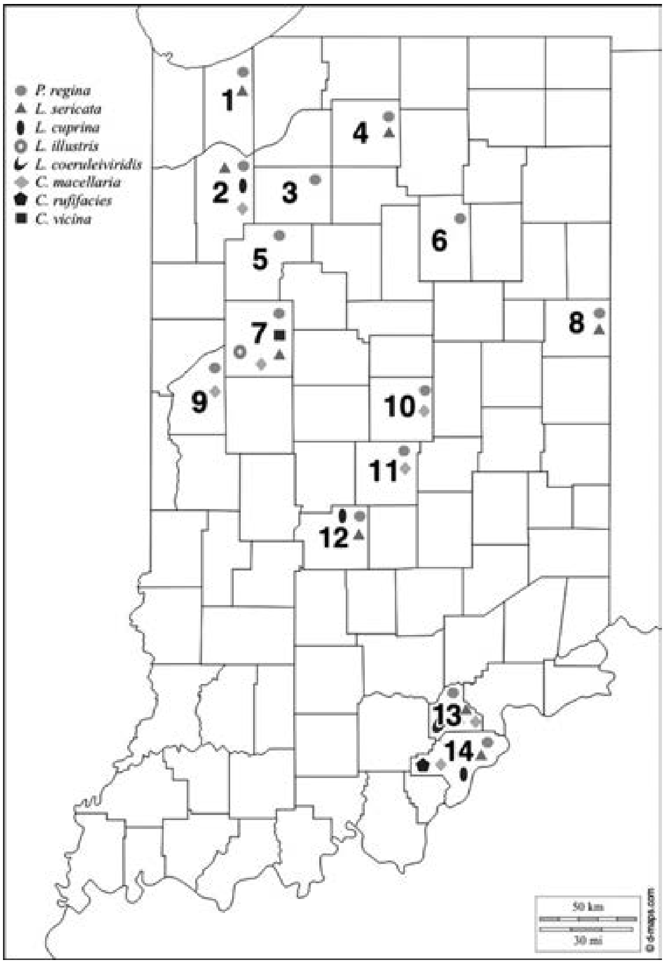


FIG. 1—*Map of Indiana, highlighting the 14 counties in which larval specimens were collected from human remains. Counties are numbered as follows: 1 = Porter, 2 = Jasper, 3 = Pulaski, 4 = Marshall, 5 = White, 6 = Wabash, 7 = Tippecanoe, 8 = Jay, 9 = Fountain, 10 = Hamilton, 11 = Marion, 12 = Morgan, 13 = Scott, and 14 = Clark. Symbols within each county indicate species collected as larvae from human remains. [Color figure can be viewed at wileyonlinelibrary.com]*

Lucilia cuprina (Wiedemann), green bottle flies, *Lucilia coeruleiviridis* (Macquart), *Lucilia illustris* (Meigen), and *Lucilia sericata* (Meigen), and the black blow fly, *Phormia regina* (Meigen). *Phormia regina* larvae were the most commonly encountered, representing 47.9% of all specimens collected, while *L. illustris* was the least encountered species, with only one individual collected (Table 2). Just under half of the cases (*N* = 10) featured two different blow fly species colonizing the remains, while some cases (*N* = 7) were colonized by three to five different species. The remaining cases with larvae present (*N* = 6) were colonized by a single species, which was most often *P. regina*. One case resulted in five unhatched pupae. DNA was successfully extracted, and approximately 650 bp of the COI was sequenced from each pupa. All sequences assigned species identity to *L. coeruleiviridis* in both NCBI and BOLD with 100% support.

Furthermore, adult blow flies were also collected from scenes, although in smaller quantities. A total of 32 adult blow flies were collected, all from indoor scenes (Table 3). More than 53% of these individuals were *P. regina*, with another 37% consisting of *L. sericata* and one individual each of *L. coeruleiviridis*, *C. macellaria* and the oriental latrine fly *Chrysomya megacephala* (Fabricius). Other specimens collected included Phoridae pupae, and adult Sarcophagidae from two indoor scenes. In addition to the larval and adult flies captured, 87 beetles were collected from two outdoor scenes. These beetle collections comprised seven families, but were predominantly composed of Silphidae (54%), followed by Staphylinidae (24%), Histeridae (14%), and Cleridae (4.5%) (Table 4). All silphids and clerids were identified down to species, with five silphid and two clerid species represented (Table 4). Due to limited keys and difficulty with identifications in the family Staphylinidae, only one common forensically important staphylinid, *Creophilus maxillosus* L., was identified to species; the seven remaining staphylinid specimens were only identified to family (Table 4).

Indoor Versus Outdoor Scenes

Thirteen cases (54%) represented indoor scenes, and eleven cases (46%) represented outdoor scenes. Indoor scenes were classified as those located within a building. In addition to being the most common species overall, *P. regina* was also the most abundant blow fly collected from outdoor scenes (73.6% of total outdoor collections), whereas *L. sericata* was the most abundant colonizer from indoor scenes (60.4% of total indoor collections; Table 2). Three blow fly species were collected exclusively from indoor scenes, *C. vicina*, *C. rufifacies*, and *L. illustris* (Table 1). *Cochliomyia macellaria*, *L. cuprina*, *L. sericata*, *L. coeruleivirdis*, and *P. regina* were collected at both indoor and outdoor scenes. Although *L. cuprina* represented only 3.7% of the specimens collected from indoor scenes, for one scene, they were the dominant colonizer (58.3% of total scene collections). One indoor scene had no colonization by blow flies (based on adult emergence and pupal characteristics). Three flesh fly (Sarcophagidae) adults emerged, and 24 sarcophagid pupae were parasitized. Two other indoor scenes, in addition to blow flies, resulted in the collection of three sarcophagids and 19 phorids.

Seasonality

The cool weather blue bottle fly, *C. vicina*, represented the most abundant colonizer (55.9%) of the specimens collected during the fall. In addition to *C. vicina*, two other species, *C. macellaria* (42.2%) and *P. regina* (2.0%), colonized remains in the fall (Fig. 2). In the spring, five species colonized remains, *C. vicina*, *C. macellaria*, *L. cuprina*, *L. sericata*, and *P. regina*, with *P. regina* representing more than 68% of specimens collected in the spring (Fig. 2). The highest blow fly diversity was observed in the summer with seven different species collected, including four *Lucilia* species, *C. macellaria*, *C. rufifacies*, and *P. regina*. Of these seven species, *P. regina* was the most abundant (46.5%), followed by *L. sericata* (44.4%), and *C. macellaria* (6.0%).

Discussion

This study is the first to analyze blow fly species colonizing human remains in Indiana. Nine species of blow flies were identified in this study, eight having been collected as larvae directly from remains. Two of these species, *C. megacephala* and *L. cuprina*, were first reported in the state within the last 10 years. Indiana's first record of *C. megacephala* originates from a study in which one female of this species was collected

TABLE 2—*Total number and proportions of larval individuals collected from indoor (*N = *13) and outdoor (*N = *11) scenes.*

Species	Total No. Collected	% of Total Collected (%)	No. of Indoor Cases Found In	No. of Outdoor Cases Found In	% of Total Collected Indoors (%)	% of Total Collected Outdoors (%)
Phormia regina	1827	47.9	8	10	30.9	73.6
Lucilia sericata	1576	41.4	8	3	60.4	12.6
Cochliomyia macellaria	240	6.0	3	6	1.9	13.0
Lucilia cuprina	92	2.4	2	1	3.7	0.5
Calliphora vicina	63	1.7	3	0	2.7	–
Chrysomya rufifacies	6	0.2	1	0	0.3	–
Lucilia coeruleiviridis	6	0.2	1	1	<0.1	0.3
Lucilia illustris	1	<0.1	1	0	<0.1	–

The second column refers to the percent of the total number of larvae in all of the collections. The last two columns (Total % indoor and Total % outdoor) refer to the percentage of total number of individuals found at that scene type.

TABLE 3—*Total number of adult blow flies collected from decedents at indoor scenes.*

Species	Total No.
Phormia regina	17
Lucilia sericata	12
Lucilia coeruleiviridis	1
Cochliomyia macellaria	1
Chrysomya megacephala	1

TABLE 4—*Total number of adult beetles collected from outdoor scenes.*

Family	No. Collected
Silphidae	
Silphinae	47
Oiceoptoma noveboracense Forster (21)	
Necrophila americana (L.) (1)	
Nicrophorinae	
Nicrophorus tomentosus Weber (14)	
Necrodes surinamensis (F.) (10)	
Nicrophorus pustulatus Herschel (1)	
Staphylinidae	
Creophilus maxillosus (14)	21
Species undetermined (7)	
Histeridae	12
Cleridae	
Necrobia ruficollis F. (3)	4
Necrobia rufipes DeGeer (1)	
Trogidae	1
Dermestidae	1
Carabidae	1
Nitidulidae	1

Number in parentheses represent the total number of that species.

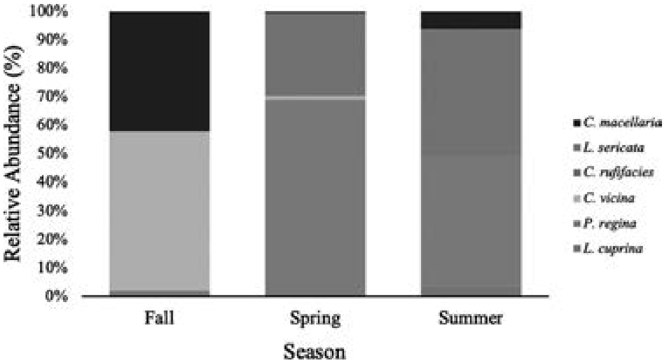


FIG. 2—*Relative Abundance of larval blow flies collected from human remains during each season (*N = *102, 458, and 3251, for fall, spring, and summer, respectively). [Color figure can be viewed at wileyonlinelibrary.com]*

using traps baited with decomposing chicken liver attractant (12). This 2013 specimen was collected in Marion county, in central Indiana, whereas the specimen from this study collected as an adult from an indoor scene was found in Clark county, in the southern portion of the state (Fig. 1). Also collected was *L. cuprina*, a blow fly species that, in addition to its forensic importance, is economically and agriculturally relevant due to its role as a facultative ectoparasite associated with myiasis of sheep in Australia (13), New Zealand (14), and Africa (15). Although *L. cuprina* has been recorded in various southern states from Virginia to southern California, it was only recently documented in Indiana in 2018 (16). In total, 28 *L. cuprina* specimens were collected using a sweep net over traps baited with decomposing chicken liver from four counties across central Indiana (16). In the current study, over 92 larvae were collected from both indoor and outdoor scenes, in three different counties, Jasper, Morgan, and Clark, located in northern, central, and southern Indiana, respectively (Fig. 1). The presence of *L. cuprina* at multiple scenes across the state highlights northern population expansion and illustrates the need of additional blow fly survey work.

This study supports seasonal differences in blow fly species composition which has been shown in other portions of the United States including California (17), South Carolina (18), and New Jersey (19). *Calliphora vicina* are commonly associated with cooler environments (18–20). In this study, they were collected in fall and spring and absent during the summer. They were collected only from indoor scenes when the outside temperature averaged between 6.11°C and 13.89°C. According to Hall (21), *C. vicina* can be found in abundance in March and April, especially in the Midwest. Although this species was collected during the spring, higher numbers were collected in the fall. This may be in part due to collecting techniques, but temperature ranges should also be considered when looking at presence/absence of species within a region.

Phormia regina and *C. macellaria* were collected across all seasons and both at indoor and outdoor scenes. As seen here, *P. regina* is commonly collected across a variety of temperatures and seasons (17,19) making it a valuable species to be studied. *Cochliomyia macellaria* is a common colonizer of remains in the southern portion of the United States; however, their southern populations have been negatively impacted by the introduction of the invasive and predatory blow fly *C. rufifacies* (21,22). This new competition and changes in temperatures may have influenced a shift in *C. macellaria* populations northward, making them another ideal species of study in Indiana. *Lucilia* spp. are commonly associated with warmer weather (17,19), and in this study, two species, *L. sericata* and *L. cuprina*, were collected during the spring, with all four *Lucilia* species being collected during the summer. These seasonal changes could have a strong correlation to temperature or other environmental parameters, and survey work across the state and throughout the year would strengthen the use of these insects in casework in this region of the United States.

Lucilia coeruleiviridis is a commonly observed forensically important fly arriving to and colonizing remains across the United States (19,23,24). However, this species is difficult to keep in colony and only one paper has discussed its development (25), and due to small sample size, it provides limited information. Slone and Gruner (24) did not analyze species composition of larval aggregations on decomposing pig remains in Indiana, but noted that they were dominated by *L. coeruleiviridis*. Given this species commonly colonizes remains and is present in Indiana, the lack of information concerning its successful development in the laboratory is a severe hinderance to the field of forensic entomology. Unsuccessful development of collected specimens can lead to a loss of important information missed during analyses. We experienced such a setback when all of the entomological evidence from one of our cases resulted in unhatched pupae, all of which were molecularly determined to be *L. coeruleiviridis*. This stressed not only the need for additional, focused development data, but also the importance of checking for uneclosed pupae when rearing entomological evidence, which can still provide valuable information.

The training levels of collectors may account for some of the variation in collections between scenes, including the quantity of blow flies and whether or not other noncalliphorid species were collected. Due to the large geographic range over which the 24 cases took place, it was not always possible for a forensic entomologist to attend and collect insect evidence. Collectors were composed of crime scene investigators, deputy coroners, and forensic pathology technicians in addition to forensic entomologists. In the event that a forensic entomologist cannot be present at the scene or autopsy, proper training of individuals responsible for collecting this evidence is imperative. Collection of adult insects, eggs, larvae, pupae, and the importance of having proper food resources, containers, and transportation of insect evidence are all key issues to highlight during such training.

Although few studies have been published pertaining to forensically relevant insects collected from human remains, such information is integral in determining which species should be studied in a given area. With two new species records of blow flies in Indiana in the last 7 years, more surveys should be conducted not only in this region, but also in many other areas across the United States where baseline information is lacking. The usefulness of the data presented here highlights the need for further baseline information and biological surveys in the field of forensic entomology for this and many other geographic regions. This information interpreted within a background of ecology and species distribution underscores the importance of updating and expanding forensic entomological surveys to strengthen insect evidence important in criminal and other investigations.

Acknowledgments

We would like to thank the Indiana State Police (ISP), Tippecanoe County Coroner's Office (TCC), and Central Indiana Forensic Associates, LLC (CIFA) for their help and support of this project. Specifically, we wish to thank Brian Olehy (ISP), Jan Kruse (ISP), Caroline Hanau (TCC) Z. Logan Miller (TCC), Donna Avolt (TCC), Matthew Durthaler (CIFA), Donald Ross (CIFA), and Dr. Thomas Sozio (CIFA) for their assistance in this research.

References

1. Catts EP, Goff L. Forensic entomology in criminal investigations. Ann Rev Entomol 1992;37(1):253–72. https://doi.org/10.1146/annurev.en.37.010192.001345
2. Campobasso CP, Gherardi M, Caligara M, Sironi L, Introna F. Drug analysis in blowfly larvae and in human tissues: a comparative study. Int J of Legal Med 2004;118(4):210–4. https://doi.org/10.1007/s00414-004-0448-1
3. Sanford MR. Insects and associated arthropods analyzed during medicolegal death investigations in Harris County, Texas, USA: January 2013-April 2016. PLoS One 2017;12(6):e0179404. https://doi.org/10.1371/journal.pone.0179404
4. Goff ML, Early M, Odom CB, Tullis K. A preliminary checklist of arthropods associated with exposed carrion in the Hawaiian Islands. Proc Hawaiian Entomol Soc 1986;26:53–7. https://doi.org/10.1093/jmedent/23.5.520
5. Anderson GS. The use of insects in death investigations: an analysis of cases in British Columbia over a five-year period. Can Soc Forensic Sci J 1995;28(4):277–92. https://doi.org/10.1080/00085030.1995.10757488
6. Sanford MR. Comparing species composition of passive trapping of adult flies with larval collections from the body during scene-based medicolegal death investigations. Insects 2017;8(2):36. https://doi.org/10.3390/insects8020036
7. Whitworth TL. Keys to the genera and species of blow flies (Diptera: Calliphoridae) of America North of Mexico. Proc Entomol Soc Wash 2006;108(3):689–725.
8. Marshall SA, Whitworth T, Roscoe L. Blow flies (Diptera: Calliphoridae) of eastern Canada with a key to Calliphoridae subfamilies and genera of eastern North America, and a key to the eastern Canadian species of Calliphorinae, Luciliinae, and Chrysomyiinae. Can J Arthropod Identif 2011;11:1–93. https://doi.org/10.3752/cjai.2011.11
9. Jones N, Whitworth T, Marshall SA. Blow flies of North America: keys to the subfamilies and genera of Calliphoridae, and to the species of the subfamilies Calliphorinae, Luciliinae and Chrysomyinae. Can J Arthropod Identif 2019;39:1–191. https://doi.org/10.3752/cjai.2019.39
10. Monk E, Hinson K, Szewczyk T, D'Oench H, McCain CM.Key to the carrion beetles (Silphidae) of Colorado & neighboring states. 2016. http://spot.colorado.edu/~mccainc/PDFs/KeytoSilphidaeofColorado.pdf (accessed July 13, 2020).
11. Brunke A, Newton A, Klimaszewski J, Majka C, Marshall S. Staphylinidae of eastern Canada and adjacent United States. Key to subfamilies: Staphylininae: tribes and subtribes, and species of Staphylinina. Can J Arthropod Identif 2011;12:1–110. https://doi.org/10.3752/cjai.2011.12
12. Picard CJ. First record of *Chrysomya megacephala* Fabricius (Diptera: Calliphoridae) in Indiana, USA. Proc Entomol Soc Wash 2013;115 (3):265–7. https://doi.org/10.4289/00138797.115.3.265
13. Tellam RL, Bowles VM. Control of blowfly sheep strike in sheep: current strategies and future prospects. Int J Parasitol 1997;27(3):261–73. https://doi.org/10.1016/S0020-7519(96)00174-9
14. Heath ACG, Bishop DM. Flystrike in New Zealand: an overview based on a 16- year study, following the introduction and dispersal of the Australian sheep blow fly, *Lucilia cuprina* Wiedemann (Diptera: Calliphoridae). Vet Parasitol 2006;137(3–4):333–44. https://doi.org/10.1016/j.vetpar.2006.01.006
15. James MT. The flies that cause myiasis in man. No. 631. Washington, DC: US Department of Agriculture, 1947.

16. Owings CG, Picard CJ. New distribution record for *Lucilia cuprina* (Diptera: Calliphoridae) in Indiana, United States. J Insect Sci 2018;18 (4):8. https://doi.org/10.1093/jisesa/iey071
17. Brundage A, Bros S, Honda JY. Seasonal and habitat abundance and distribution of some forensically important blow flies (Diptera: Calliphoridae) in Central California. Forensic Sci Int 2011;212(1–3):115–20. https://doi.org/10.1016/j.forsciint.2011.05.023
18. Tomberlin JK, Adler PH. Seasonal colonization and decomposition of rat carrion in water and on land in an open field in South Carolina. J Med Entomol 1998;35(5):704–9. https://doi.org/10.1093/jmedent/35.5.704
19. Weidner LM, Jennings DE, Tomberlin JK, Hamilton GC. Seasonal and geographic variation in biodiversity of forensically important Blow flies (Diptera: Calliphoridae) in New Jersey, USA. J Med Entomol 2015;52 (5):937–46. https://doi.org/10.1093/jme/tjv104
20. Watson EJ, Carlton CE. Insect succession and decomposition of wildlife carcasses during fall and winter in Louisiana. J Med Entomol 2005;42 (2):193–203. https://doi.org/10.1603/0022-2585(2005)042(0193:ISADOW)2.0.CO;2
21. Hall DG. Blow flies of North America. Thomas Say Foundation, Vol. IV. College Park, MD: Entomological Society of America, 1948.
22. Byrd JH, Butler JF. Effects of temperature of *Cochliomyia macellaria* (Diptera: Calliphoridae) development. J Med Entomol 1996;33(6):901–5. https://doi.org/10.1093/jmedent/33.6.901
23. Joy JE, Liette NL, Harrah HL. Carrion fly (Diptera: Calliphoridae) larval colonization of sunlit and shaded pig carcasses in West Virginia, USA. Forensic Sci Int 2006;164(2–3):183–92. https://doi.org/10.1016/j.forsciint.2006.01.008
24. Slone DH, Gruner SV. Thermoregulation in larval aggregations of carrion feeding blow flies (Diptera: Calliphoridae). J Med Entomol 2007;44(3):516–23. https://doi.org/10.1603/0022-2585(2007)44(516:tilaoc)2.0.co;2
25. Weidner LM, Tomberlin JK, Hamilton GC. Development of *Lucilia coeruleiviridis* (Diptera: Calliphoridae) in New Jersey, USA. Fla Entomol 2014;97(2):849–51. https://doi.org/10.1653/024.097.0277

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14511
Available online at: onlinelibrary.wiley.com

PAPER

PSYCHIATRY & BEHAVIORAL SCIENCE

Goran Arbanas,[1,2] *Ph.D.; Paula Marinović*,[3] *M.D.; and Nadica Buzina*,[1,4] *Ph.D.*

Psychiatric and Forensic Differences Between Men Charged with Sex Offences and Men Charged with Other Offences

ABSTRACT: Studies on differences between individuals convicted of sexual offences and nonsexual offences are sparse and there is an ongoing debate as to whether sexual offenders differ from other offenders. The primary aim of this study was to determine demographic characteristics, prevalence of mental disorders, alcohol and drug use at the time of the crime and the criminal responsibility of individuals charged with sexual offences, compared to nonsexual crimes, with the aim of bringing awareness to the similarities and differences between men charged with sex offences and those charged with other crimes. This is a single-institution retrospective study of subjects charged with sexual offences and sent for institutional psychiatric evaluation to a Forensic Psychiatric Centre in an urban, academic, tertiary-care center. The control group consisted of individuals charged with nonsexual offences referred to the same center. Results showed significant differences between individuals charged with sexual offences and nonsexual offences. Men charged with sex offences more frequently committed their crimes alone and victimized children equally as often as adults. They also less frequently pleaded guilty in court. They were more likely to be abused in childhood and more often had antisocial personality disorder and paraphilias and less often substance-related disorders. The majority were considered criminally responsible. Our results show that sex offenders are different from nonsex offenders in many characteristics of their personal history, offence characteristics and forensic evaluations and these particular differences warrant different approaches to the prevention of future re-offending, compared to nonsex offenders.

KEYWORDS: offenders, sex offence, forensic psychiatry, antisocial personality disorder, paraphilias, criminal insanity

Sexual violence is not only a widespread phenomenon, but also an affect laden criminal act that raises high public and sometimes professional emotional reactions. Especially in cases of the plea being not guilty by reason of insanity, the reactions can be even more pronounced as some research show that both public and legal institutions tend to ask for conviction and not for treatment (1,2).

Sexual crime is a wide concept that can generally be defined as any type of sexual act being forced upon an individual without his or her consent (3). Besides nonconsensual sexual intercourse, it also includes any other type of unwanted sexual attention, such as kisses or caresses. In Croatia, where this study was conducted, two chapters of the Criminal Code are dedicated to sexual crimes: chapter XVII Offences against sexual freedom and chapter XVIII Offences against sexual maltreatment and sexual exploitation of children (4). Sexual crimes cannot be attributed to any particular category of psychopathological disorders except paraphilias, but the majority of people who are convicted of sexual offences do not have a paraphilic disorder (5).

The consequences of sexual violence are as follows: physical (e.g., gynecological complications, chronic pelvic pain, and migraines), psychological (e.g., depression, PTSD, and harmful behavior), and social (strained relationships) (6).

The vast majority of people convicted of sexual offences are men (more than 97%). Men charged with sex offences make 12% of all inmates, and among those that are nonguilty by reason of insanity, the percentage of individuals who committed sexual offences is similar, between 6% and 21% (5,7).

Research has shown that people convicted of sexual offences have high rates of mental disorders, but the population studied may differ significantly. Some have chosen to focus on only the not guilty by reason of insanity (NGRI) (7), those sent for a forensic treatment (1), those in psychiatric facilities (8) and only a few researched individuals suspected of sexual offences sent for a psychiatric evaluation (9). The choice of a sample will influence the prevalence of mental disorders, as individuals who committed sexual offences found NGRI would be more likely to have more severe psychopathology than those found guilty and incarcerated or those from the general population (10). Of the studies of individuals suspected of sexual offences sent for a psychiatric evaluation, Valenca AM et al. found the lowest prevalence of mental disorders (57% of 44 individuals suspected of sexual offences had some kind of mental disorder or an intellectual disability) (9). In all the other research, the percentage was much higher (7,9,11). Also, people convicted of sexual

[1] Department of Forensic Psychiatry, University Psychiatric Hospital Vrapče, Zagreb, Croatia.

[2] Faculty of Medicine, University of Rijeka, Rijeka, Croatia.

[3] Department of Psychiatry and Psychological Medicine, University Hospital Centre Zagreb, Zagreb, Croatia.

[4] University Department of Croatian Studies, University of Zagreb, Zagreb, Croatia.

Corresponding author: Goran Arbanas, Ph.D. University Psychiatric Hospital Vrapče, Bolnička cesta 32, 10000 Zagreb, Croatia. E-mail: goran.arbanas@ka.t-com.hr

Received 25 April 2020; and in revised form 20 June 2020; accepted 23 June 2020.

murders had more and a greater variety of psychiatric disorders (the most prevalent were personality disorders, alcohol use disorders, sexual dysfunctions, and sexual sadism) than people convicted of nonhomicidal sex offences (12).

The most prevalent mental disorders among this group of offenders are personality disorders (with the highest rate of antisocial personality disorder), substance use disorders, mood disorders, and paraphilic disorders (5,8,9,11). As expected, in forensic NGRI populations, the prevalence is much higher than in individuals convicted of sexual offences sent for a forensic assessment. Also, a significant proportion of individuals convicted of sexual offences are of borderline or lower intelligence; however, some claim that people with intellectual disabilities are not overrepresented among people suspected of sexual offences (13).

The use of alcohol and other drugs at the time of the offence is present in up to one third of individuals charged with sexual offences. In addition, the individuals offending against male children have higher rates of alcohol consumption at the time of the offence compared to individuals offending against female children and against adults, while in cases of sexual murder alcohol use occurred more often in those with adult victims compared to those with child victims (14,15).

Different countries have different legal concepts regarding criminal responsibility, both in terms of philosophical and legal approaches (15). In Croatia, there is a biopsychological concept implying that full criminal responsibility can only be excluded if the individual suspected of some offence was suffering from a mental disorder (i.e., a biological cause) which made them incapable of either fully understanding his or her acts or of controlling his or her behavior (i.e., cognitive and volitional psychological consequence). The person can be assessed as fully responsible, completely nonresponsible (NGRI) or of reduced responsibility (there are two levels of reduced responsibility—reduced responsibility to a significant degree, or not to a significant degree). Complete absence of criminal responsibility is usually due to a psychotic disorder (e.g., schizophrenia and delusional disorder) in which case psychotic symptoms (hallucinations or delusions) make reality testing impossible and due to these symptoms a person is not able to understand what is really happening, and instead act according to his/her delusional ideas and interpretations. In the case of intoxication with psychoactive drugs, the ability to comprehend reality can be reduced but is not completely absent. Therefore, in cases of intoxication or withdrawal symptoms, examinees are usually assessed as of reduced criminal responsibility, but not NGRI. People with paraphilic disorders would be considered criminally responsible, unless paraphilia is coupled with some other pathology (e.g., psychosis, severe personality disorder, intoxication, and mood disorders).

People found NGRI are committed to involuntary treatment in a forensic mental institution. Those with reduced responsibility can be sent to psychiatric treatment in prison facilities or in civil institutions (if the person is not sentenced to a prison sentence), and their sentence can be reduced (16).

There is an ongoing debate on whether individuals convicted of sexual offences differ from individuals convicted of other offences, or whether antisocial characteristics themselves lead to different offences (one of the criteria for antisocial personality disorder is failure to conform to social norms with respect to lawful behaviors, as indicated by repeatedly performing acts that are ground for arrest), and people just differ in the kind of the offence they are inclined to commit (17). The fact that many individuals convicted of sexual offences have antisocial personality disorder (17–52%) supports this stance (18). Conversely, others claim that individuals convicted of sexual offences differ from other offenders and are more similar to nonoffending people with paraphilias. Some found that men convicted of sex offences showed more internalizing problems, compared to those convicted of nonsex offences, while others showed that men convicted with sex offences had less behavioral problems, but more problems in peer relationships and were more socially isolated (19–22).

Aims

The primary aim of this study was to determine demographic characteristics, prevalence of mental disorders, alcohol and drug use at the time of the crime and criminal responsibility of individuals accused of sexual offences, compared with individuals accused of nonsexual crimes. Thus, this study aims to bring awareness of the similarities and differences between individuals accused of sexual offences and of other crimes that can have implications on their treatment and legal procedures.

Materials and Methods

This is a single-institution retrospective study of subjects identified as sex offenders and sent for institutional psychiatric evaluation to a Forensic Psychiatric Centre in an urban, academic, tertiary-care center. In Croatia, if the judge, the defense or the prosecution suspect that the individual accused of any offence suffers from any kind of mental disorder (including paraphilic disorders) that can reduce criminal responsibility, they inform the judge, who in turn sends the individual for the evaluation (16). The judge can choose among individual (private) experts or can send the accused for an institutional evaluation. The latter is usually the case in very complicated or emotion-laden cases (such as sex offences).

A person sent for an institutional evaluation at the Forensic Psychiatric Centre is assessed by a psychiatrist, a psychologist, and all the necessary investigations can be done if indicated (e.g., EEG, laboratory tests, psychoactive drugs in urine or blood, and neurological investigations). If necessary, the person can be hospitalized for further observation. Finally, all the examinees are presented at the forensic meeting (usually ten to twenty psychiatrists and psychiatric trainees attend these meetings), with the person being evaluated present at the meeting. All the staff members can ask the examinee additional questions. In the end, all of the staff members participate in the decision-making of the case. The diagnoses are reached according to ICD and DSM criteria by the leading psychiatrist in each of the cases. Finally, a written report in narrative form is produced for the court.

For the purpose of this study, we used all written evaluations of people sent for institutional evaluation from 2010 to 2016. These were the people who were charged by the state attorney's office, but not yet convicted. As a result, some of them could be found not guilty during the legal process (and after the evaluation), so it is possible that some subjects were falsely accused. Nevertheless, we decided to include these as there are a very low number of falsely accused people, except during child custody processes during divorce (23,24). On the other hand, many cases not found guilty after the charge were for procedural reasons (25). This problem of sampling was dealt with similarly in other research (9). Altogether, there were 501 cases, 444 male, and 57 female. Of them, there were 57 cases of individuals charged with sexual offences, with 56 male and one female. The only female subject was excluded from the study as no statistical analysis could have been done with a single subject, leaving 56 cases. For the control sample of individuals charged with nonsexual offences, in order to approximate a random sampling process, the next male sent for the evaluation, after the person charged with sex offence, was chosen, so the nonsex offenders comprise of 56 male subjects. Nonsex offenders were charged with the following charges: homicide (27%), trading with illegal substances (17%), burglary (6%), physical assault (6%), and others. None of the individuals charged with sexual offences were charged with additional nonsex-related offences. Men charged with sex offences were charged with: rape (78%), lewd acts (7%), attempted rape (4%), use of children in creating pornography (4%), and accessory to rape (4%).

All of the subjects were referred for the evaluation of the criminal responsibility and in 18 cases (9 charged with sexual offences and 9 withn on sexual offences) the court asked for the assessment of capacity to stand trial, in addition to criminal responsibility.

In Croatia, all the individuals charged with criminal offences have to be legally represented by a lawyer, and in cases, when the person cannot afford to have a legal representative, the representative will be paid by the court.

Demographic, clinical, psychiatric, and legal data were collected from the evaluations, including the marital status and place of residence, mental disorders diagnosed and prior history of treatment, previous convictions, assessment of criminal responsibility. These written evaluations were stored in the facility and the authors read all the evaluations and collected the data for the study.

The institutional ethics committee gave the consent for the study.

Main Study Measures

Primarily, information from written assessments of people sent by the court for the evaluation of their criminal responsibility and the capacity to stand trial was utilized. Data were collected from the Centre for Forensic Psychiatry of the University Psychiatric Hospital Vrapče, Zagreb, Croatia, where the registry of all the people sent for institutional forensic assessment is stored.

Results

Demographics

Individuals charged with sexual offences (SO), compared to those charged with other (nonsexual) crimes (NSO) were of the same age (average age 37.5 ± 12.1 for SO and 39.0 ± 14.0 for NSO) and the majority were unemployed (64.3% and 55.3%), equally often single or married (Table 1.). On average, they had one child (1.04 SO; 1.18 NSO). Both SO and NSO lived most often in the capital (38.0% and 28.6%; $\chi^2 = 1.849$, $p = 0.604$).

TABLE 1—*Marital status of individuals charged with sexual offences and nonsexual offences..*

	Men Charged with Sex Offences	Men Charged with Nonsexual Offences	
Single	44.6%	42.0%	$\chi^2 = 1.953$
Married	41.1%	32.0%	$p = 0.377$
Divorced	14.3%	26.0%	

Data Regarding the Offence

In both groups, about one quarter (30.4% and 26.5%) decided to use their legal right to remain silent during legal procedures in the court.

More NSO committed the crime together with another person (22.4%), compared to SO (7.1%), $\chi^2 = 5.857$, $p = 0.015$. The victims differ significantly, as the majority of victims of SO are women (92.7%, half of them adult and half minor), while the majority of victims of NSO are adults (92.6%, two-third women). Only one child was the victim of NSO (Table 2), $\chi^2 = 52.980$, $p < 0.0001$. Among individuals charged with sexual offences, the youngest victim was two and the oldest 87. Sex offenders' victims were significantly younger (age 19.8 ± 18.3) compared to NSO's (44.3 ± 25.8) ($t = -2.923$; $p = 0.005$). In both groups, the victim was more likely to be known to the perpetrator (69.6% of SO and 81.8% of NSO). Victims were family members in 26.8% of SO and 21.4% of NSO.

In more than half of overall cases (58.2% of SO and 48.7% of NSO), the crime took place at the victim's or perpetrator's home, and in 30.8% of cases of NSO, the crime happened in the street (as compared to only in 5.5% of SO). In 9.1% of cases of sex offences, the crime took place in a car.

Close to only a third (30.2%) of SO pleaded guilty and 62.2% of NSO; $\chi^2 = 13.23$, $p = 0.001$. On the other hand, in the majority of cases other witnesses had been already questioned in the court (89.1% of SO and 79.2% of NSO).

Almost one third of SO (29.8%) and NSO (31.3%) used alcohol at the time of the offence, while drug use was extremely rare (in only one case among SO/cannabis/ and two cases among NSO/cannabis and heroin/).

Personal History

The vast majority of offenders lived with their family during their childhood, but 21.8% of SO and 10.9% of NSO were at some time during their childhood institutionalized due to juvenile offending, conduct disorder or due to parents' lack of parenting capacities.

In both groups, the majority finished high school (50.9% of SO and 60.0% of NSO), and only a minority had a university degree (9.1% of SO, none of the NSO; $\chi^2 = 8.427$; $p = 0.038$). More than one third of SO (35.2%) and only 17.0% of NSO repeated a year at school ($\chi^2 = 4.233$, $p = 0.032$).

There is no difference regarding alcohol use in personal history (78.4% of SO and 87.2% of NSO), but the NSO more often used other psychoactive drugs (most frequently cannabis)—32.1% of SO and 50.0% of NSO ($\chi^2 = 5.896$, $p = 0.053$).

Almost half of both groups were previously sentenced (42.9% and 36.2%), but SO served prison sentences in 41.1% of cases and NSO in 21.7% of cases ($\chi^2 = 4.313$, $p = 0.030$). In the

TABLE 2—*Victims of individuals charged with sexual offences and nonsexual offences.*

	Men Charged with Sex Offences	Men Charged with Nonsexual Offences	
Male child	3.6%	0	$\chi^2 = 53.650$
Female child	48.2%	2.0%	$p < 0.001$
Male adult	3.6%	15.7%	
Female adult	42.9%	33.3%	
No victim	1.8%	49.0%	

NSO group, none had been sentenced for a sex offence earlier. Earlier offences in SO were as follows: 58% nonsex offences (burglary, domestic violence, threats, attempted homicide etc.), 29% sex offences (rape and lewd and lascivious offence), while in 13% there were no data on earlier offences.

SO were more often abused in childhood than NSO (35.7% and 8.9%; $\chi^2 = 10.30$, $p = 0.016$), but if we consider only sexual abuse, there were no differences (7.1% of SO and 2.2% of NSO, $\chi^2 = 1.609$, $p = 0.205$).

In both groups, half had been psychiatrically treated previously (45.1% of SO and 61.9% of NSO, $\chi^2 = 2.140$, $p = 0.143$) and some of them were also hospitalized (30.4% of SO and 47.9% of NSO; $\chi^2 = 3.367$, $p = 0.051$). SO were most often diagnosed with substance use disorders (12.5%), anxiety disorders (10.7%), and mood disorders (7.1%). Only 3.6% were diagnosed with paraphilic disorders (pedophilic disorder) prior to the crime. NSO were most often diagnosed with substance use disorders (22.0%), neurocognitive disorder (12%), psychotic disorders (10.0%), and mood disorders (8.0%). NSO were significantly more often diagnosed with psychotic disorders ($\chi^2 = 4.757$, $p = 0.030$). The NSO group used replacement therapy for substance dependence (i.e., buprenorphine/naloxone) significantly more often ($\chi^2 = 6.665$, $p = 0.010$).

The SO group were prescribed antipsychotic drugs more often. Antipsychotic drugs were prescribed for psychotic disorders, but also in lower doses for other disorders (e.g., mood disorders, sleep disorders, conduct disorders, and personality disorders).

The Forensic Evaluation

SO were more often asked about their sexuality (or their sexuality was mentioned in the report) (78.6% compared to 15.9%), about masturbation practices (16.4% compared to zero), but not about sexual fantasies (11.1% compared to 2.2%). Interestingly, SO more often spontaneously talked about their sexuality compared with NSO (37.5% compared to 8.0%). See Table 3.

TABLE 3—*Sexual history taking in men charged with sex offences (SO) and those charged with nonsexual offences (NSO).*

	SO	NSO	
Sexual history	78.6%	15.9%	$\chi^2 = 38.716$, $p < 0.001$
Masturbation	16.4%	0	$\chi^2 = 8.092$, $p = 0.003$
Sexual fantasy	11.1%	2.2%	$\chi^2 = 3.240$, $p = 0.08$
Spontaneously	37.5%	8.0%	$\chi^2 = 15.821$, $p < 0.001$

TABLE 4—*Intelligence status of offenders..*

	SO	NSO	
Intellectual disability	3.6%	0	$\chi^2 = 3.357$
Borderline (below average) intelligence	21.4%	10.9%	$p = 0.340$
Normal range	48.2%	63.0%	
Above normal intelligence	26.8%	26.1%	

The intelligence of subjects was measured by Wechsler-Bellevue Intelligence Scale (WBII) and is shown in Table 4. There were no differences between the groups ($\chi^2 = 4.346$, $p = 0.226$).

The most prevalent mental disorders among individuals charged with sexual offences were as follows: personality disorders (55.4%), substance-related disorders (16.1%), and paraphilic disorders (10.7%). Only 7.1% of SO were assessed as having no mental disorder. The most prevalent disorders among NSO were as follows: substance-related disorders (41.2%), personality disorders (26.8%), and psychotic disorders (9.8%). Only 5.9% of NSO were assessed as having no mental disorder. SO were more often diagnosed with paraphilic disorders (as a general category of all paraphilic disorders including pedophilic disorder, but also were more often diagnosed with pedophilic disorder in particular) and personality disorders, while NSO were more often diagnosed with substance-related disorders and psychotic disorders. See Table 5.

Among personality disorders, the prevalence among SO was the following: antisocial 37.5%, narcissistic 32.1%, dependent 10.7%, borderline 8.9%, and paranoid 3.6%. Among NSO the prevalence was as follows: antisocial 27.5%, narcissistic 23.5%, borderline 7.8%, paranoid 3.5%, and dependent 2.0%.

The majority of individuals accused of sexual offences were assessed as criminally fully responsible, while the majority of NSO were of reduced responsibility. Only 1.8% (one person) of SO were assessed as criminally not responsible (not guilty by reason of insanity), in comparison with 8.5% of NSO (Table 6). Accordingly, the majority of SO were not obliged to any of the treatment measures, and a significant minority (12.5%) were sent for a psychiatric treatment. Among NSO half were not sent for any treatment, one third were sent for the treatment of a substance dependence and 14% for involuntary inpatient treatment. See Table 6.

Discussion

Demographics

Individuals charged with sexual offences and other criminal offences did not differ in any aspect of their demographic data. The same was found in the research on sex and violent offenders found NGRI (26). Subjects in both groups were in their late thirties. In the majority of research, the conclusion is that offenders are male and young. Our sample was male, but not young, similar to a sample of offenders with intellectual disabilities and other men charged with sex offences referred for psychiatric

TABLE 5—*Mental disorders diagnosed during forensic evaluation: diagnoses are coded using ICD-10.*

Mental disorders	SO	NSO	
Organic psychosyndrome (F0)	2 (4.3%)	4 (8.7%)	$\chi^2 = 0.704$, $p = 0.401$
Substance-related disorders (F1)	9 (19.6%)	21 (45.7%)	$\chi^2 = 6.556$, $p = 0.011$
Psychotic disorders (F2)	0	5 (10.9%)	$\chi^2 = 5.78$, $p = 0.022$
Mood disorders (F3)	0	1 (2.2%)	$\chi^2 = 0.343$, $p = 0.558$
Anxiety disorders (F4)	1 (2.2%)	2 (4.3%)	$\chi^2 = 0.325$, $p = 0.569$
Paraphilias (F65)	6 (13.0%)	0	$\chi^2 = 5.785$, $p = 0.018$
Paedophilia only (F65.4)	5 (10.9%)	0	$\chi^2 = 4.77$, $p = 0.036$
Mental retardation (F7)	3 (6.5%)	0	$\chi^2 = 3.083$, $p = 0.079$
Personality disorders (F60)	31 (67.4%)	15 (32.6%)	$\chi^2 = 9.444$, $p = 0.002$

TABLE 6—*Forensic evaluations of criminal responsibility and recommendations for psychiatric treatment.*

	SO	NSO	
Criminal responsibility			
Fully criminally responsible	66.1%	19.1%	χ^2 = 25.581
Reduced responsibility	30.4%	55.3%	$p < 0.001$
Reduced responsibility to a significant degree	1.8%	17.0%	
Criminally not responsible	1.8%	8.5%	
Recommendations for psychiatric treatment			
No treatment recommended	82.1%	44.9%	χ^2 = 24.83
Treatment of substance dependence	5.4%	32.7%	$p < 0.001$
Psychiatric treatment	12.5%	8.2%	
Involuntary inpatient treatment	0	14.3%	

evaluation (8,9,15,26). As more than 40% of participants had a positive criminal history, this is not their first offence, so this can, at least partly, explain why our sample is older than in other research.

The percentage of unemployed individuals was higher than in other research: More than half of our sample was unemployed; and unemployment is a known risk factor for criminal activities, but unemployment is generally high in Croatia (9,27,28). On the other hand, more subjects in our study were married (with the exception of one, in all the other research more offenders were never married; this one exception showed that 25% of men convicted of sex offences in prison were single, compared to 70% of men charged with sex offences in forensic institutions) (5,7,8).

The Offence

One fifth of the individuals who committed nonsexual crimes committed it together with other perpetrators in comparison with only 7% of individuals charged with sexual offences. This is expected, as sexual crimes are usually a very intimate type of crimes and do not include any kind of material gain (like e.g., robbery would) that warrant joint ventures.

Victims differ significantly, as the majority of victims of sex offences are women of different ages (2–87 in our study), but on average younger than the victims of nonsex offences. Almost half of the NSO committed a crime with no victims, and when victims were present, more often they were adults (one-third men and two-third women). This gender and age distribution is expected, as research conducted around the world confirms that victims of sex offenders are predominantly women, equally often underage and adult (7,9). The only exceptions are people charged with sex offences found NGRI (26). In our sample, only one was assessed as NGRI. As all the perpetrators are male (the only female perpetrator was excluded from data analysis), it is no surprise that victims are young (even underage) and female, as it is known that men are usually sexually attracted to younger women (this could be the explanation for those with victims after puberty) (26,29,30). On the other hand, in cases of victims with prepubertal children, a diagnosis of antisocial personality disorder was established in more than a third of individuals. In this subgroup of subjects (with child victims), we can assume that in some of them (with pedophilic disorder) the motive was sexual, but in a significant number of the individuals the motivation for the offence was probably not sexual (in those with antisocial personality disorder). As in other research, our study confirmed that in both sex offences and other offences the perpetrator usually knows the victim.

Home is the most dangerous place for both victims of sexual and nonsexual crimes, as half of the offences took place at home. Sex offences took place in the street in only 5.5% of cases. Therefore, it seems that in Croatia, the majority of individuals charged with sexual offences are people close to victims and the crime is not committed in dark alleys. The fact that close family members are the most prevalent offenders is confirmed by the majority of research of people convicted of sex offences, but some of the studies showed that in some countries (e.g., Brazil) half of the sex offences took place in the street (9).

Compared to individuals charged with nonsex offences, those charged with sexual offences less often pleaded guilty in court. We can explain this difference by two streams of thought. On the one hand, sex offences are looked at as dirty and stigmatizing for both the victim and the offender, so a person admitting this kind of crime would be exposed to much more stigma and embarrassment than the offenders of other types of crimes (31). On the other hand, in cases of sex offences there are usually no witnesses and there is less material evidence, so sometimes it is the word of the victim against the word of the person accused. This is less often the case in nonsexual offences. It is also possible that more individuals charged with sexual offences will be found not guilty during the court procedure, but this is contrary to the fact that sex offences are often underreported and not many people (outside divorce procedures) make false accusations (25). It is important to note that in Croatia, contrary to some other countries, the person charged with a crime needs to plead guilty or not guilty even before the psychiatric evaluation, at the beginning of the legal process (4). Similarly, witnesses can be questioned in court before the forensic evaluation.

Use of alcohol was prevalent among both groups of subjects, at the time of the offence, which is unsurprising, as we know that alcohol acts disinhibitory to human behavior and increases the risk of physical violence (32,33). But, the use of other psychoactive substances in our sample was surprisingly rare. Although other research shows that the prevalence of drug abuse in men charged with sex offences is lower than in nonsex offences, some of the research shows high percentages of substance intoxication at the time of the offence (as high as 40%) (34).

Personal History

This study confirms earlier research showing that there are some childhood factors that are related to future offending, like living outside the primary family, lower educational level, and childhood abuse (35). One fifth of our sample of individuals charged with sexual offences and 11% of those charged with nonsexual offences were institutionalized (outside of their family) during their childhood. Only 9% of SO and none of the NSO had a university degree. One third of SO (35.2%) repeated a year at school. As only a minority of SO had intellectual disability (3.6%) and 21% were of below-average intelligence (but above the threshold for intellectual disability), repeating a year was in the majority of cases not for reasons of low intellectual status, but probably due to conduct problems. This is also confirmed by the fact that 37.5% of individuals charged with sexual offences had antisocial personality disorder, and the requirement for the diagnosis of this disorder is a conduct disorder in childhood (17).

In our sample, reoffending was equally prevalent in both groups, but SO more often served prison sentences (therefore, we can conclude they had committed more serious crimes). This is contrary to an earlier study, showing that 51% of NSO and 19% of SO had reoffended. But the aforementioned study had only studied offenders with intellectual disabilities, while our sample is predominantly of people within a normal range of intelligence (36). Similarly, the study of men charged with sex offences and nonsex offences found NGRI showed fewer previous convictions in those charged with sex offences. While in the sample of people charged with sex offences found NGRI the majority of the people were diagnosed with psychotic disorders, these disorders were rarely diagnosed in our sample (26).

As mentioned earlier, almost one third of subjects used alcohol at the time of the offence and more than 75% of SO and 85% of NSO had a personal history of alcohol use. Again, less SO used other psychoactive drugs compared with NSO (32% and 50%, respectively). In those who used other drugs, cannabis was the most frequently used drug.

Individuals charged with sexual offences were more frequently physically abused in childhood (36% compared to 9%). Some other research also showed almost the same prevalence of physical abuse across the world (11,34). Sexual abuse was less frequent (7% and 2%, not statistically significant), once again in accordance with other research. We are not sure about the reasons for the high prevalence of physical abuse and the absence of sexual abuse in men charged with sex offences. What are the mechanisms that lead from physical abuse in childhood to becoming a sexual abuser in adulthood? As offenders were usually physically abused in childhood by their parents (up to 85% of those abused), and in the majority of cases the victims of these same men when grown up are close family members (partners, children, step-children, and other close people), we can speculate that the person learns about the abusive type of a relationship during childhood and then transfers it into adulthood (37,38). The low number of sexual abuses could be partly explained by a general tendency to underreport sexual violence in an attempt to forget the assault and to escape shame and embarrassment (39). It is important to note that although men convicted of sexual offences are more likely to have been abused, most people who were abused do not go on to commit sexual offences (40).

Almost half of SO and more than 60% of NSO in our sample were psychiatrically treated before the offence. This is higher compared with other research that showed that 13–46% of SO were treated before the offence (9,34). This high prevalence of a personal history of contact with mental health services can be partly explained by our sample. Our sample consisted of people who were accused of sex crimes and who were sent for forensic psychiatric evaluation. The practice of the Croatian judicial system is that if somebody was psychiatrically treated for any reason, they would be sent for an evaluation of criminal responsibility for the offence.

The most prevalent mental disorder diagnosed prior to the offence was substance dependence (12.5% of the sample), followed by anxiety disorders (10.7%). In the earlier Croatian study, the most prevalent mental disorder prior to the offence was a personality disorder (11). Among NSO the most prevalent disorder was substance abuse, unlike SO where personality disorders were significantly more prevalent.

The Forensic Evaluation

Important differences between the SO and NSO were found in the process of forensic evaluation itself. It seems that forensic psychiatrists did not feel there was a need to assess NSO sexual histories since only 15.9% of NSO were asked about their sexuality during the evaluation, none were asked about masturbation and only 2.2% about their sexual fantasies. An even more concerning finding was the fact that not many SO were asked about masturbation or their sexual fantasies either (only 16.4% and 11.1%) which raises the question about the quality of assessment of SO, especially regarding paraphilias. This should be an important factor when evaluating and treating SO and the nature of their crime. This also raises the question how paraphilic disorders were diagnosed (or excluded) without asking about sexual fantasies and masturbation-related activities. These findings definitely raise awareness of the need for improvement among Croatian physicians, especially forensic psychiatrists evaluating SO, by including sexual medicine education to help them improve the quality of forensic evaluations. This might be the result of the lack of sexual medicine education on all levels of medical education in Croatia (41).

There are many differences in terms of mental disorders diagnosed between SO and NSO. Expectedly, SO were more often diagnosed with paraphilic disorders and pedophilic disorder, but also with personality disorders. On the other hand, NSO were more often diagnosed with substance-related disorders and psychotic disorders. This is in concordance with other studies of similar samples (5,26,36).

Still, looking into each of the diagnosed personality disorders, the most prevalent one in both SO and NSO was antisocial personality disorder. This is in accordance with the fact that most criminal offenders have either antisocial personality disorder or at least antisocial personality traits, and these traits make them prone to criminal activity without feeling any remorse (42). This finding is present in all the published research and it refutes the public/lay/ opinion about patients with severe psychotic mental disorders being prone to committing crime and being generally extremely dangerous (43). In our research, none of the SO and only 9.8% of the NSO were diagnosed with a psychotic disorder, once again proving that psychosis is not the main cause of criminal activity, neither in sexual, nor nonsexual crimes.

Significant difference between SO and NSO was found in the assessment of criminal responsibility, where of SO only one person was found not guilty by reason of insanity. This can once again refute the belief that most SO are mentally ill men waiting for their victims in dark alleys, but instead, they are mentally healthy men completely aware of their actions who can be held responsible for their actions by the court. The finding that more than 80% of NSO were found of reduced responsibility is probably due to the fact that courts would usually send people for psychiatric evaluations in cases when a mental disorder is suspected to play a role in committing the crime. With sex offences the lay people more often than professionals believe that a person is mentally disturbed, so more people with no mental disorders who commit a sexual crime would be sent for the evaluation compared to people committing a nonsexual crime (1,44).

The main strength of this study is the inclusion of the control group of individuals charged with nonsexual crimes, which gives us the insight into the differences between sex offenders and offenders of other types of crimes. Also, the strength is the possibility to use data from the very extensive evaluations (on average the evaluation is longer than 20 pages of text with data from court files, medical records, and personal history, as well as detailed description of diagnosis and criminal responsibility evaluations).

The major limitation of the study is the sample, as it consists of only those individuals charged with sexual offences that were sent for the court evaluations. Therefore, these results cannot be generalized to a broader population of sex offenders. The sample was limited due to the fact that the authors did not have access to psychiatric and forensic aspects of offenders who were not sent for evaluation. In addition, the authors were not blinded to the legal status of the individuals.

Conclusions

Our results show that individuals charged with sexual offences referred for assessment of criminal responsibility were different from individuals charged with nonsexual offences referred, in many characteristics of their personal history, offence characteristics, and forensic evaluations. Referred men charged with sex offences, compared to those charged with nonsex offences, more often committed their crimes alone, victimized children as equally often as adults and committed their crimes at home. They less often pleaded guilty. Individuals charged with sexual offences were more often abused in childhood, more often had antisocial personality disorder and paraphilias and less often substance-related disorders. The majority of referred men charged with sex offences were considered criminally responsible. These differences warrant different approaches to prevention of future re-offending, compared to nonsex offenders.

References

1. Miller RD, Stava LJ, Miller RK. The insanity defence for sex offenders: jury decisions after repeal of Wisconsin's sex crimes law. Hosp Community Psychiatry 1988;39(2):186–9. https://doi.org/10.1176/ps.39.2.186.
2. Higgins PL, Heath WP, Grannemann BD. How type of excuse defence, mock juror age, and defendant age affect mock jurors' decisions. J Soc Psychol 2007;147(4):371–92. https://doi.org/10.3200/SOCP.147.4.371-392.
3. Bradford JMW, Federoff P, Firestone P. Sexual violence and the clinician. In: Simon RI, Tardiff K, editors. Violence assessment and management. Arlington, VA: American Psychiatric Publishing, 2008;441–59.
4. Croatian criminal code. Narodne Novine. 2011;125:2498. https://www.legislationline.org/download/id/7896/file/Croatia_Criminal_Code_2011_en.pdf (accessed June 23, 2020).
5. Harsch S, Bergk JE, Steinert T, Keller F, Jockusch U. Prevalence of mental disorders among sexual offenders in forensic psychiatry and prison. Int J Law Psychiatry 2006;29(5):443–9. https://doi.org/10.1016/j.ijlp.2005.11.001.
6. Bradford JM. On sexual violence. Curr Opin Psychiatry 2006;19(5):527–32. https://doi.org/10.1097/01.yco.0000238483.34894.68.
7. Novak B, McDermott BE, Scott CL, Guillory S. Sex offenders and insanity: an examination of 42 individuals found not guilty by reason of insanity. J Am Acad Psychiatry Law 2007;35(4):444–50.
8. Stinson JD, Becker JV. Sexual offenders with serious mental illness: prevention, risk, and clinical concerns. Int J Law Psychiatry 2011;34 (3):239–45. https://doi.org/10.1016/j.ijlp.2011.04.011.
9. Valenca AM, Meyer LF, Freire R, Mendlowicz MV, Nardi AE. A forensic-psychiatric study of sexual offenders in Rio de Janeiro, Brazil. J Forensic and Legal Med 2015;31:23–8. https://doi.org/10.1016/j.jflm.2015.01.003.
10. Hoertel N, Le Strat Y, Schuster JP, Limosin F. Sexual assaulters in the United States: prevalence and psychiatric correlates in a national sample. Arch Sex Behav 2012;41(6):1379–87. https://doi.org/10.1007/s10508-012-9943-5.
11. Goreta M, Peko-Čović I, Buzina N, Majdančić Ž. Aktualna pitanja forenzičko-psihijatrijskih vještačenja seksulanih delinkvenata [Current issues relating to the forensic psychiatric evaluation of sexual delinquents]. Hrvat Ljetop Kazn Pravo Praksu 2004;11(1):201–16.
12. Koch J, Berner W, Hill A, Briken P. Sociodemographic and diagnostic characteristics of homicidal and nonhomicidal sexual offenders. J Forensic Sci 2011;56(6):1626–31. https://doi.org/10.1111/j.1556-4029.2011.01933.x.
13. Langevin R, Curnoe S. Are the mentally retarded and learning disordered overrepresented among sex offenders and paraphilics? Int J Offender Ther Comp Criminol 2014;52(4):401–15. https://doi.org/10.1177/0306624X07305826.
14. Baltieri DA, de Andrade AG. Alcohol and drug consumption among sexual offenders. Forensic Sci Int 2008;175(1):31–5. https://doi.org/10.1016/j.forsciint.2007.05.004.
15. Gilbert F, Focquaert F. Rethinking responsibility in offenders with acquired paedophilia: punishment or treatment? Int J Law Psychiatry 2015;38:51–60. https://doi.org/10.1016/j.ijlp.2015.01.007.
16. Buzina N, Jukic V, Arbanas G. Kazneno-pravni aspekti forenzičke psihijatrija [Criminal law and forensic psychiatry]. In: Jukic V, editor. Hrvatska psihijatrija početkom 21. stoljeća [Croatian psychiatry at the beginning of the 21st century]. Zagreb, Croatia: Medicinska naklada, Hrvatsko psihijatrijsko društvo and Klinika za psihijatriju Vrapce, 2018;330–6.
17. American Psychiatric Association. Diagnostic and statistical manual for mental disorders, 5th edn. Arlington, VA: American Psychiatric Association, 2013.
18. Mills JF, Anderson D, Kroner DG. The antisocial attitudes and associates of sex offenders. Crim Behav Mental Health 2004;14(2):134–45. https://doi.org/10.1002/cbm.578.
19. van Wijk A, Vermeiren R, Loeber R, Hart-Kerkhoffs L, Doreleijers T, Bullens R. Juvenile sex offenders compared to non-sex offenders: a review of literature 1995–2005. Trauma Violence Abuse 2006;74 (4):227–43. https://doi.org/10.1177/1524838006292519.
20. van Wijk AP, Mali BR, Bullens RA, Vermeiren RR. Criminal profiles of violent juvenile sex and violent juvenile non sex offenders: an explorative longitudinal study. J Interpers Violence 2007;22(10):1340–55. https://doi.org/10.1177/0886260507304802.
21. Wanklyn SG, Ward AK, Cormier NS, Day DM, Newman JE. Can we distinguish juvenile violent sex offenders, violent non-sex offenders, and versatile violent sex offenders based on childhood risk factors? J Interpers Violence 2012;27(11):2128–43.https://doi.org/10.1177/0886260511432153.
22. Woodworth M, Freimuth T, Hutton EL, Carpenter T, Agar AD, Logan M. High-risk sexual offenders: an examination of sexual fantasy, sexual paraphilia, psychopathy, and offence characteristics. Int J Law Psychiatry 2013;36(2):144–56. https://doi.org/10.1016/j.ijlp.2013.01.007.
23. Clemente M, Espinosa P, Padilla D. Moral disengagement and willingness to behave unethically against ex-partner in a child custody dispute. PLoS One 2019;14(3):e0213662. https://doi.org/10.1371/journal-pone.0213662.
24. Penfol PS. Mendacious moms or devious dads? Some perplexing issues in child custody/sexual abuse allegation disputes. Can J Psychiatry 1995;40(6):337–41. https://doi.org/10.1177/070674379504000610.
25. Gunter M, du Bois R, Eichner E, Rocker D, Boos R, Klosinski G, et al. Allegations of sexual abuse in child custody disputes. Med Law 2000;19 (4):815–25.
26. Silverthorne ZA, Quinsey VL. Sexual partner age preferences of homosexual and heterosexual men and women. Arch Sex Behav 2000;29 (1):67–76. https://doi.org/10.1023/a:1001886521449.
27. Fergusson DM, McLeod GF, Horwood LJ. Unemployment and psychosocial outcomes to age 30: a fixed-effects regression analysis. Aust N Z J Psychiatry 2014;48(8):735–42. https://doi.org/10.1177/0004867414525840.
28. Botrić V. Unemployed and long-term unemployed in Croatia: evidence from labour Force Survey. Rev za Soc Politiku 2009;16(1):25–44. https://doi.org/10.3935/rsp.v16i1.807.
29. Burrows K. Age preferences in dating advertisements by homosexuals and heterosexuals: from sociobiological to sociological explanations. Arch Sex Behav 2013;42(2):203–11. https://doi.org/10.1007/s10508-012-0031-7.
30. McKay MM, Chapman JW, Long NR. Causal attributions for criminal offending and sexual arousal: comparison of child sex offenders with other offenders. Br J Clin Psychol 1996;35(1):63–75. https://doi.org/10.1111/j.2044-8260.1996.tb01162.x.
31. DeLuca JS, Vaccaro J, Rudnik A, Graham N, Giannicchi A, Yanos PT. Sociodemographic predictors of sex offender stigma: how politics impact attitudes, social distance, and perceptions of sex offender recidivism. Int J Offender Ther Comp Criminol 2018;62(10):2879–96. https://doi.org/10.1177/0306624X17723639.
32. Room R, Rossow I. The share of violence attributable to drinking. J Subst Abuse 2001;6:218–28. https://doi.org/10.1080/146598901753325048.
33. Connor J, You R, Casswell S. Alcohol-related harm to others: a survey of physical and sexual assault in New Zealand. N Z Med J 2009;122 (1303):10–6.
34. Kanyanya IM, Othieno CJ, Ndetei DM. Psychiatric morbidity among convicted male sex offenders at Kamiti prison. Kenya. East Afr Med J 2007;84(4):151–5. https://doi.org/10.4314/eamj.v84i4.9518.

35. Hakkanen-Nyholm H, Repo-Tiihonen E, Lindberg N, Salenius S, Weizmann-Henelius G. Finnish sexual homicides: offence and offender characteristics. Forensic Sci Int 2009;188(f1–3):125–30. https://doi.org/10.1016/j.forsciint.2009.03.030.
36. Lindsay WR, Smith AHW, Law J, Quinn K, Anderson A, Smith A, et al. Sexual and nonsexual offenders with intellectual and learning disabilities a comparison of characteristics, referral patterns, and outcome. J Interpers Violence 2004;19(8):875–90. https://doi.org/10.1177/0886260504266884.
37. Davis KA, Knight RA. The relation of childhood abuse experiences to problematic sexual behaviors in male youths who have sexually offended. Arch Sex Behav 2019;48(7):2149–69. https://doi.org/10.1007/s10508-108-1279-3.
38. Dargis M, Newman J, Koenigs M. Clarifying the link between childhood abuse history and psychopathic traits in adult criminal offenders. Personal Disord 2016;7(3):221–8. https://doi.org/10.1037/peroooo147.
39. Polanczyk GV, Zavaschi ML, Benetti S, Zenker R, Gammermam PW. Sexual violence and its prevalence among teenagers of Porto Alegre. Brazil. Rev Saude Publica 2003;37(1):8–14. https://doi.org/10.1590/s0034-89102003000100004.
40. Widom CS, Massey C. A prospective examination of whether childhood sexual abuse predicts subsequent sexual offending. JAMA Pediatr 2015;169(1):e143357. https://doi.org/10.1001/jamapediatrics.2014.3357.
41. Arbanas G. Who treats sexual problems in Croatian health system? Soc Psihijatr 2019;47(1):102–12. https://doi.org/10.24869/spsih.2019.102.
42. Holoyda BJ, McDermott BE, Newman WJ. Insane sex offenders: psychiatric and legal characteristics of sexual offenders found not guilty by reason of insanity. J Forensic Sci 2017;63(4):1207–14. https://doi.org/10.1111/1556-4029.13707.
43. Arbanas G, Rozman J, Bagarić S. The attitudes of medical doctors, nurses and lay people towards schizophrenia, depression and PTSD. Psychiatr Danub 2019;31(Suppl 1):84–91.
44. O’Shaughnessy RJ. Commentary: sex offenders and insanity. J Am Acad Psychiatry Law 2007;35:451–53.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14512
Available online at: onlinelibrary.wiley.com

PAPER

PSYCHIATRY & BEHAVIORAL SCIENCE

Alden J. Parker,[1,2] B.S.; Abby L. Mulay,[1] Ph.D.; and Emily D. Gottfried,[1] Ph.D.

The Personality Assessment Inventory (PAI): Treatment Scales and Interpersonal Characteristics in a Sample of Men Charged with or Convicted of a Sexual Offense*

ABSTRACT: An individual's interpersonal features are pertinent to treatment within clinical populations. The Personality Assessment Inventory (PAI) contains two scales that assess the interpersonal features of warmth (WRM) and dominance (DOM), as well as two additional measures to assess to treatment prediction, process, and rejection (RXR; TPI). The current study examined associations between these PAI scales in a sample of 92 men who underwent comprehensive evaluations of sexual behavior after being charged with or convicted of a sexual offense. Analyses indicated that RXR was positively associated with WRM and DOM, TPI was negatively associated with WRM, and the two interpersonal scales of WRM and DOM were positively correlated with each other. A significant inverse relationship was found between the two treatment scales RXR and TPI indicating that motivation for treatment may have a limited relationship with the treatment process. WRM significantly predicted scores on the TPI, and both WRM and DOM predicted individual scores on RXR. Higher scores on positive impression management (PIM) were predictive of lower TPI and higher RXR, as individuals with higher stakes cases may score higher on PIM and underreport obstacles within treatment or be unwilling to accept the need for treatment. Overall, findings suggest that interpersonal characteristics identified by the PAI scales may be advantageous in approaching treatment within this population.

KEYWORDS: sexual offending, Personality Assessment Inventory, personality pathology, interpersonal characteristics, treatment prediction, treatment motivation

There have been many personality assessment tools developed that have been utilized within evaluations of criminal defendants, such as the Personality Assessment Inventory (PAI; [1,2]). The PAI is a multi-scale, self-report questionnaire that assesses psychological and personality constructs that are applicable to anticipating treatment motivation and behavior. The PAI's clinical utilization in justice-involved persons can help identify individual characteristics to aid in institutional classification and treatment (3–5). Further, the accurate assessment of personality pathology of individuals convicted of sexual offenses is critical in order to provide proper treatment plans and eventually evaluate treatment progress for rehabilitation purposes (6–11).

The PAI additionally addresses some of the limitations of other personality assessment scales, such as the Minnesota Multiphasic Personality Inventory-2-Restructured Form (MMPI-2-RF; [12,13]) and the Millon Clinical Multiaxial Inventory-III (MCMI-III; [14]). For example, in comparison to the higher reading levels of the MCMI-III and the MMPI-2-RF, the responses of the PAI's self-report format are brief items that require fourth-grade reading skills (2). Further, in contrast to other personality measures, the PAI contains two ways of assessment developed to consider treatment process issues: The Treatment Rejection (RXR) scale and the Treatment Process Index (TPI). The RXR scale assesses treatment motivation levels, with low scores indicating high motivation for treatment and high scores indicating low motivation for treatment. However, since motivation for treatment does not always equal a good outcome, the TPI assesses difficulty of the treatment process and is composed of 12 features or obstacles that may be present during treatment, such as hostility, low motivation, defensiveness, and low perceived social support (2). These features have also been shown to be highly correlated with withdrawal, hostility, and alienation (2). The TPI is highly correlated with Borderline Features (BRD), Antisocial Features (ANT), and Paranoia (PAR) scales of the PAI (1,2). Lower scores on the TPI indicate correlates that are likely to aid in the treatment process while higher scores are indicative of items that may suggest a challenging treatment process.

Prior research has examined the RXR and TPI within a variety of samples, and these scales have been shown to be positively correlated with treatment noncompliance, treatment withdrawal, and poor treatment utilization (8,15,16). RXR has been shown to significantly differentiate individuals actively engaged in treatment and those not currently receiving treatment, in samples of clinical verses nonclinical subjects (3,17,18). Previous literature

[1] Community and Public Safety Psychiatry Division, Department of Psychiatry and Behavioral Sciences, Medical University of South Carolina, 171 Ashley Ave Suite 419, Charleston, SC, 29425.

[2] Clemson University, 105 Sikes Hall, Clemson, SC, 29634.

Corresponding author: Emily D. Gottfried, Ph.D. E-mail: gottfrem@musc.edu

*Presented at the 72nd Annual Scientific Meeting of the American Academy of Forensic Sciences, February 17–22, 2020, in Anaheim, CA.

Received 7 April 2020; and in revised form 20 May 2020; accepted 25 June 2020.

has also reported that the TPI scale predicted treatment completion rates from noncompletion rates in specialized treatment settings, such as an outpatient training clinic, an outpatient treatment center for chronic pain, and court-mandated outpatient centers for sexual behavior and substance abuse treatment (19–22). However, a 2012 study (23) reported that neither RXR nor TPI predicted treatment progress in a court-mandated residential substance abuse treatment program. Similarly, Charnas et al. (15) reported no significant findings for the TPI in a university-based outpatient psychotherapy facility. The PAI has been utilized for diagnoses and assessment in correctional settings, and the RXR scale and TPI index have been shown to be related to treatment outcome and completion. However, they have not yet been thoroughly examined within samples of individuals charged with or convicted of sexual offenses, a population that is both justice-involved as well as frequently seeking or mandated for treatment. The exploration of the RXR scale and the TPI index within assessment may be advantageous in aiding understanding of potential concerns within treatment and rehabilitation in this population.

The PAI contains additional scales that assess psychological and personality constructs that may aid in the treatment process for offenders, such as the Aggression (AGG) and Nonsupport (NON; i.e., a subject's perceived social support) scales. In prior literature, the AGG scale was significantly correlated with institutional misconduct and treatment noncompliance within men in a general correctional setting (4,24). Further, higher social support has been shown to be a positive predictor of remaining in substance abuse treatment and positive outcome of treatment for adults and adolescents (25,26). Within adult populations of individuals convicted of sexual offenses, higher levels of violence were found to be correlated with lower levels of perceived social support (27), and within a sample of juvenile sexual offenders, low levels of social support predicted higher number of offenses and antisocial behaviors (28). The AGG and ANT scales have been the focus of the majority of PAI studies conducted with offenders in recent years (e.g., Refs. [3,4]), and although those scales are important in predicting treatment process and behavior, the interpersonal functioning of offenders remains a topic that has received less attention in the literature (24,29). However, interpersonal functioning in personality pathology remains an important topic in psychological assessment (24,30–34). Hopwood et al. (35) argues that at its core, personality pathology is "fundamentally interpersonal," and that in order to significantly enhance the clinical utility of personality assessment measures, a patient's interpersonal structure and perspective must be understood (p. 13). Therefore, as the ANT and AGG scales have been specifically examined within offender populations and interpersonal functioning has not been broadly discussed in offender literature, it is important to examine these features within offender populations as the accurate assessment of interpersonal functioning is often important to the clinical utility of a measure (24,30,33,34).

Research has further shown that self-reported interpersonal features predicted later ratings of treatment collaboration, as well as therapist ratings of personality characteristics in short-term outpatient therapy (36–38). Specifically, hostile dominance was negatively associated with therapist alliance and treatment compliance, while friendly submissiveness (i.e., extreme agreeableness) was positively related to therapist alliance, openness, and treatment compliance (36,38). Additionally, an individual's observed interpersonal style has been reported to impact therapist treatment approach and lead to positive outcomes in short-term therapy for depression (37). Specific to individuals convicted of a sexual offense, Hudson and Ward (39) argued that interpersonal characteristics, such as empathy and intimacy, are critical to treatment, rehabilitation, and classification.

There are two interpersonal scales within the PAI, Warmth (WRM; degree to which an individual acts kind, empathic, and engaging in social situations) and Dominance (DOM; degree to which an individual acts dominant, assertive, and controlling in social situations), that provide information that is critical to the evolving research surrounding the treatment process within offender populations (24,29,40,41). For example, previous studies have shown that individuals convicted of child pornography offenses often score low on measures of expressed hostility, warmth, and dominance (29,40). These findings support treatment programs developed for individuals who commit internet sexual offenses to address specific needs like achieving healthy intimate relationships and building social networks (42). Additionally, studies have shown that in male criminal offender populations, low warmth and high dominance have been associated with antisocial traits, aggression, misconduct, and noncompliance (24,41). In a sample of male prison inmates, it was reported that a more dominant and cold interpersonal style statistically interacted to predict aggressive behavior and recorded misconduct while dominance was a unique predictor of treatment failure and noncompliance ratings (24). Therefore, in order to accurately assess and predict treatment compliance and outcome in justice-involved individuals, it is essential that the interpersonal scales of WRM and DOM be examined.

The application of these interpersonal characteristics in order to accurately assess, predict, and promote successful treatment processes is critical for rehabilitation of individuals convicted of or charged with a sexual offense. The present study sought to expand the developing PAI research to explore the relationship between treatment predictors and interpersonal characteristics among men who have been convicted of or charged with a sexual offense. It was hypothesized that RXR and TPI would be negatively associated with WRM and positively associated with DOM. In other words, it was expected that individuals with lower levels of warmth and higher levels of dominance would be less motivated for the treatment process and were predicted to experience a greater amount of challenges within treatment. Additionally, the present study seeks to explore relationships between the PAI scales of interest and the examinee's age at assessment, treatment history, rating of self-esteem, history of disciplinary infractions within the correctional institution, history of parental mental illness and criminal offending, and victim age and gender.

Methods

Participants

Data were collected from the forensic reports of the evaluations of 101 men located in the Southeastern United States who underwent comprehensive evaluations of sexual behavior after being charged with or convicted of a sexual offense or professional sexual boundary violation. After invalid PAIs were removed, the final sample was composed of 92 men who ranged in age from 19 to 80 years old (M = 44.84; SD = 14.79). Independent samples t-tests were conducted to determine if there were significant differences in the PAI scores of interest between those who had been convicted of the sexual offense(s) from those that had allegations. As there were no significant differences between these two groups, the analyses were conducted on the entire sample. See Table 1 for demographic information of the sample.

Measures

Personality Assessment Inventory (PAI). The PAI (1,2) is a 344-item self-report questionnaire that objectively assesses domains of adult personality and psychopathology. The PAI consists of 22 individual scales that are rated using a four-point scale from false to very true, including four validity scales, 11 clinical scales, five treatment-related scales, and two interpersonal scales. All 22 scales are nonoverlapping, and internal consistency for scores on all scales has been good to strong (e.g., $\alpha > 0.80$) in PAI normative samples (1,2). The current study utilized the four validity scales of Positive Impression Management (PIM), Negative Impression Management (NIM), Inconsistency (ICN), and Infrequency (INF) to eliminate invalid profiles. This study focused on the Treatment Rejection (RXR), Warmth (WRM), Dominance (DOM), Aggression (AGG), and Nonsupport (NON) scales, as well as the Treatment Process Index (TPI). According to the PAI manual (2), the WRM, DOM, and RXR scales and the TPI index have been demonstrated to be internally valid (0.66–0.89). Additionally, the PAI manual reports convergent and discriminant validity with over 50 measures of psychopathology; specifically, the RXR was negatively associated with the Beck Depression Inventory ([43]; $r = -0.30$ to -0.40), negatively associated with the "Poor Morale" MMPI content scale by Wiggins ([44,45]; $r = -0.78$), and positively related to measures of perceived social support in college students ($r = 0.26$–0.49). TPI has been found to be a significant predictor of treatment outcome status as well as disruptive behavior in clinical settings (19,23). However, the PAI manual does not report reliability values for TPI scores as it is an index based on scale combinations. This sample did not include individual data for scales at the item level, so analyses to determine internal reliability in this study were not conducted.

TABLE 1—*Participants.*

Demographic Characteristics			
Characteristic	*n*	Range	*M* (*SD*)
Age (years)	92	19–80	44.84 (14.79)
Age of first self-reported arrest (years)	77	9–70	27.92 (13.29)
Self-esteem rating (1–10)	86	1–10	7.54 (1.96)
	N		Percentage
Conviction/Allegation			
Allegation	39		42.4
Conviction	52		56.5
Victim gender			
Male	5		5.4
Female	54		58.7
Combination	33		35.9
Victim age			
Child ≤ 12	36		39.1
Teen 13–17	12		13.0
Adult 18+	11		12.0
Combination	33		35.9
Parental mental illness history			
No	73		79.3
Yes	6		6.5
Unknown	12		13.0
Parental criminal history			
No	67		72.8
Yes	15		16.3
Unknown	10		10.9
Prior treatment (MH/SA)			
No	47		51.1
Yes	43		46.7
Prior treatment (SB)			
No	55		59.8
Yes	35		38.0

M, mean; MH/SA, mental health/substance abuse; *N*, total number; SB, sexual behavior; *SD*, standard deviation.

Clinical interviews. Examinees underwent a clinical interview with a forensic psychiatrist or psychologist with expertise in sexual behavior in which their developmental, educational, occupational, social, general medical, psychiatric, substance use, medication, sexual behavior, legal, and treatment histories were assessed. From the report based on these interviews, the first author coded the following variables from the forensic reports authored for each case: self-reported history of arrests prior to age 18 (0 = no, 1 = yes); parental history of mental illness and criminal offending (0 = no, 1 = yes); prior mental health, substance, or sexual behavior treatment history (0 = no, 1 = yes); and self-esteem rating (on a scale from 1 to 10 with 10 representing the highest level of self-esteem).

Official documentation. The Record of Arrest and Prosecutions (RAP) sheets of each subject was examined to determine the number of convictions and age at first arrest. Police reports and arrest warrants were utilized to determine ages and sex of the victim(s). If applicable, documents from the department of corrections were examined to determine the number of recorded disciplinary infractions for each examinee.

Procedure

Male examinees who were at least 18 years of age and underwent sexual behavior evaluations from 2013 to 2018 and completed a PAI were included as subjects. Prior to their evaluation, each examinee signed a consent form agreeing to have their deidentified data used in future research projects. All procedures were approved by the university's Institutional Review Board (IRB). The following exclusionary criteria were utilized from the PAI manual (2) regarding validity: $T \geq 73$ on the Inconsistency (INC) scale (four subjects were excluded); $T \geq 75$ on the Infrequency (INF) scale (three additional subjects were excluded); $T \geq 92$ on the Negative Impression Management (NIM) scale (two additional subjects were excluded); and $T > 70$ on the Positive Impression Management (PIM) scale (no additional subjects were excluded).

Results

Descriptive Statistics

T-scores for predictor and outcome variables within the PAI are presented in Table 2.

Correlations

The bivariate and biserial correlations between the PAI scales and other measures are presented in Table 3. As predicted, DOM was significantly positively associated with RXR; however, the relationship between TPI and DOM was nonsignificant. WRM was also significantly positively associated with RXR and, as expected, was significantly negatively associated with

TABLE 2—*Results.*

	Minimum	Maximum	*M*	*SD*
NON	36.00	97.00	49.71	11.65
AGG	32.00	85.00	46.15	9.83
TPI	.00	102.00	53.24	13.73
RXR	20.00	68.00	45.37	10.48
DOM	22.00	76.00	52.49	9.57
WRM	19.00	72.00	50.86	10.62
PIM	22.00	70.00	51.60	10.58

N = 92.

AGG, Aggression Scale; DOM, Dominance Scale; *M*, mean; NON, Nonsupport Scale; RXR, Treatment Rejection Scale; *SD*, standard deviation; TPI, Treatment Process Index; WRM, Warmth Scale (1,2).

the TPI. Correlations between parental history of mental illness and criminal history were additionally nonsignificant with WRM and DOM.

Additional analyses revealed a significant inverse relationship between RXR and the TPI scales, as well as a significant positive relationship between WRM and DOM scales. Age of the examinee at the time of the evaluation was unrelated to RXR. However, age was significantly negatively related to the TPI. The RXR, WRM, and DOM scales showed significant positive correlations with the self-esteem rating, while the TPI displayed a significant negative correlation with self-esteem.

Prior treatment for sexual behavior was not significantly correlated with any variables of interest; however, prior treatment for mental health or substance abuse problems was negatively associated with RXR. Additionally, prior treatment for mental health and substance abuse problems was significantly positively related to DOM, but it was not significantly related to WRM. Number of recorded institutional disciplinary infractions was significantly positively associated with AGG, yet significantly negatively correlated with the TPI.

The AGG scale exhibited significant positive correlations with the TPI and NON scales. AGG also displayed significant negative correlations with RXR and WRM, yet it was unrelated to DOM. The NON scale was significantly positively related to the TPI and significantly negatively related to WRM, DOM, and RXR. Victim gender (female) presented a significant positive relationship with WRM and was unrelated to the other PAI scales of interest.

The PIM scale revealed a significant positive relationship with the WRM, RXR, and DOM scales. PIM was additionally significantly negatively correlated to the AGG and NON scales, as well as the TPI.

Regression Analyses

A multiple regression was conducted to determine if the interpersonal characteristics of dominance and warmth could significantly predict examinees' scores on the TPI over and above scores on PIM. The results of the regression indicated that the model was significant, $F(3, 88) = 11.72$, $p < 0.001$, and explained 28.50% of the variance. WRM ($B = -0.15$; $p < 0.001$) and PIM ($B = -0.17$; $p = 0.001$) significantly predicted scores on the TPI, but DOM did not significantly predict scores on the TPI ($B = 0.06$; $p = 0.17$).

Additionally, a multiple regression was conducted to determine if dominance and/or warmth significantly predicted examinees' scores on the RXR. The results of the regression indicated that the model was significant, $F(3, 8) = 14.97$; $p < 0.001$ and

TABLE 3—*Correlations.*

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	WRM															
2	AGG	−0.32**														
3	NON	−0.62**	0.27**													
4	RXR	0.36**	−0.35**	−0.30**												
5	TPI	−0.44**	0.44**	0.60**	−0.30**											
6	DOM	0.38**	0.03	−0.27**	0.32**	−0.11										
7	PIM	0.34**	−0.6**	−0.32**	0.52**	−0.41**	0.28*									
8	Disciplinary infractions	−0.11	0.31*	0.16	0.06	−0.31*	0.14									
9	Age	0.04	−0.2	−0.21*	−0.03	−0.28**	0.01	−0.35*								
10	Self-esteem rating	0.32**	−0.19	−0.24*	0.35**	−0.26*	0.22*	0.13	−0.04							
11	Victim gender	0.38**	−00.08	−00.13	0.17	−0.07	0.13	−0.31	−0.01	0.05						
12	Victim age	0.12	0.01	−0.11	0.23	−0.23	0.02	0.18	0.04	−0.13	−0.01					
13	Parental mental illness	−0.13	.	0.24*	−0.04	0.19	−0.02	−0.12	−0.01	0.02	−0.11	−0.01				
14	Parental criminal behavior	−0.21	0.33**	0.32**	−0.21	0.27*	0.11	0.22	−0.32**	−0.07	−0.03	−0.29*	0.02			
15	Prior treatment (MH/SA)	−0.08	0.24*	0.05	−0.23*	0.19	−0.25*	0.24	−0.14	−0.23*	−0.12	0.17	0.1	−0.01		
16	Prior treatment (SB)	−0.09	−0.1	−0.07	−0.17	−0.12	−0.08	.	0.2	0.05	−0.16	−0.22	0.06	0.01	0.1	

N = 92.

AGG, Aggression scale; DOM, Dominance scale (8); MH/SA, mental health/substance abuse; NON, Nonsupport scale; PIM, Positive Impression Management; RXR, Treatment Rejection scale; SB, sexual behavior; TPI, Treatment Process Index; WRM, Warmth scale.

*$p < 0.05$.

**$p < 0.01$

explained 33.8% of the variance. Neither WRM ($B = 0.14$; $p = 0.08$) or DOM ($B = 0.13$; $p = 0.15$) significantly predicted scores on RXR; PIM ($B = 0.45$; $p < 0.001$) significantly predicted scores on RXR.

Discussion

This study investigated the relationship between the PAI indicators relevant to treatment outcomes and interpersonal characteristics of men charged with or convicted of a sexual offense. There are few prior studies that examined the RXR and TPI and the WRM and DOM scales specific to sexual offenses, yet studies have shown the importance of interpersonal characteristics within treatment compliance and completion among justice-involved individuals (24,29,39–41). The present study builds off of this literature by examining the relationships between the treatment rejection, treatment process, warmth, and dominance scales within a sample of men charged or convicted with a sexual offense.

First, the hypothesis that there would be a negative association between the RXR scale and the WRM scale was not supported. Specifically, the present study found that those who scored higher in warmth also were shown to have less motivation for treatment, as measured by the PAI. This finding contradicts prior literature examining the PAI scales; however, there has been some concern from forensic professionals about the validity of self-report in offender populations, especially in reporting interpersonal characteristics, which could account for this finding (10,46). Second, the hypothesis that there would be a positive association between the RXR scale and the DOM scale was supported. This finding supports the notion that those less motivated for the treatment process present with higher levels of dominance. It also corresponds with the previous literature, as dominance has been shown to be negatively related to treatment compliance in clinical inpatient and outpatient populations (36,38,47). Additionally, previous studies reported that within male general offender populations, high levels of dominance were positively associated with institutional noncompliance (24,41).

Third, the hypothesis that there would be a negative association between the TPI and the WRM scale was supported, as individuals who scored higher on the TPI additionally scored lower on warmth. This finding supports the notion that individuals with higher levels of predicted obstacles in treatment may also present with lower levels of warmth. This result aligns with the prior literature within male offender populations that reported colder interpersonal styles were related to treatment failure, noncompliance, and institutional misconduct (41,47). Finally, the hypothesis that there would be a positive association between the TPI and DOM was not supported, as no significant relationship between dominance and predicted obstacles in treatment was found. Prior studies have seen dominance as a unique predictor of obstacles in treatment in male offender populations; however, our finding could offer an alternative hypothesis in that dominance is not as unique a predictor for the treatment process as previously thought (6,24).

Further analyses demonstrated there was a significant positive relationship between WRM and DOM. However, these results also contradict prior studies of male offender populations where warmth and dominance styles were not positively related (23). This finding is additionally interesting because, in interpersonal literature, dominance and warmth have been conceptualized as opposing poles on the interpersonal complex (48,49). The current study could present an alternative view on the circumplex, as individuals within this population could potentially relate differently within the poles and warmth and dominance.

An inverse relationship was found between RXR and TPI, which is not congruent with the majority of the previous literature, suggesting that the RXR and TPI may not be as similar in predicting the treatment process as previously thought. However, this inverse relationship was noted in McCredie et al.'s (16) study. From a theoretical perspective, an individual might be highly motivated to attend treatment (i.e., low scores on the RXR), but experience significant difficulties in the treatment process (i.e., high scores on the TPI). Practically speaking, an individual with symptoms or traits associated with borderline personality disorder might experience distress related to their symptoms, as well as be motivated to change their maladaptive behaviors and emotional expressions, but other aspects of the disorder (e.g., impulsive engagement in risky behaviors) might interfere with their ability to fully participate in the treatment process. In other words, motivation and treatment process difficulties are not mutually exclusive. The utility of the RXR and the TPI has been examined in other prior studies, primarily in relation to clinical populations and the validity of a self-report measure to predict treatment process and outcome (8,23,24), and some have questioned the clinical utility of the TPI (e.g., [23]). As such, some studies have begun to explore the value of staff report or other external ratings of interpersonal characteristics for validity in predicting compliance levels and treatment outcomes (e.g., [47]).

Prior sexual behavior treatment did not yield any significant relationships with the variables of interest; however, prior mental health or substance use treatment showed a negative correlation with RXR, conveying those who reported receiving prior treatment for mental health or substance use additionally presented with lower motivation for treatment. This is contrary to previous literature within samples of adults and adolescents who had received treatment for substance use and mental health, as previous studies have shown either no relation between prior treatment and motivation (50,51), or a positive correlation between prior treatment and motivation (52). This finding could be due to differences in justice-involved populations, duration, and effectiveness of prior treatment, as well as the reliance of self-report of prior treatment in our study. Prior treatment for mental health or substance use additionally was positively related to DOM, conveying those who reported receiving prior treatment for mental health or substance use also presented with higher levels of dominance. These results could potentially be impacted by experience, duration, and effectiveness of previous treatment, as prior studies within clinical and sexual offender populations have shown a correlation between poor treatment experience and hostile dominance (53). Additionally, in our sample, there was a significant negative relationship between age and TPI, suggesting that younger participants presented with higher levels of predictors of obstacles within the treatment process. This finding is similar to other research in offender populations of which age has been shown to correlate with treatment completion and aggression (4,9).

Further exploratory analyses showed that individuals scoring higher on AGG obtained lower scores on RXR, which contradicts previous literature where aggression was positively correlated with treatment noncompliance and institutional misconduct (8,23,54). This result potentially highlights the importance of the additional interpersonal measures in evaluating treatment prediction scales. Similarly, in regard to the NON scale, individuals reporting lower levels of perceived social support showed higher levels of TPI. Individuals reporting higher levels of perceived social support scored lower in WRM, DOM, and RXR. These findings are partially congruent with prior studies in that individuals with lower perceived social support have been shown to show obstacles within treatment and portray higher levels of hostility and dominance (27,28). However, our results also seemingly contradict prior literature as perceived social support has been found to increase levels of interpersonal characteristics akin to warmth as well as positive treatment outcomes in clinical settings (25,26).

Finally, an exploratory regression analysis indicated that higher degrees of WRM were predictive of lower TPI scores, which assess correlates that are likely to aid in the treatment process. This finding is consistent with research suggesting that interpersonal warmth is an important aspect of treatment outcome (e.g., Ref. [55]). Further, higher scores on PIM were predictive of lower scores on TPI, indicating those who endorsed higher levels of positive impression management endorsed fewer obstacles to treatment. This finding makes conceptual sense, as it would be expected that individuals who are involved in high-stakes evaluations might engage in positive impression management and thereby underreport any issues related to obtaining treatment. An additional exploratory regression analysis revealed DOM and WRM did not significantly predict scores on RXR. However, PIM significantly predicted scores on RXR, such that higher levels of positive impression management were associated with higher levels of treatment rejection. This finding suggests that individuals who are prone to engage in positive impression management may also be unwilling to endorse common shortfalls most are willing to acknowledge and, subsequently, may not perceive the need to engagement in treatment. Previous research suggests individuals who were convicted of sexual offenses have been shown to score significantly lower on WRM and DOM; for example, in a study by Laulik et al., (29), the authors found that, on average, individuals convicted of internet-based sexual offenses against children demonstrated *T*-scores around 40 on the DOM and WRM scales. The current study sample demonstrated *T*-scores that are consistent with the normative sample, which is not suggestive of interpersonal dysfunction. It is possible, therefore, that the RXR should not be used as the sole measure of treatment motivation, given its lack of observed relationship with DOM and WRM.

Our study could potentially be limited by our small sample size, and future studies should be conducted to test these relationships in a larger number of participants. Additionally, our study consisted only of male participants, and in order to obtain a more representative sample, the inclusivity of other genders should be met in future research. Moreover, individuals who were charged and found not guilty may not have presented results similar to those convicted of a sexual offense. However, since the data were de-identified to protect individual privacy, the sample consisted of all individuals who underwent evaluations regardless of future legal outcome. Further, the majority of PAI scales of interest obtained average mean *T*-scores, with the exception of the AGG and RXR scales, of which the mean was relatively low. Average mean scores were also shown on the PIM scale which is surprising due to the nature of the sample. That is, subject data came from forensic sexual behavior evaluations and examinees could have been motivated to portray themselves positively. There was also no measure for interpersonal characteristics or treatment prediction that was obtained from an external report, as our study relied on the participant-reported scores of the PAI. Our study utilized many external sources such as official documentation for victim gender/age; however, because of lack of access to documentation for variables, such as prior treatment or parental behavior, our study largely relied on self-report from participants. The reliability of self-report within populations of justice-involved individuals has been questioned in prior studies, and although the PAI does include scales for impression management, future research can potentially involve external sources or staff ratings of participants (10,46).

Findings suggest that the consideration of interpersonal characteristics can be advantageous to the assessment and prediction of treatment in men charged with or convicted of a sexual offense. Individuals conducting treatment can utilize results from assessment tools such as the PAI to evaluate interpersonal characteristics and treatment prediction scales in order to promote initiatives to encourage motivation, completion, and treatment alliance. Like previously mentioned, future studies should include a more representative sample in size, gender, and location, as well as focus on the expansion of the assessment of interpersonal characteristics and treatment prediction beyond self-directed measures. Further, exploring the relationships between self-reported measures and external report measures within interpersonal characteristics and treatment prediction and outcome should be conducted in future studies. Finally, because interpersonal features do show such importance within treatment and rehabilitation within justice-involved individuals, future studies should elaborate on the development of these characteristics, as well as the cultivation of features that aid the treatment and rehabilitation process.

References

1. Morey LC. The Personality Assessment Inventory professional manual. Odessa, FL: Psychological Assessment Resources, 1991.
2. Morey LC. Personality Assessment Inventory professional manual, 2nd edn. Odessa, FL: Psychological Assessment Resources, 2007;11, 27–31, 44, 46–9, 131–289.
3. Edens JF, Cruise KR, Buffington-Vollum JK. Forensic and correctional applications of the personality assessment inventory. Behav Sci Law 2001;19(4):519–43. https://doi.org/10.1002/bsl.457.
4. Gardner BO, Boccaccini MT, Bitting BS, Edens JF. Personality Assessment Inventory scores as predictors of misconduct, recidivism, and violence: a meta-analytic review. Psychol Assess 2015;27(2):534–44. https://doi.org/10.1037/pas0000065.
5. Mullen KL, Edens JF. A case law survey of the Personality Assessment Inventory: examining its role in civil and criminal trials. J Pers Assess 2008;90(3):300–3. https://doi.org/10.1080/00223890701885084.
6. Boccaccini MT, Murrie DC, Hawes SW, Simpler A, Johnson J. Predicting recidivism with the Personality Assessment Inventory in a sample of sex offenders screened for civil commitment as sexually violent predators. Psychol Assess 2010;22(1):142–8. https://doi.org/10.1037/a0017818.
7. Boccaccini MT, Turner D, Murrie DC, Rufino K. Do PCL-R scores from state or defense experts best predict future misconduct among civilly committed sexual offenders? Law Hum Behav 2012;36(3):159–69. https://doi.org/10.1037/h0093949.
8. Caperton JD, Edens J, Johnson JK. Predicting sex offender institutional adjustment and treatment compliance using the Personality Assessment Inventory. Psychol Assess 2004;16(2):187–91. https://doi.org/10.1037/1040-3590.16.2.187.
9. Clegg C, Fremouw W, Horacek T, Cole A, Schwartz R. Factors associated with treatment acceptance and compliance among incarcerated male sex offenders. Int J Offender Ther Comp Criminol 2011;55(6):880–97. https://doi.org/10.1177/0306624X10376160.
10. Edens J, Hart S, Johnson D, Johnson J, Olver M. Use of the Personality Assessment Inventory (PAI) to assess psychopathy in offender populations. Psychol Assess 2000;12(2):132–9. https://doi.org/10.1037//1040-3590.12.2.132.
11. Langevin R. Acceptance and completion of treatment among sex offenders. Int J Offender Ther Comp Criminol 2006;50(4):402–17. https://doi.org/10.1177/0306624X06286870.

12. Ben-Porath YS, Tellegen A. The Minnesota Multiphasic Personality Inventory-2 Restructured Form: manual for administration, scoring, and interpretation. Minneapolis, MN: University of Minnesota Press, 2008/2011.
13. Tellegen A, Ben-Porath YS. The Minnesota Multiphasic Personality Inventory-2 Restructured Form: technical manual. Minneapolis, MN: University of Minnesota Press, 2008/2011.
14. Millon T. The Millon clinical multiaxial inventory, 3rd edn. Minneapolis, MN: National Computer Systems, 1994.
15. Charnas JW, Hilsenroth MJ, Zodan J, Blais MA. Should I stay or should I go? Personality Assessment Inventory and Rorschach indices of early withdrawal from psychotherapy. Psychotherapy 2010;47(4):484–99. https://doi.org/10.1037/a0021180.
16. McCredie MN, Kurtz JE, Valentine L. Prediction of psychotherapy process and outcome with the Personality Assessment Inventory. Psychiatry Res 2018;269:455–61. https://doi.org/10.1016/j.psychres.2018.08.110.
17. Alterman AI, Zaballero AR, Lin MM, Siddiqui N, Brown LS, Rutherford MJ, et al. Personality Assessment Inventory (PAI) scores of lower-socioeconomic African American and Latino methadone maintenance patients. Assessment 1995;2(1):91–100. https://doi.org/10.1177/107319119500200109.
18. Boyle GJ, Ward J, Lennon TJ. Personality Assessment Inventory: a confirmatory factor analysis. Percept Mot Skills 1994;74(3):1441–2. https://doi.org/10.2466/pms.1994.79.3f.1441.
19. Hopwood C, Ambwani S, Morey L. Predicting nonmutual therapy termination with the personality assessment inventory. Psychother Res 2007;17(6):706–12. https://doi.org/10.1080/10503300701320637.
20. Hopwood CJ, Baker KL, Morey LC. Extratest validity of selected Personality Assessment Inventory scales and indicators in an inpatient substance abuse setting. J Pers Assess 2008;90(6):574–7. https://doi.org/10.1080/00223890802388533.
21. Hopwood CJ, Creech SK, Clark TS, Meagher MW, Morey LC. Predicting the completion of an integrative and intensive outpatient chronic pain treatment with the Personality Assessment Inventory. J Pers Assess 2008;90(1):76–80. https://doi.org/10.1080/00223890701693785.
22. Percosky AB, Boccaccini MT, Bitting BS, Hamilton PM. Personality Assessment Inventory scores as predictors of treatment compliance and misconduct among sex offenders participating in community-based treatment. J Forensic Psychol Res Pract 2013;13(3):192–203. https://doi.org/10.1080/15228932.2013.795425.
23. Magyar MS, Edens JF, Lilienfeld SO, Douglas KS, Poythress NG, Skeem JL. Using the Personality Assessment Inventory to predict male offenders' conduct during and progression through substance abuse treatment. Psychol Assess 2012;24(1):216–25. https://doi.org/10.1037/a0025359.
24. Edens JF. Interpersonal characteristics of male criminal offenders: personality, psychopathological, and behavioral correlates. Psychol Assess 2009;21(1):89–98. https://doi.org/10.1037/a0014856.
25. Dobkin PL, Civita MD, Paraherakis A, Gill K. The role of functional social support in treatment retention and outcomes among outpatient adult substance abusers. Addiction 2002;97(3):347–56. https://doi.org/10.1046/j.1360-0443.2002.00083.x.
26. Richter SS, Brown SA, Mott MA. The impact of social support and self-esteem on adolescent substance abuse treatment outcome. J Subst Abuse Treat 1991;3(4):371–85. https://doi.org/10.1016/S0899-3289(10)80019-7.
27. Gutiérrez-Lobos K, Eher R, Grünhut C, Bankier B, Schmidl-Mohl B, Frühwald S, et al. Violent sex offenders lack male social support. Int J Offender Ther Comp Criminol 2001;45:70–82. https://doi.org/10.1177/0306624X01451005.
28. Borduin C, Schaeffer C, Heiblum N. A randomized clinical trial of multisystemic therapy with juvenile sexual offenders: effects on youth social ecology and criminal activity. J Consult Clin Psychol 2009;77(1):26–37. https://doi.org/10.1037/a0013035.
29. Laulik S, Allam J, Sheridan L. An investigation into maladaptive personality functioning in Internet sex offenders. Psychol Crime Law 2007;35(5):523–35. https://doi.org/10.1080/10683160701340577.
30. Bender DS, Morey LC, Skodol AE. Toward a model for assessing level of personality functioning in DSM-5, Part I: a review of theory and methods. J Pers Assess 2011;93(4):332–46. https://doi.org/10.1080/00223891.2011.583808.
31. Morey LC. Development and initial evaluation of a self-report form of the DSM–5 level of Personality Functioning Scale. Psychol Assess 2017;29(10):1302–8. https://doi.org/10.1037/pas0000450.
32. Pincus AL. An interpersonal perspective on Criterion A of the DSM-5 Alternative Model for Personality Disorders. Curr Opin Psychol 2018;21:11–7. https://doi.org/10.1016/j.copsyc.2017.08.035.
33. Pincus AL. Some comments on nomology, diagnostic process, and narcissistic personality disorder in the DSM-5 proposal for personality and personality disorders. Personal Disord 2011;2(1):41–53. https://doi.org/10.1037/a0021191.
34. Pincus AL, Gurtman MB. Interpersonal theory and the interpersonal circumplex: evolving perspectives on normal and abnormal personality. In: Strack S, editor. Differentiating normal and abnormal personality. New York, NY: Springer Publishing, 2006;83–111.
35. Hopwood CJ, Wright AGC, Ansell EB, Pincus AL. The interpersonal core of personality pathology. J Pers Disord 2013;27(3):270–95. https://doi.org/10.1521/pedi.2013.27.3.270.
36. Gurtman MB. Interpersonal problems and the psychotherapy context: the construct validity of the Inventory of Interpersonal Problems. Psychol Assess 1996;8(3):241–55. https://doi.org/10.1037/1040-3590.8.3.241.
37. Hardy GE, Stiles WB, Barkham M, Startup M. Therapist responsiveness to client interpersonal styles during time-limited treatments for depression. J Consult Clin Psychol 1998;66(2):304–12. https://doi.org/10.1037//0022-006x.66.2.304.
38. Muran JC, Segal ZV, Samstag LW, Crawford CE. Patient pretreatment interpersonal problems and therapeutic alliance in short-term cognitive therapy. J Consult Clin Psychol 1994;62(1):185–90. https://doi.org/10.1037/0022-006X.62.1.185.
39. Hudson SM, Ward T. Interpersonal competency in sex offenders. Behav Modif 2000;24(4):494–527. https://doi.org/10.1177/0145445500244002.
40. Magaletta PR, Faust E, Bickart W, McLearen AM. Exploring clinical and personality characteristics of adult male internet-only child pornography offenders. Int J Offender Ther Comp Criminol 2014;58(2):137–53. https://doi.org/10.1177/0306624X12465271.
41. Magyar MS, Edens JF, Lilienfeld SO, Douglas KS, Poythress NG. Examining the relationship among substance abuse, negative emotionality and impulsivity across subtypes of antisocial and psychopathic substance abusers. J Crim Justice 2001;39(3):232–7. https://doi.org/10.1016/j.jcrimjus.2011.02.013.
42. Middleton D. From research to practice: the development of the Internet sex offender treatment programme (i-SOTP). Ir Probat J 2008;5:49–64. https://doi.org/10.1080/13552600802673444.
43. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. Arch Gen Psychiatry 1961;4:561–71. https://doi.org/10.1001/archpsyc.1961.01710120031004.
44. Morey LC. The Personality Assessment Inventory (PAI). In: Maruish ME, editor. The use of psychological testing for treatment planning and outcomes assessment: instruments for adults. Mahwah, NJ: Lawrence Erlbaum Associates, 2004;509–51.
45. Butcher JN, Dahlstrom WG, Graham JR, Tellegen A, Kaemmer B. MMPI– 2: Minnesota Multiphasic Personality Inventory– 2: manual for administration and scoring. Minneapolis, MN: University of Minnesota Press, 1989/2001.
46. Poythress NG, Edens JF, Lilienfeld SO. Criterion-related validity of the Psychopathic Personality Inventory in a prison sample. Psychol Assess 1998;10(4):426–30. 1037/1040-3590.10.4.426.
47. Doyle M, Dolan M. Evaluating the validity of anger regulation problems, interpersonal style, and disturbed mental state for predicting inpatient violence. Behav Sci Law 2006;24(6):783–98. https://doi.org/10.1002/bsl.739.
48. Alden LE, Wiggins JS, Pincus AL. Construction of circumplex scales for the Inventory of Interpersonal Problems. J Pers Assess 1990;55(3–4):521–36. https://doi.org/10.1080/00223891.1990.9674088.
49. Gurtman MB, Pincus AL. Interpersonal adjective scales: confirmation of circumplex structure from multiple perspectives. Pers Soc Psychol Bull 2000;26(3):374–84. https://doi.org/10.1177/0146167200265009.
50. Ball SA, Carroll KM, Canning-Ball M, Rounsaville BJ. Reasons for dropout from drug abuse treatment: symptoms, personality, and motivation. Addict Behav 2006;31(2):320–30. https://doi.org/10.1016/j.addbeh.2005.05.013.
51. Battjes RJ, Gordon MS, O'Grady KE, Kinlock TW, Carswell MA. Factors that predict adolescent motivation for substance abuse treatment. J Subst Abuse Treat 2003;24(3):221–32. https://doi.org/10.1016/S0740-5472(03)00022-9.
52. Boyle K, Polinsky ML, Hser YI. Resistance to drug abuse treatment: a comparison of drug users who accept or decline treatment referral assessment. J Drug Issues 2000;30(3):555–74. https://doi.org/10.1177/002204260003000304.

53. Hill CE, Kellems IS, Kolchakian MR, Wonnell TL, Davis TL, Nakayama EY. The therapist experience of being the target of hostile versus suspected-unasserted client anger: factors associated with resolution. Psychother Res 2003;13(4):475–91. https://doi.org/10.1093/ptr/kpg040.

54. Edens JF, Buffington-Vollum JK, Colwell KW, Johnson DW, Johnson JK. Psychopathy and institutional misbehavior among incarcerated sex offenders: a comparison of the Psychopathy Checklist-Revised and the Personality Assessment Inventory. Int J Forensic Ment Health 2002;1 (1):49–58. https://doi.org/10.1080/14999013.2002.10471160.

55. Moors F, Zech E. The effects of psychotherapist's and clients' interpersonal behaviors during a first simulated session: a lab study investigating client satisfaction. Front Psychol 2017;8:1868. https://doi.org/10.3389/fpsyg.2017.01868.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14539
Available online at: onlinelibrary.wiley.com

PAPER

PSYCHIATRY & BEHAVIORAL SCIENCE

Eveline E. Schippers,[1] *M.Sc.; Larissa M. Hoogsteder,*[1] *Ph.D.; and Geert Jan J.M. Stams,*[2] *Ph.D.*

Responsive Aggression Regulation Therapy (Re-ART) Improves Executive Functioning in Adolescents and Young Adults with Severe Aggression Problems: A Pilot Study

ABSTRACT: This pilot study ($N = 25$) compared the effects of a short, four-month version of Responsive Aggression Regulation Therapy Outpatient (Re-ART Compact) and the entire, ten-month intervention (Re-ART Complete) on specific executive functioning (EF) and the risk of violent recidivism in adolescents and young adults (13–23 years). Re-ART is a cognitive behavioral-based intervention for adolescents and young adults with severe aggression problems. The Re-ART Compact and Re-ART Complete groups were comparable on the EF measures inhibition, flexibility, emotion regulation, self-evaluation, and self-control, but the Re-ART Complete group showed more improved risk of violent recidivism. We conclude that Re-ART Compact can be used as a compact, short intervention for EF, which is a valuable addition to the field of forensic mental health care where many problem behaviors relate to poor EF.

KEYWORDS: adolescent, young adult, forensic mental health, aggression, executive functioning, treatment

Violent offenses are, after property crime, the most common type of offenses among adolescents and young adults (1–3). Untreated violent adolescents have a higher risk of violent recidivism and persistent aggression in adulthood than adolescents who received any type of treatment (medium effect) (4). Current interventions might be improved by paying extra attention to executive functioning (EF) (5,6), because deficits in EF are associated with both the development and maintenance of aggressive behavior (7,8).

Executive functioning concerns cognitive processes that enable goal-oriented complex behaviors through the control of thoughts, emotions, and behavior (9,10). Three subfunctions are generally acknowledged as most important in successful self-regulation of aggression: working memory, cognitive flexibility, and inhibition (11,12). Working memory enables the production and evaluation of alternative responses, such as assertive behavior (13,14). Cognitive flexibility includes flexible creation of alternative thoughts; the absence thereof can lead to rigid thinking and cognitive distortions (13,14). Inhibition refers to the control of impulsivity and suppression of dominant responses and enables alternative behavior that is more appropriate in certain situations (11,13). Furthermore, inhibition concerns appropriate emotion regulation, self-evaluation, and self-control. Stress can hamper EF in people with aggression, which, among other things, has been shown to increase impulsivity and decrease adaptive emotion regulation (15,16). It is thus relevant to address EF and stress regulation in the treatment of aggression problems.

There is empirical evidence showing that deficits in EF are associated with the development and maintenance of aggressive behavior. For instance, impulsivity and impaired inhibition in childhood were found to predict aggression later in life (medium effect) (17,18). Impaired inhibition was found to relate to both reactive and proactive aggression in delinquent adolescents (19). Also, poor emotion regulation predicted aggression in an adult forensic psychiatric sample (20). Additionally, poor inhibition and impulse control have been related to disorders that have a relatively high prevalence in aggressive youth, such as ADHD, autism spectrum disorder, and conduct disorder (21–23), and predicted juvenile criminal recidivism (23). Overall, impulsive decision making and poor inhibitory control are important factors related to aggression in forensic patients (7).

Responsive Aggression Regulation Therapy (Re-ART) is a promising cognitive behavioral intervention that aims to decrease criminal recidivism in delinquent adolescents and young adults with aggressive behavior (24,25). Re-ART has a strong focus on improving EF using modules that target stress reduction, impulse control, anger control, and emotion regulation. In aiming for behavioral change, Re-ART adheres to the risk–need–responsivity principles, which have been shown to increase treatment effectiveness (26,27). These principles require that the intensity of treatment matches the recidivism risk; that treatment focuses on the individual dynamic criminogenic risk factors (needs); and that treatment is adjusted to the characteristics of the specific

[1] De Waag, Forensic Care Specialist, Oudlaan 9, Utrecht, 3515 GA, The Netherlands.

[2] Faculty of Social and Behavioral Sciences, Forensic Child and Youth Care, University of Amsterdam, Postbus 15776, Amsterdam, Noord-Holland, 1001 NG, The Netherlands.

Corresponding author: Eveline E. Schippers, M.Sc. E-mail: eschippers@dfzs.nl

Received 25 June 2020; and in revised form 14 July 2020; accepted 17 July 2020.

target group (general responsivity) as well as the characteristics of the individual adolescent (specific responsivity).

Re-ART applies techniques known to be effective in the treatment of aggressive adolescents and young adults: It uses an individual approach combined with cognitive behavioral techniques and attention to the adolescent's social system. This includes recognition of irrational thoughts and the use of helping thoughts (28,29) and problem-solving skills (26,28). The intervention uses drama-therapeutic techniques such as role-play (30) and exercises such as mindfulness and experimental techniques, which improve relaxation and attention and reduce feelings of anger (31).

Given that outpatient treatment is often bound by time restraints imposed by judicial conviction or client motivation, compact interventions are very much desired. The current pilot study investigates the effects of Re-ART Compact (a short, four-month version of Re-ART) on EF and risk of violent recidivism compared to Re-ART Complete (the entire, ten-month intervention). Re-ART Compact is specifically aimed at improving EF and consists of the start module (with attention to motivation and psychoeducation) and the modules controlling anger, stress reduction, and impulse control. Studies show that Re-ART is effective in reducing aggressive behavior and attention deficits, improving self-control (32), and decreasing criminal recidivism (24). However, it is not clear to what extent improvements in EF can be seen after only the first modules of the intervention. Previous research shows that it is possible to improve EF to some degree in adolescents using different treatment modalities (33,34). Improved EF seems to be a prerequisite for behavioral change in other target areas, such as social skills or problem-solving skills. We expect that youths show improved inhibition, flexibility, emotion regulation, self-evaluation, and self-control after both Re-ART Compact and Re-ART Complete. Additionally, we expect that these improvements will be larger after Re-ART Complete than after Re-ART Compact, because its longer duration might stimulate continued behavioral change.

Methods

Sample

The sample (N = 25, 68% male) consisted of adolescents and young adults aged 13–23 (M = 17.44, SD = 2.80). All subjects received Re-ART at an outpatient treatment center for forensic psychiatry. This center provides mental health treatment for clients with delinquent behavior, including violent, sexual, and property crimes, in order to reduce the risk of future delinquent behavior. All subjects had an average moderate-to-high pretreatment risk of violent recidivism, as assessed by the RAF MH (see Table 1 for more detailed information). Subjects were randomly assigned to the shorter Re-ART Compact (n = 14) or to the longer Re-ART Complete (n = 11). Chi-square tests and t-tests compared Re-ART Compact and Re-ART Complete on various background variables, such as age, gender, intelligence, ethnic background, and diagnosis (see Table 1). The groups only differed significantly in risk of violent recidivism (p = 0.02, d = 0.97), with the Re-ART Complete group showing a higher pretreatment risk of violent recidivism (M = 4.18, SD = 0.75) than the Re-ART Compact group (M = 3.50, SD = 0.65). The average treatment duration of Re-ART Complete was 43.7 weeks (SD = 14.3), while the average duration of Re-ART Compact was 18.3 weeks (SD = 6.4).

Procedure

Data were collected in 2015 at three departments of an outpatient treatment facility for forensic psychiatry in the Netherlands. All adolescents and young adults, and for those younger than 16 years their legal guardian as well, signed the informed consent form for participation in scientific research. All subjects complied with Re-ART's indication criteria. In short, all adolescents and young adults showed medium-to-high risk of violent

TABLE 1—*Descriptive information (M[SD] or %[*n*]) of Re-ART Compact (*n *= 15) and Complete (*n *= 11).*

	Re-ART Compact	Re-ART Complete	*p*	*d*
Gender (male)	64% (9)	73% (8)	0.65	0.18
Age (start treatment)	17.71 (2.20)	17.09 (3.51)	0.59	0.25
IQ (clinical estimation)	90.23 (6.80)	88.45 (6.86)	0.53	0.37
Ethnic Background (Dutch)	79% (11)	55% (6)	0.43	0.40
DSM-5 Diagnosis				
Conduct Disorder NOS	43% (6)	46% (5)	0.90	0.06
APD/CD	36% (5)	46% (5)	0.62	0.20
ADHD	21% (3)	27% (3)	0.73	0.14
PTSD	14% (2)	18% (2)	0.79	0.11
ODD	14% (2)	27% (3)	0.42	0.33
Other	29% (4)	36% (4)	0.68	0.17
Problematic Substance Use	50% (7)	36% (4)	0.50	0.28
Problem Behavior				
Violent Behavior	100% (14)	100% (11)		
Property Crime Behavior	43% (6)	46% (5)	0.90	0.06
Sex Offense Behavior	7% (1)	9% (1)	0.86	0.07
Risk Violent Recidivism	3.50 (.65)	4.18 (.75)	0.02	0.97
Intensity of Treatment				
<1 session per week	7% (1)	9% (1)	0.40	0.39
1 session per week	71% (10)	46% (5)		
1.5 sessions per week	21% (3)	46% (5)		
Optional Modules				
Stress Reduction	71% (10)	82% (9)	0.55	0.25
Impulse Control	57% (8)	64% (7)	0.74	0.13

APD, antisocial personality disorder for subjects aged 18+; CD, conduct disorder for subjects aged 18−. Risk of violent recidivism was assessed with the RAF MH, where 1 = low and 5 = high.

recidivism; had a diagnosis of oppositional defiant disorder, conduct disorder, conduct disorder not otherwise specified, or antisocial personality disorder; and showed severe aggressive behavior that caused disruptions in two or more areas of life (such as school, home, work) and moderate or high risk of violent recidivism. In both groups, the clients and therapists filled out the instruments RAF MH, BRIEF-A, and Re-ART List before starting treatment (T0) and after completing the intervention (T1).

Description of Re-ART

Re-ART Complete consists of a set of standard and optional modules, which can be used to create a custom-made intervention that matches individual risk levels and problem behaviors (24). Standard modules include start module (regarding motivation for treatment and analysis of the aggression chain), controlling anger, Influence of Thinking, and Assertive Behavior. The optional modules include stress reduction (contraindicated if there is no high stress), impulse control (if the degree of impulsivity is very high and the module controlling anger does not suffice), Emotion Regulation, Observation and Interpretation, Conflict Management, and a Family module. Duration and intensity of the intervention depend on the individual's risk level: Adolescents receive a minimum of one and a maximum of three individual sessions a week. Re-ART Compact consists of the start module and the controlling anger module. When indicated, the optional modules stress reduction and impulse control were also offered.

Treatment Integrity

A sufficient degree of program integrity, or adherence to the intervention's standardized treatment protocol, is a necessary precondition to draw valid conclusions about the effectiveness of a treatment (35,36). Treatment integrity across the entire treatment (both groups) was assessed using evaluation forms, reported by both the therapists and the adolescents. These forms assessed whether the treatment sessions incorporated the intervention's effective elements, for instance, whether the risk–need–responsivity principles and specific Re-ART techniques had been applied. This study only included cases with sufficient program integrity (i.e., the protocol elements were followed for more than 70%), which means that the intervention was provided as intended (37).

Instruments

RAF MH

The risk of violent recidivism was assessed by means of the Risk Assessment for outpatient Forensic Mental Health (RAF MH) youth version, a structured professional judgment for risk of recidivism in adolescents (38,39). The instrument consists of 94 static and dynamic items in 12 domains: 1. Previous and current offenses, (8 items), 2. School/work (9 items), 3. Finances (3 items), 4. Living environment (3 items), 5. Family (12 items), 6. Social network (5 items), 7. Leisure (3 items), 8. Substance use (7 items), 9. Emotional/Personal (16 items), 10. Attitude (5 items), 11. Risk management (8 items), and 12. Sexual problems (15 items). All therapists were trained in the use of the RAF MH and made structured judgments about the risk level at each domain on a 6-point Likert scale ranging from 0 = "very unsatisfying" to 5 = "very satisfying." In conclusion, the therapist classified a general risk level for each offense in the adolescent's offense history as either "low," "low-moderate," "moderate," "moderate-high," or "high." The RAF MH has shown good inter-rater reliability (ICC = 0.73–0.88) and predicted general and violent recidivism within one year after outpatient treatment (AUC = 0.77) (39).

BRIEF-A

The Dutch self-report Behavior Rating Inventory of Executive Function-Adult version (BRIEF-A) was used to assess EF on a behavioral level (40,41). Since the adolescent and adult version of the BRIEF are very similar, for sake of homogeneity we used the adult version (BRIEF-A) for the entire sample. The current pilot study uses the scales Inhibition, Flexibility, Emotion Regulation, and Self-Evaluation. Sample items are as follows: "I can't sit still" (Inhibition), and "I get emotionally upset easily" (Emotion Regulation). Subjects rate the occurrence of these behaviors in the past month using the options: "never" (1 point), "sometimes" (2), and "often" (3). A higher total score on each scale indicates more problematic EF. The BRIEF-A has shown good construct validity and convergent validity with other measures of EF (42,43). In the current study, all scales of the BRIEF-A showed satisfactory reliability (Cronbach's $\alpha \geq 0.60$ reflects good reliability [44]), with Cronbach's alpha (α) for Inhibition T0 = 0.68, T1 = 0.70, Flexibility T0 = 0.68, T1 = 0.78, Emotion Regulation T0 = 0.89, T1 = 0.79, and Self-Evaluation T0 = 0.69, T1 = 0.87.

Re-ART List

Self-control was assessed with both the therapist-report and the adolescent and young adult self-report (hereafter: adolescent-report) version of the Re-ART List (45). The Re-ART List assesses self-control and assertiveness skills related to aggression. For the current purposes, only the self-control scale was used. Sample items are as follows: "The adolescent has influence on his behavior" (therapist version), and "I have insight in the consequences of my behavior" (adolescent version). Each item is rated on a 5-point Likert scale ranging from 1 ("not true at all") to 5 ("completely true"). Higher sum scores indicate more self-control. The construct validity of both therapist and adolescent versions of the list was found to be satisfactory (45). In the current study, the self-control scale showed good reliability, with Cronbach's α for therapist version T0 = 0.90, and T1 = 0.79, and adolescent version T0 = 0.90, and T1 = 0.88.

Statistical Analyses

The assumptions for analyses of variance were met (46). First, paired-samples t-tests were used to evaluate any within-subject change between T0 and T1 on inhibition, flexibility, emotion regulation, self-evaluation (assessed with the BRIEF-A), therapist-reported self-control, adolescent-reported self-control (assessed with the Re-ART List), and risk of violent recidivism (assessed with the RAF MH). Second, t-tests and chi-square tests were used to check for any a priori differences between the groups on any of the background variables (reported in Table 1). We conducted analyses of covariance (ANCOVA) to test for differences between Re-ART Compact and Re-ART Complete at T1, controlling for T0 scores and for any a priori differing background variables by including them as covariates (47). Effect sizes were computed according to Cohen where 0.02> d < 0.49 means a small effect, 0.50> d < 0.79 a moderate effect, and d > 0.80 a large effect (48). Adjusted d's were calculated by subtracting the pretreatment effect sizes from the post-treatment effect sizes for all experimental variables.

Results

Within-Subjects Change

Paired-samples t-tests showed that in the Re-ART Complete group, all seven experimental variables (inhibition, flexibility, emotion regulation, self-evaluation, therapist-reported self-control, adolescent-reported self-control, and risk of violent recidivism) improved significantly from T0 to T1. See Table 2 for means and effect sizes. In the Re-ART Compact group, five experimental variables improved significantly, whereas flexibility and adolescent-reported self-control did not. The Re-ART Complete group showed larger effect sizes than the experimental group on all experimental variables with exception of emotion regulation and self-evaluation.

Difference Between Groups

With regard to a priori differences between the group, the Re-ART Complete group showed a higher pretreatment risk of violent recidivism than the Re-ART Compact group, $t(23) = -2.43$, $p = 0.02$, $d = 0.97$ (see Table 1). Risk of recidivism was therefore included as a covariate in the following analyses. ANCOVAs could not detect any significant differences between the Re-ART Compact group and the Re-ART Complete group at T1 with regard to EF variables (see Table 2). Small effect sizes were found, with the exception of a large effect size in adolescent-reported self-control in favor of Re-ART Complete. The Re-ART Complete group showed a significantly lower risk of violent recidivism than the Re-ART Compact group, $F(1,22) = 34.649$, $p = 0.000$, $d^{\text{adjusted}} = -2.13$, a large effect size.

Discussion

This pilot study investigated whether the short, four-month intervention Re-ART Compact yielded similar results regarding executive functioning (EF) as the entire intervention Re-ART Complete in aggressive adolescents and young adults. After completing Re-ART Compact, subjects showed significantly improved EF skills regarding inhibition, emotion regulation, self-regulation, and therapist-reported self-control. Flexibility and adolescent-reported self-control did not significantly improve after completing Re-ART Compact. After completing Re-ART Complete, all skills were improved. We conclude that Re-ART Compact is adequate to improve certain EF, and if time is limited, these modules may be a good option for the treatment of aggressive adolescents. However, whenever possible, it is still recommended to complete the entire intervention in order to secure sustainable behavioral change and a larger reduction in recidivism. Re-ART Complete targets other risk areas than EF, such as cognitive distortions, social skills, or coping skills, which might add to the larger reduction in risk of violent recidivism.

Re-ART Compact did not improve flexibility. It is possible that more time is needed to achieve changes in this domain, given that inflexible behavior can be shaped by years of negative life experiences and childhood trauma (49,50). Re-ART Complete addresses rigid thinking and cognitive distortions in later modules, using both cognitive behavioral techniques and mindfulness-based exercises to support flexible thinking and broaden response options (49,51).

We found that adolescents and young adults reported much higher self-control before and after treatment than the therapists. An explanation could be that adolescents and young adults with impulse control problems overestimate their EF ability (52) and that adolescents with lower EF scores show less self-awareness (53). Also, they may show more social desirable responding before treatment. Additionally, therapists might underestimate their client's self-control when they base their pretreatment scoring on less information than their four-month evaluation. Another possible explanation is that therapists show more post-treatment social desirable responding as a reflection of their own treatment efficacy (54).

Our findings have some important implications. First, the results show that change might already occur in an early stage of the Re-ART intervention. Seeing early results might add to the sense of self-efficacy, or belief in self-control over one's actions (55), which may prove fruitful for maintaining motivation and completing the intervention program in a population that is generally reluctant to engage in treatment (25). Second, based on the difference between therapist-report and adolescent-report, it is recommended that therapists use multiple sources of information, such as clients and parents, to obtain a better picture of the adolescent's behavior. Third, improved EF might be

TABLE 2—*Means, SDs, and ANCOVA results for Re-ART Compact and Re-ART Complete.*

	Re-ART Compact (n = 14)					Re-ART Complete (n = 11)							
	T0		T1			T0		T1					
	M	SD	M	SD	d	M	SD	M	SD	d	F(1,21)	d	d^{adjusted}
Inhibition	14.57	2.92	12.21*	2.42	0.88^{L}	15.91	2.51	12.64*	2.62	1.27^{L}	.096	0.17^{S}	-0.31^{S}
Flexibility	10.14	2.68	8.78	2.22	0.55^{M}	12.27	2.37	9.64**	2.06	1.19^{L}	.018	0.40^{S}	-0.43^{S}
Emotion Regulation	17.57	3.83	13.35*	3.12	1.21^{L}	22.00	5.73	16.82*	2.56	1.17^{L}	2.503	1.19^{L}	0.23^{S}
Self-Evaluation	11.07	2.16	9.00*	2.63	0.86^{L}	12.09	2.55	10.45*	2.54	0.64^{M}	.252	0.56^{M}	0.12^{S}
Self-Control Adolescent	54.75	8.01	58.58	12.33	0.37^{S}	44.18	8.55	60.73**	3.88	2.49^{L}	1.801	-0.22^{S}	-1.50^{L}
Self-Control Therapist	28.69	8.27	36.69*	6.07	1.10^{L}	26.09	7.62	38.09**	2.95	2.08^{L}	2.534	-0.29^{S}	0.03^{S}
Risk Violent Recidivism	3.50	.65	3.21*	.69	0.43^{S}	4.18	.75	2.36**	.81	2.35^{L}	34.649**	-1.15^{L}	-2.13^{L}

Note. Annotations S, M, L indicate small, medium, or large effect size. Risk violent recidivism df = (1,22). d^{adjusted} = Cohen's d controlled for pretreatment scores.

*p < .05

**p < .001

beneficial in more life areas than the clinical context (56). This may boost the clients' positive life goals, which improves general well-being and receptivity for treatment (57). Fourth, Re-ART Compact can be used autonomously to improve EF in adolescents with aggression problems or in addition to various other treatment programs. Improving EF is beneficial for many problem behaviors, since not only aggression, but many psychiatric problems relate to poor EF (21,22). Finally, modules focusing on stress reduction, impulse control, and inhibition skills seem to improve EF, which emphasizes the role of stress and the lack of impulse control in aggressive behavior.

Since adolescent aggression is an equally severe problem in other countries and Re-ART has shown to be equally effective for adolescents with Dutch and other ethnic backgrounds (24), we assume that results may be generalized to non-Dutch samples of aggressive adolescents and young adults. By adhering to the risk, need, responsivity principles for effective forensic treatment (26), the intervention adapts to the target population's needs as well as the individual's needs. Interventions that are sensitive to the individual's criminogenic needs have shown to be as effective in reducing youth delinquency as interventions that are culturally specific (58,59). Also, Re-ART's responsive approach means that individual learning style, motivational factors, and therapeutic relation factors are acknowledged. It is not a one-size-fits-all, but a tailor-made intervention which accounts for individual differences, which is important when treating cultural minority youths (60,61). Moreover, cognitive behavioral interventions for disruptive behavior in youth can be implemented without loss of effectiveness in most Western countries (62). We assume that Re-ART might be a successful intervention for aggressive youths in different cultures and countries as well; however, follow-up research to the effectiveness of EF interventions should compare the effects between various ethnic backgrounds.

This pilot study has several limitations, such as its small sample size. Yet, we assume that if we discover substantial changes within only four months, it is worthwhile to further examine Re-ART Compact. Further research might demonstrate whether changes are sustainable in larger groups and whether the intervention is effective if compared to a control group receiving treatment as usual or waitlist controls. Furthermore, the Re-ART Complete group did show a higher pretreatment risk of recidivism than the Re-ART Compact group. Even though we controlled for these confounding effects in our analyses, it might have affected the results, since high-risk individuals may show greater change in treatment.

Conclusion

In conclusion, Re-ART Compact, a brief cognitive behavioral intervention for EF, may be a valuable addition to the field of forensic mental health care, where many problem behaviors relate to poor EF. Modules focusing on stress reduction, impulse control, and inhibition skills improved EF in aggressive adolescents and young adults. Improvement of EF seems to be an important part of a cognitive behavioral intervention that has shown to be effective in reducing aggression problems.

Acknowledgments

The authors like to thank Daphne Bonnes for her help during data acquisition and Anna Laura Huckelba for her help in proofreading the article.

Conflict of Interest

L.M. Hoogsteder is the developer and owner of RE-ART and a Re-ART trainer and consultant.

References

1. Office of Juvenile Justice and Delinquency Prevention. Law enforcement & juvenile crime: juvenile arrest rates. OJJDP statistical briefing book. 2017. https://www.ojjdp.gov/ojstatbb/crime/JAR.asp(accessed July 13, 2019).
2. Youth Justice Board, Ministry of Justice. Youth justice statistics 2016/17 England and Wales. London, U.K: Youth Justice Board/Ministry of Justice, 2018.
3. Statistics Netherlands [CBS]. Verdachten; delictgroep, geslacht, leeftijd en migratieachtergrond [Suspects: offense type, age, and migration background]. https://opendata.cbs.nl/statline/#/CBS/nl/dataset/81947NED/table?ts=1581343921244 (accessed February 13, 2020).
4. Sawyer AM, Borduin CM, Dopp AR. Long-term effects of prevention and treatment on youth antisocial behavior: a meta-analysis. Clin Psychol Rev 2015;42:130–44. https://doi.org/10.1016/j.cpr.2015.06.009
5. Denson TF. Four promising psychological interventions for reducing reactive aggression. Curr Opin Behav Sci 2015;3:136–41. https://doi.org/10.1016/j.cobeha.2015.04.003
6. Ogilvie JM, Stewart AL, Chan RCK, Shum DHK. Neuropsychological measures of executive function and antisocial behavior: a meta-analysis. Criminology 2011;49(4):1063–107. https://doi.org/10.1111/j.1745-9125.2011.00252.x
7. Tonnaer F, Cima M, Arntz A. Executive (dys)functioning and impulsivity as possible vulnerability factors for aggression in forensic patients. J Nerv Ment Dis. 2016;204(4):280–6. https://doi.org/10.1097/NMD.0000000000000485
8. Van Nieuwenhuijzen M, Van Rest MM, Embregts PJCM, Vriens A, Oostermeijer S, Van Bokhoven I, et al. Executive functions and social information processing in adolescents with severe behavior problems. Child Neuropsychol 2017;23(2):228–41. https://doi.org/10.1080/09297049.2015.1108396
9. Klenberg L, Korkman M, Lahti-Nuuttila P. Differential development of attention and executive functions in 3- to 12-year-old Finnish children. Dev Neuropsychol 2001;20(1):407–28. https://doi.org/10.1207/S15326942DN2001_6
10. Schoemaker K, Mulder H, Deković M, Matthys W. Executive functions in preschool children with externalizing behavior problems: a meta-analysis. J Abnorm Child Psychol 2013;41(3):457–71. https://doi.org/10.1007/s10802-012-9684-x
11. Diamond A. Executive functions. Annu Rev Psychol 2013;64:135–68. https://doi.org/10.1146/annurev-psych-113011-143750
12. Hofmann W, Schmeichel BJ, Baddeley AD. Executive functions and self-regulation. Trends Cogn Sci 2012;16(3):174–80. https://doi.org/10.1016/j.tics.2012.01.006
13. Davidson MC, Amso D, Anderson LC, Diamond A. Development of cognitive control and executive functions from 4 to 13 years: evidence from manipulations of memory, inhibition, and task switching. Neuropsychologia 2006;44(11):2037–78. https://doi.org/10.1016/j.neuropsychologia.2006.02.006
14. Garon N, Bryson SE, Smith IM. Executive function in preschoolers: a review using an integrative framework. Psychol Bull 2008;134(1):31–60. https://doi.org/10.1037/0033-2909.134.1.31
15. Maier SU, Makwana AB, Hare TA. Acute stress impairs self-control in goal-directed choice by altering multiple functional connections within the brain's decision circuits. Neuron 2015;87(3):621–31. https://doi.org/10.1016/j.neuron.2015.07.005
16. Sandi C, Haller J. Stress and the social brain: behavioural effects and neurobiological mechanisms. Nat Rev Neurosci 2015;16(5):290–304. https://doi.org/10.1038/nrn3918
17. Sarkisian K, Van Hulle C, Lemery-Chalfant K, Goldsmith HH. Childhood inhibitory control and adolescent impulsivity and novelty seeking as differential predictors of relational and overt aggression. J Res Pers 2017;67:144–50. https://doi.org/10.1016/j.jrp.2016.07.011
18. Youn C, Meza JI, Hinshaw SP. Childhood social functioning and young adult intimate partner violence in girls with and without ADHD: response inhibition as a moderator. J Atten Disord 2019;23(12):1486–96. https://doi.org/10.1177/1087054718778119
19. Zhang Z, Wang Q, Liu X, Song P, Yang B. Differences in inhibitory control between impulsive and premeditated aggression in juvenile

inmates. Front Hum Neurosci 2017;11:373. https://doi.org/10.3389/fnhum.2017.00373

20. Bousardt AMC, Noorthoorn EO, Hoogendoorn AW, Nijman HLI, Hummelen JW. On the link between emotionally driven impulsivity and aggression: evidence from a validation study on the Dutch UPPS-P. Int J Offender Ther Comp Criminol 2018;62(8):2329–44. https://doi.org/10.1177/0306624X17711879
21. Lawson RA, Papadakis AA, Higginson CI, Barnett JE, Wills MC, Strang JF, et al. Everyday executive function impairments predict comorbid psychopathology in autism spectrum and attention deficit hyperactivity disorders. Neuropsychology 2015;29(3):445–53. https://doi.org/10.1037/neu0000145
22. Puiu AA, Wudarczyk O, Goerlich KS, Votinov M, Herpertz-Dahlmann B, Turetsky B, et al. Impulsive aggression and response inhibition in attention-deficit/hyperactivity disorder and disruptive behavioral disorders: findings from a systematic review. Neurosci Biobehav Rev 2018;90:231–46. https://doi.org/10.1016/j.neubiorev.2018.04.016
23. Wibbelink CJM, Hoeve M, Stams GJJM, Oort FJ. A meta-analysis of the association between mental disorders and juvenile recidivism. Aggress Violent Behav 2017;33:78–90. https://doi.org/10.1016/j.avb.2017.01.005
24. Hoogsteder LM, Stams G-JJM, Schippers EE, Bonnes D. Responsive Aggression Regulation Therapy (Re-ART): an evaluation study in a Dutch juvenile justice institution in terms of recidivism. Int J Offender Ther Comp Criminol 2018;62(14):4403–24. https://doi.org/10.1177/0306624X18761267
25. Hoogsteder LM, Van Horn JE, Stams GJJM, Wissink IB, Hendriks J. The relationship between the level of program integrity and pre- and post-test changes of Responsive-Aggression Regulation Therapy (Re-ART) outpatient: a pilot study. Int J Offender Ther Comp Criminol 2016;60(4):435–55. https://doi.org/10.1177/0306624X14554828
26. Bonta J, Andrews DA, Bonta J, Andrews DA, Bonta J. The psychology of criminal conduct, 6th edn. New York, NY: Routledge, 2017.
27. Lowenkamp CT, Latessa EJ. Increasing the effectiveness of correctional programming through the risk principle: Identifying offenders for residential placement. Criminol Public Policy 2005;4(2):263–90. https://doi.org/10.1111/j.1745-9133.2005.00021.x
28. McCart MR, Sheidow AJ. Evidence-based psychosocial treatments for adolescents with disruptive behavior. J Clin Child Adolesc Psychol 2016;45(5):529–63. https://doi.org/10.1080/15374416.2016.1146990
29. Oostermeijer S, Smeets KC, Jansen LMC, Jambroes T, Rommelse NNJ, Scheepers FE, et al. The role of self-serving cognitive distortions in reactive and proactive aggression. Crim Behav Ment Heal 2017;27(5):395–408. https://doi.org/10.1002/cbm.2039
30. Lipsey MW. The primary factors that characterize effective interventions with juvenile offenders: a meta-analytic overview. Vict Offenders 2009;4(2):124–47. https://doi.org/10.1080/15564880802612573
31. Sukhodolsky DG, Smith SD, McCauley SA, Ibrahim K, Piasecka JB. Behavioral interventions for anger, irritability, and aggression in children and adolescents. J Child Adolesc Psychopharmacol 2016;26(1):58–64. https://doi.org/10.1089/cap.2015.0120
32. Hoogsteder LM, Kuijpers N, Stams GJJM, Van Horn JE, Hendriks J, Wissink IB. Study on the effectiveness of responsive aggression regulation therapy (Re-ART). Int J Forensic Ment Health 2014;13(1):25–35. https://doi.org/10.1080/14999013.2014.893711
33. Franco C, Amutio A, López-Gonález L, Oriol X, Martínez-Taboada C. Effect of a mindfulness training program on the impulsivity and aggression levels of adolescents with behavioral problems in the classroom. Front Psychol 2016;7:1385. https://doi.org/10.3389/fpsyg.2016.01385
34. Morley JL. Impact of group cognitive behavior therapy on adolescents with deficits in inhibition [dissertation]. Philadelphia, PA: Philadelphia College of Osteopathic Medicine, 2015.
35. Goense PB, Assink M, Stams GJ, Boendermaker L, Hoeve M. Making 'what works' work: a meta-analytic study of the effect of treatment integrity on outcomes of evidence-based interventions for juveniles with antisocial behavior. Aggress Violent Behav 2016;31:106–15. https://doi.org/10.1016/j.avb.2016.08.003
36. Duwe G, Clark V. Importance of program integrity: outcome evaluation of a gender-responsive, cognitive-behavioral program for female offenders. Criminol Public Policy 2015;14(2):301–28. https://doi.org/10.1111/1745-9133.12123
37. Durlak JA, DuPre EP. Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. Am J Community Psychol 2008;41(3–4):327–50. https://doi.org/10.1007/s10464-008-9165-0
38. van Horn JE, Wilpert J, Bos M, Mulder J. Risicotaxatie-instrument voor de Ambulante Forensische GGZ – RAF GGZ jeugd. Handleiding [RAF MH youth: structured risk assessment instrument for outpatient delinquents. Manual]. Utrecht, The Netherlands: De Waag, 2012.
39. Van Horn JE, Wilpert J, Bos MGN, Eisenberg M, Mulder J. WaagSchaal jeugd: de psychometrische kwaliteit van een gestructureerd klinisch risicotaxatie-instrument voor de ambulante forensische psychiatrie [RAF MH youth: the psychometric quality of a structured clinical risk assessment instrument for outpatient forensic psychiatry]. Panopticon 2009;30(2):23–34.
40. Roth RM, Gioia GA. Behavior rating inventory of executive function–adult version. Lutz, FL: Psychological Assessment Resources, 2005.
41. Scholte E, Noens I. Brief-A: Vragenlijst over executieve functies bij volwassenen. Handleiding [BRIEF-A: questionnaire about executive functions in adults. Manual]. Amsterdam, the Netherlands: Hogrefe, 2011.
42. Ciszewski S, Francis K, Mendella P, Bissada H, Tasca GA. Validity and reliability of the Behavior Rating Inventory of Executive Function—Adult Version in a clinical sample with eating disorders. Eat Behav 2014;15(2):175–81. https://doi.org/10.1016/j.eatbeh.2014.01.004
43. Roth RM, Lance CE, Isquith PK, Fischer AS, Giancola PR. Confirmatory factor analysis of the behavior rating inventory of executive function-adult version in healthy adults and application to attention-deficit/hyperactivity disorder. Arch Clin Neuropsychol 2013;28(5):425–34. https://doi.org/10.1093/arclin/act031
44. Streiner DL. Starting at the beginning: An introduction to coefficient alpha and internal consistency. J Pers Assess 2003;80(1):99–103. https://doi.org/10.1207/S15327752JPA8001_18
45. Hoogsteder LM, Hendriks J, Van Horn JE, Wissink IB. Agressie Regulatie op Maat: Een evaluatie studie in een justitiële jeugdinrichting [Responsive Aggression Regulation Therapy: an evaluation study in a juvenile detention center]. Orthop Onderz Prakt 2012;51(11):481–95.
46. Field A. Discovering statistics using SPSS. London, U.K.: SAGE, 2009.
47. Rausch JR, Maxwell SE, Kelley K. Analytic methods for questions pertaining to a randomized pretest, posttest, follow-up design. J Clin Child Adolesc Psychol 2003;32(3):467–86.
48. Cohen J. A power primer. Psychol Bull 1992;112(1):155–9. https://doi.org/10.1037/0033-2909.112.1.155
49. Chou WP, Lee KH, Ko CH, Liu TL, Hsiao RC, Lin HF, et al. Relationship between psychological inflexibility and experiential avoidance and internet addiction: mediating effects of mental health problems. Psychiatry Res 2017;257:40–4. https://doi.org/10.1016/j.psychres.2017.07.021
50. Moffitt TE. Life-course-persistent and adolescence-limited antisocial behavior: a 10-year research review and a research agenda. In: Lahey BB, Moffitt TE, Caspi A, editors. Causes of conduct disorder and juvenile delinquency. New York, NY: Guilford Press, 2003;49–75.
51. Merwin RM, Timko CA, Moskovich AA, Konrad Ingle K, Bulik CM, Zucker NL. Psychological inflexibility and symptom expression in anorexia nervosa. Eat Disord 2011;19:62–82. https://doi.org/10.1080/10640266.2011.533606
52. Steward KA, Tan A, Delgaty L, Gonzales MM, Bunner M. Self-awareness of executive functioning deficits in adolescents with ADHD. J Atten Disord 2017;21(4):316–22. https://doi.org/10.1177/1087054714530782
53. Zlotnik S, Toglia J. Measuring adolescent self-awareness and accuracy using a performance-based assessment and parental report. Front Public Heal 2018;6:15. https://doi.org/10.3389/fpubh.2018.00015
54. Cuijpers P, Cristea IA. How to prove that your therapy is effective, even when it is not: a guideline. Epidemiol Psychiatr Sci 2016;25(5):428–35. https://doi.org/10.1017/S2045796015000864
55. Bandura A. Self-efficacy: the exercise of control. New York, NY: Freeman, 1997.
56. Diamond A, Ling DS. Conclusions about interventions, programs, and approaches for improving executive functions that appear justified and those that, despite much hype, do not. Dev Cogn Neurosci 2016;18:34–48. https://doi.org/10.1016/j.dcn.2015.11.005
57. Garland EL, Fredrickson B, Kring AM, Johnson DP, Meyer PS, Penn DL. Upward spirals of positive emotions counter downward spirals of negativity: insights from the broaden-and-build theory and affective neuroscience on the treatment of emotion dysfunctions and deficits in psychopathology. Clin Psychol Rev 2010;30(7):849–64. https://doi.org/10.1016/j.cpr.2010.03.002
58. Wilson SJ, Lipsey MW, Soydan H. Are mainstream programs for juvenile delinquency less effective with minority youth than majority youth? A meta-analysis of outcomes research. Res Soc Work Pract 2003;13:3–26. https://doi.org/10.1177/1049731502238754

59. Miranda J, Bernal G, Lau A, Kohn L, Hwang W-C, LaFromboise T. State of the science on psychosocial interventions for ethnic minorities. Annu Rev Clin Psychol 2005;1:113–42. https://doi.org/10.1146/annurev.clinpsy.1.102803.143822

60. Vergara AT, Kathuria P, Woodmass K, Janke R, Wells SJ. Effectiveness of culturally appropriate adaptations to juvenile justice services. J Juv Justice 2016;5(2):85–103.

61. Huey S Jr, Polo A. Evidence-based psychosocial treatments for ethnic minority youth. J Clin Child Adolesc Psychol 2008;37(1):262–301. https://doi.org/10.1080/15374410701820174

62. Leijten P, Melendez-Torres GJ, Knerr W, Gardner F. Transported versus homegrown parenting interventions for reducing disruptive child behavior: a multilevel meta-regression study. J Am Acad Child Adolesc Psychiatry 2016;55(7):610–7. https://doi.org/10.1016/j.jaac.2016.05.003

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14543
Available online at: onlinelibrary.wiley.com

PAPER

PSYCHIATRY & BEHAVIORAL SCIENCE

Selena McKay-Davis,[1] *M.F.S.; Tharinia Robinson*,[2] *Ph.D.; Ismail M. Sebetan*,[3] *M.D., Ph.D.; and Paul Stein*,[3] *Ph.D.*

Civilian Forensic Technician and Sworn Police Officer Job-Related Stress

ABSTRACT: Forensic Technicians provide crime scene investigation services and are exposed to stressful violent crimes, motor vehicle accidents, biological or chemical hazards, and other appalling imagery. Forensic Technicians would likely experience physical and psychological stress after exposure to trauma, and security vulnerabilities similar to Sworn Police Officers. The perceived availability of mental health resources, job-related physical, psychological stress, and traumatic experiences of both Forensic Technicians and Sworn Police Officers from California law enforcement agencies were investigated using a self-reported survey. Responses were evaluated for any significant differences in the perceived stress, job-related physical stress, and resulting psychological impact affecting the participants. The survey contained a mix of True/False, Circle/Check the Appropriate Box, or Likert Scale (1–5) responses. The results were evaluated statistically and discussed. Results indicated Sworn Police Officers and Forensic Technicians have different on-duty stress levels, but similar off-duty stress levels. Nearly two-thirds of 54 job-related stressors were not significantly different between the two occupations. However, Forensic Technicians reported more adverse effects in 17 physical and psychological job-related activities compared with Sworn Police Officers. Forensic Technicians reported lower awareness levels and availability of agency mental health support services than were reported by Sworn Police Officers. This study reports for the first time an unexpected outcome that perceived and job-related psychological stress is greater for Forensic Technicians than Sworn Police Officers. Possible reasons for this disparity will be discussed as well as stress management tools that should be implemented to reduce health risk factors for both career professionals as well as increase public safety.

KEYWORDS: forensic psychology, crime scene investigator, sworn police officer, forensic civilian technician, occupational stress, stress management

Sworn Police Officers are citizens whose job duties help maintain order by enforcing the laws of the land, detecting crime, and arresting violators. Sworn Police Officers can operate under various titles including Police Officer, Deputy Sheriff, Marshall, Special Agent, State Trooper and Correctional Deputy. Job-related stress in Sworn Police Officers has been studied for several decades and has revealed that law enforcement-related stress has profound and damaging influence on a Sworn Police Officer's psychological, physical, and familial well-being (1,2). The trauma and chronic stress experienced by law enforcement officers can be additive in certain individuals. This subconscious accumulation of stressful and/or traumatic experiences can affect one's ability to make necessary split-second decisions that would otherwise protect themselves or fellow officers from injury or death (3).

A recent survey (4) of Police Officers' cognitive performance and stress levels showed that even over a two-week duty cycle performance on the working memory test was significantly decreased by stress. Unfortunately, this would not always allow the Sworn Police Officer to correlate job stress experienced with specific activities executed during the course of their duty cycle and impacted their ability to accurately recall the event.

Forensic Technicians and other similar positions, such as Crime Scene Investigators, have primary duties that include identifying, collecting, preserving, and documenting physical and other evidence in criminal investigations (5). Their efforts can directly affect the direction and depth of an investigation. Forensic Technicians are civilian responders reporting after the scene has been secured; but often while the aftermath of the traumatic event is being managed. Despite media and television portrayals, most Forensic Technicians do not carry firearms, interview, or apprehend suspects. Nor do they generally run into uncontrolled or violent situations as a Sworn Police Officer is expected to do (5). They are distinct from many other civilian law enforcement professionals such as dispatchers or criminalists because their crime scene duties are performed in view of the public, and often at scenes that involved violent activities and possibly surrounded by dangerous individuals.

Few studies have evaluated the conditions and impact of occupational stress on civilian forensic personnel. Most studies have focused on job satisfaction and burnout. It has been reported that moderate levels of stress in Forensic Scientists were similar to other traditional criminal justice professionals (6,7). Physiological measurements (heart rates) over a one-week work shift for United Kingdom civilian crime scene investigators were evaluated for the cardiovascular effects relating to stress from job-

[1] Riverside Police Department-Forensics Unit, 4102 Orange St, Riverside, CA 92501.

[2] Criminal Justice Program, Piedmont College, 595 Prince Avenue, Athens, GA 30601.

[3] Forensic Sciences Program, National University, 11255 North Torrey Pines Road, La Jolla, CA, 92037.

Corresponding author: Ismail Sebetan, M.D., Ph.D. E-mail: isebetan@nu.edu

Received 11 April 2020; and in revised form 21 July 2020; accepted 24 July 2020.

related duties. Despite the routine nature of the calls (burglary and vehicle examinations), the increase in heart rates was attributed to the physical activity and psychological stressors (8).

The stress reaction can be viewed as a subjective perception and evaluation by the individual of the task at hand and whether it exceeds the individual's ability to achieve the demands of the task (9). The determination of what is stressful along with the physical and emotional impact is highly dependent on personality, perception of the danger, coping strategies, social support networks, intervention tools, and temperament (10–13). The effect of training, postevent debriefings, and counseling is also not to be overlooked.

The purpose of this study was to compare the various types of job-related stressors, dangers, physical and psychological stress experienced by Forensic Technicians and Sworn Police Officers. Participant ratings of a core group of identified stressors were used to evaluate significant similarities or differences between Forensic Technicians and Sworn Police Officers.

Stress factors that can accumulate over one's professional career were compared with identify the positive or negative emotional and physical impact on Forensic Technicians and Sworn Police Officers. The research also investigated the participants' awareness of stress management resources and the frequency that participants utilized resources provided by their respective agencies.

Materials and Methods

Survey

A self-reporting survey was developed consisting of 25 questions related to the stress and trauma experienced by members of the professions responding to the survey from California Law Enforcement Agencies (Appendix S1). The listed stressors were developed from personal experience and the Spielberger Stress Survey (14). The survey questions comprised scaled scores (1–5), Yes/No, and fill-in-the-blank type questions. Preliminary questions provided demographic information from the Sworn Police Officers and Forensic Technicians. Sworn Police Officers responding to the survey were comprised of titles that including Police Officer, Deputy Sheriff, and District Attorney Investigator. Technicians that process crime scenes go by a wide range of titles including Forensic Technician, Evidence Technician, Police Lab Technician, Crime Scene Investigator, Crime Scene Specialist, and Forensic Specialist.

This demographic information was later used for data analysis comparisons between the Forensic Technician and the Sworn Police Officer study groups. In this report not all survey questions were utilized, only those relevant to the study objectives. Remaining questions will be addressed in a follow-up publication. Respondents were recruited by personal contact and e-mails from California law enforcement agencies. The participants were required to have at least one year of experience in their position, and in processing at least two major types of crime scenes as a primary job duty on a weekly basis. The study and survey (Appendix S1) received Institutional Review Board approval.

Data Analysis

Participant Qualification

Sworn Police Officers and Forensic Technicians provided their current crime scene investigation experience and duties (Q10, 11) which provided key information for qualifying the respondents.

Demographic Comparisons

Sworn Police Officers and Forensic Technicians provided demographic and professional background information (Q1, 3, 5). These data were analyzed with the chi-square test, significance determined at p value < 0.05.

Perception of Stress

Survey participants rated their perception of on and off-duty stress levels using a scale from one (minimal) to five (maximum) stress (Q13, 14). The descriptive statistics: mean, standard deviation (SD) were determined and results were analyzed using the Student t-test for mean scores. Significance was determined at p value < 0.05.

Stressors Common to Sworn Police Officers2

Participants (Sworn Police Officers and Forensic Technicians) rated 54 stressors common to law enforcement personnel (Q12). The mean scores were analyzed by Student t-test (p value < 0.05). These respondents also reported their perception of their "well-being" in 17 physical and psychological areas since beginning their professional careers (Q15). These results were analyzed using chi-square test (p value < 0.05).

Stress Management

The survey addressed whether agencies provided mental health stress management resources (Q20) and how frequently they were made available (Q21) for both occupations. These data were analyzed using chi-square test and Student t-test, respectively (p value < 0.05).

Results

Participant Survey Submissions

A total of 86 participant surveys were received. Thirteen Sworn Police Officers who participated in the survey were disqualified because of either insufficient crime scene experience (Q11), submission of an incomplete survey, or an incorrectly filled out survey.

Demographic Comparisons

The results from the demographic data indicated that there was no significant difference in ethnicity of the Sworn Police Officers and Forensic Technicians, who were predominantly Caucasian. There were more female Forensic Technicians than males ($p < 0.05$), and more male than female Sworn Police Officers ($p < 0.05$). There was a significant difference in agency type with more respondents from Police than Sheriff Departments. The experience and age range of the participants were not statistically different (Table 1).

Perception of Stress

Survey respondents provided their perceived stress levels experienced in 54 stressors common to law enforcement (Q12).

TABLE 1—*Survey participant demographics (Survey Questions 1, 3, 5).**

Demographics	*N* Forensic Technicians	%	*N* Sworn Police Officers	%
Ethnicity (ns)				
Caucasian	26	70.3	20	55.6
African Amer.	1	2.7	2	5.6
Hispanic	8	21.6	6	16.7
Asian	1	2.7	4	11.1
Other/ Unknown	1	2.7	4	11.1
Gender ($p < 0.05$)				
Male	14	40.5	29	80.6
Female	23	59.5	7	19.4
Agency ($p < 0.05$)				
Police Dept.	34	91.9	34	94.4
Sheriff Dept.	3	8.1	2	5.6
Experience (ns)				
1–5 years	8	21.6	4	11.1
6–10 years	8	21.6	8	22.2
11–15 years	6	16.2	7	19.4
16–20 years	6	16.2	9	25.0
21+ years	9	24.3	8	22.2
Age range (ns)				
18–20 years	0	0.0	0	0.0
21–25 years	1	2.7	2	8.3
26–30 years	4	8.1	3	8.3
31–35 years	5	13.5	7	19.4
36–40 years	7	21.6	5	13.9
41–45 years	5	13.5	7	19.4
46–50 years	5	13.5	6	13.9
51–55 years	6	16.2	3	8.3
56+ years	4	10.8	3	8.3
N	37		36	

*Chi-square data analysis. *p* value > 0.05 = (ns).

The mean ratings were organized from highest to lowest stress level ratings by profession, and the shared responses indicated (bold). The top 20 out of these 54 self-reported stressors are listed for Forensic Technicians and Sworn Police Officers (Table 2). Approximately 33 (61%) of participant ratings out of the 54 stressors were similar (no statistical difference noted) between Forensic Technicians and Sworn Police Officers. The stressors are organized from highest to lowest of their rating scores (Table 3).

The perceived stress levels on and off duty (Q13, 14) were reported on a scale, one (minimal) to five (maximum), and were analyzed for significance. The results indicated that Forensic Technician on-duty stress level (Mean = 3.50, SD = 0.99) was significantly different ($p < 0.05$) and higher than Sworn Police Officers (Mean = 2.99, SD = 0.97). Forensic Technician mean self- reported off-duty stress level (Mean = 2.11, SD = 0.91) was higher than the self-reported off-duty stress level of Sworn Police Officers (Mean = 1.96, SD = 0.85); but was not significantly different.

Regarding gender-related differences, results also indicated that there was a significant difference (*p* value < 0.05) in the mean rating of on-duty stress levels between female Forensic Technicians and female Sworn Police Officers, with female Forensic Technicians reporting higher stress ratings. No significant difference was found in the mean stress levels of female and male Forensic Technicians, nor female and male Sworn Police Officers (Q13). There was also no significant difference in the on-duty stress levels related to age of the respondents (data not shown).

Survey participants detailed the perception of their behavioral, physical, psychological, and lifestyle changes experienced since the beginning of their employment (Q15). The majority of Forensic Technicians indicated an overall positive impact in only three of the 17 survey categories (f, g, j; 17.6%). In contrast, most Sworn Police Officers expressed a positive impact in nine of the 17 categories (b, c, d, f, g, i, n, p, q; 53%). This implies that Forensic Technicians reported negative perceptions in the survey (Q15) at a frequency approximately three times that of Sworn Police Officers. Forensic Technicians and Sworn Police Officers shared positive responses on only two of the questions: Do you feel supervision realistically gives you enough time to complete your assigned tasks (f), and Do you feel your job has a positive impact on society (g).

TABLE 2—*Top 20 self-reported stressors for scaled response ratings (Q12): Forensic Technicians versus Sworn Police Officers.*

Rank	Stressor	Forensic Technicians Mean ± SD	Stressor	Sworn Police Officers Mean ± SD
1	Active Scenes (OCC)	3.54 ± 1.26*	**Staffing** (ORG)	3.56 ± 1.25
2	Promotability (ORG)	3.46 ± 1.32*	**Negative Perception** (OCC)	3.33 ± 1.20
3	No Defense Weapon (ORG)	3.36 ± 1.36*	Physical Attack (OCC)	3.25 ± 1.48*
4	**Staffing** (ORG)	3.32 ± 1.31	**Line of Duty Death** (OCC)	3.25 ± 1.78
5	Biohazard Exposure (OCC)	3.32 ± 1.36*	Killing Someone (OCC)	3.14 ± 1.69*
6	**Line of Duty Death** (OCC)	3.29 ± 1.66	**Insufficient Support** (ORG)	2.97 ± 1.46
7	Child Abuse (OCC)	3.19 ± 1.41*	**Department Politics** (ORG)	2.97 ± 1.34
8	**Subpoena** (OCC)	3.19 ± 1.29*	Family Security (OCC)	2.86 ± 1.29
9	**Court Testimony** (OCC)	3.16 ± 1.24*	Physical Injury (OCC)	2.78 ± 1.27
10	Mistaken ID (OCC)	3.16 ± 1.52*	Excessive Calls (ORG)	2.78 ± 1.38
11	**Negative Perception** (OCC)	3.08 ± 1.40	Use of Force (OCC)	2.75 ± 1.20*
12	**Insufficient Support** (ORG)	3.08 ± 1.40	**Personal Field Security** (OCC)	2.69 ± 1.19
13	Inadequate Pay (ORG)	3.03 ± 1.44*	Angry Civilians (OCC)	2.69 ± 1.06
14	**Personal Field Security** (OCC)	3.00 ± 1.37	Balance Work/Home (ORG)	2.64 ± 1.38
15	**Expect Zero Errors** (OCC)	3.00 ± 1.35	Supervision (ORG)	2.64 ± 1.36
16	On-Call (ORG)	2.97 ± 1.42	Lenient Court System (OCC)	2.64 ± 1.42
17	**Off Duty Security** (OCC)	2.95 ± 1.47	**Court Testimony** (OCC)	2.58 ± 1.05*
18	**Department Politics** (ORG)	2.95 ± 1.37	**Off Duty Security** (OCC)	2.50 ± 1.34
19	Partner Conflicts (ORG)	2.95 ± 1.20*	**Expect Zero Errors** (OCC)	2.47 ± 1.40
20	Never Feel Off Duty (OCC)	2.94 ± 1.45	**Subpoena** (OCC)	2.47 ± 1.03*

Highlighted (bold) stressors were common to both occupations.
OCC, Occupational stressors; ORG, Organizational stressors.
*Significant difference: Student *t*-test, $p < 0.05$ for the common stressors.

TABLE 3—*Top 20 stressors ranked by similarity: Forensic Technicians versus Sworn Police Officers (Q12).*

Rank	Stressors	*p*-Value
1.	Physical Injury on the Job (OCC)	0.94
2.	Collision in Assigned Unit (OCC)	0.94
3.	Department Politics (ORG)	0.93
4.	Friends Killed in Line of Duty (OCC)	0.93
5.	Excessive Calls for Service (ORG)	0.92
6.	Lenient Court System (OCC)	0.88
7.	Family Security (OCC)	0.86
8.	Insufficient Department Support (ORG)	0.75
9.	Pressure to Solve Cases Quickly (ORG)	0.71
10.	Crimes Reported to You (OCC)	0.70
11.	Supervision Expectations/Discipline (ORG)	0.65
12.	Shift Work (ORG)	0.60
13.	Disagreeable Duties/Laws (OCC)	0.58
14.	Service Request Deadlines (ORG)	0.50
15.	Angry Civilians (OCC)	0.49
16.	Insufficient Staffing (ORG)	0.44
17.	Negative Public/Press Perception (OCC)	0.41
18.	Balance Work/Home (ORG)	0.35
19.	Overtime (OCC)	0.34
20.	Personal Security in Field (OCC)	0.32

Student *t*-test, no significant difference, *p*-value > 0.05.
OCC, Occupational stressors, ORG, Organization stressor.

Chi-square analysis was completed to determine if there were any significant differences ($p < 0.05$) between the responses of Forensic Technicians and Sworn Police Officers in their perceptions of the 17 categories. The analysis indicated that there was a significant difference for seven categories (b, c, d, k, n, p, q). The analysis also identified responses in 10 of the 17 categories (a, e, f, g, h, i, j, l, m, o) that were not significantly different.

Stressors Common to Sworn Police Officers

Ten of the top 20 highest mean scores for shared law enforcement stressors (Q12) were shared (Table 2, "bold") by Forensic Technicians and Sworn Police Officers. These included: insufficient staffing, friends killed in line of duty, subpoenaed during time off, court testimony, negative public/press perception, insufficient department support, personal field security, expectation of zero errors, off-duty personal security, and department politics. Eight of the ten shared mean stressor ratings were not significantly different. The two significantly different ratings between the professions were related to the stress of "subpoenas" and "court testimony." Forensic Technicians ranked occupational type stressors (OCC) as twelve of the top 20 stressors. Whereas Sworn Police Officers ranked these occupational type stressors (OCC) as 14 of the top 20 stressors, which resulted in similar ranking for the 2 professions for OCC.

When asked if they felt less depressed (Q15-n), depression was reported in 68% of the Forensic Technicians as opposed to 41% of the Sworn Police Officers. This was a statistically significant difference (p value < 0.03).

Stress Management

Forensic Technicians ($N = 36$) and Sworn Police Officers ($N = 36$) reported on knowledge about the availability (Q20) of mental health support services at their agency and the frequency to which they were made available (Q21). Fewer Forensic Technicians reported availability of such resources than the Sworn Police Officers. Approximately one-third of the Forensic Technicians reported that they did not know if mental health resources were available at their agency, whereas all the Sworn Police Officers were aware of the availability of mental health resources (p value < 0.01). Forensic Technicians reported (Q21) a mean score of 2.81 for the availability of mental health resources; whereas Sworn Police Officers reported a mean of 3.5, which was not statistically significant (p value > 0.05).

Discussion

The current research compared the perception of stress and stress factors affecting job responsibilities, behavior, and general well-being, between Forensic Technicians and Sworn Police Officers. It is commonly accepted that law enforcement officers have an elevated rate of "burn-out" and suicide, and this may correlate with their stressful occupation (15,16). Whether similar issues exist for Forensic Technicians and other law enforcement civilians that process crime scenes has only been hypothesized; but not supported by any earlier research. The following parameters were investigated through a self-reporting survey: perceptions of stress levels, physical, behavioral, and psychological job stress factors common to both occupations, and availability and utilization stress management resources. Correlations between these parameters and various demographic factors (ethnicity, gender, age, and experience) were also determined.

The paucity of resources and lack of availability of these for Forensic Technicians is apparent. Law Enforcement professionals are the primary authority for investigating and preventing all types of crimes, improving public safety, and protecting victims. This engenders the development of a strong bond of camaraderie and support, unlike that afforded to Forensic Technician, who generally work alone and often outside established law enforcement agencies. The nature of activities that Sworn Police Officers engage in requires the purchase of special equipment (bullet proof vests, firearms, tasers, and body cameras). The use of these requires training each year and appropriate funding to a much greater extent than provided to Forensic Technicians. Law Enforcement agencies have strong political support groups and powerful union bargaining organizations. This guarantees better vacations and welfare and retirement benefits than afforded to Forensic Technicians. There is also the perception that Forensic Technicians do not earn the same level of recognition as do Sworn Police Officers, and in the author's view, create added stress and the impression that mental health resources are limited. All these differences create a separate culture for these two groups that are both tasked with seeing that crime does not pay; but Sworn Police Officers appear to operate in a higher tier and landscape than Forensic Technicians.

Demographic Comparisons

The Sworn Police Officer and Forensic Technician participants were predominantly Caucasian. The major demographic findings indicated that more females were surveyed in Forensic Technician careers than males and the reverse true for Sworn Police Officers. The ages ranged between 21 and 56 or more years old with no difference between the Sworn Police Offices and Forensic Technicians. Examining years of experience on the job indicated no significance difference. A predominant difference was observed in that most participants were employed in police departments as opposed to sheriff departments; but not in the distribution between the two groups of professionals.

Perception of Stress

It was reported that there was an overall higher mean on-duty stress score for Forensic Technicians compared with Sworn Police Offices. This was not an expected outcome for the on-duty stress experience. Off-duty stress between the two occupations was not significantly different.

The survey data indicated a serious perceived job-related stress on Forensic Technicians. A negative score was reported in 14 of 17 questions (Q15). Sworn Police Officers reported negative results in only eight of the 17 questions, almost half that as the Forensic Technicians. This finding could be a result of the consequences of working violent crime scenes, long hours, shift work, and frequent operation in unsafe environments. Another unexplored potential cause could be the initial expectations and the actual experienced reality of the job and its impact on the Forensic Technician.

The areas where there were significant differences between Forensic Technicians and Sworn Police Officers were for the following categories in descending order: feel less depressed, in better physical shape, less trouble sleeping, equal part of the team, exercise more frequently, feel more sympathetic, feel more trusting. In the 17 physical and psychological impact areas, Forensic Technicians did not feel as if they were treated as an equal part of the team with Sworn Police Officers. It is somewhat surprising that compared to when they first entered their profession the Forensic Technicians considered themselves: more easily aggravated, less sympathetic toward victims and suspects, less trusting of people, did not feel civilians and officers were equally part of the team, were more anxious about their safety in their personal life, used more time off, consumed more alcohol, had more trouble sleeping, felt more exhausted, had a harder time communicating job issues with family/friends, felt more depressed, felt the job had more control over their personal life, were in worse physical shape, and exercised less.

The self-reported negative effects are consistent with prior research that identified Sworn Police Officers are prone to psychological ailments like anxiety, loss of self-esteem, loss of confidence, a loss of the feeling of control over one's own life, depression, suicide, and post-traumatic stress disorder (13,17). In comparing the self-reported physical and psychological responses of Forensic Technicians to prior research, the Forensic Technicians exhibited indications of four major effects of job-related stress and trauma.

The first major effect was stress exhaustion, which can lead to poor work performance, post-traumatic stress disorder, early retirement or attrition, absenteeism, job burn-out, marital discord, depression, substance abuse, suicide, and premature death (1,11,12,15,18). The second major effect was burnout potential which was based upon their reported perceptions of unfairness in the work environment and work-life conflict, each of which is considered a reliable predictor of burnout (16).

The third major effect of cynicism and hypervigilance are consistent with the generally negative Forensic Technician responses in that they became aggravated easier, were less sympathetic to victims and suspects, displayed less trust of people, suffered increased exhaustion, and experienced increased anxiousness about their personnel safety. These characteristics can be caused by, or result from, cynicism and hypervigilance (19,20,21). The fourth major effect of increased depression was reported by 68% of Forensic Technicians respondents, as opposed to 41% of Sworn Peace Officer participants.

Forensic Technicians and Sworn Police Officers shared positive responses in two of the same questions: Do you feel supervision realistically gives you enough time to complete your assigned tasks (f), and Do you consider your job has positive influence on society (g).

Stressors Common to Sworn Police Officers

The fact that ten of the top 20 mean stressor ratings were shared by Forensic Technicians and Sworn Police Officers, and that eight of these were not significantly different between the two professions, was a good indication of similar occupational effect. This finding was augmented by the fact that nearly two-thirds of the total 54 stressors evaluated were not significantly different between Forensic Technicians and Sworn Police Officers. Forensic Technicians reported that their top three stressors were processing active scenes, limited promotability, and lacking a defensive weapon. For Sworn Police Officers, staffing was the topmost stressor.

The two stressors common to both professions, yet found to be significantly different, were being subpoenaed during time off and providing court testimony. These two experiences were significantly more stressful for Forensic Technicians than for Sworn Police Officers. This may suggest a lack of training, or higher expectations, for delivering courtroom testimony. Of the significantly similar job-related stressor categories, between these two professions, departmental politics was ranked third highest; and excessive calls for service were reported as fifth highest.

A common concern about physical danger and safety dominated the top five similar stressors, including: physical injury on the job, involvement in traffic accidents, and colleagues killed in the line of duty. Similar to Sworn Police Officers, Forensic Technician stressors include regular physical stress sustained during the task of processing a crime scene, moving equipment or objects and other strenuous physical activities (11). Safety concerns were also indicated by Forensic Technicians rating the lack of a defensive weapon as the third highest stressor and biohazard exposure the fifth highest stressor. Biohazard exposure concern was unique to Forensic Technicians, as it was ranked thirtieth overall by Sworn Police Officers.

Stress Management

Forensic Technicians and Sworn Police Officers are under occupational stress that may negatively affect their job decision-making, physical health, family life, risk of suicide and may especially influence Sworn Police Officers, creating a toxic work environment. The result of which will likely compromise public safety. Obviously, stress management is important for the well-being of these professionals and their anticipated job performance. The elevated stress levels reported in this research and impact scores for the Forensic Technicians may have been influenced by the low reported availability of mental health resources. In the author's experience, many Forensic Technicians are also unaware that mental health support exists, like peer support, or training outside of the general employee assistance programs. Increased awareness and availability of these resources, in addition to practicing coping strategies, may reduce job-related stress in Forensic Technicians.

Implementation of stress management tools and peer support programs within Forensics Units, along with the incorporation of stress management awareness modules into the training manuals or educational programs are two methods that would improve the mental health of Forensic Technicians (22). A recent trend toward the introduction of Mindfulness-Based Resilience Training (MBRT) for those involved with community policing activities (Sworn Police Officers) has similar beneficial potential to help

Forensic Technicians cope with job-related stress (23). Law Enforcement Officers (Sworn Police Officers) have the highest suicide risk than any other profession (24). There is the potentially elevated suicide risk for Forensic Technicians and Sworn Police Officers in light of the negative job impact for these two occupations. This is another serious effect of job-related stress that could be mitigated through the implementation of MBRT programs.

Limitations

In an effort to identify commonalities between the Sworn Police Officer and Forensic Technician crime scene-related experiences, the self-reporting survey omitted stressors that would not generally be experienced by Forensic Technicians. Examples of omitted situations included: response to a felony in progress, death notifications, and high-speed chases. Although each of these situations is stressful and traumatic for police officers, they do not apply to Forensic Technicians and therefore they could not rate or evaluate their impact.

Self-reporting surveys carry the inherent potential risk that respondents may consciously lie about their opinions or experiences (25). Respondents in this study may have lowered the stress ratings to cast a more formidable persona. Law enforcement personnel is expected to project confidence and authority in the execution of their duties, which many officers may perceive as threatened by admitting job-related stress. Respondents in this study may have lowered their ratings due to concerns for any negative consequences if their agencies or a defense attorney become aware of the elevated stress ratings.

Conclusions

Forensic Technicians and Sworn Police Officers are often required to make sacrifices in order to fulfill their professional duties. This does not absolve their respective agencies from providing the needed guidance and support tools necessary to ameliorate the potential physical, mental, and personal stress that these professionals experience due to frequent violence and suffering. Forensic Technicians should receive stress management training, traumatic incident recovery, communication skills, fitness training, and social support networks as they are available for Sworn Police Officers. It is also recommended that they attend incident debriefings and stress management training programs. These are realistic ways to mitigate the stress and trauma related to their professional occupation and related health concern.

It would be a significant contribution to well-being and health of Forensic Technicians and Sworn Police Officers if research is done about job expectations and how they fit with reality. Poor correlation with the real-life situations and job duties would make for additional stress that can impact job performance and family life.

References

1. Abdollahi K. Understanding police stress research. J Forensic Psychol Pract 2002;2(2):1–24. https://doi.org/10.1300/J158v02n02_01
2. Ayers RM, Flanagan GS, Ayers MB. Preventing law enforcement stress: the organization's role. Alexandria, VA: National Sheriffs' Association, 1990;3–10.
3. Pearlin LI. The study of coping: problems and directions. In: Eckenrode J, editor. The social context of coping. New York, NY: Plenum Press, 1991;261–76.
4. Gutshall CL, Hampton DP Jr, Sebetan IM, Stein PC, Broxtermann T. The effects of occupational stress on cognitive performance in police officers. Police Pract Res 2017;18(5):463–77. https://doi.org/10.1080/15614263.2017.1288120
5. Young TJ, Ortmeier PJ. Crime scene investigation: the forensic technician's field manual. Upper Saddle River, NJ: Pearson Prentice Hall, 2011;5–13.
6. Holt TJ, Blevins KR. Examining job stress and satisfaction among digital forensic examiners. J Contemp Crim Justice 2011;27(2):230–50. https://doi.org/10.1177/1043986211405899
7. Holt TJ, Blevins KR, Foran DR, Smith RW.Examination of the conditions affecting forensic scientists' workplace productivity and occupational stress. United States Department of Justice Award Number: 2011-DN-BX-0006. September 2016. https://www.ncjrs.gov/pdffiles1/nij/grants/250233.pdf (accessed July 24, 2020).
8. Adderley R, Smith LL, Bond JW, Smith M. Physiological measurement of crime scene investigator stress. Int J Police Sci Manag 2012;14 (2):166–76. https://doi.org/10.1350/ijps.2012.14.2.274
9. Eden D. Acute and chronic job stress, strain, and vacation relief. Organ Behav Hum Decis Process 1990;45(2):175–95. https://doi.org/10.5465/ambpp.1987.17534117
10. Amendola KL. Stress in police work. In: Greene J, editor. Encyclopedia of police science, 3rd edn. vol. 2. New York, NY: Routledge, 2007;1208–14.
11. Anderson GS, Litzenberger R, Plecas D. Physical evidence of police officer stress. Policing 2002;25(2):399–20. https://doi.org/10.1108/13639510210429437
12. Anshel M. A conceptual model and implications for coping with stressful events in police work. Crim Justice Behav 2000;27(3):375–400. https://doi.org/10.1177/0093854800027003006
13. Violanti JM, Aron F. Police stressors: variations in perception among police personnel. J Crim Justice 1995;23(3):287–94.
14. Spielberger CD, Westberry LG, Grier KS, Greenfield G. The police stress survey: sources of stress in law enforcement (Monograph Series Three, No. 6). Tampa, FL: Human Resources Institute, University of South Florida, 1981.
15. Waters JA, Ussery W. Police stress: history, contributing factors, symptoms, and interventions. Policing 2007;30(2):169–88.
16. McCarty WP, Skogan WG. Job-related burnout among civilian and sworn police personnel. Police Q 2012;16(1):66–84. https://doi.org/10.1177/1098611112457357
17. Maguen S, Metzler TJ, McCaslin SE, Inslict SS, Henn-Haase C, Neylan TC, et al. Routine work environment stress and PTSD symptoms in police officers. J Nerv Ment Dis 2009;197(10):754–60. https://doi.org/10.1097/NMD.0b013e3181b975f8
18. Morash M, Haarr R, Kwak D. Multilevel influences on police stress. J Contemp Crim Justice 2006;22(1):26–43. https://doi.org/10.1177/1043986205285055
19. Brink JD. Police suicide: living between the lines. In: Sheehan D, Warren J, editors. Suicide and law enforcement. Quantico, VA: Behavioral Science Unit FBI Academy, 2001;305–13.
20. Gilmartin KM. Emotional survival for law enforcement: a guide for officers and their families. Tucson, AZ: E-S Press, 2002;33–46.
21. Violanti JM, Marshall JR. Police stress process. J Police Sci Admin 1983;11(4):389–94.
22. Jeanguenat AM, Dror IE. Human factors effecting forensic decision making: workplace stress and well-being. J Forensic Sci 2018;63 (1):258–61. https://doi.org/10.1111/1556-4029.13533
23. Christopher MS, Hunsinger M, Goerling RJ, Bowen S, Rogers BS, Gross CR, et al. Mindfulness-based resilience training to reduce health risk, stress reactivity, and aggression among law enforcement officers: a feasibility and preliminary efficacy trial. Psychiatry Res 2018;264:104–15. https://doi.org/10.1016/j.psychres.2018.03.059
24. Ramchand R, Saunders J, Osilla KC, Ebener P, Kotzias V, Thornton E, et al. Suicide prevention in U.S. law enforcement agencies: a national survey of current practices. J Police Crim Psychol 2019;34:55–66. https://doi.org/10.1007/s11896-018-9269-x
25. Baldwin W. Information no one else knows: the value of self-report. In: Stone AA, Turkkan JS, Bachrach CA, Jobe JB, Kurtzman HS, Cain VS, editors. The science of self-report: implications for research and practice. Mahwah, NJ: Lawrence Erlbaum Associates Publishers, 2000;3–7.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Participant survey — impact of forensic technician job stress to that of sworn police officers

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14514
Available online at: onlinelibrary.wiley.com

PAPER

QUESTIONED DOCUMENTS

Ning Zhang,[1,†] *Ph.D.; Peng Jiang,*[2,†] *B.S.; Weixin Wang,*[1] *M.S.; Chengming Wang,*[3] *Ph.D.; Lanchi Xie,*[1] *M.S.; Zhigang Li,*[1] *B.S.; Wei Huang,*[1] *Ph.D.; Gaojun Shi,*[1] *B.S.; Lei Wang,*[1] *M.S.; Yuwen Yan,*[1] *M.S.; and Shuhui Gao,*[2] *Ph.D.*

Initial Study for the Determination of the Sequence of Intersecting Lines between Gel Pens and Seals by Optical Coherence Tomography*

ABSTRACT: Determining the sequence of intersecting lines is a significant issue in the forensic document examination that can reveal the fraud or distinguish between different allegations. Optical coherence tomography (OCT) is a high-resolution cross-sectional imaging technique that has been introduced into forensic science field recently. The potential of OCT as a novel method to determine the sequence of intersecting lines was examined for the first time. In this study, a spectral-domain OCT system with a center wavelength of 900 nm was employed to perform nondestructive examination on determining the sequence of 18 heterogeneous intersecting line samples produced using three types of gel pens and three brands of stamp pad ink seals. Two-dimensional (2D) cross-sectional, and three-dimensional (3D) volumetric images of the intersecting lines were obtained by the OCT system. Several features were noted and analyzed to successfully determine the sequence of all the 18 samples. Blind tests were also conducted to demonstrate the effectiveness of OCT technique. The results illustrate that OCT technology can provide an effective and accurate method for sequencing intersecting lines of gel pen ink and seal ink, which may complement the conventional methods used in the examination of questioned documents.

KEYWORDS: questioned document examination, optical coherence tomography, optical coherence tomography, sequence of intersecting lines, cross-sectional imaging, 3D imaging

In recent years, as the criminal behaviors become increasingly intelligent and concealed, an examination of the sequence of intersecting lines has attracted more and more attention since it is a significant method of assessing the authenticity and validity of documents (1). For example, in Far East Asia, the perpetrators often add additional text content onto important documents such as contracts and office paperwork with genuine signatures or preaffixed genuine seals in order to make an altered document for illegal purposes (2,3). If the additional text overlaps or intersects with the signature or seal, the relative chronological sequence may be an important clue or evidence to reveal the fraud or distinguish between different allegations. However, the determination of the sequence of intersecting lines still remains a challenge for forensic document examiners, and it has become one of the most difficult technical problems in the forensic science field (4,5).

A number of techniques have been developed to examine the intersecting line samples. In general, these techniques may be divided into destructive and nondestructive methods. The destructive techniques mainly include powder adsorption, scraping and reducing layers, section examination, taping, and decolorization (6), which would consume the specimens and cause irreversible damage to the samples, making it difficult to re-examine. Nondestructive techniques are preferred over the destructive ones as they are conducive to retaining the originality and integrity of the evidence. Optical examination techniques are the most widely used nondestructive methods to examine line crossings, which include stereomicroscopy (6), laser confocal microscopy (7), fluorescence microscopy (8,9), Raman spectroscopy (10), microspectrophotometry (11) and Fourier transform infrared (FTIR) spectral imaging (12). Some ultra-high-resolution imaging techniques such as scanning electron microscopy (SEM) (4,13) and atomic force microscopy (AFM) (14) have been also adopted for this application. Each of these techniques has its advantages and disadvantages. For example, the resolution of stereomicroscopy is insufficient to obtain the fine features of the intersection lines and mainly depends on the experience of the examiners. The fluorescence microscope requires a careful selection of excitation light source, magnification, excitation, and cutoff filter under different scenarios and might need to be

[1] National Engineering Laboratory for Forensic Science, Institute of Forensic Science, Ministry of Public Security, Beijing, 100038, China.

[2] School of Forensic Science and investigation, People's Public Security University of China, Beijing, 100038, China.

[3] Nuctech Company Limited, Beijing, 100084, China.

Corresponding author: Ning Zhang, Ph.D. E-mail: zhangning@cifs.gov.cn; Shuhui Gao, Ph.D. E-mail: gaoshuhui@ppsuc.edu.cn

*This research was sponsored in part by the Beijing Nova Program of Science and Technology Z191100001119039 and by the National Key Research and Development Program 2018YFC0807304.

[†] Authors contributed equally.

Received 11 May 2020; and in revised form 24 June 2020; accepted 26 June 2020.

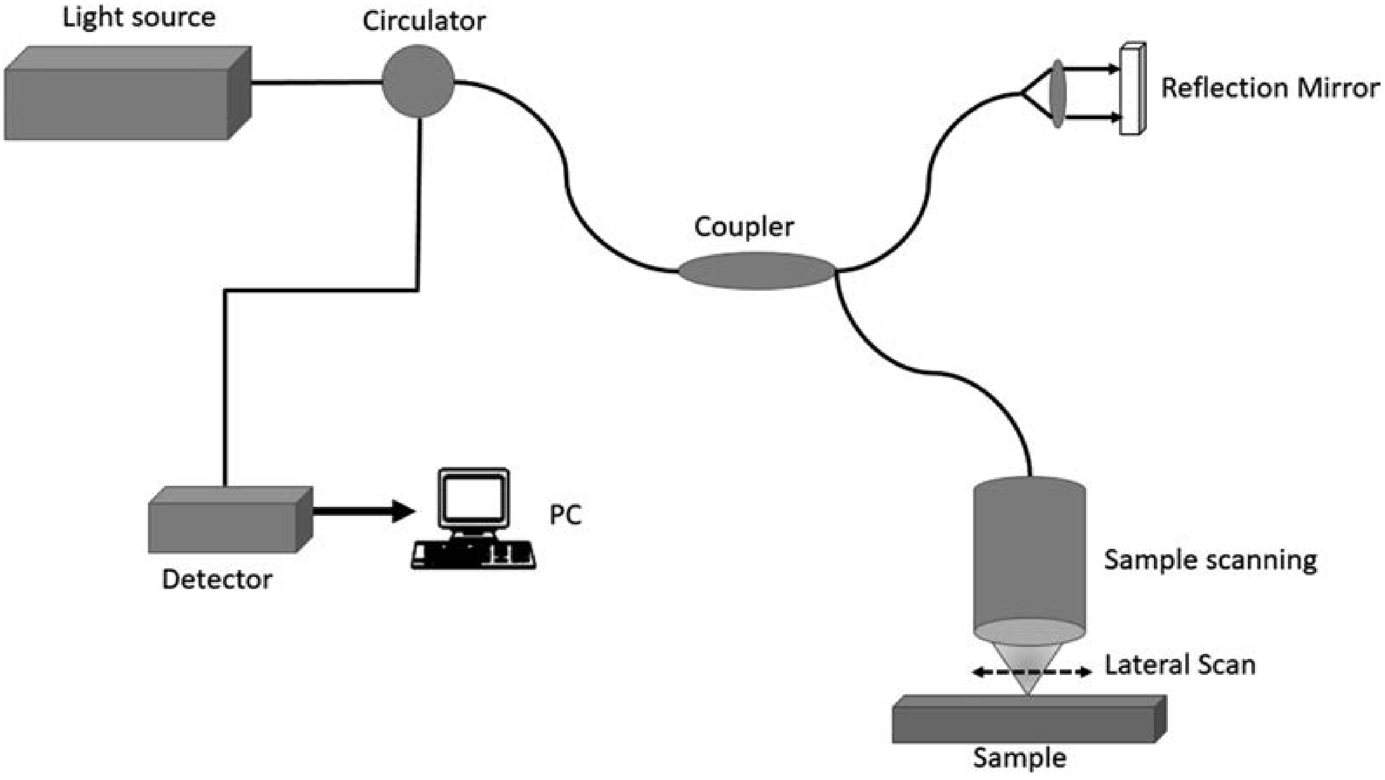


FIG. 1—*The typical optical setup of OCT system. [Color figure can be viewed at wileyonlinelibrary.com]*

combined with confocal microscopy and image processing techniques to obtain high quality images. The SEM and AFM methods are very complicated, time-consuming and expensive. In addition, most of the mentioned nondestructive techniques only obtain surface information, instead of depth information. Due to the superimposition of the lines, the cross-sectional images and the spatial structure of the intersecting parts may contain useful information for the determination of sequence, but the conventional methods still need to slice the sample before carrying out the examination to extract the information from the cross section, which would damage the evidence.

Optical coherence tomography (OCT) is an emerging optical imaging technology that performs *in situ*, noncontact, nondestructive, high-resolution, cross-sectional, and three-dimensional volumetric imaging (15). The principle of OCT is low coherence interferometry. The optical setup typically consists of a Michelson interferometer as shown in Fig. 1. OCT is very similar to ultrasound imaging technology, but it uses light instead of sound to detect the intensity and delay of backscattered light based on the low coherence interferometry method (16). OCT technology has been widely used in medical imaging as a diagnosis technique and has been applied to many branches such as ophthalmology, cardiovascular systems, dermatology, and cancer (17–20). OCT can be seen as a type of optical biopsy, with the advantages of causing no damage to the sample and acquiring micro-meter level tomographic images. Therefore, this technique is a powerful tool for forensic purposes. Recently, OCT has been used in a few forensic applications such as counterfeit banknote detection (21), *Calliphora vicina pupae* age estimation (22), bloodstain volume determination (23), automotive paint characterization (24,25), and hidden fingerprint detection (26).

TABLE 1—*Type of gel pens and seals used to make the intersecting line samples.*

Type	Brand	Color	Model
Gel pen 1	Chenguang	Black	Large capacity APG13604
Gel pen 2	Deli	Black	S34
Gel pen 3	Baile	Black	BL-G2-10
Seal 1	Deli	Red	Water-based stamp pad ink NO.9875
Seal 2	Qixin	Red	Water-based stamp pad ink B3748
Seal 3	Deli	Red	Oil-based stamp pad ink NO. 9864

However, to the best of our knowledge, there has been no related study on determining the sequence of intersecting lines by OCT. In this study, OCT technology is proposed to determine the sequence of intersecting lines for the first time, which provides a novel and reliable technique for document examination. This study focuses on the examination of the sequence of intersecting lines between stamp pad seal ink and gel pen ink that remains a difficult problem because of the characteristics of water-based ink. In our experiments, three types of seals made of water-based and oil-based stamp pad inks, and three types of gel pens (water-based ink) were chosen. A total of 18 intersecting line samples in both sequences were prepared for obtaining the OCT images. A spectral-domain OCT (SD-OCT) system with the center wavelength of 900 nm was employed, achieving ~3 μm axial resolution. The selection of center wavelength needs to be carefully considered because there is a trade-off between axial resolution and penetration depth. OCT instruments working at longer wavelengths (e.g., 1.3 μm) can achieve deeper penetration, but the axial resolution would be degraded according to the principle of OCT imaging. Choosing the center wavelength of 900 nm can balance this trade-off in this study. Two-dimensional (2D) and three-dimensional (3D) OCT images were implemented, and the subsurface images of each sample at a

TABLE 2—*Type of intersecting line samples.*

Sample Number	Order
1	Gel pen 1 over Seal 1
2	Gel pen 1 over Seal 2
3	Gel pen 1 over Seal 3
4	Gel pen 2 over Seal 1
5	Gel pen 2 over Seal 2
6	Gel pen 2 over Seal 3
7	Gel pen 3 over Seal 1
8	Gel pen 3 over Seal 2
9	Gel pen 3 over Seal 3
10	Seal 1 over Gel pen 1
11	Seal 2 over Gel pen 1
12	Seal 3 over Gel pen 1
13	Seal 1 over Gel pen 2
14	Seal 2 over Gel pen 2
15	Seal 3 over Gel pen 2
16	Seal 1 over Gel pen 3
17	Seal 2 over Gel pen 3
18	Seal 3 over Gel pen 3

certain depth were presented. The results demonstrated that OCT can rapidly distinguish intersecting lines created by gel pens and seals with different sequences while maintaining the original physical and chemical state of the sample. OCT may be a powerful supplement to the conventional techniques and further improve the accuracy of the examination when combined with the conventional techniques.

Materials and Methods

Samples

This study focused on the examination of intersecting lines produced by signatures and seals. Three common types of red seals and black gel pens were prepared to produce heterogeneous intersecting line samples, as shown in Table 1. Two types of the seals were made with water-based stamp pad ink, and one was made with oil-based stamp pad ink. The main components of oil-based stamp pad ink are nonpolar solvents, mineral oils, vegetable oils, surfactants, pigments, resin oils, and oil-soluble dyes. The water-based stamp pad ink is composed of polar solvents, water, water-soluble resin, and water-soluble dyes. All intersecting line samples (including blind testing) were formed on A4 size white papers with a basis weight of 70 g/m^2 (Asia Symbol Paper Co., Ltd., Guangdong, China).

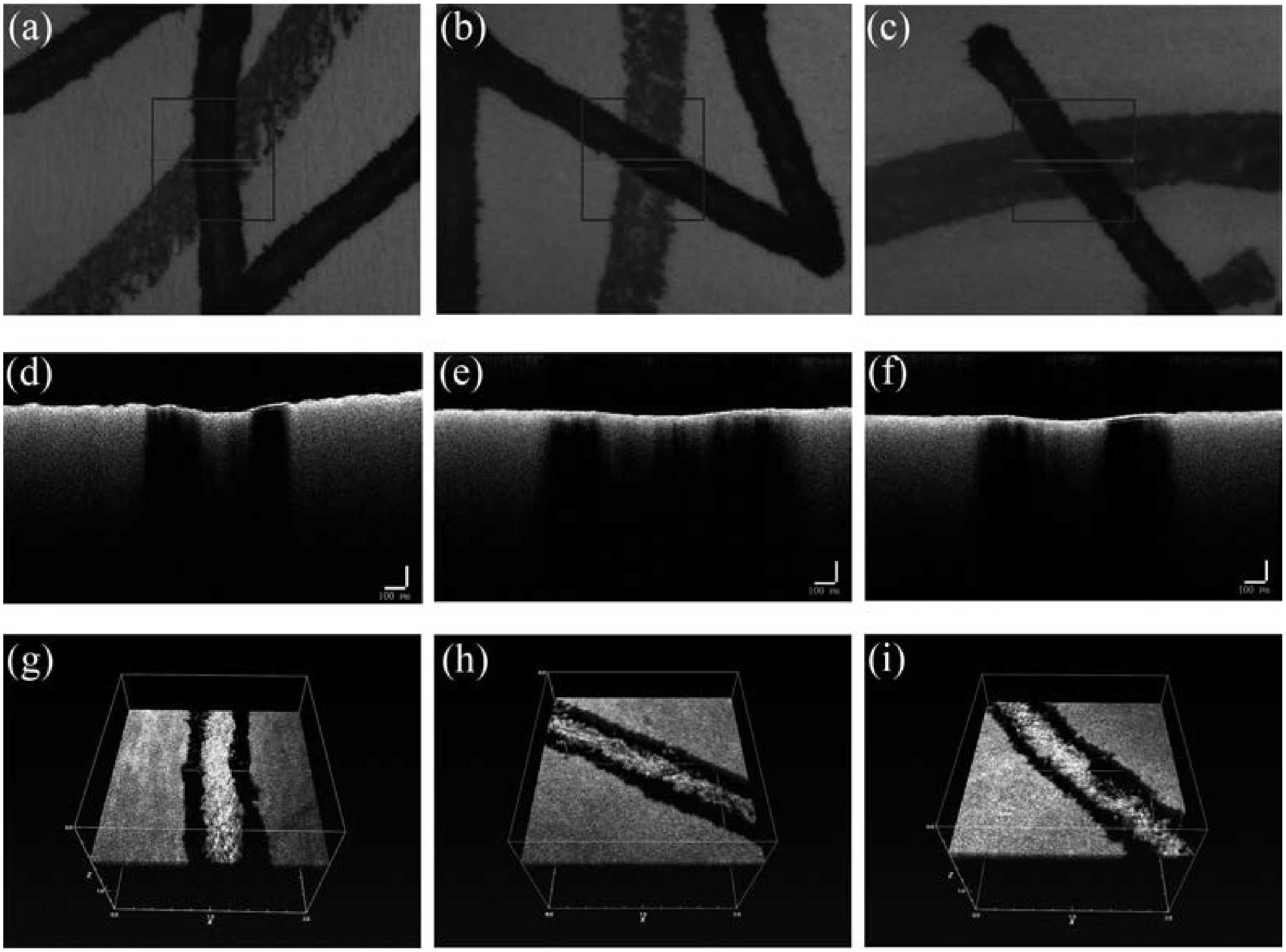


FIG. 2—*The images of the intersecting line samples when the gel pen 1 were present over three different types of stamp pad ink seals. The Photograph of the ROI of (a) Sample 1, (b) Sample 2, and (c) Sample 3; The 2D cross-sectional OCT images of the intersecting parts of (d) Sample 1, (e) Sample 2, and (f) Sample 3; the 3D volumetric OCT images of the intersecting parts of (g) Sample 1, (h) Sample 2, and (i) Sample 3. The 1D lateral scanning paths for 2D OCT images and the 2D lateral scanning ranges for 3D OCT images were indicated by the red arrows and the blue rectangular shown in (a), (b), and (c). The surfaces of the 3D images show the sub-layer images extracted at the depth of ~120 μm. [Color figure can be viewed at wileyonlinelibrary.com]*

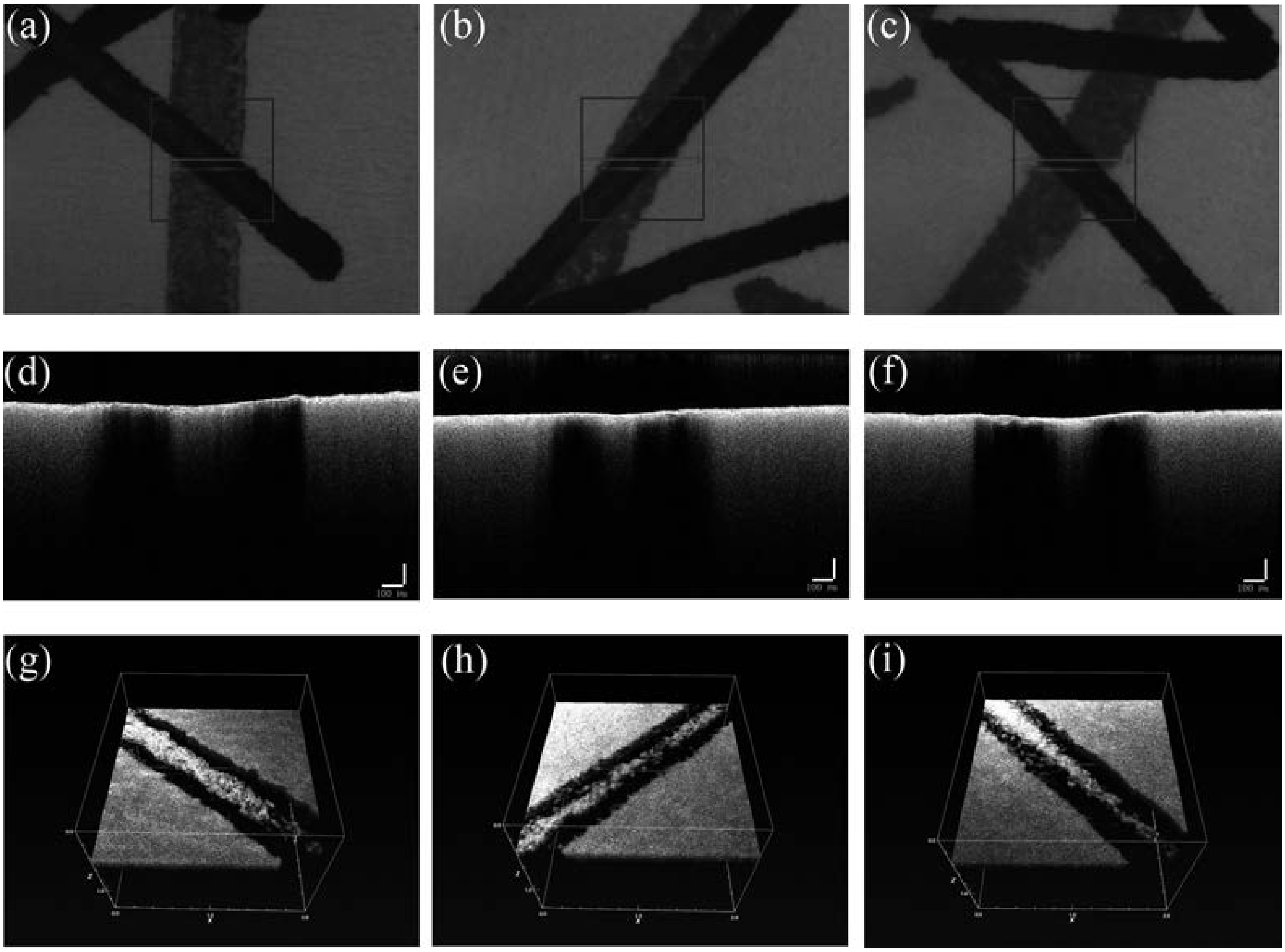


FIG. 3—*The images of the intersecting line samples when the gel pen 2 were present over three different types of stamp pad ink seals. The Photograph of the ROI of (a) Sample 4, (b) Sample 5 and (c) Sample 6; The 2D cross-sectional OCT images of the intersecting parts of (d) Sample 4, (e) Sample 5, and (f) Sample 6; the 3D volumetric OCT images of the intersecting parts of (g) Sample 4, (h) Sample 5, and (i) Sample 6. The 1D lateral scanning paths for 2D OCT images and the 2D lateral scanning ranges for 3D OCT images were indicated by the red arrows and the blue rectangular shown in (a), (b), and (c). The surfaces of the 3D images show the sub-layer images extracted at the depth of ~120 μm. [Color figure can be viewed at wileyonlinelibrary.com]*

Samples were prepared in the following manner. In the first set of samples, three different bands of seals were stamped first on each of the three sheets of white copy paper, respectively. In other words, each of these three sheets contained three seals made of different stamp pad inks. After the seals dried, some writings were made on top of these seals with one type of gel pen on each of the sheets respectively, and thus the gel pen inks overlapped on top of the stamp pad inks. In the second set of samples, three different bands of gel pens were used to first write on three sheets of white copy paper, respectively, then each type of the seals was placed on top of the writings, and thus the stamp pad inks were superimposed on top of the gel pen inks. All samples were written and stamped by the same person with similar and normal pressure as their usual habits. In this way, a total of 18 specimens were prepared and numbered for this study, as shown in Table 2.

Apparatus and Operations

As the gel pen ink and stamp pad ink are only an ultra-thin layer of material, the OCT system is required to have a high axial resolution (the minimum distinguishable distance in the depth direction). A high-resolution spectral-domain OCT (SD-OCT) system (GAN220, Thorlabs Inc., Newton, NJ) was employed to perform the examinations. SD-OCT modality enables an *in situ*, noninvasive, high-resolution, high-speed, high-sensitivity, and stable cross-sectional imaging. In a SD-OCT system, the IR light emitted from the broadband light source was divided into two parts by the optical coupler. One part was focused on the sample after passing through the mirror galvanometers (motorized mirror mounts and systems used for light beam steering or scanning) and objective lens and then returned after being backscattered and backreflected from the sample (sample light). The other part entered the reference arm and returned after being reflected back from a plane mirror (reference light). The interference between the sample light and the reference light was detected and recorded by a spectrometer consisted of a CCD as the detector. The A-line (A-scan) signal along the depth direction of the sample (*Z*-axis) was obtained after Fourier transform and a series of image processing to the original spectral interferogram. For a two-axis galvanometer, there are two orthogonal scanning

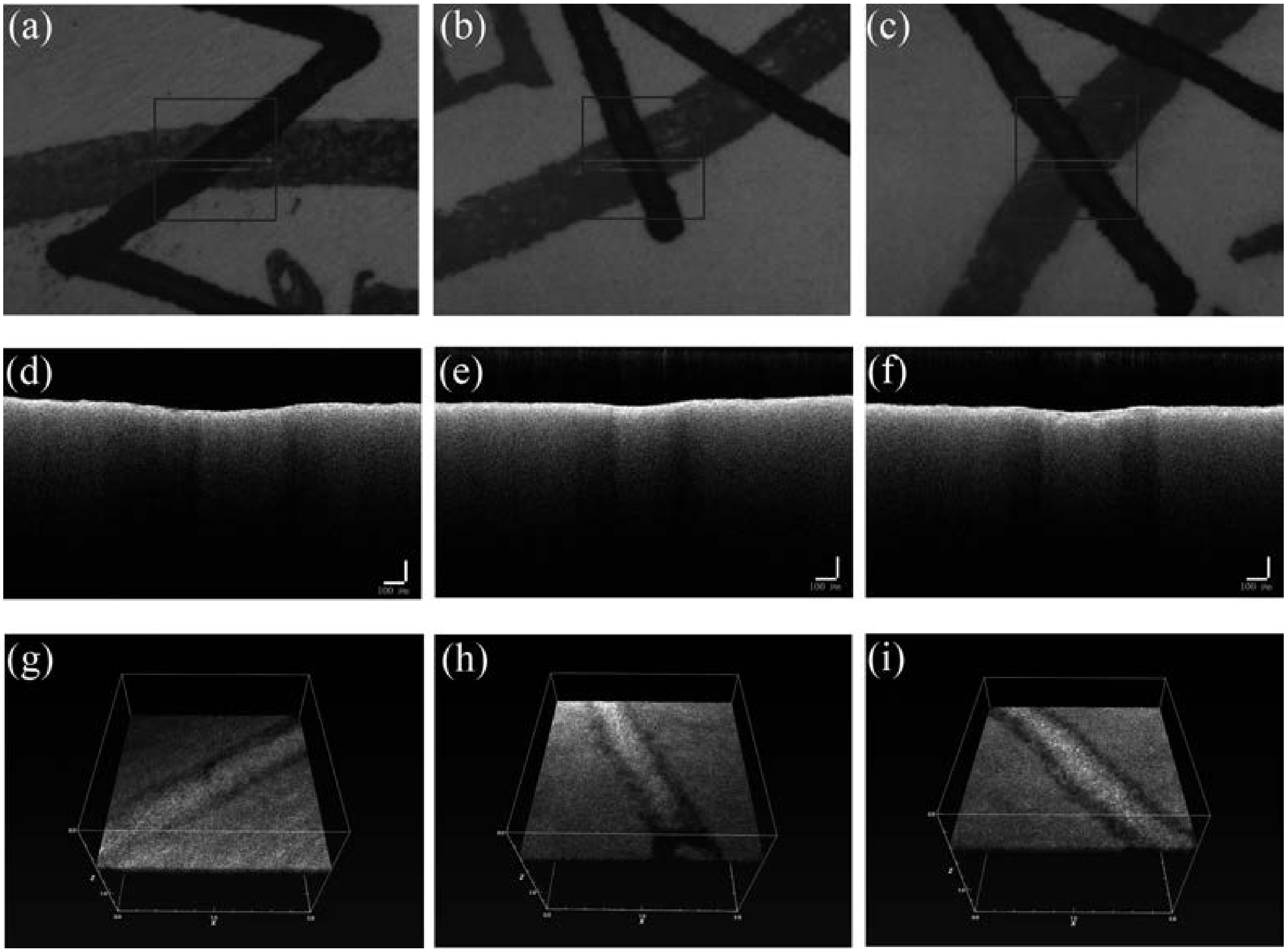


FIG. 4—*The images of the intersecting line samples when the gel pen 3 were present over three different types of stamp pad ink seals. The Photograph of the ROI of (a) Sample 7, (b) Sample 8, and (c) Sample 9; The 2D cross-sectional OCT images of the intersecting parts of (d) Sample 7, (e) Sample 8, and (f) Sample 9; the 3D volumetric OCT images of the intersecting parts of (g) Sample 7, (h) Sample 8, and (i) Sample 9. The 1D lateral scanning paths for 2D OCT images and the 2D lateral scanning ranges for 3D OCT images were indicated by the red arrows and the blue rectangular shown in (a), (b), and (c). The surfaces of the 3D images show the sub-layer images extracted at the depth of ~180 μm. [Color figure can be viewed at wileyonlinelibrary.com]*

directions to produce a raster scanning, which is achieved by inputting sawtooth signals and step signals into the fast axis and slow axis of the galvanometer, respectively. The 2D cross-sectional OCT image (B-scan) was generated by combining those A-scans acquired with the lateral scanning with a fast axis along one direction (*X*-axis). 3D volumetric OCT images were composed of those B-scans with the lateral scanning of a slow axis along the other direction (*Y*-axis).

The SD-OCT system adopted a broadband light source SLD (Super-Luminescent Diode) with the center wavelength of 900 nm and bandwidth of 150 nm. The optical power on the sample was 1–2 mW. The axial resolution of the system was tested to be 3 μm (in the air) and the lateral resolution was 4 μm. The maximum imaging depth in the air was 1.9 mm. The imaging frequency (A-line rate) of this system was 36 kHz, which could achieve a real-time 2D cross-sectional imaging and high signal-to-noise ratio (SNR) of 93 dB. The working distance of the system from objective lens (LSM02-BB, Thorlabs Inc.) was 18 mm. The size of the 2D OCT image captured in this experiment was 2.0 mm (lateral) × 1.2 mm (depth). The operation of this system was extremely convenient, the examiners only need to place the sample on the sample stage and adjust the focus and reference light intensity, and then the OCT images of the region-of-interest could be collected and analyzed in a fast manner.

Gel pen ink contains dyes or pigments, solvents, resins, viscosity adjustors, and lubricants, which is a water-based ink and the composition may vary by manufacturers (8). Two types of seals in our experiment are made of water-based stamp pad inks, and one is made of an oil-based stamp pad ink. The OCT system utilized near-infrared light as the light source that could penetrate into the intersecting line samples and obtain the inner structures of the samples. When different gel pen inks and stamp pad inks formed the intersecting line samples, the depth of light penetration into the intersecting parts could be varied due to the different optical attenuation properties of these parts. In OCT images, the bright areas indicate that there is light backscattering and backreflecting from the sample, and there is no light returning when dark areas are shown. When gel pen ink blends with the stamp pad ink as well as the fibers of paper, it may result in distinguishing optical features in OCT images.

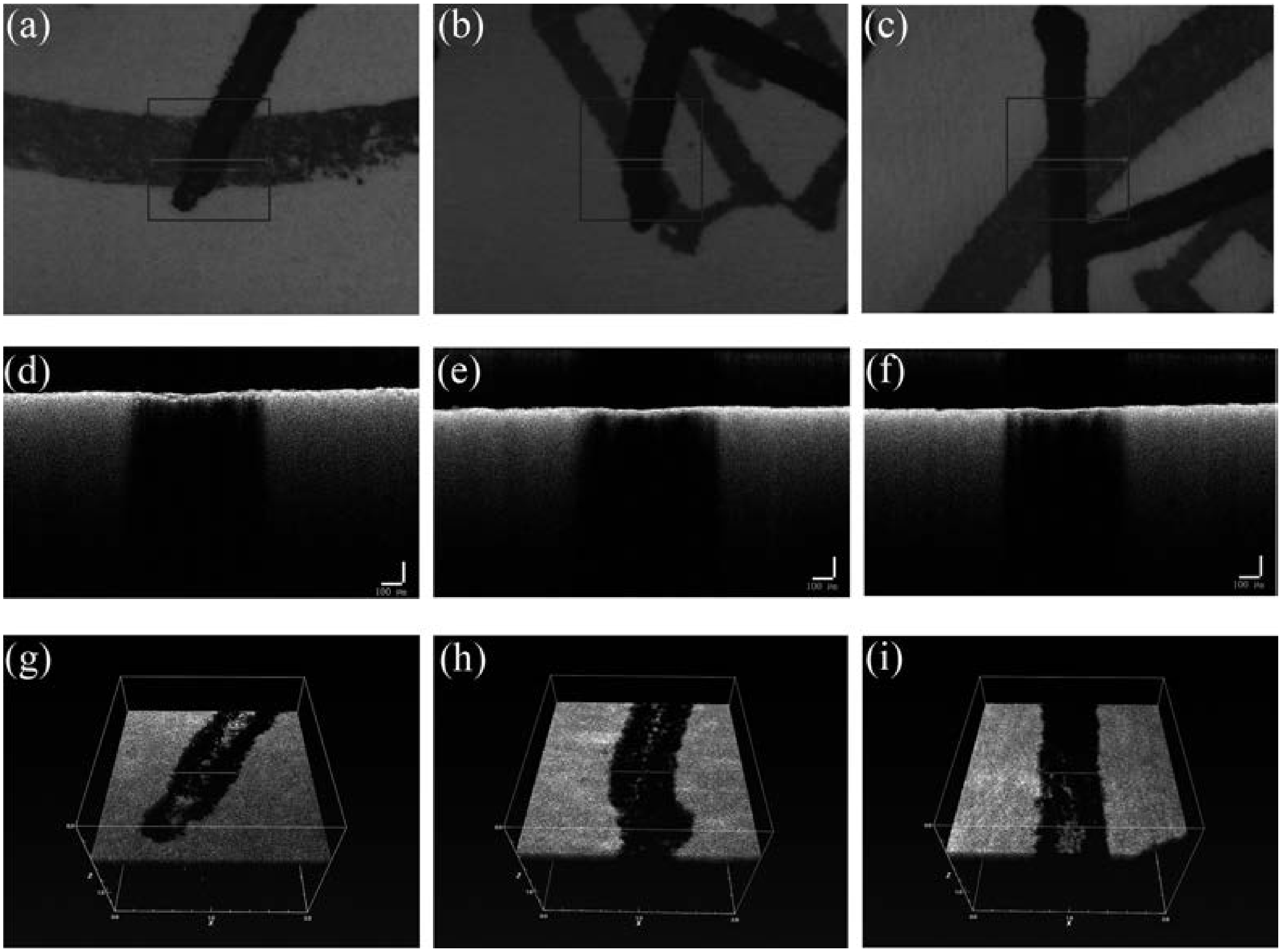


FIG. 5—*The images of the intersecting line samples when three different types of stamp pad ink seals were present over the gel pen 1. The Photograph of the ROI of (a) Sample 10, (b) Sample 11, and (c) Sample 12; The 2D cross-sectional OCT images of the intersecting parts of (d) Sample 10, (e) Sample 11, and (f) Sample 12; the 3D volumetric OCT images of the intersecting parts of (g) Sample 10, (h) Sample 11, and (i) Sample 12. The 1D lateral scanning paths for 2D OCT images and the 2D lateral scanning ranges for 3D OCT images were indicated by the red arrows and the blue rectangular shown in (a), (b), and (c). The surfaces of the 3D images show the sub-layer images extracted at the depth of ~120 μm. [Color figure can be viewed at wileyonlinelibrary.com]*

Results and Discussion

The OCT Images of Intersecting Line Samples: The Gel Pen Ink Over Stamp Pad Ink

The first set of experiments performed OCT imaging of the intersecting line samples when the gel pen ink strokes were present over the stamp pad ink seals, as shown in Fig. 2 (ink of gel pen 1 over three different seals, Sample 1–Sample 3), Fig. 3 (ink of gel pen 2 over three different seals, Sample 4–Sample 6), and Fig. 4 (ink of gel pen 3 over three different seals, Sample 7–Sample 9). In each figure, (*a*) (*d*) (*g*) represented the photograph of the region-of-interest (ROI), the 2D cross-sectional OCT image, and the 3D volumetric OCT image of the sample, respectively. The image size of each 2D cross-sectional OCT image was 2 mm (lateral) × 1.2 mm (depth). The lateral scanning path was indicated by the red arrows shown in (*a*), (*b*), and (*c*). From the cross-sectional images as shown in (*d*), (*e*), and (*f*), it can be seen that the light penetration depth at the intersecting position (the central part of the image) was relatively low as compared to the other parts due to a strong light absorption of the ink of black gel pen. 3D volumetric OCT images were presented as 3D rendering views in (*g*), (*h*), and (*i*), with a corresponding lateral scanning range of 2 mm × 2 mm as indicated by the blue rectangular in (*a*), (*b*), and (*c*). The surface of these 3D rendering images actually displayed the sub-layer images extracted inside the sample at a certain depth, which demonstrated the capability of 3D OCT to precisely obtain a tomographic image at a specific depth.

The OCT Images of Intersecting Line Samples: The Stamp Pad Ink Over Gel Pen Ink

The second set of experiments performed OCT imaging of the intersecting line samples when the stamp pad ink seals were present over the gel pen ink strokes, as shown in Fig. 5 (ink of gel pen 1 under three different seals, Sample 10–Sample 12), Fig. 6 (ink of gel pen 2 under three different seals, Sample 13–Sample 15), and Fig. 7 (ink of gel pen 3 under three different seals, Sample 16–Sample 18). The photographs of the ROI, the 2D cross-sectional OCT images, and the 3D volumetric OCT images were also obtained and shown as the same in the first set of experiments. The surface of the 3D volumetric OCT images displayed the sub-layer images extracted at a certain depth.

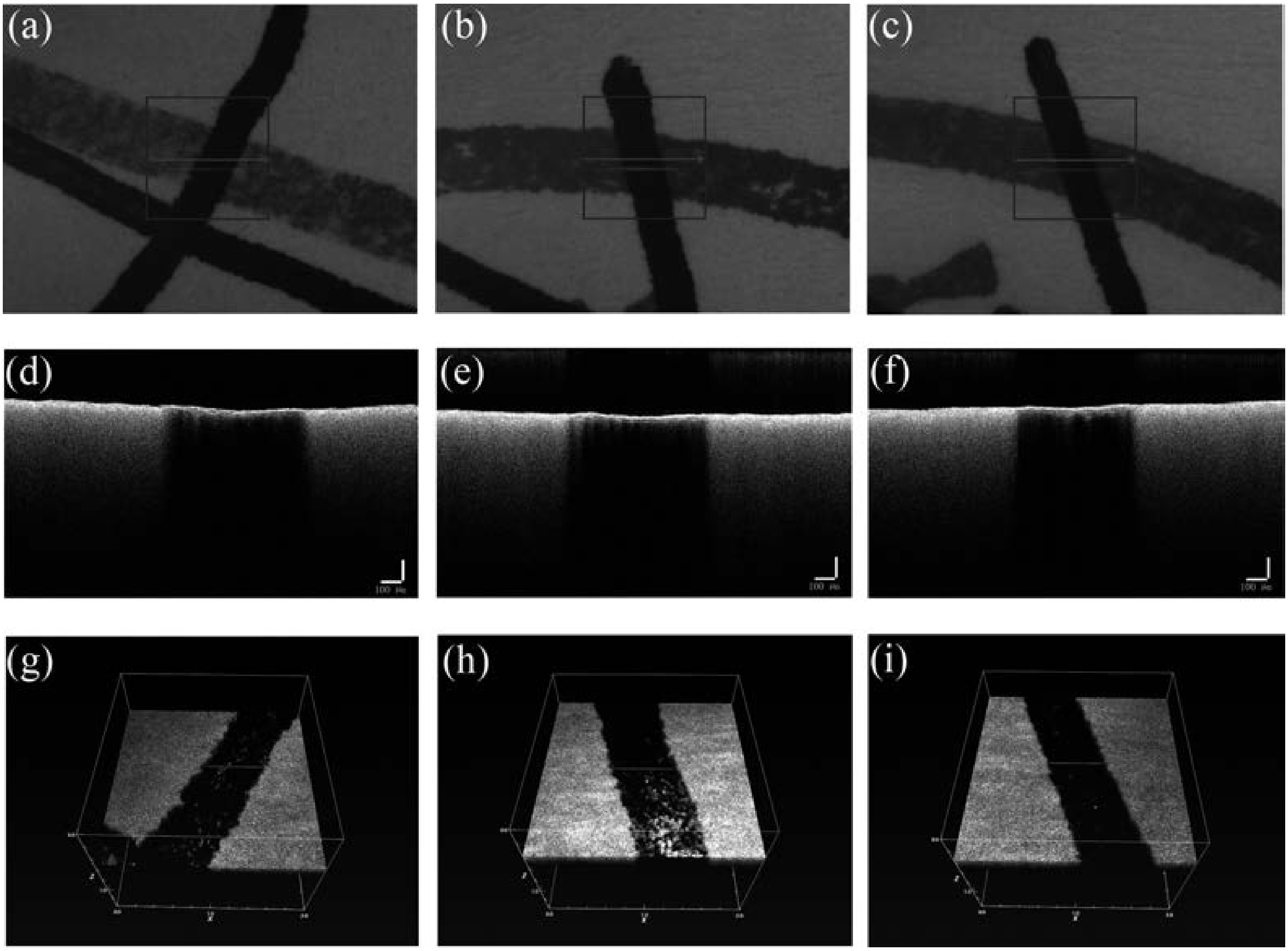


FIG. 6—*The images of the intersecting line samples when three different types of stamp pad ink seals were present over the gel pen 2. The Photograph of the ROI of (a) Sample 13, (b) Sample 14, and (c) Sample 15; The 2D cross-sectional OCT images of the intersecting parts of (d) Sample 13, (e) Sample 14, and (f) Sample 15; the 3D volumetric OCT images of the intersecting parts of (g) Sample 13, (h) Sample 14, and (i) Sample 15. The 1D lateral scanning paths for 2D OCT images and the 2D lateral scanning ranges for 3D OCT images were indicated by the red arrows and the blue rectangular shown in (a), (b), and (c). The surfaces of the 3D images show the sub-layer images extracted at the depth of ~120 μm. [Color figure can be viewed at wileyonlinelibrary.com]*

Discussion

When comparing the 2D cross-sectional OCT images with the opposite sequences, several distinguishing features could be clearly observed. In the middle of the intersecting part where the gel pen ink was present over the stamp pad ink, there was a region that had relatively large penetration depth and curved surface, as shown in Figs 2–4. However, it was difficult to observe these features in the other sequence, as shown in Figs 5–7. These features could be also observed through the 3D volumetric OCT images that the backscattering or backreflecting intensity from the central portion of the intersecting strokes when the gel pen ink was placed over stamp pad ink was higher than that from the both sides of the center. In contrast, such phenomenon could not be observed when the gel pen ink was present below the stamp pad ink.

These distinguishing features observed in the above two sets of experiments were likely caused by the lower light absorption of particles in the central part of the intersecting strokes when the gel pen ink was present over the stamp pad ink. When writing with a gel pen over the stamp pad ink seal, the central part of the gel pen-tip was applied with more pressure than both sides from the center. Therefore, the gel pen ink was integrated with the stamp pad ink in the central part of the stroke but it only covered on top of the stamp pad ink at the edges of the pen trough, leading to a lower absorption and larger penetration depth in the central part but the opposite at the edges of pen tract. In contrast, when affixing a seal to the gel pen ink stroke, the intersecting strokes were evenly stressed and the gel pen ink was overlaid with the stamp pad ink instead of an integration, resulting in a lower penetration depth in the whole intersecting part.

The cross-sectional OCT images in Figs 4 and 7 show that the light penetration depth in the intersecting parts of these samples was significantly larger than that of the other samples. The ink of gel pen 3 presented a weak light absorption as compared to gel pen 1 and 2, which resulted in a larger penetration depth in the intersecting parts. Meanwhile, from those 3D volumetric OCT images, we could observe a higher image contrast between the central and peripheral portion of the intersecting strokes when the gel pen ink was present over the stamp pad ink than the counterpart in the opposite sequential order. Although the observed features in those samples related to gel pen 3 were not as obvious as that related to the other gel pens due to the relatively weak light absorption in gel pen ink, they might be still helpful to determine the sequence of intersecting lines.

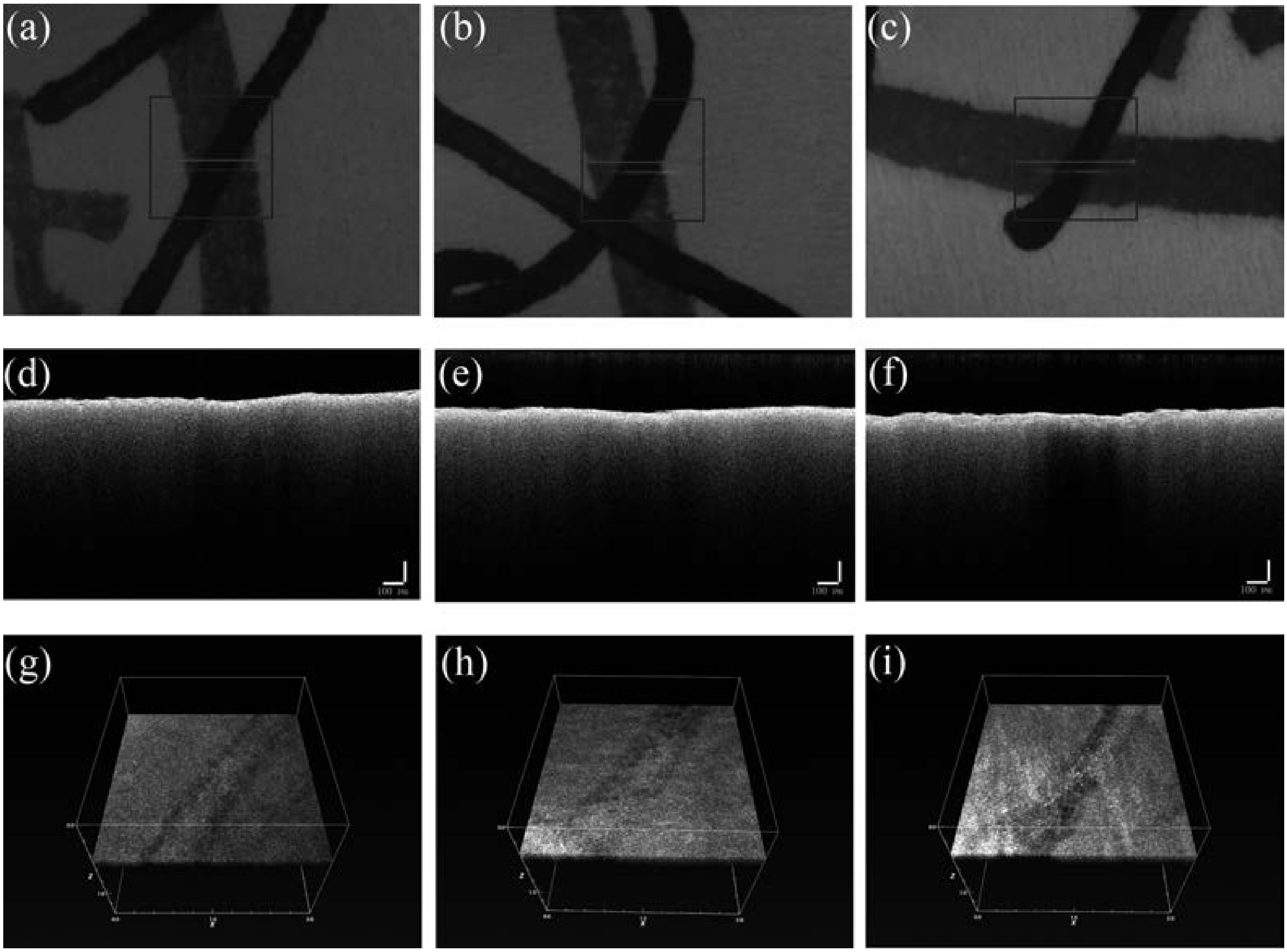


FIG. 7—*The images of the intersecting line samples when three different types of stamp pad ink seals were present over the gel pen 3. The Photograph of the ROI of (a) Sample 16, (b) Sample 17, and (c) Sample 18; The 2D cross-sectional OCT images of the intersecting parts of (d) Sample 16, (e) Sample 17, and (f) Sample 18; the 3D volumetric OCT images of the intersecting parts of (g) Sample 16, (h) Sample 17, and (i) Sample 18. The 1D lateral scanning paths for 2D OCT images and the 2D lateral scanning ranges for 3D OCT images were indicated by the red arrows and the blue rectangular shown in (a), (b), and (c). The surfaces of the 3D images show the sub-layer images extracted at the depth of ~180 μm. [Color figure can be viewed at wileyonlinelibrary.com]*

TABLE 3—*Blind testing results.*

Sample Number	Order	DE1*	DE2*	DE3*	DE4*	DE5*
BTS1*	Gel pen 1 over Seal 1	✓	✓	✓	✓	✓
BTS2*	Seal 1 over Gel pen 1	✓	✓	✓	✓	✓
BTS3*	Gel pen 2 over Seal 1	✓	✓	✓	✓	✓
BTS4*	Seal 1 over Gel pen 2	✓	✓	✓	✓	✓
BTS5*	Seal 3 over Gel pen 1	✓	✓	✓	✓	✓
BTS6*	Gel pen 1 over Seal 3	✓	✓	✓	✓	✓
BTS7*	Gel pen 2 over Seal 2	✓	✓	✓	✓	✓
BTS8*	Seal 3 over Gel pen 2	✓	✓	✓	✓	✓
BTS9*	Gel pen 3 over Seal 1	✓	✓	✓	✓	✓
BTS10*	Seal 1 over Gel pen 3	✓	Inconclusive	✓	Inconclusive	✓
BTS11*	Gel pen 3 over Seal 2	✓	✓	Inconclusive	✓	✓
BTS12*	Seal 2 over Gel pen 3	✓	✓	✓	Inconclusive	✓

*BTS represents blind testing sample; DE represents document examiners performing blind testing with OCT.

Compared with the existing optical methods, the advantage of OCT technology is that it can obtain high-resolution structural information inside the sample, especially the internal layering information, which is critical for determining the sequence of intersecting lines. However, most of the traditional optical methods can only obtain surface information. In addition, the OCT system is easy to operate. The operator only needs to adjust the focus and the sample position on the sample stage for a real-time imaging due to the high imaging speed. Finally, fiber-based OCT system can become portable and miniaturized, which is promising to carry out a flexible on-site imaging. Due to the nondestructive, high-speed, high-resolution imaging characteristics, OCT technology can be applicable as a pre-examination method before any destructive or complicated methods in the workflow of a standard forensic science laboratory.

Blind Testing

Five questioned document examiners took part in the blind testing. A total of 12 blind tests were conducted where the sequence of the intersecting strokes was unknown to these examiners. The OCT methods were explained to these participants, including the basic principles of OCT technology, the distinguishing features and the judging criteria. The five questioned document examiners operated the OCT system independently to obtain the OCT images. Then, they were asked to determine the order of all the blind test samples individually and carefully. As listed in Table 3, the results indicated that five examiners correctly determine the sequence of blind testing samples when using the gel pen 1 or gel pen 2 over different stamp pad seals, achieving 100% accuracy. However, some inconclusive results were given when the gel pen 3 was one of the materials forming the samples due to the similar image contrast between the central and peripheral portion of the intersecting strokes. In these cases, the accuracy could still reach 80%. In other words, it may be more difficult (but not impossible) to distinguish the intersecting lines when using the gel pen ink with a low absorption of infrared light.

Conclusion

In this study, OCT technology was proposed for the first time to determine the sequence of intersecting lines. A total of 18 intersecting line samples produced with 3 different types of seals and 3 different kinds of gel pens were imaged by a high-resolution SD-OCT system. 2D cross-sectional and 3D volumetric OCT images were obtained to perform the feature comparison and analysis. A total of 12 blind tests were conducted to demonstrate the effectiveness of OCT technique and the light absorption of the ink would affect the accuracy of examination. The results from this preliminary study demonstrated that OCT can be very useful in determining the sequence of intersecting lines between gel pens and stamp pad ink seals, which provided a novel method for questioned document examination. Although OCT method still had some limitations such as relatively low penetration depth when the sample had a strong absorption of infrared light, this use of OCT technology was a preliminary study which offers a promising means to determine the sequence of intersecting lines between gel pens and stamp pad ink seals. It is promising to combine OCT technology with traditional methods to improve the accuracy of the examination. In the future, further tests will be conducted to test the effectiveness of the OCT method to determine the sequence of intersecting lines formed by other writing and printing materials.

References

1. Shiver FC. Intersecting lines: documents. In: Jamieson A, Moenssens A, editors. Wiley encyclopedia of forensic science. Chichester: Wiley, 2009;1594–600. https://doi.org/10.1002/9780470061589.fsa328.
2. Li B, Ouyang GL. An examination of the sequence of intersecting seal and laser printing toner line. J Forensic Sci 2017;62(2):476–82. https://doi.org/10.1111/1556-4029.13283.
3. Bao R, Li B, Xie P. Stained document examination. Beijing, China: Police Education Press, 2014.
4. Wang Y, Li B. Determination of the sequence of intersecting lines from laser toner and seal ink by Fourier transform infrared microspectroscopy and scanning electron microscope/energy dispersive X-ray mapping. Sci Justice 2012;52(2):112–8. https://doi.org/10.1016/j.scijus.2011.10.001.
5. Saini K, Kaur R, Sood NC. Determining the sequence of intersecting gel pen and laser printed strokes—a comparative study. Sci Justice 2009;49(4):286–91. https://doi.org/10.1016/j.scijus.2009.07.003.
6. Poulin G. Establishing the sequence of strokes: the state of the art. Int J Forensic Doc Exam 1996;2(1):16–32.
7. Cheng K, Chao C, Jeng B, Lee S. A new method of identifying writing sequence with the laser scanning confocal microscope. J Forensic Sci 1998;43(2):348–52. https://doi.org/10.1520/JFS16144J.
8. Li B, Ouyang G, Zhao P. Preliminary study on determining the sequence of intersecting Lines by fluorescence technique. J Forensic Sci 2018;63(2):577–82. https://doi.org/10.1111/1556-4029.13572.
9. Wu X, Fang F, Li B. Determination of the sequence of intersecting lines between toners and seals by laser fluorescence microscope. J Forensic Sci 2019;64(6):1761–8. https://doi.org/10.1111/1556-4029.14097.
10. Braz A, López-López M, García-Ruiz C. Raman imaging for determining the sequence of blue pen ink crossings. Forensic Sci Int 2015;249:92–100. https://doi.org/10.1016/j.forsciint.2015.01.023.
11. Li B. An examination of the sequence of intersecting lines using microspectrophotometry. J Forensic Sci 2016;61(3):809–14. https://doi.org/10.1111/1556-4029.13022.
12. Bojko K, Roux C, Reedy BJ. An examination of the sequence of intersecting lines using attenuated total reflectance–Fourier transform infrared spectral imaging. J Forensic Sci 2008;53(6):1458–67. https://doi.org/10.1111/j.1556-4029.2008.00796.x.
13. Kim J, Kim MJ, An W, Kim Y. Determination of the sequence of intersecting lines using Focused Ion Beam/Scanning Electron Microscope. J Forensic Sci 2016;61(3):803–8. https://doi.org/10.1111/1556-4029.13076.
14. Kasas S, Khanmy-Vital A, Dietler G. Examination of line crossings by atomic force microscopy. Forensic Sci Int 2001;119(3):290–8. https://doi.org/10.1016/S0379-0738(00)00458-8.
15. Huang D, Swanson EA, Lin CP, Schuman JS, Stinson WG, Chang W, et al. Optical coherence tomography. Science 1991;254(5035):1178–81. https://doi.org/10.1126/science.1957169.
16. Fercher AF, Drexler W, Hitzenberger CK, Lasser T. Optical coherence tomography-principles and applications. Rep Prog Phys 2003;66(2):239. https://doi.org/10.1088/0034-4885/66/2/204.
17. Cense B, Chen TC, Park BH, Pierce MC, de Boer JF. In vivo birefringence and thickness measurements of the human retinal nerve fiber layer using polarization-sensitive optical coherence tomography. J Biomed Opt 2004;9(1):121–6. https://doi.org/10.1117/1.1627774.
18. Chen Z, Zhao Y, Srinivas SM, Nelson JS, Prakash N, Frostig RD. Optical doppler tomography. IEEE J Sel Top Quant 1999;5(4):1134–42. https://doi.org/10.1109/2944.796340.
19. Welzel J, Lankenau E, Birngruber R, Engelhardt R. Optical coherence tomography of the human skin. J Am Acad Dermatol 1997;37(6):958–63. https://doi.org/10.1016/S0190-9622(97)70072-0.
20. Wang ZG, Durand DB, Schoenberg M, Pan YT. Fluorescence guided optical coherence tomography for the diagnosis of early bladder cancer in a rat model. J Urol 2005;174(6):2376–81. https://doi.org/10.1097/01.ju.0000180413.98752.a1.
21. Choi WJ, Min GH, Lee BH, Eom JH, Kim JW. Counterfeit detection using characterization of safety feature on banknote with full-field optical coherence tomography. J Opt Soc Korea 2010;14(4):316–20. https://doi.org/10.3938/jkps.57.79.
22. Brown K, Harvey M. Optical coherence tomography: age estimation of Calliphora vicina pupae in vivo? Forensic Sci Int 2014;242:157–61. https://doi.org/10.1016/j.forsciint.2014.07.001.
23. Laan N, Bremmer RH, Aalders MCG, Bruin KG. Volume determination of fresh and dried bloodstains by means of optical coherence tomography. J Forensic Sci 2014;59(1):34–41. https://doi.org/10.1111/1556-4029.12272
24. Zhang N, Wang C, Sun Z, Mei H, Huang W, Xu L, et al. Characterization of automotive paint by optical coherence tomography. Forensic Sci Int 2016;266:239–44. https://doi.org/10.1016/j.forsciint.2016.06.007.
25. Wang C, Zhang N, Sun Z, Li Z, Li Z, Xu X. Recovering hidden sublayers of repainted automotive paint by 3D optical coherence tomography. Aust J Forensic Sci 2019;51(3):331–9. https://doi.org/10.1080/00450618.2017.1367418.
26. Zhang N, Wang C, Sun Z, Li Z, Xie L, Yan Y, et al. Detection of latent fingerprint hidden beneath adhesive tape by optical coherence tomography. Forensic Sci Int 2018;287:81–7. https://doi.org/10.1016/j.forsciint.2018.03.030.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14547
Available online at: onlinelibrary.wiley.com

PAPER

QUESTIONED DOCUMENTS

Cami Fuglsby,[1] *M.S.; Christopher Saunders,*[1] *Ph.D.; Danica M. Ommen,*[2] *Ph.D.; and Michael P. Caligiuri,*[3] *Ph.D.*

Use of an Automated System to Evaluate Feature Dissimilarities in Handwriting Under a Two-Stage Evaluative Process*,†

ABSTRACT: The two-stage evaluative process is an established framework utilized by forensic document examiners (FDEs) for reaching a conclusion about the source(s) of handwritten evidence. In the second, or discrimination, stage, the examiner attempts to estimate the rarity of observations in a relevant background population. Unfortunately, control samples from a relevant background population are often unavailable, leaving the FDE to reach this determination based on subjective experience. Automated handwriting feature recognition systems are capable of performing both feature comparison and discrimination, yet these systems have not been subjected to empirical validation studies. In the present study, we repurposed a commercially available automated system to generate empirical distributions for ranking feature dissimilarity scores among pairs of handwritten phrases. The blinded results of this automated process were used to survey an international cohort of 36 FDEs regarding their strength of support for same- and different-writer propositions. The survey served to cross-validate FDE decision-making under the two-stage approach. Results from the survey demonstrated a clear pattern of response consistent with ground truth. Predictive regression analyses indicated that the automated feature dissimilarity scores and the log of their cumulative distribution functions accounted for 72% of the variability in FDE opinions. This study demonstrated that feature dissimilarity scores acquired using automated processes and their distributions are closely aligned with FDE decision-making processes supporting the heuristic value of the two-stage evaluative framework.

KEYWORDS: automated systems, handwriting, handprinting, two-stage approach, evidence interpretation, questioned documents

A common approach to evidence interpretation in handwriting examination involves two stages (1–5). The first stage in this process is described as the *match or comparison* stage and relies upon the examiners' observational skills to determine whether characteristics or features of the two sets of suspect evidence are indistinguishable. A set of suspect samples are deemed *distinguishable* if they share few features in common and have a number of discriminating elements. A *discriminating element* is "a relatively discrete characteristic or feature of writing that varies observably or measurably across writers and may contribute reliably to distinguishing between samples from different individuals, or conversely, support the contention of sameness within a common writer" (6). If the suspect samples are deemed distinguishable, the evidence suggests that they share two different sources. Alternatively, if the suspect samples are deemed *indistinguishable*, they share many features and characteristics between them, and few (or no) discriminating elements are observed. The term indistinguishable does not indicate that the two samples do share a common source, but only describes the characteristics of the two samples with respect to the proposition that they share a common but unknown source. If the suspect samples are deemed indistinguishable, the examination proceeds to the second stage, described as the *discrimination or significance* stage; if two sets of evidence are considered indistinguishable from one another, the examiner attempts to estimate the rarity of the observed characteristics in a relevant background population. In a slightly more formal sense, the second stage describes the likelihood of a chance match (as described in Found and Bird [7]), or random match probability. Thus, under the two-stage approach, the probative value of finding that the suspect samples are indistinguishable is strengthened by the findings that shared combinations of features and characteristics between two individuals are extremely uncommon among members of a relevant population.

Prior research offers broad support for document examiners' proficiency for reaching accurate writership decisions based on handwriting feature analyses in closed-set laboratory experiments (8–13). However, research on the impact of population-level information to the FDE evaluative process under a two-stage framework is lacking for several reasons. With few exceptions (14,15), the population distribution of specific combinations of

[1] Department of Mathematics and Statistics, South Dakota State University, AME Building Box: 2225, Brookings, SD, 57007.

[2] Department of Statistics, Iowa State University, 2438 Osborn Dr, Ames, IA, 50011.

[3] Department of Psychiatry (0603), University of California at San Diego, 9500 Gilman Drive, La Jolla, CA, 92093.

Corresponding author: Michael P. Caligiuri, Ph.D. E-mail: mcaligiuri@ucsd.edu

*This research was supported by a grant from the National Institute of Justice, 2017-DN-BX-0148.

†The preliminary findings were presented at the 77th Annual Meeting of the American Society of Questioned Document Examiners, 3–7 August 2019, in Cary, ND; and at the 72nd Annual Scientific Meeting of the American Academy of Forensic Sciences, 17–22 February 2020, in Anaheim, CA.

Received 30 April 2020; and in revised form 29 June 2020; accepted 28 July 2020.

handwriting features and characteristics is currently insufficient to assist the examiner in meeting the goals of the discrimination stage. Because population-level frequency distributions for handwriting characteristics relevant to a particular case are usually unavailable to assist the examiner in this discrimination stage, in cases where samples are distinguishable, writership conclusions under the two-stage approach are largely inferential. This introduces a critical challenge to admissibility of evidence involving handwriting often prompting Daubert hearings (16–18). As in many forensic pattern-matching disciplines, handwriting is high-dimensional and complex. Handwriting is comprised of a sequence of individual movements each with multiple temporal (e.g., movement duration and speed), spatial (e.g., horizontal and vertical size), and geometric (e.g., slant or loop area) attributes that distinguish one writing segment from another to convey meaning. The complexity of handwriting is borne out by the interactions among these attributes driven by context variability, physical constraints, and individual's natural variation. Taken together, estimating the prevalence of specific feature differences within a population of writers can be a herculean undertaking. Even if such data were publicly available to examiners, statistical procedures are needed to quantify the atypicality of these feature differences in the population distribution.

Automated feature extraction programs such as FLASH ID® (Sciometrics LLC, Chantilly, VA, USA) have advantages over their human counterparts particularly with respect to estimating the likelihood that a questioned handwriting sample came from a candidate residing within a population reference set. Moreover, with careful reprogramming and armed with a large database of features and feature differences, automated systems can be repurposed to generate population distribution functions for a numeric estimation of the rarity of feature dissimilarity scores from within a large reference set. In this way, the automated system would inform the discrimination stage of the two-stage process.

Two experiments were conducted for the present study. The purpose of the first experiment was to deploy an automated feature extraction program to generate feature dissimilarity scores and population distribution functions for ranking these feature dissimilarity scores among pairs of handwritten phrases across different phrases and styles of handwriting. We designed a specialized algorithm that customizes the output of an automated feature extraction program designed for closed-set identification (19) to first calculate feature dissimilarities between 81,180 sample pairs of print and cursive handwriting from known between- and within-writer sources. The terms between-writer and within-writer also refer to different and common source, respectively. The latter terms are more general in application and can refer to source specimens other than handwriting. Population-level distribution functions were then created from the dissimilarity scores for each unique phrase. In this way, feature dissimilarity scores are ranked according to their placement within the dissimilarity score population. Sample pairs with dissimilarity scores in the tails of the distribution would be considered "rare" for a relevant population of samples.

The aim of the second experiment was to utilize these dissimilarity scores and distribution functions to design a series of difficult-case scenarios for FDEs to evaluate. Sample pairs falling along the tails of their respective population distributions were submitted to an international cohort of FDEs to demonstrate the utility of an automated feature-based process within the two-stage evaluative framework. The second experiment served as a cross-validation of FDE decision-making under the two-stage approach.

Methods and Procedures

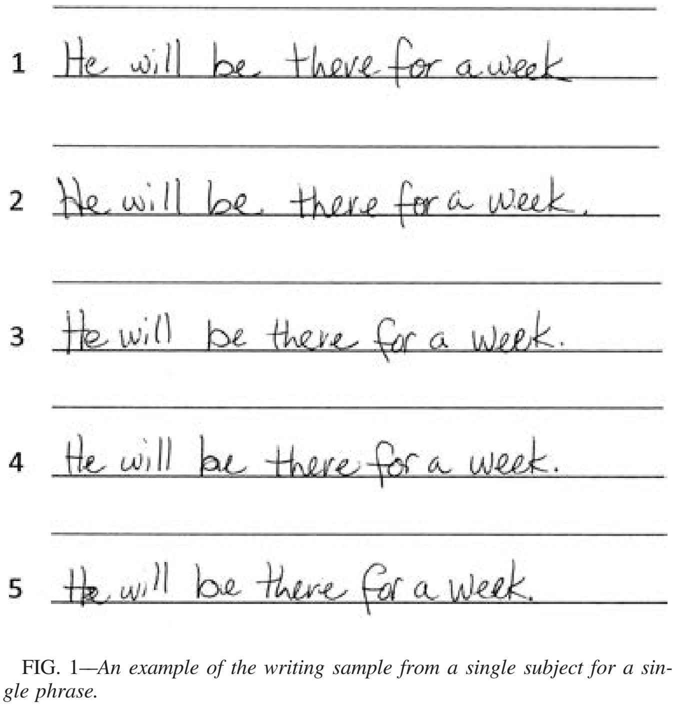
Writers and Handwriting Samples

The study recruited 33 individuals from the San Diego Sheriff's Crime Laboratory who were asked to write six phrases from the London Letter and to repeat each phrase five times using both print and cursive writing styles. This provided 60 phrases per individual writer. The six phrases from the London Letter were as follows: (i) Our London business is good; (ii) but Vienna and Berlin are quiet; (iii) Mr. Lloyd has gone to Switzerland; (iv) and I hope for good news; (v) He will be there for a week; and (vi) and then goes to Turin. Subjects wrote each of the phrases five times with an inking pen on lined paper placed on a Wacom (Intuos Pro, model PTH-660) digitizing tablet. The stimulus phrase was shown on the top of each page, and repetitions were written vertically, five per page. Seven subjects returned to the laboratory two weeks later and repeated the writing experiment. Figure 1 shows a page layout and writing sample from a single subject for a single phrase.

While the digitized samples were subjected to analyses of kinematic features to be used in the predictive modeling component of this research (not included here), the ink copies were used in the present study. Each page of the hard copy ink samples was scanned at 600 dpi, cropped into individual repetitions, and saved as separate 16-bit TIFF files for automated analyses. With 33 writers, each writing six phrases five times each in both cursive and print, there was a total of 13,530 sample pairs available per phrase and writing style that were used to calculate phrase-dependent population-level dissimilarity scores and their respective dissimilarity functions.

FLASH ID® Feature Dissimilarity Scores and Population Density Curves

Due to the practical constraints of examining thousands of pairs of writing samples to identify cases of interest for inclusion



1 He will be there for a week

2 He will be there for a week.

3 He will be there for a week.

4 He will be there for a week.

5 He will be there for a week.

FIG. 1—*An example of the writing sample from a single subject for a single phrase.*

in an FDE survey, it is desirable to use the help of an automated system to identify these pairs. All paper samples obtained during the collection process were scanned into FLASH ID® for the purpose of obtaining a univariate score for each pair of samples representing the level of dissimilarity between the writing contained in the samples (the higher the score is, the more dissimilar the two samples are).

FLASH ID®, an automated feature extraction program, generally serves the purpose of closed-set biometric identification (19). FLASH ID® provides a ranking (with respect to a reference set of writing samples from 50 known writers) of candidate writers based on similarities between combinations of features from a single questioned sample and the reference set of known writing samples. For the present study, we leveraged this capability to develop a univariate omnibus dissimilarity score for comparing features from two questioned handwriting samples. Our procedure involved obtaining the Euclidean distance between the two vectors of scores (from the FLASH ID® output) to provide a univariate score reflecting the feature distance between two samples. This univariate score thus represents the dissimilarity in features between all possible pairs of samples with larger scores reflecting greater feature dissimilarity. The output of this customized algorithm consisted of dissimilarity scores for all possible pairs across the six phrases, five repeats, and two writing styles for 33 writers, leading to a total of 81,180 possible pairs for each writing style. We then derived the distributions for each phrase and writing style from the available sample pool. The cumulative density score or function served as an index of the rarity of the dissimilarity score within the relevant population of dissimilarity scores and is referred to as the empirical cumulative distribution function or ECDF.

The statistical programming language R (20) was used to compute all pairwise comparison scores for all possible between- and within-writer sample pairs and population distributions. Twenty-four separate population distributions (six phrases × two styles × two writership sources) were calculated. Forty difficult-case scenarios were then identified using the scores, ordered from largest to smallest, from the 24 distributions and included in the FDE survey.

Two types of difficult cases were included in the survey: (1) pairs that were written by the same writer, but were associated with high dissimilarity scores, and (2) pairs that were written by different writers but were associated with low dissimilarity scores. The first set of difficult cases are characterized by an unusually large dissimilarity score compared to other dissimilarity scores from within-writer pairs. The second set of difficult cases are characterized by an unusually small dissimilarity score compared to other dissimilarity scores from between-writer pairs. Thirty between-writer pairs and ten within-writer pairs were selected from this larger pool for inclusion in the survey to increase the difficulty of the survey and challenge the examiners. Figure 2 shows examples of population density curves for between-writer and within-writer pairs, respectively, for cursive handwriting along with their corresponding FLASH ID® dissimilarity and ECDF scores (shaded area). In the examples shown in Fig. 2, the dissimilarity and ECDF scores for the between-writer sample (A) were 1.83 and 0.00008, respectively. The dissimilarity and ECDF scores for the within-writer sample (B) were 5.91 and 0.96, respectively. Sample pairs for the survey were selected to represent uncommon feature dissimilarities for their respective sources. That is, low dissimilarity scores are unusual for between-writer samples, whereas higher dissimilarity scores are unusual for within-writer samples. Forty such pairs having FLASH ID® dissimilarity scores residing near the tails of their respective distributions were used in the FDE survey.

Writership Survey and Forensic Document Examiners

We designed a writership survey consisting of difficult-case scenarios to obtain FDE strength of support for same-writer and different-writer propositions. For the purpose of this study, we considered two difficult-case scenarios: (1) when the probability of observing a small dissimilarity score between two samples of handwriting from unknown sources drawn from a relevant population is low for samples from different writers and (2) probability of observing a large dissimilarity score between two samples of handwriting from unknown sources drawn from a relevant population is low for samples from the same writer.

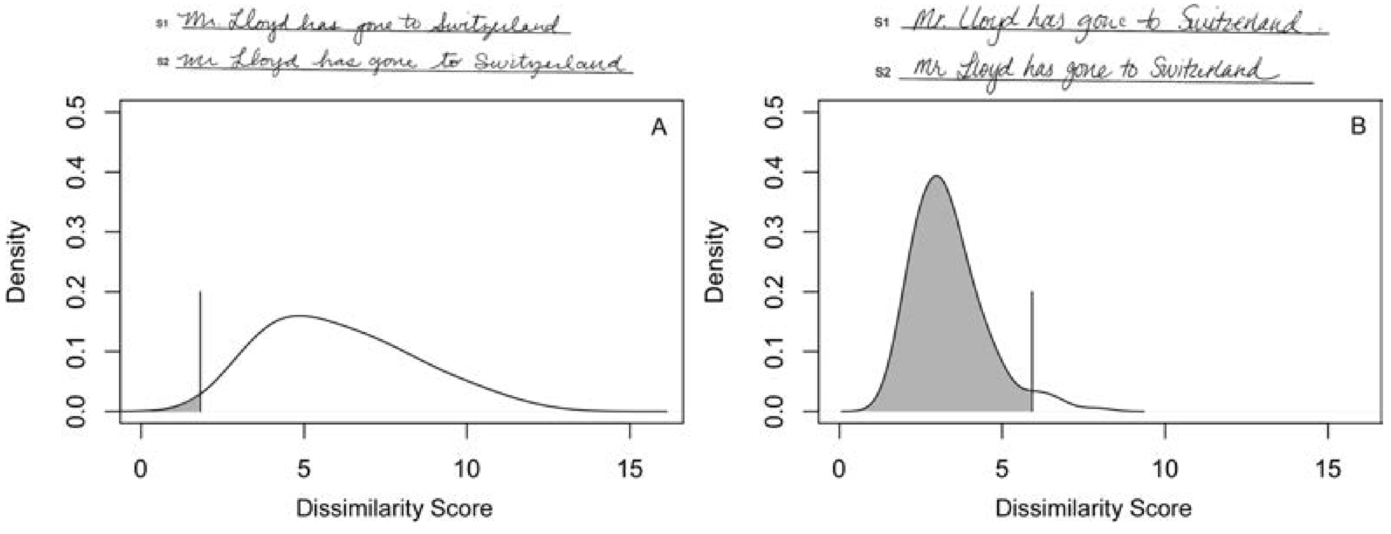


FIG. 2—*Population density distribution plots from between-writer pairs (A) and within-writer pairs (B) for the phrase "Mr. Lloyd has gone to Switzerland." Vertical lines identify the dissimilarity scores on the X-axis, while the shaded areas represent the cumulative density scores for the sample pairs displayed above each plot.*

The survey consisted of 40 sample pairs: 20 print pairs (15 between-writer and five within-writer) and 20 cursive pairs (15 between-writer and five within-writer). Pairs were presented in the survey in random order, and examiners were blinded to the writer source(s) for each pair. Five of the 40 pairs were repeated in the survey with sample order reversed for the purpose of testing examiner repeatability. Each survey item required examiners to score their strength of support for two propositions. Proposition 1 (H1) pertained to the samples being written by the same writer. Proposition 2 (H2) pertained to the samples being written by different writers. Examiners indicated their strength of support using a 7-point scale rating from extremely strong support "7" to extremely low support "1." An example of a survey pair with the scoresheet is shown in Fig. 3. For each respondent to the survey, there were 90 strength-of-support scores available for analysis (two from each of 40 sample pairs and two from each of the five repeated pairs).

Email requests were sent to 60 FDEs from North America, Europe, and Australia or New Zealand to participate in a writership survey. Of the 60, 41 FDEs submitted responses to the survey (68.3% response rate). Six were from North America, nine from Australia/New Zealand, and 26 from European countries. In addition to writership judgments, the FDEs provided de-identified information about their experience and work environment to the study. Of the 41 examiners participating in the survey, 37 (90.2%) worked in government laboratories; 33 (80.5%) reported that at least 75% of their casework involved handwriting; and 31 (75.6%) reported having been in practice for at least 10 years. This component of the research was reviewed and approved by the University of California San Diego Institutional Review Board. The average time to complete the 90-item survey was 66 min.

Five randomly selected survey items were duplicated to test FDE repeatability. Absolute difference scores between FDE strength of support for duplicated sample pairs were calculated for each of the 41 FDEs. The distribution of the average absolute differences from all 41 FDEs revealed a bimodal distribution with a cut-point located at 1.5. Based on this profile, we considered scores of 1.5 or larger to reflect inconsistent performance across repeated items of the survey. Five FDEs had scores of 1.5 or larger and were therefore excluded from further analyses. There were no differences in demographic characteristics between the five excluded examiners and the remaining 36 examiners.

Statistical Analyses

In order to explore patterns in FDE response, the survey included both cursive and print sample pairs written by the same writer and sample pairs written by different writers. For 36 FDEs, we collected two strength-of-support scores from each survey item: one registering support for the same-writer proposition and one registering support for the different-writer proposition. The scores were averaged to create a summary statistic for each examiner. We then calculated a sample mean and sample standard deviation of the summary statistics across the examiners. These steps were followed for the 20 cursive and 20 print pairings, each consisting of five within-writer and 15 between-writer pairs. Paired t-tests were conducted to test significance of a difference in examiner average strength-of-support scores arising from known same-writer versus different-writer survey items (for both print and cursive items) for each proposition. We found the distributions to be non-normal; however, due to the boundedness of the sample space of the observations, and for the sample sizes we have for this experiment, the paired t-test is robust to departures from normality.

To examine whether FDE responses were linked in any way to the dissimilarity and probability scores derived from the automated system, we used multiple linear regression models. The regression models were tested for estimating FDE strength-of-support scores for the same-writer (H1) and different-writer (H2) proposition separately for each style of handwriting. Due to the small number of within-writer samples in the survey, models were run using only the between-writer sample pairs (n = 15 for each style). Two explanatory variables were tested in each of the four models: the FLASH ID® dissimilarity score and the log of the between-writer ECDF.

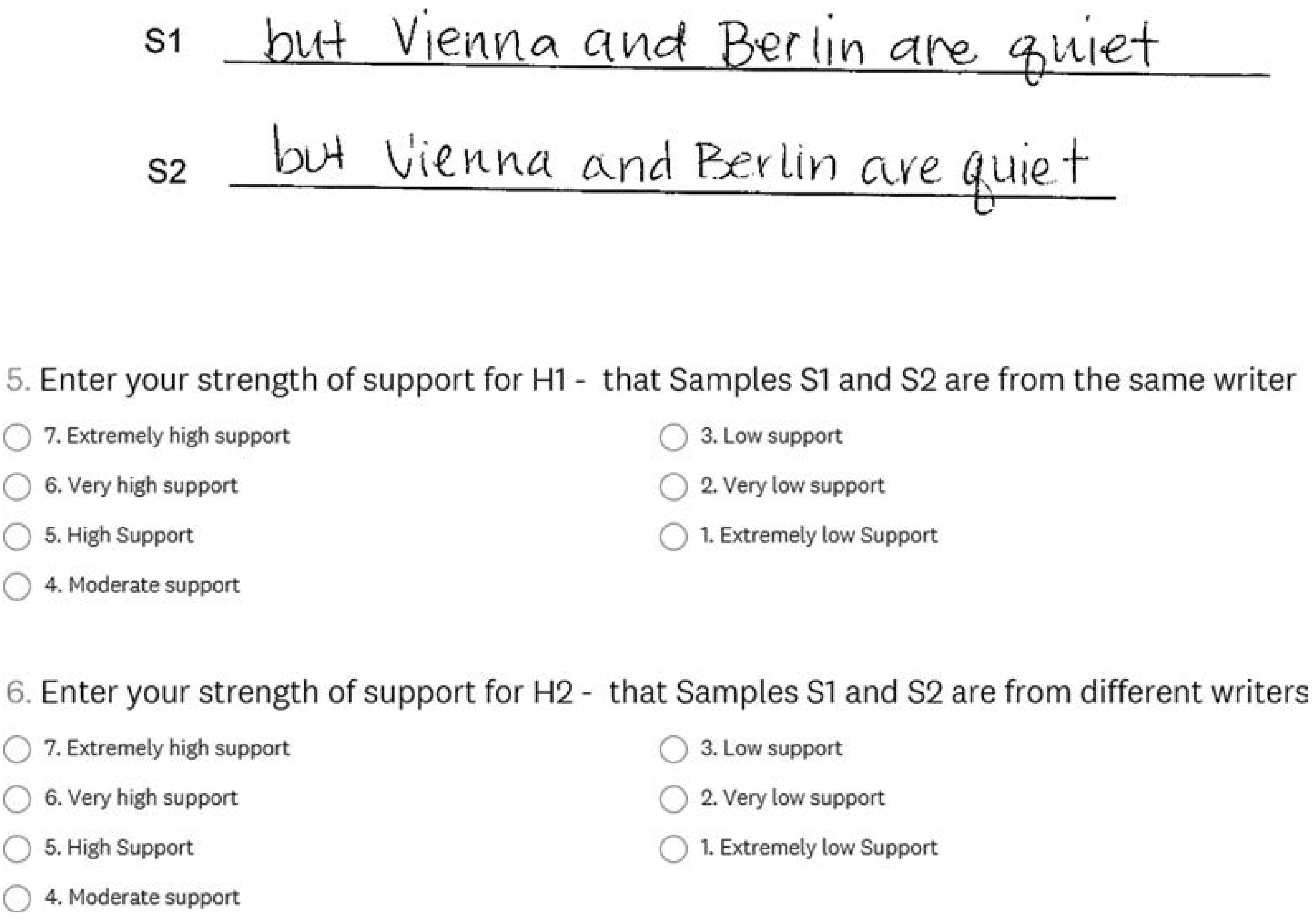
We hypothesized that examiners would register stronger support for the common source proposition when the sample pairs came from the same writer compared to different writers and register stronger support for the different source proposition when the sample pairs came from different writers compared to the same writer. We also hypothesized that stronger FDE support for a given proposition would be associated with more extreme dissimilarity scores and lower ECDFs within the distribution corresponding to the considered proposition. Support for these hypotheses would demonstrate that a feature-based automated system could be deployed to perform a two-stage evaluation of handwriting evidence.

Results

Table 1 shows the means of examiner averages (with standard deviations) for FDE strength of support for the same-writer and different-writer propositions when presented with samples written by the same writer or different writers. When asked to express strength of support for the same-writer proposition, examiners expressed significantly stronger support when the sample pair came from the same writer than from different writers. When asked to express strength of support for the different-writer proposition, examiners expressed significantly stronger support when the sample pair came from the different writers than from the same writer. These patterns held for both print and cursive handwriting. Analogous results were observed from using a nonparametric Wilcoxon signed rank-sum test on the medians.

Results from the multiple linear regression models were statistically significant only for FDE responses to the proposition that the samples came from different writers (H2) when presented with samples from different writers (i.e., FDE strength of support for ground truth) and only for printed samples ($F_{2,12} = 15.58$; $p < 0.001$; $R^2 = 0.72$). Table 2 shows the results of the regression analysis for known different-writer samples. This two-factor model yielded estimated coefficients (±SE) of 4.69 (0.86) and − 1.16 (0.24) for the dissimilarity score and log ECDF for the between-writer distribution, respectively. The predictive model indicates that larger dissimilarity scores and lower log ECDF scores predict stronger support for the proposition that two samples with unknown writer sources likely came from different writers.

Because selection of pairings included in the survey was based on dissimilarity and ECDF scores and not writer, the possibility existed that samples from one writer (albeit a different trial for a given phrase) might be used more than once when paired with samples from another writer. To reduce the effect of this sampling bias on the statistical results, we considered incorporating the two writers of each pair as random effects. To the best of our knowledge, there is no natural way to incorporate these effects within a standard random-effects model. Therefore,



S1 but Vienna and Berlin are quiet

S2 but Vienna and Berlin are quiet

5. Enter your strength of support for H1 - that Samples S1 and S2 are from the same writer

○ 7. Extremely high support

○ 6. Very high support

○ 5. High Support

○ 4. Moderate support

○ 3. Low support

○ 2. Very low support

○ 1. Extremely low Support

6. Enter your strength of support for H2 - that Samples S1 and S2 are from different writers

○ 7. Extremely high support

○ 6. Very high support

○ 5. High Support

○ 4. Moderate support

○ 3. Low support

○ 2. Very low support

○ 1. Extremely low Support

FIG. 3—*A sample scoresheet from the survey.*

TABLE 1—*Mean of examiner average responses (sd of examiner average responses) scores for 36 FDEs representing strength of support for writership determinations under the same-writer and different-writer propositions for handwriting pairs written in print and cursive style.*

	Same-writer Proposition		Different-writer Proposition	
Source	Cursive	Print	Cursive	Print
Same writer	4.25 (0.86)	3.89 (0.73)	2.91 (0.68)	3.14 (0.66)
Different writers	3.17 (0.68)	3.16 (0.64)	3.90 (0.64)	3.78 (0.64)
Difference	1.08	0.73	−1.00	−0.64
Paired *T*-statistic	9.34	7.73	−10.85	−8.06
p-value	<0.0001	<0.0001	<0.0001	<0.0001

we fit two different random-effects models: one that treats each writer as a block effect and a second that treats each pair of writers as a distinct random effect. Both models effectively gave the same results for tests concerning the fixed effects with the magnitude of the T-statistics ranging from 5.1 to 6.7.

The between-writer model for cursive sample pairs written by different writers was not statistically significant ($F_{2,12}$ = 0.36; p> 0.10; R^2 = 0.06). Models estimating FDE support for the same- or different-writer propositions when presented with samples from within-writer had only five pairs per writing style were not considered due to insufficient statistical power.

TABLE 2—*ANOVA results from the regression analysis estimating FDE strength of support for the proposition that samples within a pair were written by different writers for sample pairs written by different writers.*

	df	SS	MS	*F*	*p*
Dissimilarity score	1	0.32	0.32	8.51	0.01
Log ECDF*	1	0.85	0.85	22.66	< 0.001
Residuals	12	0.45	0.04		

*Empirical cumulative distribution function: the cumulative value from the probability density curve that reflects the rarity of a dissimilarity score for a given phrase and writing style.

Discussion

The present study had two main objectives. The first objective was to repurpose an automated handwriting feature extraction program to yield output scores that parallel the two-stage evaluative process consisting of a dissimilarity stage and discrimination stage. To accomplish this, we leveraged the powerful capabilities of FLASH ID® to develop a univariate omnibus feature dissimilarity score for comparing two questioned handwriting samples. Feature dissimilarity scores were calculated for all possible within- and between-writer pairings, producing 81,180 possible pairs each for print and cursive handwriting. This large pool of dissimilarity scores was used to generate densities and cumulative distribution functions for each phrase and style of handwriting. This enabled the reliable assignment of how rare a dissimilarity score from any handwriting pair was when compared against a population of dissimilarity scores. To our knowledge, this is the first study capable of quantifying the rarity of an observed difference in handwriting features between two samples within a population of writers for the explicit purpose of validating the significance stage within the two-stage framework.

The second objective of the study was to utilize these distribution functions to design a series of difficult-case scenarios for FDEs to evaluate. Sample pairs falling along the tails of their respective population distributions were submitted to an international cohort of 36 FDEs to demonstrate the utility of an automated feature-based process within the two-stage evaluative framework. The results from the survey supported our hypotheses that average examiner strength of support for a given proposition between pairs of samples of different writership (between or within) are consistent with ground truth for difficult-case scenarios. Further results from multiple regression analyses indicated that the feature dissimilarity score and the log ECDF for the between-writer distribution combine to account for 72% of the variability in FDE strength of support for the proposition that two questioned handwriting samples came from different writers. Specifically, these results indicate that larger dissimilarity scores and lower log ECDF scores for the between-writer distribution predict stronger support for the proposition that two samples with unknown writer sources likely came from different writers.

The findings of the present study inform the ongoing controversy over evidence interpretation within the forensic science community. One of the attributes of the two-stage approach is that it allows examiners to reject the proposition that two suspect samples arose from the same source without necessarily supporting the proposition that the two suspect samples share the same source. This allows an examiner to make statements along the line of “given the observed degree of similarities and dissimilarities, I am comfortable concluding that these two suspect samples do not share a common but unknown source.” Depending on a number of factors, this attribute can be considered either a strength or a weakness of this approach. To address this perceived limitation of the two-stage approach, the most commonly used alternative is a likelihood ratio-based approach. A likelihood ratio-based approach evaluates the subjective likelihood of the two suspect samples given the first proposition relative to the likelihood of the two suspect samples given the second proposition (21). The likelihood ratio-based approaches require a well-specified background population and evidence concerning how samples arise from that population. Conversely, if the results of the examination are determined solely within the first stage, then a strength of the two-stage approach is that there is no need for a specified background population. This means fewer evidential resources are needed to provide useful information to the decision-makers in this scenario. When failing to conclude that the two suspect samples *do not share a common source,* the two-stage approach requires the use of evidence from a specified background population to assess the evidential support for the proposition that the suspect samples do share a common source. For an overview on the comparison of likelihood-based and two-stage approaches, see (22); for a discussion as it applies to handwriting examination, see (21); and for a rigorous statistical discussion on two-stage approaches for Bayesian model selection as it pertains to specific-source propositions, see (23).

The present study contributes to the body of research on handwriting evidence interpretation in two ways. First, we successfully repurposed FLASH ID® from a feature extraction program designed to estimate the likelihood that a questioned handwriting sample was written by each writer in a list of candidate writers residing in a reference database to population distributions for a precise numeric estimation of the rarity of feature dissimilarities. In this way, the output of the automated system characterized handwriting feature dissimilarities and their distributions rather than writer identification. This is an important step toward automating the two-stage process, particularly with respect to the discrimination stage. Conventionally, an FDE might observe similarities among questioned documents and conclude that they were written by a single individual. Such an opinion would be strengthened based on the classical premise and estimating the rarity combinations of features and characteristics or their variability, leading to the conclusion that the chance of observing this combination of features and characteristics from samples in the background population is extremely low. Unfortunately, examiners reach conclusions about the relevant population based on experience alone with little or no support from actual prevalence data. With few exceptions (8,9), estimates of population variance in handwriting features or estimates of the dissimilarity in features between two writers are unavailable.

Moreover, there is growing support cautioning against reaching a conclusion based on handwriting evidence that identifies an individual writer (7). In its current form, FLASH ID® outputs a ranking of the reference set of candidate writers based on feature extraction. Our approach generates a score of the rarity of feature dissimilarities without any identification inference.

The second contribution of this study to the body of research on handwriting evidence interpretation stems from the results of the FDE survey. While it was not possible to know whether FDEs applied a two-stage or likelihood approach when reaching decisions in support for or against a particular proposition, their scores and decision patterns were consistent with what would be expected in a difficult-case scenario. For example, in the discrimination stage of the two-stage evaluative process, if there are many similarities in the writing pairs, then that pair may be considered typical of the within-writer population and one would expect the FDE to respond with strong to extremely strong support for the same-writer proposition. On the other hand, if the opposite were true, with many differences in the writing pairs, then that pair may be considered typical of the between-writer population and one would expect the FDE to respond in strong to extremely strong support for the different-writer proposition. Neither of these outcomes was evident from the survey results. However, if there are both similarities and differences in the writing pair, then that pair is somewhere in the overlap of the two populations and you would expect the FDE to respond with only moderate support of the propositions. Indeed, this was the most frequent response pattern observed from the survey results. The average strength of support for the same-writer proposition when presented with cursive sample pairs from the same writer was 4.25 (3.89 for print), while the average strength of support for the different-writer proposition when presented with cursive sample pairs from the different writers was 3.90 (3.78 for print). This is not surprising considering the survey was designed to include item pairs drawn from the tails of their respective population distributions where the magnitude of the feature dissimilarity would be uncommon considering the ground truth of that pair.

The value of an automated feature-based program is underscored further by the results from our predictive regression analyses. Two feature-based scores derived from the repurposed FLASH ID® program representing each stage of the two-stage process combined to account for 57% of the variability in FDE strength of support for the proposition that a pair of handprinted samples were produced by different writers are a blinded test of ground truth. While these findings are limited to handprinting, they support our overall hypothesis that a feature-based automated system could be deployed to perform a two-stage evaluation of handwriting evidence.

It is important to recognize that the results from the FDE survey cannot be used to evaluate proficiency. The questions posed to the examiners focused on strength of support or confidence that a specific writership proposition is correct. These scores cannot be converted to proficiency with respect to ground truth. A further limitation of this study is that it is unclear which interpretation paradigm the examiner used (e.g., likelihood-based or two-stage). We observed a number of examiners whose measures of support for the prosecution hypothesis were near perfectly negatively correlated with their measures of support for the defense hypothesis, as well as a number of examiners whose measures of support were minimally or positively correlated. This indicates that different examiners are using different methods of evidence interpretation. A natural extension of this research is to focus on alternative survey designs to shed light on this issue.

In conclusion, the present study demonstrated that with careful reprogramming and armed with a large database of features and feature differences, automated systems can be repurposed to generate population density curves for a numeric estimation of the rarity of feature dissimilarity scores from within a large reference set. In this way, the automated system would inform the discrimination stage of the two-stage framework and support FDE examination process.

Acknowledgments

The authors acknowledge Dr. Linton Mohammed, Dr. Carolyne Bird, and Jonathan Morris for their contributions in the design and piloting of the survey and in recruiting volunteers who participated in the survey and Brenda Lanners and Gina Hunter for their assistance in collecting the handwriting samples. We wish to thank the volunteers who participated providing the handwriting samples or participated in the survey.

References

1. Kirk PL, Thornton JI. Crime investigation, 2nd edn. New York, NY: John Wiley and Sons Ltd., 1974;9–17.
2. Parker J. A statistical treatment of identification problems. J Forensic Sci Soc 1966;6:33–9.
3. Parker J. The mathematical evaluation of numerical evidence. J Forensic Sci Soc 1967;7(3):134–44.
4. Parker J, Holford A. Optimum test statistics with particular reference to a forensic science problem. J R Stat Soc Ser C Appl Stat 1968;17 (3):237–51. https://doi.org/10.2307/2985461.
5. Evett IW, Berger CEH, Buckleton JS, Champod C, Jackson G. Finding the way forward for forensic science in the US – a commentary on the PCAST report. Forensic Sci Int 2017;273:16–23. https://doi.org/10.1016/j.forsciint.2017.06.018.
6. Huber RA, Headrick AM. Handwriting identification: facts and fundamentals. Boca Raton, FL: CRC Press LLC., 1999;33–59.
7. Found B, Bird C. The modular forensic handwriting method. J Forensic Doc Exam 2016;26:7–83. https://doi.org/10.31974/jfde26-7-83.
8. Kam M, Wetstein J, Conn R. Proficiency of professional document examiners in writer identification. J Forensic Sci 1994;39(1):5–14. https://doi.org/10.1520/JFS13565J.
9. Kam M, Gummadidala Fielding G, Conn R. Signature authentication by forensic document examiners. J Forensic Sci 2001;46(4):884–8.
10. Kam M, Lin E. Writer identification using hand-printed and non-hand-printed questioned documents. J Forensic Sci 2003;48:1391–5. https://doi.org/10.1520/JFS15062J.
11. Galbraith O, Galbraith CS, Galbraith NG. The principle of the 'Drunkards" search as a proxy for scientific analysis: the misuse of handwriting test data in a law journal article. Int J Forensic Doc Exam 1995;1:7–17.
12. Sita J, Found B, Rogers DK. Forensic handwriting examiners' expertise for signature comparison. J Forensic Sci 2002;47(5):1117–24. https://doi.org/10.1520/JFS15521J.
13. Bird C, Found B, Rogers D. Forensic document examiners' skill in distinguishing between natural and disguised handwriting behaviors. J Forensic Sci 2010;55(5):1291–5. https://doi.org/10.1111/j.1556-4029.2010.01456.x.
14. Johnson ME, Vastrick TW, Schuetzner E. Measuring the frequency of occurrence of handwriting and handprinting characteristics. J Forensic Sci 2017;62(1):142–63. https://doi.org/10.1111/1556-4029.13248.
15. Vastrick TW, Schuetzner E, Osborn K. Measuring the frequency occurrence of handwritten numeral characteristics. J Forensic Sci 2018;63 (4):1215–20. https://doi.org/10.1111/1556-4029.13678.
16. U.S. v Lewis, 220 F. Supp. 2d 548 (S.D. W. Va. 2002).
17. U.S. v Johnsted, 30 F. Supp. 3d 814 (W.D. Wis. 2003).
18. U.S. v Saelee, 549 U.S. 1147, 127 S. Ct. 1016 L. Ed. 2d 766 (2007).
19. Miller JJ, Patterson RB, Gantz DT, Saunders CP, Walch MA, Buscaglia J. A set of handwriting features for use in automated writer identification. J Forensic Sci 2017;62(3):722–34. https://doi.org/10.1111/1556-4029.13345
20. Core R, Team R. a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2017.
21. Expert Working Group for Human Factors in Handwriting Examination. Forensic handwriting examination and human factors: improving the practice through a systems approach. NISTIR 8282. Gaithersburg, MD: U.S. Department of Commerce, National Institute of Standards and Technology, 2020. https://doi.org/10.6028/NIST.IR.8282.
22. The Statistical and Applied Mathematical Sciences Institute. Statistics and applied mathematical science aspects of forensic science – Part 1. 2015. https://www.samsi.info/news-and-media/29-sep-christopher-saunders-samsi/ (accessed June 27, 2020).
23. Ausdemore MA, Neumann C, Saunders CP, Armstrong D, Muehlethaler C. Two-stage approach for the inference of the source of high-dimensional and complex chemical data in forensic science. J Chemom 2020:1–16. https://doi.org/10.1002/cem.3247

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14537
Available online at: onlinelibrary.wiley.com

PAPER

TOXICOLOGY

Ahmed I. Al-Asmari,[1] *Ph.D.*

Postmortem Liver and Kidney Tissue Concentrations of Heroin Biomarkers and Their Metabolites in Heroin-Related Fatalities*,†

ABSTRACT: A method was developed and validated for analyzing 6-monoacetylmorphine, morphine, 6-acetylcodeine, and codeine in routine postmortem liver and kidney specimens using liquid chromatography–tandem mass spectrometry. Samples were prepared with a Stomacher instrument followed by solid-phase extraction. All calibration curves [0.5–1000 ng/g] were linear with coefficients of determination greater than 0.99 and limits of quantification of 1.0 ng/g. Within-run precision ranged between 2.0% and 8.0%, between-run precision ranged between 1.0% and 9.0%, and accuracy ranged between −5.0% and +3.0%. Matrix effects ranged from −18% to +9%. After matrix effects were excluded, analytical recoveries ranged from 76% to 94%. The distributions of 6-monoacetylmorphine, morphine, 6-acetylcodeine, and codeine were investigated in 31 postmortem cases in which heroin was the primary cause of death. In the current study, the median free morphine ratios were calculated for liver to blood and kidney to blood, which were 2.2 and 4.0, respectively. The current report highlights the importance of testing multiple specimens, including liver and kidney, in heroin-related deaths, especially if no blood samples are available. Furthermore, this work presents new information regarding the distribution of heroin metabolites in liver and kidney.

KEYWORDS: forensic toxicology, opiates, opioids, 6-monoacetylmorphine, morphine, heroin, LC-MS-MS, postmortem, kidney, liver

Although heroin-related deaths have been well studied, opportunities remain to better understand heroin metabolites for interpreting heroin-related deaths. In forensic toxicology, it is very important to understand the distribution of heroin metabolites in bodily fluids and tissues to design a suitable method for extraction and analysis (1). In postmortem opioid (s) analysis, it is important to detect the metabolite's precursor compound or one of its biomarkers, which can clarify the source of the administered opioids. In many cases, tedious interpretation and calculation of free and total morphine and codeine are required to identify the particular opioid(s) present in the body at the time of death (2,3). In particular, clarifying the source of morphine is a challenging process because morphine can result from illegal heroin, clinically used and controlled morphine, or (in some countries, such as Saudi Arabia) over-the-counter codeine-containing medication (4,5). In heroin-related fatalities, heroin is unstable in biological specimens with a half-life of only 2–6 min; instead, the identification of heroin biomarkers is adequate for proving that heroin was the opioid administered and clarify the timing of death relative to administration. Within minutes of intake, heroin is converted to 6-monoacetylmorphine (6-MAM), which has a short half-life of only 6–25 min. 6-MAM is converted to morphine, and morphine's half-life is only 2–3 h (6,7). 6-acetylcodeine (6-AC) is not a heroin metabolite; it is formed due to manufacturing impurities during heroin synthesis. 6-AC is considered to be a heroin biomarker, indicating that street heroin was administered, as pure pharmaceutical heroin and medications that contain codeine and morphine are free from 6-AC. The presence of 6-AC in a specimen gives clear evidence that street heroin was used (8–10).

Blood is the most common specimen tested in forensic toxicology, especially in postmortem examination. However, biological fluids such as blood, urine, and vitreous humor may not be available in cases with decomposition, burns, destruction to the body, or blood transfusion occurs. In such cases, solid tissues may be analyzed to interpret the cause of death and the source of opioids used (11,12). However, very limited research exists on the drug distribution in solid tissues, making toxicological interpretation very difficult (12–18).

Most studies of postmortem heroin metabolites have used bodily fluids, especially blood, urine, and bile, with only few studies having examined solid tissues. Felby et al. (19) developed a method using gas chromatography coupled with a flame ionization detector (GC-FID) for analyzing morphine in liver tissue. Using nalorphine as an internal standard, morphine recovery

[1] Laboratory Department, King Abdul-AzizHospital-Jeddah Health Affair, Ministry of Health, P.O. BOX 4670, Jeddah, Makkah AL-Mukharmah, 21442, Saudi Arabia.

Corresponding author: Ahmed Al-Asmari, Ph.D. E-mail: Ahmadalasmari@yahoo.com

*This work was supported by Poison Control and Forensic Medical Chemistry Center, Ministry of Health, Jeddah Health Affair, Jeddah, Kingdom of Saudi Arabia.

†Presented in part at the 57th International Association of Forensic Toxicologists Meeting, September 2-6, 2019, in Birmingham, U.K.

Received 23 May 2020; and in revised form 25 June 2020, 15 July 2020; accepted 16 July 2020.

ranged from 60% to 70%. In 1987, Spiehler and Brown (20) developed a method for analyzing morphine in liver specimens. In that method, acid hydrolysis was performed using concentrated hydrochloric acid followed by multiple steps of liquid–liquid extraction, and the final extract was derivatized using trifluoroacetic anhydride before injection into the gas chromatography-mass sepectrometry (GC-MS) instrument. They used deuterated morphine and codeine as internal standards, and four quality control standards ranging in concentration from 5.0 to 100.0 ng/mL were used to investigate extraction recoveries, which ranged from 115% to 140% indicating matrix interference. However, an acceptable intra-assay and inter-assay precisions were obtained with coefficients of variation (%CVs) less than 15%. In 1988, Steentoft et al. (21) reported a liquid–liquid extraction and gas chromatography method for analyzing morphine in liver specimens obtained from heroin-related fatalities, and this method was a modification of that reported by Felby et al. (19), but validation parameters were not detailed. Moriya and Hashimoto (16) studied the distribution of heroin metabolites in different bodily fluids and tissues, including liver and kidney specimens. They used acid hydrolysis to analyze total morphine. In this method, levallorphan was used as an internal standard for their liquid–liquid extraction and GC-MS. Unfortunately, no validation parameters were detailed; however, the authors referred to a previous method reported by Spiehler and Brown (20). 6-MAM, free morphine, and free codeine were measured in liver specimens in a previous report by Wyman and Bultman (22). Liver specimens were extracted using mixed-mode solid-phase extraction cartridges and GC-MS; no hydrolysis procedure was used. Limits of quantification (LOQ) ranged from 1.0 to 2.0 ng/g, and assay precision % CV was better than 7.5%. Marglho et al. (5) described a method for analyzing free morphine in liver specimens. Their limits of detection (LODs) were 4.0 ng/g for morphine, 5.0 ng/g for codeine, and 4.0 ng/g for 6-MAM. In addition, their extraction recovery was better than 70.0%. Other reported methods have measured total morphine after converting glucuronide conjugates of morphine to their free form with enzymatic hydrolysis (15,23) or have used direct determination of morphine and its glucuronide conjugates without hydrolysis using LC-MS-MS (24). In fact, few studies have reported the level of heroin metabolites in solid tissue specimens. Most previous studies focused on detecting morphine, with little information provided regarding heroin metabolite distribution in solid tissues (5,15,16). Heroin metabolites in kidney tissue have not yet been reported using LC-MS-MS. Most of the procedures detailed above used GC. Most of the important metabolites are found in their conjugated form, which requires hydrolysis and derivatization procedures to prepare samples for GC-MS, which is time consuming, tedious, and dangerous (5,15,16,19,20,22,23,25).

A method for analyzing morphine and their glucuronides in liver and kidney tissues has been reported using high-performance liquid chromatography coupled with a photodiode array detector (26). In that study, the LOD ranged 0.7–4.5 ng/g, and the LOQ ranged 2.0–4.4 ng/g. Accuracy and precision were acceptable, with %CVs less than 15%, and the extraction recoveries for free morphine were 85.4% for liver and 89% for kidney. However, the method was not applied to real liver specimens collected from postmortem cases. To the author's knowledge, only one method has been reported for analyzing postmortem liver tissues with LC-MS-MS (24). In that method, liver concentrations were determined using calibration curves derived from plasma extraction. The LOD was 4.0 ng/mL, and the LOQ was 10.0 ng/mL. Two quality controls consisting of spiked animal specimens were used to determine method accuracy and bias, which were within the acceptable range (±15%). Solid-phase extraction with C18 SPE columns was used in both studies (24,26). To retain morphine's polar metabolites, suppression of matrix effects and recovery were calculated in liver samples and found to be 2% and 38%, respectively. In one of these studies (24), non-human tissues were used for validation, and only morphine, morphine-3-glucuronide, and morphine-6-glucronide were analyzed. Neither study examined other heroin related-metabolites, that is, 6-MAM, 6-AC, or codeine.

The current, published literature has only limited methods for solid tissue analysis with complete method validation, such as selectivity, linearity, limit of detection, limit of quantification, accuracy, precision, recovery, and matrix effects. The aim of this work was to develop and validate extraction and LC-MS-MS methods for analyzing heroin biomarkers and their metabolites (morphine and codeine) in postmortem liver and kidney tissues and apply this validated method to routine postmortem forensic toxicology analyses to provide novel insights into the distributions of target compounds in postmortem specimens. This method has been previously validated in biological fluids (1) and is now being adapted to solid tissues.

Methods and Materials

Reagents and Standards

Methanol (HPLC grade), acetonitrile (HPLC grade), ammonium carbonate, formic acid, and ammonium hydroxide were obtained from BDH (Poole, UK). Ammonium formate was obtained from Sigma Aldrich (USA).

Morphine, morphine-D3, 6-MAM, 6-monoacetylmorphine-D3 (6-MAM-D3), codeine, codeine-D3, and 6-AC were purchased from Lipomed (Arlesheim, Switzerland). All standards and internal standards were obtained as solutions in methanol at concentrations of 0.1 or 1.0 mg/mL, and each had a purity of more than 99%. Clean Screen® (CSDAU203) cartridges were purchased from United Technology Company (Bristol, USA). A Stomacher and Stomacher 80 bags with 101 × 152 mm filters were obtained from Seward limited (UK).

Solid-Phase Extraction and Chromatography Conditions

The previously reported extraction procedure for heroin biomarkers and their metabolites from postmortem bodily fluids was modified for solid tissue samples (1). One gram of each tissue sample was diluted 2:1 w:w (aqueous 1% sodium fluoride: tissue) and then placed into a Stomacher bag for 5 min of homogenization in the Stomacher. One-half gram of tissue homogenate was subsequently transferred to a 15 mL glass tube. Then, 50 ng/g of internal standard solution (6-MAM-D3, morphine-D3, and codeine-D3) was added to each tube and vortexed. Two mL of phosphate buffer (0.1 M, pH 6) was then added to each tube. A Clean Screen® (CSDAU203) cartridge was preconditioned with 2 mL of methanol, 2 mL of deionized water and 2 mL of phosphate buffer (0.1 M, pH 6), and then, the samples were loaded onto the column. Each column was subject to two washing steps using 1 mL of deionized water following by 1 mL of acetic acid (0.1 M), and a full vacuum was applied for 5 min to dry the column. Each column was subjected to two washes with

1 mL of hexane. Fraction A was eluted using 2 mL of hexane/ethyl acetate (1:1, V/V); after that, the elution tube was removed. Then, each column was washed with 3 mL of methanol and dried for 2 min with a full vacuum. Fraction B was then eluted using 3 mL of dichloromethane/isopropanol/ammonium hydroxide (78:20:2, V/V/V). Fractions A and B were combined and then evaporated to dryness under a nitrogen stream and reconstituted in 200 μL of the initial mobile phase. Of this, 1 μL was injected into the LC-MS-MS instrument.

Instrumentation

Opioid analysis was carried out using a Shimadzu LCMS-8050 triple quadrupole mass spectrometer (Kyoto, Japan) equipped with a Shimadzu Nexera UHPLC system. During analysis, the auto-sampler was maintained at 4°C and the column oven at 40°C. The LC column consisted of a Raptor Biphenyl column (50 × 3.0 mm, 2.7 μm) coupled with a SecurityGuard Raptor Biphenyl column (5.0 × 3.0 mm, 2.7 μm, Restek, USA). Separation was obtained using a gradient elution based on a mobile phase consisting of 10 mM ammonium formate adjusted to pH 3 (A) and methanol (B) at a flow rate of 0.3 mL/min. The elution started with 97% of solution A for 1 min, which decreased to 95% at 2 min, then 5% at 15 min, where it was maintained for the next 1 min before returning to 97% at 16 min. It was maintained at 97% for the next 4 min prior to the next injection. The total run time was 20 min.

Analytes and their internal standards were identified and quantified based on their retention times and the presence of two product ions or one product ion in multiple reaction monitoring (MRM) mode as follows: morphine m/z = 285.65–165 and 285.65–153, morphine-D3 m/z = 288.6–165, codeine m/z = 299.75–165 and 299.75–44, codeine-D3 m/z = 302.9–165, 6-MAM m/z = 327.5–165 and 327.5–221, 6-MAM-D3 m/z = 331–165, and 6-AC m/z = 341.9–164.9 and 341.9-225.2. Labsolution (version 5.75) was used for data processing.

Method Validation

Scientific Working Group for Forensic Toxicology (SWGTOX) method validation guidelines were followed (27). Accordingly, the following method parameters were measured: linearity, sensitivity, matrix effects, recovery, accuracy, precision, selectivity, dilution integrity, and carryover. The validation protocols for these parameters are detailed in the publication of this method in bodily fluids (1). The current study was similarly validated for liver and kidney tissues.

Case Samples

To confirm the applicability of the presented method, liver and kidney specimens from investigations into drug-related fatalities were analyzed (ethical approval no. A00188). During the study period of April 19th, 2014, until July 31st, 2018, more than 1000 routine postmortem whole blood samples were investigated. Of these, 62 cases (6.2%) were determined to have involved heroin. In many of these cases, blood, vitreous humor, urine, stomach contents, or bile tested positive for 6-MAM and/or 6-AC, confirming that heroin had been used before death. The focus of this report is to relate the importance of analyzing complementary specimens in heroin-related fatalities. Among the 62 heroin-related deaths, liver and/or kidney tissues were available in 31 cases (50%), and these were included in this study. Full toxicology testing was conducted, which included immunoassays, alcohol testing, carbon monoxide, general unknown screening, and confirmation using GC-FID, GC-MS, and LC-MS-MS.

Results

Method Validation

Linearity was determined for each analyte of interest, and linear regressions were calculated for each analyte of interest in liver and kidney specimens over the calibration range (0.5–1000.0 ng/g). Correlations of determination were greater than 0.999 for all analytes of interest. The LODs were determined theoretically and ranged from 0.21 to 0.38 ng/g. The LOQs were 1.0 ng/g for all analytes of interest. At the LOQ, accuracies ranged from 4.0% to 7.0% in liver specimens and 1.0%–9.0% in kidney specimens (median: 4.5%). In liver specimens with the LOQ concentration, within-run precision ranged 8.0%–13.0%, and between-run precision ranged 5%–12.0%. In kidney specimens with the LOQ concentration, within-run precision ranged 6.0%–10.0%, and between-run precision ranged 4.0%–12.0%. LOD, LOQ, and ULOQ values are detailed in Table 1.

Method precision was investigated by measuring within-run and between-run precisions using three concentrations at low, medium, and high concentrations (5 ng/g, 100.0 and 800.0 ng/g) within the linear dynamic range (LDR). Acceptable precisions were obtained for all analytes of interest with %CV values ranging in liver from 2.0% to 8.0% for within-run precision and 1.0% to 9.0% for between-run precision. In kidney, %CVs ranged from 2.0% to 6.0% for within-run precision and 2.0%–8.0% for between-run precision. Accuracy was investigated using three concentrations within the LDR: 5.0, 100.0, and 800.0 ng/g and ranged in liver from −5.0 to +3.0% and in kidney from −4.0 to +3.0%. Precision and accuracy results are detailed in Table 2 and were within the acceptable ranges provided by SWGTOX guidelines for method validation.

Six different sources of human postmortem specimens were used to investigate matrix effects and recovery for each specimen type. Matrix effects and recoveries were determined using three different concentrations (5.0, 100.0, and 800.0 ng/g). Ion suppression and ion enhancement were detected, and the matrix effects ranged in liver specimens from less than −18 to +7.0% and in kidney specimens from less than −17.0 to +9.0%, which were within the acceptable range for method validation guidelines (±25%) (27). After excluding matrix effects, recoveries for analytes of interest in liver ranged from 76.0% to 94.0%, and recoveries in kidney ranged from 78.0% to 93.0%. Matrix effects and recoveries were assessed by %CVs. In liver, the matrix effects' %CV ranged 2.0–10.0%, and the recoveries' %CV ranged 1.0%–14.0%. In kidney, the matrix effects' %CV ranged 3.0%–15.0%, and the recoveries' %CV ranged 4.0%–8.0%.

When analyte concentrations outside the calibration range were obtained, analyses were repeated after sample dilution. To evaluate dilution integrity, five replicates of two different concentrations were analyzed: 10 ng/g (QC low dilution, 1:10 dilution of 100 ng/g) and 75 ng/g (QC high dilution, 1:100 dilution of 7500 ng/g). Dilution integrity was found to be robust and reproducible, being within the acceptable range for method validation (± 15%). In liver, measured concentrations of the 10.0 ng/g dilution yielded a STDV of ±0.2 ng/g, and the 75.0 ng/g dilution yielded a STDV of ±2.6 ng/g. Accuracies for both concentrations ranged from −4.0% to +4.0%. Within-run and between-run precisions for dilution integrity experiments

TABLE 1—*Validation results including linear coefficient determination, limit of detection, limit of quantification, and upper limit of quantification (*n = *5).*

Specimens	Analytes	R^2	LDR (ng/g)	Intercept	STDEV (intercept)	Slope	LOD	LOQ	ULOQ	Unit
Liver	6-Monoacetylmorphine	0.999	0.5–1000.0	0.416449	0.003206	0.044316	0.24	1.0	1000.0	ng/g
	Morphine	0.999	0.5–1000.0	−0.00623	0.000386	0.005176	0.25	1.0	1000.0	ng/g
	6-Acetylcodine	0.999	0.5–1000.0	−0.04405	0.001675	0.014387	0.38	1.0	1000.0	ng/g
	Codeine	0.999	0.5–1000.0	0.085888	0.001006	0.013706	0.24	1.0	1000.0	ng/g
Kidney	6-Monoacetylmorphine	0.999	0.5–1000.0	0.033691	0.00352	0.045319	0.26	1.0	1000.0	ng/g
	Morphine	0.999	0.5–1000.0	−0.0062	0.000448	0.00523	0.28	1.0	1000.0	ng/g
	6-Acetylcodine	0.999	0.5–1000.0	0.047963	0.00144	0.014312	0.33	1.0	1000.0	ng/g
	Codeine	0.999	0.5–1000.0	0.085828	0.000882	0.013706	0.21	1.0	1000.0	ng/g

LDR, linear dynamic range; LOD, limit of detection; LOQ, limit of quantification; R^2, coefficient of determination; STDEV, standard deviation; ULOQ, upper limit of quantification.

ranged 2.0%–10.0% and 3.0%–9.0%, respectively. In kidney, measured concentrations of the 10.0 ng/g dilution produced a STDV of ±0.1 ng/g, and the 75.0 ng/g dilution produced a STDV of ±4.0 ng/g. Accuracies of both concentrations ranged from −8.0% to +3.0%. Within-run and between-run precisions for dilution integrity experiments ranged 3.0%–8.0% and 2.0%–9.0%, respectively. The effects of dilution on bias and precision values are detailed in Table 2. Similar to what has been reported in previous work (1), no interference from commonly encountered compounds, blank postmortem specimens, or carryover effects from the last injection were observed in this study.

Case Studies

The study involved 31 postmortem cases, and all but two were male. The mean age of the deceased was 30 years (median: 33, range: 18–70 years old). In all of these cases, heroin was the primary cause of death. History of the deceased, manner of death, age, sex, and time interval between death and sample collection (PMI) are provided in Table S1. The concentrations of the analytes of interest are listed in Table S2. The time interval between death and sample collection differed between cases and ranged from 1 to 28 days with a median of 2 days. The specimens collected at autopsy included blood with sodium fluoride preservative (BN) in 25 cases (80.6%), blood without preservative (B) in 22 cases (71.0%), liver in 30 cases (96.8%), and kidney in 23 cases (74.2%).

The median free morphine concentration in BN was 98.0 ng/mL (range: 4.0–1222.0 ng/mL) and in B was 151 ng/mL (range: 16.0–1927.0 ng/mL). 6-MAM was detected in 15 of the BN samples (48.4%) and 8 of the B samples (25.8%). The median 6-MAM concentration in BN was 5.0 ng/mL (range: 1.0–55.0 ng/mL) and in B was 9.0 ng/mL (range: 1.0–17.0 ng/mL). 6-AC was detected in BN in 5 cases (16.1%) and in B in one case (3.2%, Case 6, 2.0 ng/mL). The median 6-AC concentration in BN was 2.0 ng/mL (range: 2.0–12.0 ng/mL). The median codeine concentration in BN was 13.0 ng/mL (range: 1.0–76.0 ng/mL) and in B was 20.0 ng/mL (range: 3.0–94.0 ng/mL). The median B to BN percentage of detected morphine concentrations was 133% (range: 66.8%–400%). The median free morphine to free codeine ratio was 10.6 in B (range: 3.3–30.6) and 9.4 in BN (range: 1.2–37.0). The ratio of free morphine to free codeine was higher than 1 in all B and BN samples.

Liver specimens were available for 30 cases, with 6-MAM detected in four of these (cases 4, 17, 18, and 31) at concentrations ranging from 1.2 to 71.0 ng/g. The median morphine concentration was 230.0 ng/g (range: 76.0–2095.0 ng/g). Codeine was detected in 27 liver specimens (90.0%) with a median concentration of 13.0 ng/g (range: 1.3–50.0 ng/g). 6-AC was not detected in any of the samples. The median free morphine to free codeine ratio in liver specimens was 20.0 (range: 5.9–511.5). The median free morphine in liver was compared to that in BN and found to be 2.2-fold higher (range: 0.2–38.8-fold).

Kidney specimens were available for 23 cases (74.2%), with only two testing positive for 6-MAM (Case 7, 2.0 ng/g and Case 18, 58.0 ng/g). The median morphine concentration was 338.0 ng/g (range: 40.0–3485.0 ng/g). Codeine was detected in 21 kidney specimens (91.3%), with a median concentration of 24.0 ng/g (range: 2.0–95.0 ng/g). 6-AC was not detected in any of the samples. The median free morphine to free codeine ratio in kidney specimens was 14.6 (range: 5.4–101.6). The median free morphine in kidney was compared to that in BN and found to be 4.0-fold higher (range: 0.3–24.7-fold).

In the current study, heroin-related fatalities were classified as solely involving heroin (no other drugs were detected) or heroin-related (at least another drug was detected). Using that criteria, thirteen cases (42.0%) solely involved heroin (cases 1, 2, 10, 11, 14, 16, 17, 18, 22, 24, 27, 28, and 30). Heroin was found in combination with another drug in 18 fatalities (cases 3, 4, 5, 6, 7, 8, 9, 12, 13, 15, 19, 20, 21, 23, 25, 26, 29, and 31). These cases also tested positive for cocaine (3 cases), amphetamine (9 cases), methamphetamine (2 cases), tramadol (3 cases), pregabalin (2 cases), alprazolam (6 cases), diazepam and its metabolites (one case), and alcohol (one case). The concentrations of analytes of interest and other detected drugs are listed in Table S1.

The time of death following heroin intake can be divided into rapid or delayed. Rapid deaths can be identified by the detection of 6-MAM in blood samples, indicating that death occurred within 3 h of intake. In contrast, negative blood findings for 6-MAM have been suggested to indicate a longer time between death and heroin intoxication (delayed death). This conclusion benefits from detecting 6-MAM in other specimens, for example, blood, vitreous humor, or urine. According to classification suggested in previous reports (20,28), in the current study, sixteen cases were rapid deaths (cases 1, 2, 5, 6, 7, 9, 11, 13, 14, 17, 18, 21, 23, 29, 30, and 31) with death occurring between 40 min and 3 h after heroin administration, while fifteen cases were identified to be delayed deaths (cases 3, 4, 8, 10, 12, 15, 16, 19, 20, 22, 24, 25, 26, 27, and 28), with death occurring more than 3 h after heroin administration. In this study, putrefaction was evident in 15 of the 31 cases (51.6%), which were categorized as either heavy or some putrefaction. Only four cases were heavily putrefied (13.0%), and 10 cases were mildly putrefied (32.3%). The remaining 17 cases showed no signs of putrefaction. Differences in the concentrations of the analytes of interest based on type of death, time until sampling, and the degree of putrefactions are detailed in Table S3.

TABLE 2—*Accuracy, within-run precision, and between-run precision results* (n = 5).

		Precision			Dilution	
Analytes	LOQ	Low	Medium	High	D1ψ	D2ψ
Specimens						
Target concentration						
Nominal concentration	1.0	5.0	100.0	800.0	10.0	75.0
		Measured Concentration ng/g				
Liver						
6-Monoacetylmorphine						
Mean (ng/g)	1.05	5.1	99.4	804.1	9.9	76.4
STDEV±	0.14	0.4	8.5	22.0	0.3	5.1
% E	5.0	2.0	-1.0	1.0	-1.0	2.0
Within Run %CV	12.0	8.0	3.0	3.0	4.0	5.0
Between Run %CV	12.0	7.0	9.0	3.0	3.0	3.0
Morphine						
Mean (ng/g)	1.04	5.2	102.9	805.1	10.1	72.0
STDEV±	0.11	0.3	6.8	28.7	0.4	3.9
% E	4.0	3.0	3.0	1.0	1.0	-4.0
Within Run %CV	11.0	5.0	5.0	2.0	2.0	5.0
Between Run %CV	8.0	6.0	7.0	4.0	4.0	6.0
6-Acetylcodine						
Mean (ng/g)	1.07	5.0	94.6	804.1	9.7	74.1
STDEV±	0.14	0.3	5.1	19.5	0.5	6.9
% E	7.0	1.0	−5.0	1.0	−3.0	−1.0
Within Run %CV	13.0	4.0	6.0	2.0	5.0	10.0
Between Run %CV	10.0	3.0	5.0	2.0	6.0	6.0
Codeine						
Mean (ng/g)	1.06	4.8	102.3	795.7	9.8	77.9
STDEV±	0.1	0.4	7.6	16.9	0.4	6.8
%E	6.0	-3.0	2.0	−1.0	−2.0	4.0
Within Run %CV	8.0	6.0	3.0	2.0	3.0	8.0
Between Run %CV	5.0	8.0	5.0	1.0	4.0	9.0
Kidney						
6-Monoacetylmorphine						
Mean (ng/g)	1.03	5.0	97.1	792.3	9.8	77.1
STDEV±	0.05	0.2	5.4	11.0	0.4	6.2
% E	3.0	-4.0	-3.0	−1.0	−2.0	3.0
Within Run %CV	6.0	4.0	4.0	2.0	3.0	6.0
Between Run %CV	4.0	3.0	2.0	3.0	5.0	7.0
Morphine						
Mean (ng/g)	1.0	5.2	98.2	808.4	9.9	77.5
STDEV±	0.1	0.3	5.9	18.5	0.7	6.9
% E	1.0	3.0	2.0	1.0	−1.0	3.0
Within Run %CV	7.0	5.0	3.0	2.0	6.0	8.0
Between Run %CV	11.0	5.0	2.0	2.0	7.0	9.0
6-Acetylcodine						
Mean (ng/g)	1.1	5.0	102.3	809.8	9.8	69.1
STDEV±	0.1	0.3	8.3	28.7	0.4	4.4
% E	6.0	1.0	2.0	1.0	−2.0	−8.0
Within Run %CV	10.0	4.0	4.0	4.0	4.0	5.0
Between Run %CV	12.0	7.0	4.0	3.0	4.0	2.0
Codeine						
Mean (ng/g)	1.1	4.8	101.1	793.8	9.9	73.9
STDEV±	0.1	0.4	6.7	15.6	0.6	5.2
% E	9.0	−3.0	1.0	−1.0	−1.0	−1.0
Within Run %CV	9.0	6.0	4.0	2.0	4.0	7.0
Between Run %CV	6.0	8.0	2.0	2.0	3.0	3.0

%E, relative error; %CV, the coefficient of variation; D1ψ, dilution factor 1, D2 ψ, dilution factor 2; LOQ, limit of quantification.

Discussion

Liver and kidney tissue are the most common samples collected at autopsy and are valuable for testing heroin, especially in cases of putrefaction or when blood samples are not available (14,28). Some studies have shown that free morphine is stable in liver tissue, even when putrefied for two years (20,28). Kidney samples, on the other hand, are appropriate for analyzing free morphine but not morphine conjugates because the conjugates are quickly excreted into urine. While free morphine is stable in liver and kidney samples at temperatures from 4.0 to 37.0°C, the morphine conjugates turn into free morphine at temperatures from 18 to 37 in less than 10 days (16). Kidney has a higher affinity for free morphine compared to conjugated morphine. In contrast, the majority of the morphine found in liver is conjugated. The time span between deaths and analysis can cause some of the conjugated morphine in the liver to be degraded to its free form. This was supported by the current work's finding of free morphine in the liver being 4-fold greater than that in the blood, which can be explained by the postmortem redistribution and/or morphine conjugated degradation to free morphine in liver specimens (24). Therefore, the free morphine is more logically detected in cases that have decomposed.

Little information regarding the concentration of morphine in postmortem liver and kidney has been reported. One of the earliest studies of the concentration of morphine in solid tissues was reported by Felby et al. in 1974 (19). In the 14 heroin deaths included in their study, the median morphine concentration was 400.0 ng/mL in blood (range: 200.0–2800.0 ng/mL) and 1050.0 ng/g in liver (range: 400.0–18,000.0 ng/g). Since then, multiple studies have reported morphine concentrations in liver and blood simultaneously with a wide range of concentrations reported, and morphine concentration liver to blood ratios ranging from 0.4 to 2.4 (16,22,29,30). Previous studies reported median total morphine concentrations in blood of 110.0 ng/mL (range: 10.0–570.0 ng/mL) (22), 530 ng/mL (range: 20.0–3540.0 ng/mL) (23), and 600 ng/mL (range: 70.0–1600.0 ng/mL) (15). Liver has higher values with medians between 320.0 and 1270.0 ng/g, minimums of 40 ng/g, and maximums of 5728 ng/g (15,23) To provide greater insight, Spiehler and Brown (20) subdivided their cases into rapid (56 cases) or delayed deaths (145 cases). In rapid deaths, the median morphine concentrations were higher than in delayed deaths, and in contrast, in delayed deaths, morphine concentrations in liver were higher than in rapid deaths. In both types of death, liver showed higher concentrations than blood. In blood, rapid deaths had a median of 360.0 ng/mL (range: 80.0–1650.0 ng/mL), and delayed deaths had a median of 125.0 ng/mL (range: 3.0–2200.0 ng/mL). In liver, rapid deaths had a median of 1280.0 ng/g (range: 410.0–5500.0 ng/g), and delayed deaths had a median of 1000.0 ng/g (range: 40.0–6000.0 ng/g). Several groups have specifically reported free morphine concentrations, which have ranged in blood from 10.0 to 1350.0 ng/mL, and in liver from 10.0 to 2550.0 ng/g (15,16,31). All of these studies reported free morphine concentrations within the range measured by the current study.

In Spiehler and Brown's study of heroin- and/ or morphine-related fatalities (20), the liver to blood ratio of morphine concentrations were 2.1 in rapid deaths and 1.6 in delayed deaths. Subsequent studies have reported this ratio to be between 1.5 and 2.76 (19,21,22). In the current study, free morphine concentrations were 2.2-fold higher in liver than in blood, indicating that liver is an excellent tissue for analysis when no blood is available.

Wyman and Bultman (22) studied the distribution of 6-MAM and codeine in blood and liver in heroin-related death. They did not detect 6-MAM in any of the liver tissues analyzed in their study. In all of Wyman and Bultman's cases, the ratio of codeine to morphine was less than one in liver specimens, with a mean ratio of 0.06 (range: 0.03–0.14) and a median codeine concentration of 10.0 ng/mL in blood (range: 10.0–40.4 ng/mL) and 15.0 ng/g in liver (range: 10.0–90.0 ng/g). As such in the current work, the median liver to blood ratio of the free codeine concentration was 2.0 (range: 1.0–9.0), which is consistent with Wyman and Bultman study (22).

6-MAM is rarely detected in liver or kidney specimens. 6-MAM has been detected in kidney specimen in two previous reports at concentrations of 2.0 and 1790.0 ng/g (16,29). In the current study, 6-MAM was detected in four liver specimens and two kidney specimens. Published kidney to blood ratios of morphine concentrations range from 1.3 to 3.7 (16,29). In the current study, free morphine concentrations were 4.0-fold higher in kidney than in blood, making kidney an ideal tissue for analysis when no blood is available.

In heroin- and/or morphine-related fatalities, morphine concentrations in postmortem kidney and liver are rarely reported together. To the author's knowledge, only three such case reports are available in the literature (16,29,30). Reed et al. (30) reported two cases of heroin overdose; they found that morphine concentrations in the kidney (Case 1: 1510.0 ng/g and Case 2: 700.0 ng/g) were almost double those in the liver (Case 1: 660.0 ng/g and Case 2: 350.0 ng/g). In Moriya et al., (16) a high ratio between free and total morphine was obtained in liver tissue, with only a slight little difference in kidney tissue. Total morphine was three times greater than free morphine in the liver (free morphine: 1440.0 ng/g, total morphine: 4200.0 ng/g), and free morphine constituted 94% of the total morphine in kidney tissues (free morphine: 1790.0 ng/g, total morphine: 1900.0 ng/g). In addition, Goldberger et al. (29) presented two heroin-related fatalities that had both liver and kidney available for analysis. In that study, free morphine concentrations of 216.0 ng/g for Subject V and 91.0 ng/g for Subject W were reported. Total morphine concentrations were much higher, at 482.0 ng/g for Subject V and 359.0 ng/g for Subject W. Their findings are in agreement with the current study's free morphine concentration range of 40–3485 ng/g in kidney and 76.0–2095 ng/g in liver. It has been noticed that these values appear to be dependent on the time of death relative to when heroin was administered. In rapid deaths, the median morphine concentration in the kidney was almost double of that in the liver (1.9-fold). Delayed deaths, in contrast, had similar median free morphine concentrations in kidney as in the liver. This is consistent with data from one case study (14) in which a patient was given an overdose of morphine (150.0 mg) and died three days later at the hospital. The free morphine concentration in their liver was 157.0 ng/g, and in their kidney was 87 ng/g. The total morphine concentration in their kidney was 535.0 ng/g and in their liver was 479 ng/g. As such, the liver to kidney ratio of free morphine was 0.6, and the liver to kidney ratio of total morphine was 0.9.

Few studies discuss the differences of the concentrations of heroin metabolites in solid tissues between solely heroin-related deaths versus those in which another drug was present, or between rapid versus delayed deaths. Steentoft et al. reported blood and liver morphine concentrations in 12 heroin-related deaths with a short interval between injection and death (21). The median concentration of morphine in the blood was 328.0 ng/mL (range: 114.0–2284.4 ng/mL) and in the liver was 485.0 ng/g (range: 114.0–1169.0 ng/g); the ratio between liver and blood was 1.1 (range: 0.5–4.7). Morphine was the sole drug detected in 98 of their 245 cases (40%), with median concentrations in the blood of 200.0 ng/mL (range: 29.0–5420.0 ng/mL) and in the liver of 342.0 ng/g (range: 86.0–4850.0 ng/g). This produced a ratio of liver to blood concentrations of 0.7. The remaining 147 cases involved at least one other drug: ethanol alone (32%), other drugs with no ethanol (19%), or ethanol and other drugs (9%). When ethanol was the only co-drug, the median concentration of morphine in blood was 114.0 ng/mL (range: 29.0–2853.0 ng/mL) and in liver was 314.0 ng/g (range: 57.0–3994.0 ng/g), yielding a liver to blood ratio of 2.8. When other drugs were the only co-drug, the median concentration of morphine was 200.0 ng/mL in the blood (range: 86.0–2568.0 ng/mL) and 257.0 ng/g in the liver (range: 29.0–2054.0 ng/g), yielding a liver to blood ratio of 0.8. When both ethanol and other drugs were the co-drug, the median concentration of morphine was 114.0 ng/mL in the blood (range: 57.0–913.0 ng/mL) and 314.0 ng/g in the liver (range: 114.0–8559.0 ng/g), yielding a liver to blood ratio of 1.6. Of the 29 cases reported by Duflou et al., (15) morphine was the only drug detected in only four cases, which had a liver to blood ratio of 2.4. In the remaining 25 cases with at least one other drug detected, the liver to blood ratio was 3.4. The current study's liver to BN and kidney to BN morphine ratios were consistent with the published ratios for cases in which heroin was the only drug detected and cases in which multiple drugs were detected.

Blood samples with preservatives and blood samples without preservatives are sent to toxicology laboratories in Jeddah for forensic toxicology examination. In the first part of this research (1), the effect of preservatives on data for heroin-related deaths was assessed. Despite the fact that heroin and its biomarkers are greatly reduced in blood samples without preservatives, these compounds can be still detected. The use of blood with preservative is highly recommended but not mandatory. In some cases, the use of blood without fluoride preservatives is recommended, such as if organophosphorus compounds are suspected (32,33). In a previous report (1), relative concentrations in blood samples without preservative compared to those with sodium fluoride preservative were 75% for 6-MAM, 129% for morphine, and 131% for codeine. Similar findings were reported in the current study. The relative average concentrations in blood samples without versus with sodium fluoride were 69% for 6-MAM, 162% for morphine, and 156% for codeine. This decrease in 6-MAM and increase in morphine and codeine concentrations in blood samples without preservative could be due to the stability of 6-MAM in sodium fluoride and hydrolysis of 6-MAM and 6-AC to morphine and codeine in blood without preservative. Similar findings were reported in blood without preservative after one month of storage with morphine and codeine concentrations increasing by 45% and 48%, respectively (31).

Conclusion

Only limited data exist concerning the utility of complimentary specimens in heroin-related deaths. The current report highlights the importance of testing multiple specimens in heroin-related fatalities, especially when no or limited blood samples are available. Additionally, the valuable information available from blood, liver, and kidney provides insight into which opioids were administered and the survival time following intoxication. The uniqueness of this study is the full optimization and validation of target analytes in each matrix. This method was applied to routine postmortem analysis of cases involving heroin and used as a routine method for almost three years, yielding reducible and accurate results. The distributions of heroin biomarker and their metabolites were investigated in 31 postmortem cases.

Acknowledgments

The author would like to thank all staff at Forensic Toxicology department-Jeddah Poison Control and Forensic Medical Chemistry Center for supporting this work.

References

1. Al-Asmari AI. Postmortem fluid concentrations of heroin biomarkers and their metabolites. J Forensic Sci 2020;65(2):570–9. https://doi.org/10.1111/1556-4029.14200
2. Graziano S, Anzillotti L, Mannocchi G, Pichini S, Busardo FP. Screening methods for rapid determination of new psychoactive substances (NPS) in conventional and non-conventional biological matrices. J Pharm Biomed Anal 2019;163:170–9. https://doi.org/10.1016/j.jpba.2018.10.011
3. Al-Asmari AI, Anderson RA. Comparison of nonhydrolysis and hydrolysis methods for the determination of buprenorphine metabolites in urine by liquid chromatography-tandem mass spectrometry. J Anal Toxicol 2008;39(9):744–53. https://doi.org/10.1093/jat/32.9.744
4. Nedahl M, Johansen SS, Linnet K. Brain-blood ratio of morphine in heroin and morphine autopsy cases. Forensic Sci Int 2019;301:388–93. https://doi.org/10.1016/j.forsciint.2019.06.007
5. Margalho C, Franco J, Corte-Real F, Vieira DN. Illicit drugs in alternative biological specimens: a case report. J Forensic Leg Med 2011;18 (3):132–5. https://doi.org/10.1016/j.jflm.2010.12.006
6. Baselt RC. Disposition of toxic drugs and chemicals in man, 8th edn. Foster City, CA: Biomedical Publications, 2008;730–4.
7. Al-Asmari A, Anderson RA, Kidd S, Thomson AH. Method for the quantification of diamorphine and its metabolites in pediatric plasma samples by liquid chromatography-tandem mass spectrometry. J Anal Toxicol 2010;34(4):177–95. https://doi.org/10.1093/jat/34.4.177
8. Brenneisen R, Hasler F, Wursch D. Acetylcodeine as a urinary marker to differentiate the use of street heroin and pharmaceutical heroin. J Anal Toxicol 2002;26(8):561–6. https://doi.org/10.1093/jat/26.8.561
9. O'Neal CL, Poklis A. The detection of acetylcodeine and 6-acetylmorphine in opiate positive urines. Forensic Sci Int 1998;95(1):1–10. https://doi.org/10.1016/S0379-0738(98)00074-7
10. Staub C, Marset M, Mino A, Mangin P. Detection of acetylcodeine in urine as an indicator of illicit heroin use: method validation and results of a pilot study. Clin Chem 2001;47(2):301–7.
11. Ketola RA, Kriikku P. Drug concentrations in post-mortem specimens. Drug Test Anal 2019;11(9):1338–57. https://doi.org/10.1002/dta.2662
12. Thaulow CH, Oiestad AML, Rogde S, Karinen R, Brochmann W, Anderson JM, et al. Metabolites of heroin in several different post-mortem matrices. J Anal Toxicol 2018;42(5):311–20. https://doi.org/10.1093/jat/bky002
13. Thaulow CH, Oiestad AML, Rogde S, Anderson JM, Hoiseth G, Handal M, et al. Can measurements of heroin metabolites in post-mortem matrices other than peripheral blood indicate if death was rapid or delayed? Forensic Sci Int 2018;290:121–8. https://doi.org/10.1016/j.forsciint.2018.06.041
14. Kudo K, Ishida T, Nishida N, Yoshioka N, Inoue H, Tsuji A, et al. Simple and sensitive determination of free and total morphine in human liver and kidney using gas chromatography-mass spectrometry. J Chromatogr B Analyt Technol Biomed Life Sci 2006;830(2):359–63. https://doi.org/10.1016/j.jchromb.2005.10.049
15. Duflou J, Darke S, Easson J. Morphine concentrations in stomach contents of intravenous opioid overdose deaths. J Forensic Sci 2009;54 (5):1181–4. https://doi.org/10.1111/j.1556-4029.2009.01123.x
16. Moriya F, Hashimoto Y. Distribution of free and conjugated morphine in body fluids and tissues in a fatal heroin overdose: is conjugated morphine stable in postmortem specimens? J Forensic Sci 1997;42(4):736–40.
17. Hargrove VM, Molina DK. Morphine concentrations in skeletal muscle. Am J Forensic Med Pathol 2014;35(1):73–5. https://doi.org/10.1097/paf.0000000000000046
18. Vandenbosch M, Somers T, Cuypers E. Distribution of methadone and metabolites in skeletal tissue. J Anal Toxicol 2018;42(6):400–8. https://doi.org/10.1093/jat/bky014
19. Felby S, Christensen H, Lund A. Morphine concentrations in blood and organs in cases of fatal poisoning. Forensic Sci 1974;3(1):77–81. https://doi.org/10.1016/0300-9432(74)90010-7
20. Spiehler V, Brown R. Unconjugated morphine in blood by radioimmunoassay and gas chromatography/mass spectrometry. J Forensic Sci 1987;32(4):906–16.
21. Steentoft A, Worm K, Christensen H. Morphine concentrations in autopsy material from fatal case after intake of morphine and or heroin. J Forensic Sci Soc 1988;28(2):87–94. https://doi.org/10.1016/s0015-7368(88)72812-1
22. Wyman J, Bultman S. Postmortem distribution of heroin metabolites in femoral blood, liver, cerebrospinal fluid, and vitreous humor. J Anal Toxicol 2004;28(4):260–3. https://doi.org/10.1093/jat/28.4.260
23. Mercurio I, Ceraso G, Melai P, Gili A, Troiano G, Agostinelli F, et al. Significance of morphine concentration in bile, liver, and blood analysis of 52 cases of heroin overdoses. Am J Forensic Med Pathol 2019;40 (4):329–35. https://doi.org/10.1097/paf.0000000000000508
24. Maskell PD, Wilson NE, Seetohul LN, Crichton M, Beer L, Drummond G, et al. Postmortem tissue distribution of morphine and its metabolites in a series of heroin-related deaths. Drug Test Anal 2019;11(2):292–304. https://doi.org/10.1002/dta.2492
25. Al-Asmari AI. Method for the identification and quantification of sixty drugs and their metabolites in postmortem whole blood using liquid chromatography tandem mass spectrometry. Forensic Sci Int 2020;309:110193. https://doi.org/10.1016/j.forsciint.2020.110193
26. Oliveira A, Carvalho F, Pinho PG, Remiao F, Medeiros R, Dinis-Oliveira RJ. Quantification of morphine and its major metabolites M3G and M6G in antemortem and postmortem samples. Biomed Chromatogr 2014;28(9):1263–70. https://doi.org/10.1002/bmc.3158
27. Scientific Working Group for Forensic Toxicology. Scientific working group for forensic toxicology (SWGTOX) standard practices for method validation in forensic toxicology. J Anal Toxicol 2013;37(7):452–74. https://doi.org/10.1093/jat/bkt054
28. Stevens HM. The stability of some drugs and poisons in putrefying human-liver tissues. J Forensic Sci Soc 1984;24(6):577–89. https://doi.org/10.1016/s0015-7368(84)72350-4
29. Goldberger BA, Cone EJ, Grant TM, Caplan YH, Levine BS, Smialek JE. Disposition of heroin and its metabolites in heroin-related deaths. J Anal Toxicol 1994;18(1):22–8. https://doi.org/10.1093/jat/18.1.22
30. Reed D, Spiehler VR, Cravey RH. Two cases of heroin-related suiside. Forensic Sci 1977;9(1):49–52. https://doi.org/10.1016/0300-9432(77)90066-8
31. Al-Asmari AI, Anderson RA. Method for quantification of opioids and their metabolites in autopsy blood by liquid chromatography-tandem mass spectrometry. J Anal Toxicol 2007;31(7):394–408. https://doi.org/10.1093/jat/31.7.394
32. Dinis-Oliveira RJ, Vieira DN, Magalhaes T. Guidelines for collection of biological samples for clinical and forensic toxicological analysis. Forensic Sci Res 2016;1(1):42–51. https://doi.org/10.1080/20961790.2016.1271098
33. Skopp G. Preanalytic aspects in postmortem toxicology. Forensic Sci Int 2004;142(2–3):75–100. https://doi.org/10.1016/j.forsciint.2004.02.012

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Demographics of the 20 heroin-related fatalities Included in this study.

Table S2. Tissue distribution of heroin biomarkers in 20 fatalities.

Table S3. Differences in the concentrations of analyses of interest based on survival time, cause of deaths, and the degree of putrefaction.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14550
Available online at: onlinelibrary.wiley.com

TECHNICAL NOTE

ANTHROPOLOGY

Richard L. Jantz,[1] Ph.D.; Lee Meadows Jantz,[1] Ph.D.; and Stephen D. Ousley,[1] Ph.D.

Measuring the Tibia: Trotter's Error Revisited

ABSTRACT: In Trotter and Gleser's (*Am J Phys Anthropol* 1952;10:463) classic study of stature estimation, a definition of the tibia length measurement is given that agrees with the standard condylar-malleolar length. That Trotter did not in fact measure according to her definition, but rather omitted the malleolus, has been well documented by Jantz et al. (*J Forensic Sci* 1995;40:758). Both the Terry collection and the World War 2 (WW2) samples were affected, although questions remain about the latter that cannot be resolved directly because it is no longer available for examination. Trotter's data from hundreds of servicemen are especially important because the statures were measured by technicians, rather than based on cadaver lengths or forensic statures. The questions examined in this note are as follows: Was WW2 measured uniformly in the same way as Terry; are there differences between Terry and WW2 that could influence estimation of the adjustment; and is the 10 millimeter (mm) adjustment proposed by Jantz et al. (*J Forensic Sci* 1995;40:758) still appropriate. Our analysis relies on a measurement taken by Trotter that is clearly and uniquely defined, what she called "ordinary length". This measurement was used to create expectations about how Trotter measured what she called maximum length of the tibia. Results provide no evidence that WW2 was measured any differently than Terry, with the exception of one small series. They also show slight morphological differences on the distal and/or proximal end of the tibia between Terry and WW2. Despite the slight difference, the adjustment to account for the malleolus is still valid.

KEYWORDS: forensic anthropology, tibia measurements, stature estimation, methodology, postcranial, regression

For reasons unknown, Trotter (1) omitted the malleolus from her tibia length measurement, despite providing a definition that indicated it was included. The error was incorporated into the popular stature estimation equations and persisted for over 40 years. It went undetected for that time because the error had a minor effect on the stature estimate, about 25–30 millimeters (mm) (2). The discovery of the error and its implications for stature estimation has been described (2), and those interested in the background can consult other works (2–5).

Jantz et al. (2) proposed adding 10 millimeters (mm) to Trotter's measurement to make it comparable to condylar-malleolar length, which is the definition given in Trotter and Gleser (1). Whether the 10 mm adjustment is appropriate has been called into question by Lynch et al. (6) during the course of their investigation of the *USS Oklahoma* bone lengths. They found that the 10 mm adjustment in Jantz et al. (2) produced height estimates systematically different when compared to other bones, and that an adjustment of 6 mm was more appropriate. Lynch et al. (6) concluded this difference may be attributable either to a difference between how Trotter measured the Terry collection or to how the World War 2 (WW2) tibiae were measured. They suggest two possibilities. One is that some WW2 tibiae may have been measured including the malleolus, as was shown to be the case in the 1958 study (2,7). Another is that there are morphological differences between the Terry collection tibiae, from which the correction was derived, and the WW2 tibiae. The answer to these questions, particularly the first, is critical to our understanding of what Trotter did and to the continued use of the WW2 data for stature estimation and research.

These questions cannot be addressed directly because the WW2 tibiae are not available for examination. It is, however, possible to address them using Trotter's own data. These data have been previously used to address various questions (8,9). The purpose of this paper is to address questions raised in Lynch et al. (6) and to re-evaluate the adjustment proposed in Jantz et al. (2). Specifically, analyses will be designed to address questions as follows:

- Did Trotter measure Terry and WW2 the same way, as is clearly implied by her statement that she measured both series herself? In other words, was the malleolus included in some WW2 tibia measurements?
- Are there differences between Terry and WW2 tibiae that might affect estimation of the adjustment?
- Is the 10 mm adjustment set forth in Jantz et al. (2) still appropriate?

Materials and Methods

Data used in these analyses are Trotter's original data as described in Jantz et al. (2). The analyses will be limited to White males because it is WW2 White males and the crew of the *Oklahoma* that are at issue. The data therefore consist of the 255 White males from the Terry collection and the 545 complete military sample described in Trotter and Gleser (1). The Terry sample contained only the average values of right and left bones but the WW2 data contained right and left separately, so were averaged to make the two data sets comparable for some comparisons.

The core sample of 545 (referred to hereafter as WW2 core) met Trotter and Gleser's (1) criteria of having all elements present and being 18 years old or older. An additional

[1]Department of Anthropology, University of Tennessee-Knoxville, 505 Strong Hall, Knoxville, TN, 37996.

Corresponding author: Richard L. Jantz, Ph.D. E-mail: rjantz@utk.edu

Received 30 May 2020; and in revised form 15 July 2020; accepted 24 July 2020.

approximately 580 individuals missing one or more bones were not included in Trotter and Gleser (1). The full WW2 sample is described in Meadows Jantz (10). Because of the missing data, sides were not averaged but rather analyzed separately and referred to as WW2-L and WW2-R. These samples also appear to have outliers that represent errors. Outliers having T^2 probabilities <0.01, based on the two tibia measurements, were excluded. This resulted in eight from WW2-L and seven from WW2-R being omitted.

In addition, we identified a small sample numbering about 120 individuals with a large amount of missing data that has not been included in any of the previous research using the WW2 data. Meadows Jantz assigned them numbers 7001–7124. This sample will be referred to as the 7000 series and will also be examined for evidence of measurement variation.

Here it may be useful to review the measurements relevant to the present inquiry. Table 1 presents Trotter's measurements in the five papers dealing with bone lengths, along with earlier definitions that are similar in name or description. What one realizes is that Trotter did not rely on established authority for her tibia measurements. The measurement she called "Ordinary Length" in her 1952 (1) paper was used in the two 1951 papers (11,12) but no definition reference was provided, and it was not called Ordinary Length until the 1952 paper. In that paper, her authority is a personal communication from Krogman. Ordinary length, as described in (1), measured with spreading calipers from the center of the articular surface of the lateral condyle to the center of the distal articular surface, differs from what Krogman (13) called Physiological Length, "from the depression on the top of the medial half of the condylar surface to the lower articular surface near the medial malleolus, but excluding it." Either Trotter misunderstood Krogman, or Krogman made a distinction between ordinary length and physiological length, providing the former definition to Trotter.

The measurement Trotter called maximum length of the tibia, Tibia$_m$ in Trotter and Gleser (1) is not actually a maximum length. Her definition corresponds to condylar-malleolar length as defined by Martin (14). What she actually measured, however, was from the lateral condyle to the most distal of the anterior or posterior trochlear facet, defined in Turley et al. (15), excluding the malleolus, if she in fact placed the tibia on its dorsal surface as described in Trotter and Gleser (1), and shown in Jantz et al. (2). Importantly, Trotter's definition mentions placing the end of the malleolus against the vertical board and applying a freely movable block at the other end. As such it is essentially the same as technique B in Lynch et al. (16) where the sliding end of the osteometric board is used as a block. As nearly as we can determine what Trotter actually measured is unique, never defined or used before or since.

Some of the definitions in Table 1 approximate what Trotter measured, but none is an exact description of what she actually measured. Martin's (14) physiological length and Wilder's (17) physiological length differ from Trotter's ordinary length in measuring from the medial rather than the lateral condyle. Hrdlicka (18) defined an ordinary length, but it is quite different from Trotter's. It is measured using Broca's osteometric board, from the most proximal points on each condyle, excluding the spine, to the distal tip of the malleolus. The closest approximation to Trotter's ordinary length is the more recent biomechanical length, the average of measurements from the center of the distal articular surface to each midpoint of the lateral and medial condyles (19).

For purposes of this paper, we will call what Trotter measured tibia malleolar-trochlear length (MTL), abbreviated tibia$_{mtl}$. We will call Trotter's ordinary length tibia$_{ol.}$ These variables are tibia$_m$ and tibia respectively in Trotter and Gleser (1). Trotter's use of "maximum length", tibia$_m$, is enough of a misnomer to be avoided.

Three analyses were conducted to address the following questions:

- Were WW2 tibiae measured in such a way that some included the malleolus, as Trotter directed, and some omitted it? To address this question we rely on tibia$_{ol.}$ The value of this smaller measurement is there can be little doubt about how it was taken, and the observer variation will be minimal. The difference between tibia$_{mtl}$ and tibia$_{ol}$ was calculated for the Terry and WW2 samples. If the WW2 bones were not measured consistently and the malleolus was included on some of tibia$_{mtl}$ measurements, it should increase the mean length and therefore the difference between tibia$_{mtl}$ and tibia$_{ol}$. This test was conducted separately for the WW2 core sample and WW2-L and WW2-R. Whether the 7000 series was measured in the same way as WW2 core was evaluated by comparing their means and calculating t-tests.
- Are there morphological differences between Terry and WW2 apart from possible measurement methods? The analysis described above also addresses this question.

TABLE 1—*Trotter's measurements of the tibia, along with other measurement definitions. Trotter could have consulted all but the last in deciding how to measure.*

Reference	Name	Proximal	Distal	Instrument
T&G 1951 (12)	Length described, no reference	Center lat condyle articular surface	Center distal articular surface	Large spreading caliper
T&G 1951 (13)	Length described, no reference	Center lat condyle articular surface	Center distal articular surface	Large spreading caliper
T&G 1952 (1)	Max Length, no reference, Tibia$_m$	Most prox pt lat border lat condyle	End of malleolus	Osteometric board
T&G 1952 (1)	Ordinary Length, cites Krogman, personal comm.	Center lat condyle articular surface	End of malleolus	Osteometric board
T&G 1958 (7)	Max Length. Refers to (1)	Most prox pt lat border lat condyle	End of malleolus	Osteometric board
Martin and Knussmann (14)	*Länge*, #1b, Condylar-malleolar length	Articular surface lat condyle	Center distal articular surface	Osteometric board
Martin and Knussmann (14)	*Condylo-astragal Länge*, #2, Physiological length	Center med condyle	Deepest pt distal articular surface	Large spreading caliper
Wilder 1920 (17)	Physiological length	Deepest point med condyle	Center of distal articular surface	Large spreading caliper
Hrdlicka 1920 (18)	Ordinary length	Most prox pt using Broca's opening	End of malleolus	Broca osteometric board
Ruff and Hayes 1983 (19)	Biomechanical length	Center of lat and med condyles	Center of distal articular surface	Large spreading caliper

lat, lateral; med, medial; prox, proximal; pt, point.

- Is the 10 mm. adjustment proposed by Jantz et al. (2) and based on the Terry collection appropriate for WW2? This was evaluated using Trotter's data as follows:
  - ○ Estimate the stature equation for tibia$_{ol}$ from WW2 core sample and use this equation to estimate stature for WW2-L and WW2-R. We are defining WW2 core as the calibration sample, and WW2-L and WW2-R as test samples.
  - ○ Add 10 mm to tibia$_{mtl}$ in the WW2 core sample and get the stature estimation equation. Use this equation on tibia$_{mtl}$ in WW2-L and WW2-R with incremental adjustments and compare to the estimate obtained in (a) above.
  - ○ All calculations were conducted using NCSS 2020 (20).

Results

Table 2 presents the summary statistics for the difference between tibia$_{mtl}$ and tibia$_{ol}$ for WW2 core, WW2-L, WW2-R, and Terry. Intergroup differences are based on t-test results. The most obvious mean difference concerns Terry vs all WW2 samples. Terry tibia lengths have the largest difference from all other groups and a standard deviation larger than all but WW2-L. This is the opposite of what would be expected if WW2 were measured with some including the malleolus and some not. The evidence therefore does not support the hypothesis that some WW2 tibiae measurements included the malleolus.

It is unexpected to see some variation among the WW2 samples. WW2-R, possessing the lowest value, differs from WW2 core, with the highest of the WW2 values. To what this difference might be attributed is unclear. We do not think it likely that it reflects measurement difference, because, although statistically significant, it is minor compared to the Terry-WW2 core difference.

Table 3 presents the comparison of the WW2 7000 series to WW2 core series. The 7000 series presents a puzzling pattern of missing data. Sample size for stature is 120, and most bones have sample sizes ranging from 78 to 109. The two exceptions are maximum femur length and ordinary tibia length that have sample sizes in the mid-thirties. In both cases, it would presumably have been possible to measure more maximum lengths of the femur because bicondylar lengths were measured, and likewise for ordinary length of the tibia because malleolar-trochlear length (tibia$_{mtl}$) was measured. What is also clear is that all bones in the 7000 series except the tibia have means quite similar to the WW2 core sample. Tibia measurements all differ significantly from the WW2 core sample at $p \leq 0.05$.

There are two possible explanations for why the Terry data produce a larger difference between tibia$_{mtl}$ and tibia$_{ol.}$ It could suggest that the malleolus was included in some Terry measurements, but this is unlikely because we did not encounter it in the 178 tibiae we did measure (2). Rather, it is better viewed as a difference between the mainly 19th century Terry sample and early 20th century WW2 sample. Morphologically, looking at tibia$_{mtl}$ minus tibia$_{ol}$ indicates that Terry distal articular surfaces or the proximal articular surfaces, or both, are deeper, or more concave, than is the case in WW2.

TABLE 2—*Means and standard deviations for difference between Tibia$_{mtl}$ and Tibia$_{ol}$.*

Group/Diff	*N*	Mean (mm)	SD (mm)	Different from
WW2 core	545	9.7	1.6	Terry, WW2-R
WW2 L	486	9.4	2.1	Terry
WW2 R	482	9.2	1.8	Terry, WW2 core
Terry	255	10.3	2.1	WW2 core, WW2-R, WW2-L

TABLE 3—*Comparison of the means of the 7000 series with WW2 core sample. Only tibia measurements are significantly different.*

	WW2 7000 series			WW2 core (*N* = 545)			
	N	Mean	SD	Mean	SD	Diff	*T*
Height	120	1741.3	65.4	1739.0	66.3	−2.3	−0.3
Rhum	106	335.2	16.4	336.4	16.9	1.2	0.7
Lhum	99	333.7	16.6	336.0	16.7	2.3	1.2
Rrad	98	251.4	13.2	252.3	12.9	0.9	0.6
Lrad	82	250.0	13.3	250.6	12.7	1.5	0.5
Ruln	88	269.0	13.6	271.3	13.0	2.4	1.6
Luln	78	269.0	13.1	269.4	12.9	0.4	0.3
Rfem	109	467.5	21.1	468.5	23.2	1.1	0.4
Lfem	106	469.1	22.1	469.6	23.2	0.5	0.2
Rmaxfem	35	471.5	21.6	472.3	23.6	0.8	0.2
Lmaxfem	35	473.0	21.3	472.9	23.6	−0.1	0.0
Rtibia$_{mtl}$	103	383.3	22.0	377.8	21.3	−5.4	−2.3*
Ltibia$_{mtl}$	100	384.3	21.4	378.4	21.6	−5.9	−2.5*
Rtibia$_{ol}$	36	376.1	19.7	368.3	21.0	−7.8	−2.2*
Ltibia$_{ol}$	33	375.8	20.1	368.6	21.4	−7.1	−2.0*
Rfib	80	381.5	19.6	381.2	20.8	−0.3	0.1
Lfib	93	379.7	20.8	381.5	21.1	1.8	0.8

*$p \leq 0.05$.

Table 4 presents the stature estimates based on tibia$_{ol}$ from WW2 core and applied to the two test samples. This provides a reliable estimate of stature because there is little question about how ordinary length was measured. Table 4 also presents the stature estimates using tibia$_{mtl}$ plus a 10 mm adjustment, again from WW2 core. When applied to the test samples, the estimates differ significantly from the tibia$_{ol}$ estimates. The adjustment was then raised by 0.1 mm increments until the highest *t*-test probability was attained. The best adjustments were 10.4 mm for WW2-L and 10.3 mm for WW2-R.

Discussion

Trotter's measurement of the tibia in a way that differs from the definition provided in Trotter and Gleser (1) continues to cause problems. The foregoing results should settle some of these problems. Our results do not support the suggestion by Lynch et al. (6) that some of WW2 may have been measured inconsistently, some including the malleolus and some not. Small morphological differences between WW2 and Terry collection in tibia morphology occur, but this difference does not affect estimates of the adjustment required to make Trotter's measurement comparable to the now commonly used condylar-malleolar length. The adjustment estimated from WW2 agrees with the adjustment obtained from the Terry collection.

The one exception to the above conclusions is found in the 7000 series. It has a peculiar pattern of missing data, suggesting that it might have been measured by technicians who were experimenting with different measurement definitions. That both ordinary length and "maximum length" were longer than WW2 core suggests that the former was measured as Trotter measured "maximum length", and "maximum length" was actually condylar-malleolar length, perhaps guided by Trotter's actual definition. This small series offers support for Lynch et al.'s (6) suggestion that some WW2 were measured to include the malleolus, but it has no bearing on stature estimation because there is no evidence it was ever used for that.

TABLE 4—*Re-evaluation of 10 mm adjustment using tibia$_{ol}$ criterion.*

Group	Mean tibia$_{ol}$	SD tibia$_{ol}$	Stature from tibia$_{ol}$	Stature from tibia$_{mtl}$ + 10	Adjusted Stature (Adjustment)
WW2 core Left	368.6	21.4	174.0	174.0	
WW2 core Right	368.3	21.0	174.0	174.0	
WW2-L	369.5	21.0	174.2	174.1*	174.3 (10.4)
WW2-R	369.1	20.6	174.2	174.1*	174.2 (10.3)
Stature = 0.265(Ltibia$_{ol}$) + 76.3163; stature = 0.2705(Rtibia$_{ol}$) + 74.3703					
Stature = 0.2632(Ltibia$_{mtl}$ + 10) + 71.608; stature = 0.2672(Rtibia$_{mtl}$ + 10) + 70.3702					

*Differs significantly different from tibia$_{ol}$ estimate by paired *t*-test.

During the course of this investigation, additional errors were discovered. One in particular deserves special attention because it occurs in Trotter and Gleser's (1) WW2 core sample and was incorporated into their stature estimation equations. One individual had a tibia$_{mtl}$ right value of 498 mm, the same value as the individual's femur, so clearly a data entry error. The left value is 469, probably a typographical error for 409, the value on the handwritten card. Correcting the errors and recomputing the stature equation yields 2.67*Tib$_m$ + 73.14 ± 3.26 (using Trotter's variable name). This equation corrects the one given in table 9 and table 13 in Trotter and Gleser (1) for White males. Correcting the error increases the slope (2.67 vs. 2.52), and decreases the intercept (73.14 vs. 78.62). It also decreases the root mean square error (3.26 vs. 3.37). However, the effect on stature estimation is minimal, being most pronounced for tibiae above the mean. For tibiae, one standard deviation above the mean Trotter and Gleser's equation underestimates height by about 5 mm, for tibiae two standard deviations above the mean it underestimates by about 10 mm.

The question for which there is not an obvious answer is why a 6 mm correction produced better results than 10 mm for the *Oklahoma* tibiae (6). Antemortem height for the WW2 sample is 173.99 cm (68.5 inches [in]), essentially identical to 68.53 in reported for the *Oklahoma* in Lynch et al. (6). The average tibia$_{mtl}$ for WW2 is 378.13 mm. If we take it as 378, add 10 mm and estimate stature with Fordisc (21) using the Trotter data, the point estimate is 68.5 in, identical to the WW2 antemortem mean. Using a 6 mm correction returns 68.1 in, underestimating WW2 stature by 0.4 in. Because WW2 and *Oklahoma* heights are essentially the same, the 6 mm adjustment could mean that *Oklahoma* tibiae average 4 mm longer than WW2.

The forensic anthropology community has accepted condylar-malleolar length as the standard method of measuring the tibia. This acceptance is likely due to Trotter and Gleser's (1) large sample based on antemortem heights that was best suited for purposes of living stature estimation. Unfortunately, their definition can be implemented in different ways that can result in average differences of almost 3 mm and maximum differences up to 6 mm (16). In view of variation that exists even when observers are ostensibly following the same definition, the most prudent course for forensic practitioners is to avoid the tibia, if possible, in favor of the fibula as an indication of lower leg length.

References

1. Trotter M, Gleser G. Estimation of stature from long bones of American Whites and Negroes. Am J Phys Anthropol 1952;10(4):463–514. https://doi.org/10.1002/ajpa.1330100407
2. Jantz RL, Hunt DR, Meadows L. The measure and mismeasure of the tibia: implications for stature estimation. J Forensic Sci 1995;40(5):758–61. https://doi.org/10.1520/JFS15379J
3. Jantz RL. Modification of the Trotter and Gleser female stature estimation formulas. J Forensic Sci 1992;37(5):1230–5. https://doi.org/10.1520/JFS13310J
4. Giles E. Modifying stature estimation from the femur and tibia. J Forensic Sci 1993;38(4):758–63.
5. Iscan MY. A comparison of techniques on the determination of race, sex and stature from the Terry and Hamann-Todd collections. In: Gill GW, Rhine S, editors. Skeletal attribution of race. Albuquerque, NM: Maxwell Museum of Anthropology, 1990;73–81.
6. Lynch JJ, Brown C, Palmiotto A, Maijanen H, Damann F. Reanalysis of the Trotter tibia quandary and its continued effect on stature estimation of past-conflict service members. J Forensic Sci 2019;64(1):171–4. https://doi.org/10.1111/1556-4029.13806
7. Trotter M, Gleser GC. A re-evaluation of estimation of stature taken during life and of long bones after death. Am J Phys Anthropol 1958;16 (1):79–123. https://doi.org/10.1002/ajpa.1330160106
8. Jantz RL, Meadows Jantz L. Limb bone allometry in modern Euro-Americans. Am J Phys Anthropol 2017;163(2):252–63. https://doi.org/10.1002/ajpa.23203
9. Meadows L, Jantz RL. Allometric secular change in the long bones from the 1800s to the present. J Forensic Sci 1995;40(5):762–7. https://doi.org/10.1520/JFS15380J
10. Meadows Jantz L. Secular change and allometry in the long limb bones of Americans from the mid 1700s through the 1970s [dissertation]. Knoxville, TN: University of Tennessee, 1996. https://trace.tennessee.edu/utk_graddiss/4039 (accessed July 24, 2020).
11. Trotter M, Gleser GC. Trends in stature of American whites and Negroes born between 1840 and 1924. J Phys Anthropol 1951;9(4):427–40. https://doi.org/10.1002/ajpa.1330090404
12. Trotter M, Gleser G. The effect of aging on stature. Am J Phys Anthropol 1951;9(3):311–24. https://doi.org/10.1002/ajpa.1330090307
13. Krogman WM. The human skeleton in forensic medicine. Springfield, IL: C.C. Thomas, 1962;322.
14. Martin R, Knußmann R. Anthropologie: handbuch der vergleichenden biologie des menschen. [Handbook of comparative human biology]. Stuttgart, Germany: Gustav Fischer, 1988;221.
15. Turley K, Guthrie EH, Frost SR. Geometric morphometric analysis of tibial shape and presentation among Catarrhine taxa. The Anat Rec 2011;294(2):217–30. https://doi.org/10.1002/ar.21307
16. Lynch JJ, Maijanen H, Prescher A. Analysis of three commonly used tibia length measurement techniques. J Forensic Sci 2019;64(1):181–5. https://doi.org/10.1111/1556-4029.13868
17. Wilder HH. A laboratory manual of anthropometry. Philadelphia, PA: P. Blakiston's Son & Co., 1920;145.
18. Hrdlicka A. Anthropometry. Philadelphia, PA: The Wistar Institute of Anatomy and Biology, 1920;129.
19. Ruff CB, Hayes WC. Cross-sectional geometry of Pecos Pueblo femora and tibiae—a biomechanical investigation: 1. Method and general patterns of variation. Am J Phys Anthropol 1983;60(3):359–81. v10.1002/ajpa.1330600308
20. NCSS Statistical Software. NCSS 2020. ncss.com/software/ ncss (accessed July 24, 2020).
21. Jantz RL, Ousley SD. FORDISC 3.1: Computerized forensic discriminant functions. Knoxville, TN: University of Tennessee, Department of Anthropology, 2005.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14535
Available online at: onlinelibrary.wiley.com

TECHNICAL NOTE

ANTHROPOLOGY

Barbara Bertoglio,[1] M.Sc., Ph.D.; Sofia Corradin,[1] MS; Annalisa Cappella,[1,2] M.Sc., Ph.D.; Debora Mazzarelli,[1,3] B.Sc.; Lucie Biehler-Gomez,[1] M.Sc., Ph.D.; Carmelo Messina,[4,5] M.D.; Grazia Pozzi,[4] M.D.; Luca Maria Sconfienza,[4,5] M.D., Ph.D.; Francesco Sardanelli,[5,6] M.D.; Chiarella Sforza,[2] M.D., Ph.D.; Danilo De Angelis,[1] D.D.S., Ph.D.; and Cristina Cattaneo,[1] M.D., Ph.D.

Pitfalls of Computed Tomography 3D Reconstruction Models in Cranial Nonmetric Analysis*

ABSTRACT: Many studies in the literature have highlighted the utility of virtual 3D databanks as a substitute for real skeletal collections and the important application of radiological records in personal identification. However, none have investigated the accuracy of virtual material compared to skeletal remains in nonmetric variant analysis using 3D models. The present study investigates the accuracy of 20 computed tomography (CT) 3D reconstruction models compared to the real crania, focusing on the quality of the reproduction of the real crania and the possibility to detect 29 dental/cranial morphological variations in 3D images. An interobserver analysis was performed to evaluate trait identification, number, position, and shape. Results demonstrate a false bone loss in 3D models in some cranial regions, specifically the maxillary and occipital bones in 85% and 20% of the samples. Additional analyses revealed several difficulties in the detection of cranial nonmetric traits in 3D models, resulting in incorrect identification in circa 70% of the traits. In particular, pitfalls included the detection of erroneous position, error in presence/absence rates, in number, and in shape. The lowest percentages of correct evaluations were found in traits localized in the lateral side of the cranium and for the infraorbital suture, mastoid foramen, and crenulation. The present study highlights important pitfalls in CT scan when compared with the real crania for nonmetric analysis. This may have crucial consequences in cases where 3D databanks are used as a source of reference population data for nonmetric traits and pathologies and during bone-CT comparisons for identification purposes.

KEYWORDS: nonmetric traits, computed tomography, CT accuracy, anatomical variants, cranial traits, forensic anthropology

Radiological imaging has become increasingly important and even essential in the forensic field. Many studies report methods and techniques based on radiology currently used in forensic anthropology and legal medicine to support the analyses of both the living (1–8) and the dead (9–11).

In particular, over the last years several studies have highlighted the potential of clinical radiological data in anthropological research. This is quite common in countries where the lack of contemporary skeletal collections necessarily leads to the creation of virtual databases collecting computed tomography (CT) images from known living patients. Generally, these records were used as population-specific reference data to assess the reliability of methods important for building a biological profile, especially for the estimation of sex (12–16) or age-at-death (17,18). More recently, virtual skeletal collections have been used as a basis for population analyses. Specifically, three-dimensional (3D) models were used to describe the expression and variability of nonmetric traits and to calculate population frequencies (19,20) with the aim to provide useful data for anthropological investigations, especially for identification purposes and population studies.

However, in contrast to the increasing literature on this topic, few studies evaluated the accuracy of the radiological material and its suitability as a substitute for skeletal collections. These studies focused mainly on metrics and geometrical precision (21–25), while nonmetric traits were not considered. Nonetheless, analyses including nonmetric traits should be performed considering the use of clinical radiological records in the development of methods and collection of data which are then used in anthropological analyses of skeletal remains. This shortcoming would have even greater effects on the personal identification process. In this context, radiological records of the missing person are used and compared with those of the unknown remains to find a match and reach

[1]LABANOF, Dipartimento di Scienze Biomediche per la Salute, Sezione di Medicina Legale, Università degli Studi di Milano, Via Luigi Mangiagalli 37, Milan, 20133, Italy.

[2]Laboratorio di Anatomia Funzionale dell'Apparato Stomatognatico (LAFAS), Dipartimento di Scienze Biomediche per la Salute, Università degli Studi di Milano, Via Luigi Mangiagalli 31, Milan, 20133, Italy.

[3]Fondazione Isacchi Samaja ONLUS, Via Nino Bixio 30, Milan, 20129, Italy.

[4]IRCCS Istituto Ortopedico Galeazzi, Via Riccardo Galeazzi 4, Milan, 20161, Italy.

[5]Dipartimento di Scienze Biomediche per la Salute, Università degli Studi di Milano, Via Luigi Mangiagalli 31, Milan, 20133, Italy.

[6]IRCCS Policlinico San Donato, Piazza Edmondo Malan 2, Milan, 20097, Italy.

Corresponding author: Barbara Bertoglio, M.Sc., Ph.D. E-mail: barbara.bertoglio@unimi.it

*Presented at the FASE Advanced Course in Forensic Anthropology Biomechanics, September 14, 2019, in Bruxelles, Belgium.

Received 10 April 2020; and in revised form 27 May 2020, 9 July 2020; accepted 17 July 2020.

identification. Generally, several morphological variants are considered, including normal anatomical variations (e.g., sinuses, suture and trabecular patterns), skeletal anomalies, or nonmetric traits (e.g., supernumerary bones, accessory foramina, and nonfusion anomalies), and pathological conditions or skeletal changes related to repetitive activities (26). Many studies in literature clearly describe and support the usefulness of morphological traits analysis in personal identification (27–43). However, despite the great number of studies in the field, no research has investigated the accuracy of radiological data compared to the real bone samples in nonmetric analysis and how the traits, especially skeletal anomalies, appear in the radiological record and specifically in CT images. In forensic scenarios, knowing which traits are easily recognizable in radiological images to reach a positive match is of paramount importance.

With this purpose, 20 skulls and their corresponding CT images were compared to determine the accuracy of the reproduction of anatomical features in radiological images. In addition, 29 dental/cranial morphological variations were analyzed by two observers on the original bone samples and their 3D CT models to verify the ease with which they can be identified in radiological records.

Materials and methods

Samples

Twenty crania were selected from contemporary commingled remains of the skeletal collection of unknown individuals of the LABANOF (Laboratory of Forensic Anthropology and Odontology), University of Milan, Italy. The selected crania belonged to unclaimed unidentified adults including African migrants, which were all complete and in a good state of preservation. Only mandibles were excluded.

CT Scans and Reconstructions

CT examinations were performed at the IRCCS Istituto Ortopedico Galeazzi (Milan, Italy) using a 64-slice CT scan (Somatom Emotion, Siemens Medical Solutions, Erlangen, Germany). The crania were scanned by placing them in *norma basalis* (i.e., with the cranial vault leaning on the table and the cranial base facing upwards). Image acquisition was carried out using the following parameters: slice thickness of 0.60 mm, 100 kV, 64 mA, and with a field of view size of 213 mm.

CT images were analyzed with Slicer 4.10.1 software (44,45). In particular, 3D models were built using the volume rendering tool, selecting the rendering option "VTK GPU Ray Casting" and adjusting window setting in order to limit the visualization to the bone tissue. Material properties of the volume were changed with the following parameters: ambient 0.30, diffuse 0.36, specular 0.40, and power 10. Further changes could be performed during the analysis to optimize bone visualization.

Bone Appearance

To determine the accuracy of CT scan 3D reconstructions, the quality of the reproduction of the real crania on CT images was evaluated. The analyses were limited to the main bones forming the human cranium, namely the frontal, nasal, maxillary, zygomatic, temporal, parietal, occipital, sphenoid, and palatine bones. Each bone was described based on its appearance as complete or with slight bone loss, in both real crania and the 3D models. Bone loss for the real crania was intended as minimal bone defects, that is, an area of loss of bone tissue smaller than 0.5 mm. 3D models were analyzed applying the options reported in the previous paragraph. When bone loss was observed, changes in the window setting were carried out shifting levels to lower values to verify the persistence of bone loss. However, in these cases the visualization of the sample was limited due to the inclusion of units corresponding to air levels. Under such conditions, observation was limited to the area of interest (where bone loss was visible in the 3D model), but no considerations could be performed concerning the other bone elements. 3D models from CTs and crania were subsequently compared and differences were recorded.

Morphological Variations Analysis

Twenty-nine commonly observed morphological variations were included in the study. Among them, 28 corresponded to dental/bone normal traits and skeletal anomalies. We not only evaluated the possibility to detect bony traits (presence/absence score), but we also considered the number, position, and shape when applicable depending on the anatomical variations. In addition to these, *cribra orbitalia* was evaluated to investigate the detection of pathological signs in the radiological record (46).

The full list of traits and the different evaluations are reported in Table 1. Some modifications to the standard methodological approach described in literature were performed depending on the available bone material.

The skulls and their corresponding CT reconstructions were analyzed by two anthropologists (both with 10 years of experience in the field) blinded one to the other. Bilateral traits were evaluated on each side and dental traits on each molar tooth available. 3D models were always analyzed before the "real" sample to avoid any influence on CT evaluation.

Interobserver agreement was calculated with Cohen's kappa coefficient (51) using the online VassarStats calculator (52). Evaluations assessed the reproducibility of trait detection and classification in the real crania. Cohen's kappa values were interpreted according to Landis and Koch (53), with values as < 0 indicating no agreement, 0–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as almost perfect agreement.

To evaluate the accuracy of the reproduction of the traits on CT images, the real crania were used as reference sample for comparison. For each trait, true positives were calculated as the percentage where the traits were correctly identified in 3D models based on the number detected in real crania by each observer. For traits with a bilateral expression and dental traits, percentages were calculated based on the number of sides/molar teeth showing the traits. Results were then grouped in three categories depending on the percentage of correct identifications, as: (i) high accuracy, when a trait was correctly identified in 81%–100% of the 3D models; (ii) intermediate accuracy, when a trait was correctly identified in 51%–80% of the 3D models; and (iii) low accuracy, when a trait was correctly identified in 50% or less of the 3D models.

Results

Bone Appearance

All crania included in this study are in a good state of preservation (i.e., all bones were present and generally complete). Signs of bone resorption or bone loss related to pathological or taphonomical events were observed in maxillary and temporal

TABLE 1—*List of the morphological variations included in the study.*

Morphological variations	Position	Definition	P/A	Number	Position	Shape	Reference
Nasal foramen (1)	Nasal bones	One or more foramina perforating the nasal bones.	x	x	On the bone On the nasomaxillary suture		(47)
Supraorbital foramen (2)	Frontal bone	One or more foramina on the supraorbital margin.	x	x			(47)
Supraorbital notch (3)	Frontal bone	Notch on the supraorbital margin.	x				(47)
Infraorbital suture (4)	Maxillary bones	Suture marking externally the course of the infraorbital canal on the orbital floor and on the facial surface of the maxilla.	x				(47)
Infraorbital foramen (5)	Maxillary bones	Foramen located on the anterior surface of the maxillary bones below the infraorbital margin and above the canine fossa.	x			Single infraorbital foramen Bipartite infraorbital foramen	(47)
Secondary infraorbital foramen (6)	Maxillary bones	Supernumerary foramen localized next to (superiorly or medially) the infraorbital foramen.	x	x			(47)
Zygomatic foramen (7)	Zygomatic bones	Foramen localized 5-8 mm below the orbital margin and aligned vertically to the zygomatic tubercle.	x				(47)
Secondary zygomatic foramen (8)	Zygomatic bones	Supernumerary foramen localized in the same concentric arc around the orbital margin.	x	x	On the corpus On the frontal process		(47)
Os japonicum (9)	Zygomatic bones	Division of the zygomatic bone by one or more sutures (generally horizontal) in two or more parts.	x			Partial division Total division	(47)
Pterion shape (10)	Pterion	Pattern of contact between parietal and sphenoid bones at the pterion region (point of articulation of the coronal suture, sagittal suture, squamous suture and the greater wing of the sphenoid).				Fronto-temporal articulation Presence of epipteric bone H-shape variant	(47)
Epipteric bone (11)	Pterion	Supernumerary ossicle localized at pterion (point of articulation of the coronal suture, sagittal suture, squamous suture and the greater wing of the sphenoid).	x		Ossicle along the temporal squama Ossicle along the frontal bone True large single bone		(47)
Squamous ossicle (12)	Squamous suture	Supernumerary ossicle localized between the temporal squama and the parietal bone.	x	x			(47)
Coronal suture ossicle (13)	Coronal suture	Supernumerary ossicle localized on the coronal suture.	x	x			(47)
Parietal foramen (14)	Parietal bones	One or more foramina near the sagittal suture or in the sagittal suture in the obelion area (3.5 cm above the lambda).	x	x			(47)
Sagittal suture ossicle (15)	Sagittal suture	Supernumerary ossicle localized on the sagittal suture.	x	x			(47)
Ossicle at lambda (16)	Lambda	Supernumerary ossicle localized at lambda (point of junction of the sagittal and lambdoid sutures).	x				(47)
Os inca (17)	Occipital bone	Superior part of the occipital squama divided by a transverse suture running at the highest nuchal line. Several bone variations can be observed.	x			Complete Bipartite, tripartite or multipartite Incomplete	(47)
Lambdoid suture ossicle (18)	Lambdoid suture	Supernumerary ossicle localized on the lambdoid suture.	x	x			(47)
Parietal notch bone (19)	Parietal notch	Supernumerary ossicle localized on the notch of the temporal bone.	x				(47)
Sutura mendosa (20)	Occipital bone	Suture originating from asterion (point of junction of the lambdoid, occipito-mastoid and parieto-mastoid sutures), or close to it, and usually located little above the transverse groove of the cerebral surface.	x				(47)

TABLE 1—*Continued.*

Morphological variations	Position	Definition	P/A	Number	Position	Shape	Reference
Ossicle at asterion (21)	Asterion	Supernumerary ossicle localized at asterion (point of junction of the lambdoid, occipito-mastoid and parieto-mastoid sutures)	x				(47)
Occipito-mastoid ossicle (22)	Occipito-mastoid suture	Supernumerary ossicle localized at the occipito-mastoid suture	x	x			(47)
Mastoid foramen (23)	Temporal/ occipital bones	Foramen localized in the posterior region of the mastoid process; it could be visible even in the occipital bone or at the occipito-mastoid suture.	x	x	Temporal foramen Sutural foramen Occipital foramen		(47)
Posterior condylar canal (24)	Occipital bone	Canal localized posteriorly to the occipital condyles.	x			Complete Blind	(47)
Secondary anterior condylar canal (25)	Occipital bone	Canal localized at the base of the occipital condyles.	x			Single Bipartite	(47)
Occipital condyle shape (26)	Occipital bone	Shape of the occipital condyles.				As classified in the reference	(48)
Reduced hypocone (27)	Molar teeth	Reduced size of the hypocone (disto-lingual cusp) of the superior molar teeth.	x			Reduced hypocone Normal hypocone	(49)
Crenulation (28)	Molar teeth	Curved fissures surrounding the primary ridges of the main cusps of the molar teeth.	x				(50)
Cribra Orbitalia (29)	Frontal bone	Porous hypertrophic lesions on the roof of the orbits	x				(46)

For each one, the evaluations performed are reported (P/A: presence/absence of the trait, number, position, and shape).

bones in 80% (16/20) and 5% (1/20) of the crania, respectively. In particular, in the maxillary bone, lesions were limited to the region close to the dental alveoli and, specifically, involved the entire dental socket when pathological signs were detected, or a small portion close to the root in taphonomic cases.

The 3D cranial models generally revealed a good reproduction of the bone preservation for the majority of the bones involved (Fig. 1). Fifty-six percent (5/9) of the selected bones, specifically the nasal, zygomatic, parietal, sphenoid, and palatine bones, showed a perfect representation of the “real” skulls (100% of the samples [20/20] showed the same appearance in both 3D models and the real crania). This number increased to 78% (7/9 of the selected bones), including the frontal and temporal bones, when window settings adjustments were carried out.

Differences between CT scans and crania were particularly important in the maxillary and occipital bones, which were only accurate in 5% (1/20) and 10% (2/20) of cases, respectively. These differences were related to bone loss visible in 3D models but not in the real crania and to bone loss detected in the real crania but not in the 3D models. In the former, bone loss was observed especially in the infraorbital region close to the infraorbital foramen (Fig. 2*a–c*), in dental alveoli (Fig. 2*a-c*), more frequently at the anterior teeth sockets, and in the occipital condyles (Fig. 2*d–f*). After window adjustment, image accuracy increased to 15% (3/20) and 80% (16/20) of the samples in the two regions, respectively.

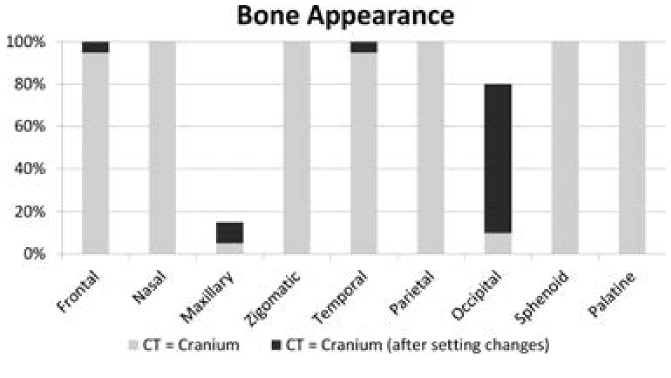


FIG. 1—*Bone appearance. Percentages of cases showing a perfect reproduction of each bone element before and after setting adjustments. [Color figure can be viewed at wileyonlinelibrary.com]*

Bone loss observed in the maxillary region of the real crania was detected in 3D models in 60% (21/35) of the affected dental sockets, showing a moderate reproduction of bone defects. Generally, difficulties arose for small bone defects which were not recognized in 3D models.

Morphological Variations Analysis

Skeletal variations were not identified in all crania. In particular, a frequency of less than 10 was observed for the following traits: *os japonicum*, squamous ossicle, coronal suture ossicles, sagittal suture ossicles, ossicle at lambda, *os inca*, parietal notch bone, *sutura mendosa*, occipito-mastoid ossicle, and secondary anterior condylar canal by both observers and ossicle at asterion, crenulation, and *cribra orbitalia* by one of the two observers. A summary of the traits observed on crania is reported in Table 2.

Since differences were detected in cranial observations between observers, interobserver agreement was assessed. High agreement between the two observers was noted for 69% of the traits (20/29), indicating a high reproducibility for most. In the remaining cases (supraorbital foramina, secondary infraorbital foramina, squamous ossicle, ossicle at asterion, mastoid foramen, posterior condylar canal, occipital condyle shape, reduced hypocone, and crenulation), interobserver reproducibility was

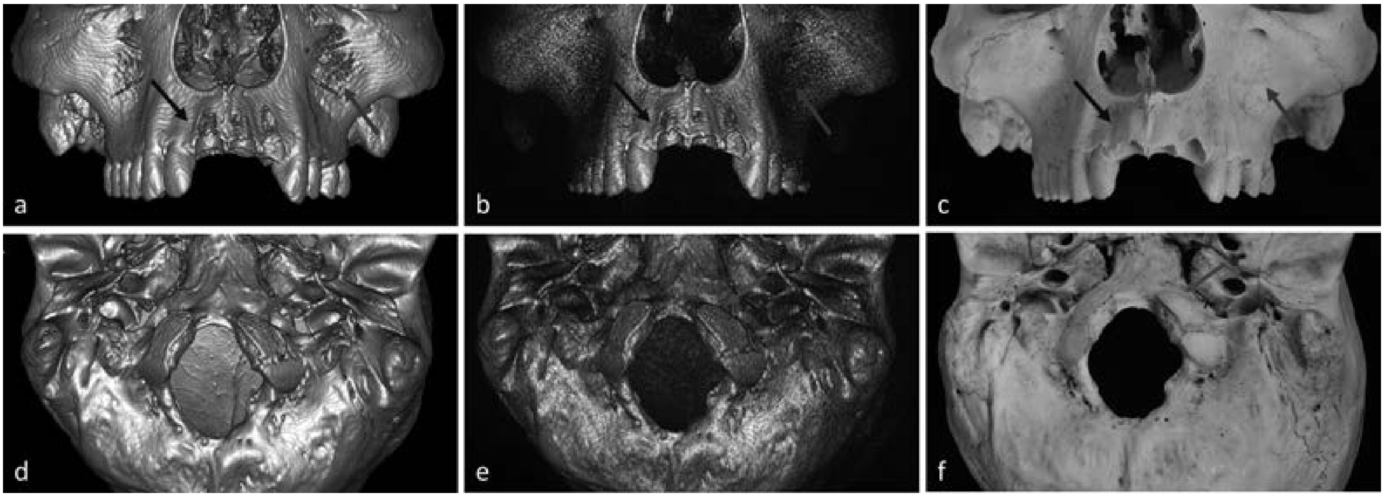


FIG. 2—*Bone appearance in CT scan. Differences between CT scan images before (a, d) and after (b, e) optimization of the images and skulls (c, f) in: (a–c) the region close to the infraorbital foramen (red arrow) and dental alveoli (blue arrow) and (d–f) the occipital condyles (green arrow). Figures (a) and (d) show bone loss in 3D models, not visible in the real crania, in the regions indicated by arrows. Bone loss disappeared after setting changes in the occipital condyles (e), while it persists at dental sockets and below the infraorbital foramen (b). [Color figure can be viewed at wileyonlinelibrary.com]*

moderate (7/29, 24%, $k = 0.41–0.60$) or lower (2/29, 7%, $k = 0–0.40$). All results are reported in Table 2.

Looking at CT data of each observer separately, a similar trend was detected. Around 30% of the traits were in fact identified correctly with high accuracy, 40% with intermediate accuracy, and 30% with low accuracy (28%, 38%, and 34% for observer A and 34%, 42%, and 24% for observer B). Circa 70% of the traits were therefore wrongly detected in the 3D reconstructions by both observers. Eleven traits were detected with different percentages: supraorbital notch, secondary zygomatic foramen, *os japonicum*, parietal notch bone, *sutura mendosa*, ossicle at asterion, occipito-mastoid ossicle, posterior condylar canal, secondary anterior condylar canal, occipital condyle shape, and *cribra orbitalia*. These variations may be explained by different detection rates in 3D models and/or in the real crania. However, since a disagreement was highlighted between observers, they were not included in the following considerations.

Considering the traits which were classified similarly (62%, 18/29), high accuracy was recorded in only 21% (6/29). Specifically, a perfect match (100% of the crania showing the trait correctly identified in 3D models) was detected for the sagittal suture ossicle, ossicle at lambda, *os inca,* and for normal traits (infraorbital and zygomatic foramina), while a lower percentage (86%–94%) was detected for the parietal foramen. The highest differences (traits with the lowest accuracy of identification in 3D models) were seen for the variations located on the lateral side of the cranium (epipteric bone, [Fig. 3*a,b*], squamous ossicles, and coronal suture ossicles) and for the infraorbital suture (Fig. 3*c,d*), mastoid foramen, and crenulation. The remaining traits (nasal foramen, supraorbital foramen, secondary infraorbital foramen, pterion shape, lambdoid suture ossicles, and reduced hypocone) showed intermediate accuracy with values ranging from 55% to 78%.

Analyzing the traits erroneously classified in 3D models, differences could be seen for each. Generally, an inability to detect the trait was identified for most, particularly for supernumerary ossicles, foramina, sutures (infraorbital suture, *os japonicum,* and *sutura mendosa*), *cribra orbitalia,* and dental traits. Concerning foramina and lambdoid suture ossicles, differences in the quantities were detected. Specifically, smaller or greater numbers of foramina/ossicles were identified on 3D models. Among the foramina, the mastoid foramen was the trait with the greatest variation recorded, both in number and in position (foramen located on the mastoid process, on the occipito-mastoid suture or on the occipital bone). Moreover, shape differences were also observed, in particular for traits described morphologically, such as pterion, posterior condylar canal, secondary anterior condylar canal, occipital condyles, and reduced hypocone. In addition, false positives (i.e., crania where traits were identified on 3D models but not in the real sample) were also detected. In particular, c*ribra orbitalia* was the variant with the greatest number of false positives by both observers, followed by the supraorbital foramen and dental traits.

All results are reported in Table 2 and Figs 4 and 5.

Discussion

The present study provided a general insight on how the human cranium and its variations appeared in Slicer 3D models based on CT images and how they differentiate from the real samples. For this, nine of the main bones forming the human cranium and twenty-nine of the most commonly observed cranial variations were considered in the analyses. Issues include the following: i) inaccurate reproduction of some cranial components; ii) difficulties in the detection of skeletal variants; and iii) erroneous identification of skeletal variants (false-positive results).

Over the last few years, a number of studies have described the accuracy of 3D models from CT scans and evaluated the reliability of radiological images as a substitute for skeletal collections in anthropological research (21–25). Among these, the main focus was metric analysis. No information was provided concerning the accuracy of the rendering of bone samples for morphological variations. Considering the important role of radiological techniques in forensic medicine and especially in personal identification, as well as the increasing use of 3D datasets in anthropological research, these data should be implemented. Indeed, if anatomical variations cannot be correctly identified on 3D models, should we continue to recommend the use of virtual osteological collections for anthropological research?

TABLE 2—*Summary of the observations and analyses performed in crania and CT scan images by the two observers.*

		Observer 1									Observer 2								
					Differences									Differences					
						Quantity									Quantity				
	Cohen's kappa	Crania/ Sides	CT = C	CT ≠ C	Position	>	<	0	Shape	False Positive	Crania/ Sides	CT = C	CT ≠ C	Position	>	<	0	Shape	False Positive
Nasal foramina (1)	0.76	32	25 (78%)	7 (22%)	1	2	1	3	0	2	32	20 (62%)	12 (38%)	0	5	4	3	0	1
Supraorbital foramina (2)	0.55*	17	11 (65%)	6 (35%)	0	3	1	2	0	9	25	16 (64%)	9 (36%)	0	7	0	2	0	5
Supraorbital notch (3)	0.65	21	15 (71%)	6 (29%)	0	0	0	6	0	2	24	20 (83%)	4 (17%)	0	0	0	4	0	3
Infraorbital suture (4)	0.90	23	6 (26%)	17 (74%)	0	0	0	17	0	1	23	11 (48%)	12 (52%)	0	0	0	12	0	2
Infraorbital foramina (5)	1.00	40	40 (100%)	0 (0%)	0	0	0	0	0	0	40	40 (100%)	0 (0%)	0	0	0	0	0	0
Secondary infraorbital foramina (6)	0.34*	14	10 (71%)	4 (29%)	0	0	0	4	0	4	21	12 (57%)	9 (43%)	0	4	1	4	0	4
Zygomatic foramina (7)	0.89	35	35 (100%)	0 (0%)	0	0	0	0	0	2	34	34 (100%)	0 (0%)	0	0	0	0	0	3
Secondary zygomatic foramina (8)	0.61	22	13 (59%)	9 (41%)	0	1	1	7	0	2	18	16 (89%)	2 (11%)	0	1	1	0	0	2
Os japonicum (9)	1.00	5	2 (40%)	3 (60%)	0	0	0	3	0	0	5	3 (60%)	2 (40%)	0	0	0	2	0	0
Pterion shape (10)	0.77	40	23 (58%)	16 (40%)†	0	0	0	0	16	0	40	24 (60%)	16 (40%)	0	0	0	0	16	0
Epipteric bone (11)	0.96	16	4 (25%)	12 (75%)	1	0	0	11	0	1	16	5 (31%)	11 (69%)	2	0	0	9	0	2
Squamous ossicles (12)	0.09*	2	0 (0%)	2 (100%)	0	0	0	2	0	0	7	0 (0%)	7 (100%)	0	0	0	7	0	2
Coronal suture ossicles (13)	0.64	4	1 (25%)	3 (75%)	0	0	0	3	0	0	5	1 (20%)	4 (80%)	0	0	0	4	0	1
Parietal foramen (14)	0.68	31	29 (94%)	2 (6%)	0	0	1	1	0	2	29	25 (86%)	4 (14%)	0	1	2	1	0	2
Sagittal suture ossicles (15)	0.85	2	2 (100%)	0 (0%)	0	0	0	0	0	0	2	2 (100%)	0 (0%)	0	0	0	0	0	0
Ossicle at lambda (16)	1.00	5	5 (100%)	0 (0%)	0	0	0	0	0	0	5	5 (100%)	0 (0%)	0	0	0	0	0	0
Os inca (17)	0.71	4	4 (100%)	0 (0%)	0	0	0	0	0	0	4	4 (100%)	0 (0%)	0	0	0	0	0	0
Lambdoid suture ossicles (18)	0.74	21	16 (76%)	5 (24%)	0	1	4	0	0	1	23	14 (61%)	9 (39%)	0	2	6	1	0	3
Parietal notch bone (19)	0.73	8	4 (50%)	4 (50%)	0	0	0	4	0	0	5	4 (80%)	1 (20%)	0	0	0	1	0	0
Sutura mendosa (20)	0.79	3	2 (67%)	1 (33%)	0	0	0	1	0	1	2	2 (100%)	0 (0%)	0	0	0	0	0	1
Ossicle at asterion (21)	0.47*	8	5 (62%)	3 (38%)	0	0	0	3	0	0	12	6 (50%)	6 (50%)	0	0	0	6	0	1
Occipito-mastoid ossicle (22)	0.70	6	5 (83%)	1 (17%)	0	0	0	1	0	3	8	5 (62%)	3 (38%)	0	0	0	3	0	0
Mastoid foramen (23)	0.41*	35	14 (40%)	21 (60%)	6	7	9	1	0	3	37	10 (27%)	27 (73%)	13	19	4	1	0	2
Posterior condylar canal (24)	0.55*	31	25 (81%)	6 (19%)	0	0	0	1	5	2	30	18 (60%)	12 (40%)	0	0	0	3	9	1
Secondary anterior condylar canal (25)	0.75	8	6 (75%)	2 (25%)	0	0	0	0	2	3	7	7 (100%)	0 (0%)	0	0	0	0	0	3
Occipital condyle shape (26)	0.51*	40	17 (43%)	23 (57%)	0	0	0	0	23	0	40	21 (52%)	19 (48%)	0	0	0	0	19	0
Reduced hypocone (27)	0.48*	22	12 (55%)	9 (41%)†	0	0	0	4	5	8	15	9 (60%)	6 (40%)	0	0	0	5	1	4
Crenulation (28)	0.52*	12	3 (25%)	6 (50%)†	0	0	0	6	0	4	6	2 (33%)	2 (33%)†	0	0	0	2	0	7
Cribra orbitalia (29)	0.63	13	4 (31%)	9 (69%)	0	0	0	9	0	10	9	5 (56%)	4 (44%)	0	0	0	4	0	8

Interobserver agreement calculated on the data recorded on crania are shown in the first column ("Cohen's kappa"). In the following, the number of sides/molar teeth showing the traits in crania (Crania/Sides) and number and percentage of times in the corresponding 3D models where the traits were (CT = C) and were not (CT ≠ C) identified are reported. Percentages were calculated based on the observations reported in the column "Crania/Sides." Dental traits were evaluated for each molar tooth available and the percentage was calculated on the base of the number of positive molar teeth present in each cranium. Differences highlighted are summarized depending on the position, quantity (greater:>, lower: <, not detected: 0) and shape. False-positive 3D models are also reported. For morphological variations (pterion and occipital condyle shape), the number of crania (i.e., 40) corresponds to the number of portions evaluable in the real samples.

*The trait has an interobserver agreement equal or lower than 0.60, classified as moderate (0.41–0.60), fair (0.21–0.40) or slight (0.00–0.20) by (53).

†The trait in some samples was not determinable in 3D reconstructions and the number of 3D models which differed from the real crania was therefore lower than the expected).

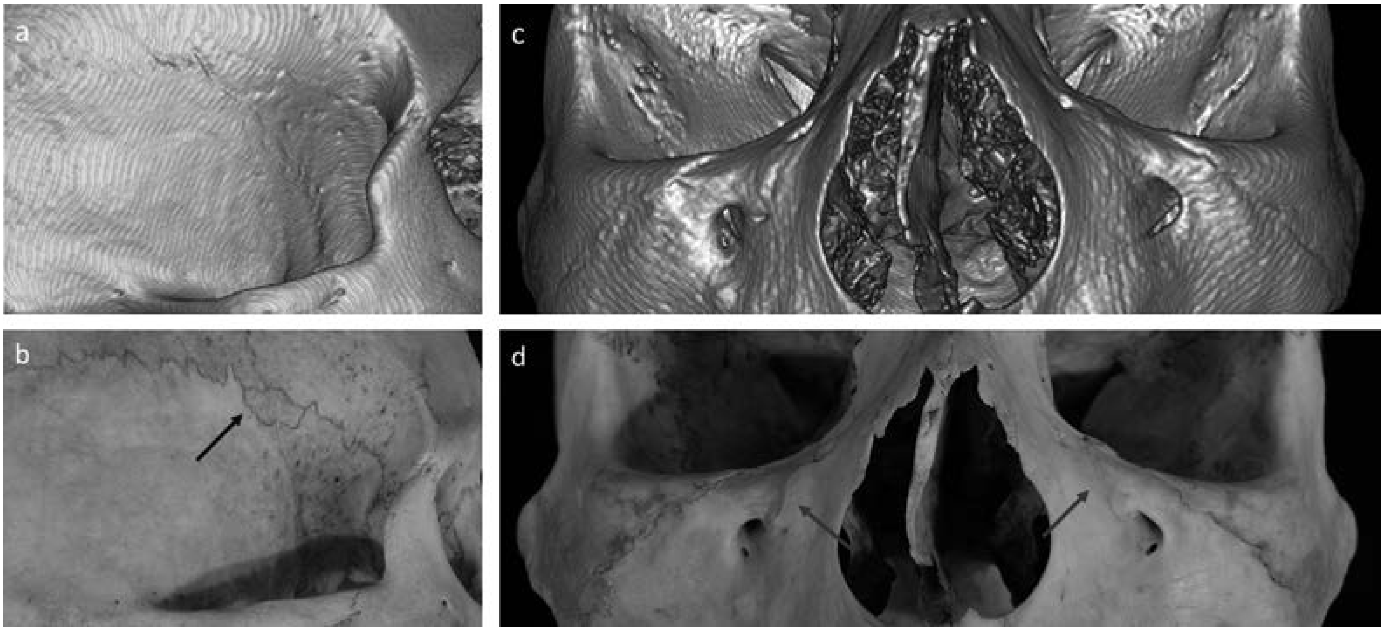


FIG. 3—*Skeletal anomalies identification in CT scan. Differences between CT scan images and the "real" skulls in: (a-b) the epipteric bone in the right side (blue arrow), that was difficult to detect in the corresponding 3D model; (c-d) the infraorbital suture, clearly visible in bone in both sides (red arrows) but poorly defined in the right side and not visible in the left side of the corresponding 3D model. [Color figure can be viewed at wileyonlinelibrary.com]*

The analysis performed on bone appearance showed a generally good reproduction of the bone samples. However, incompleteness and difficulties in morphological investigation were highlighted for specific cranial regions, causing confusion during 3D model examination. Pitfalls arose for the areas characterized by a thin layer of compact bone, specifically the maxillary bone and occipital condyles, where a loss of bone substance was detected in the 3D models not present in the real samples. These signs could be confused with taphonomical lesions or bone resorption related to pathological conditions

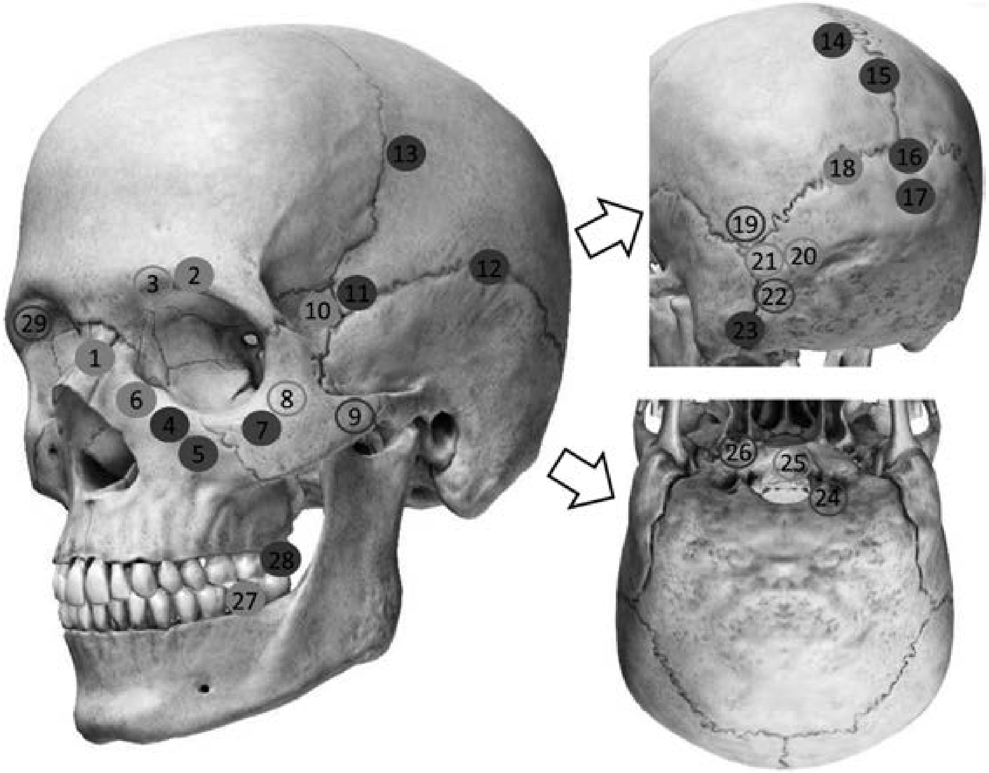


FIG. 4—*Success of identification of anatomical variations in CT images compared to crania in Observer 1. Numbers refer to anatomical variations listed in Table 1. Green: 81%–100%; orange: 51%–80%; red: 0%–50%. [Color figure can be viewed at wileyonlinelibrary.com]*

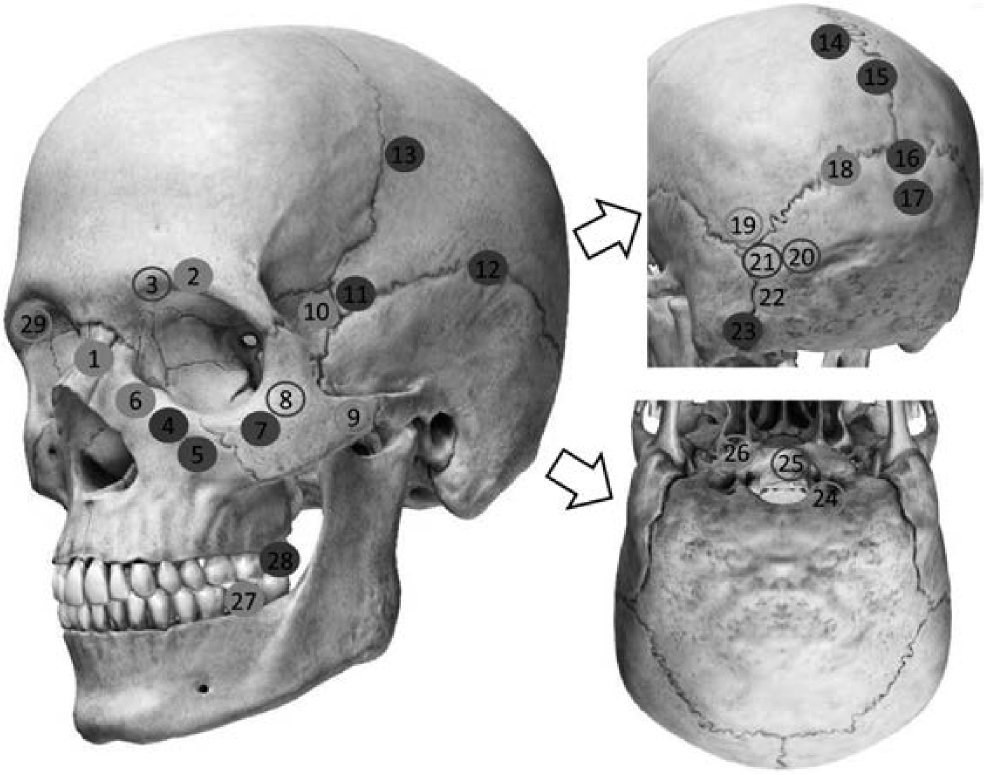


FIG. 5—*Success of identification of anatomical variations in CT images compared to crania in Observer 2. Numbers refer to anatomical variations listed in Table 1. Green: 81%–100%; orange: 51%–80%; red: 0%–50%. [Color figure can be viewed at wileyonlinelibrary.com]*

(the latter especially on alveolar sockets), which could lead to incongruous results in a forensic identification scenario. The possibility to optimize the images by lowering the window level cannot be considered a reliable solution to assess whether bone loss was really present. In fact, this adjustment failed in the majority of the 3D models for regions where loss was detected (especially in the maxillary bones). In addition, the difficulty to detect actual small bone defects around the dental alveoli on CT reconstruction suggests that particular attention should be paid to the evaluation of 3D models in the maxillary region.

Focusing on skeletal variants, the blind test showed many difficulties in the ability to detect and correctly classify skeletal anomalies on 3D images. Only about 30% of the traits were correctly identified, and 21% were correctly scored by both anthropologists. These corresponded to the variants with medium–large size (i.e., parietal foramen, sagittal suture ossicle, ossicle at lambda, and *os inca*) and to normal traits (infraorbital and zygomatic foramina), that appeared more marked and clearly defined in the radiological records. A high percentage of traits was instead classified with intermediate and low accuracy, the latter with a percentage similar to the first category (21%). No relation was observed between the category of the nonmetric trait and its difficulty of identification on 3D models, since an inability to detect the trait was seen for most of the variants: supernumerary ossicles, foramina, sutures, *cribra orbitalia*, and dental traits. This inability is more likely related to the reduced size of these traits (in particular for small ossicles and foramina) and their location (particularly on the lateral side). Additional problems concerned the detection of erroneous positions and numbers in foramina and lambdoid suture ossicles and the incorrect identification of the shape in morphological variations. Low accuracy was noted for the traits positioned on the lateral side. In addition, the infraorbital suture, mastoid foramen, and crenulation collectively corresponded to the most challenging traits to identify, as they were correctly detected in less than 50% of the positive samples.

To exclude any difference between observers' evaluations, interobserver reproducibility was tested on crania. Results clearly support CT inaccuracy for those cases where high agreement was reached, and incorrect evaluations were similarly performed by both observers on 3D models. However, even the traits with low agreement but categorized similarly during crania-CT comparisons provided interesting data on CT accuracy. In fact, in such cases, despite the need to reassess and to standardize the technical method of scoring, both observers had similar difficulties in trait detection on 3D models. In particular, these difficulties may be more likely ascribed to the inaccuracy of the CT images than to possible scoring issues. In fact, looking at the differences between crania and CT images, an inability to identify the traits was highlighted, especially in two cases where a frequency three times lower (crenulation) and even null (squamous ossicle) than with the real crania was detected. In addition, the loss of bone substance recorded in thin cranial bones and misleading alterations of bone components may be the reasons behind the lower number of foramina correctly detected. This was evident especially for secondary infraorbital foramina, for which the loss of bone substance in the infraorbital region may have made the correct identification difficult, and for supraorbital and mastoid foramina, for which the misleading changes in pore size may have been confused with real foramina. All these observations therefore supported the inaccuracy of CT images and their unreliability for correctly identifying nonmetric traits (and even pathology) in radiological images even in those cases where low interobserver agreement was detected.

The results of our study recommend caution in cases where CT scan images and reconstructions are considered. In particular, the absence of a trait in a radiological image may not signify its absence in the real sample. Similarly, considerations have to be made in cases of the presence of small traits such as foramina, small sutures, and ossicles. We showed that some traits were wrongly detected on 3D images of crania considered negative by the same observer (false-positive crania). Such data demonstrated the low reliability of 3D reconstructions from CTs, in some cases. Therefore, caution is needed during the collection of population data from virtual skeletal collections and the comparisons between crania and antemortem CT reconstructions for identification purposes. In such cases, erroneous analyses would cause the creation of false reference data or could lead to incorrect judgments during personal identification. However, future improvement of technologies and software may provide superior image quality, thus enhancing the identification of nonmetric variants.

Our results were different to those summarized in the previous studies where CTs were considered an acceptable source for anthropological research (21–23). This difference may be related to the main topics of the analyses since the latter focused solely on metric investigation and no morphological evaluation was included.

The present study has limitations. 3D images came from CT scans of dry bones only, and no comparisons were carried out using CTs from bones covered by soft tissue to simulate clinical data. Since acquisitions were performed using more radiations than clinical CTs and filters specific for bone samples, images had a better quality and higher resolution. Therefore, this study described the scenario in the best conditions and worse results could be expected with clinical data. This was supported by the conclusions outlined by Colmar et al. (25) who showed that clinical CTs are less accurate than dry CTs and optical 3D models in *ossa coxae*. An additional limitation lies in the uneven numbers of positive and negative crania for each trait. Although no comparison could be performed among traits, this study provided a general overview on the difficulties and the accuracy of 3D models in nonmetric traits analysis.

In conclusion, the present study evaluated the accuracy of Slicer 3D models coming from dry CT scans of human crania compared to the real bone samples, providing data on how crania and morphological variations compare to radiological images. The results highlight some pitfalls in CT scan evaluation, especially difficulties in correctly detecting nonmetric traits in 3D reconstructions from CT scans. In the future, different software for 3D reconstruction may shift whether the use of different algorithms can influence the accuracy of 3D volume-rendered images. In the meantime, the data obtained in this study suggest interpretations and analyses from 3D models should be considered with care to avoid erroneous conclusions, such as in population studies of anatomical variants collected from radiological records or in the standardization of anthropological methods for the analysis of real cases.

References

1. Greulich WW, Pyle SI. Radiographic atlas of skeletal development of the hand and wrist. Redwood City, CA: Stanford, University Press, 1959.
2. Kvaal SI, Kolltveit KM, Thomsen IO, Solheim T. Age estimation of adults from dental radiographs. Forensic Sci Int 1995;74(3):175–85. https://doi.org/10.1016/0379-0738(95)01760-g
3. Schmeling A, Shulz R, Reisinger W, Mulher M, Wernecke KG, Geserick G. Studies on the time frame for ossification of the medial clavicular epiphyseal cartilage in conventional radiography. Int J Legal Med 2004;118(1):5–8. https://doi.org/10.1007/s00414-003-0404-5
4. Cameriere R, Ferrante L, Belcastro MG, Bonfiglioli B, Rastelli E, Cingolani M. Age estimation by pulp/tooth ratio in canines by peri-apical X-rays. J Forensic Sci 2007;52(1):166–70. https://doi.org/10.1111/j.1556-4029.2006.00336.x
5. Offiah A, van Rijn RR, Perez-Rossello JM, Kleinman PK. Skeletal imaging of child abuse (non-accidental injury). Pediatr Radiol 2009;39(5):461–70. https://doi.org/10.1007/s00247-009-1157-1
6. AlQahtani SJ, Hector MP, Liversidge HM. Brief communication: the London atlas of human tooth development and eruption. Am J Phys Anthropol 2010;142(3):481–90. https://doi.org/10.1002/ajpa.21258
7. Matteoli M, Piacentino D, Kotzalidis GD, Serata D, Rapinesi C, Angeletti G, et al. The clinical and radiological examination of acute intimate partner violence injuries: a retrospective analysis of an Italian cohort of women. Violence Vict 2016;31(1):85–102. https://doi.org/10.1891/0886-6708.VV-D-14-00107
8. Russo A, Reginelli A, Pignatiello M, Cioce F, Mazzei G, Fabozzi O, et al. Imaging of violence against the elderly and the women. Semin Ultrasound CT MR 2019;40(1):18–24. https://doi.org/10.1053/j.sult.2018.10.004
9. Dirnhofer R, Jackowski C, Vock P, Potter K, Thali MJ. VIRTOPSY: minimally invasive, imaging-guided virtual autopsy. Radiographics 2006;26(5):1305–33. https://doi.org/10.1148/rg.265065001
10. Bolliger SA, Thali MJ. Imaging and virtual autopsy: looking back and forward. Philos Trans R Soc Lond B Biol Sci 2015;370(1674):20140253. https://doi.org/10.1098/rstb.2014.0253
11. Decker SJ, Braileanu M, Dey C, Lenchik L, Pickup M, Powell J, et al. Forensic radiology: a primer. Acad Radiol 2019;26(6):820–30. https://doi.org/10.1016/j.acra.2019.03.006
12. Decker SJ, Davy-Jow SL, Ford JM, Hilbelink DR. Virtual determination of sex: metric and nonmetric traits of the adult pelvis from 3D computed tomography models. J Forensic Sci 2011;56(5):1107–14. https://doi.org/10.1111/j.1556-4029.2011.01803.x
13. Franklin D, Flavel A, Kuliukas A, Cardini A, Marks MK, Oxnard C, et al. Estimation of sex from sternal measurements in a Western Australian population. Forensic Sci Int 2012;217(1–3):230.e1–5. https://doi.org/10.1016/j.forsciint.2011.11.008
14. Franklin D, Cardini A, Flavel A, Marks MK. Morphometric analysis of pelvic sexual dimorphism in a contemporary Western Australian population. Int J Legal Med 2014;128(5):861–72. https://doi.org/10.1007/s00414-014-0999-8
15. De Angelis D, Gibelli D, Gaudio D, Cipriani Noce F, Guercini N, Varvara G, et al. Sexual dimorphism of canine volume: a pilot study. Leg Med (Tokyo) 2015;17(3):163–6. https://doi.org/10.1016/j.legalmed.2014.12.006
16. Ekizoglu O, Inci E, Palabiyik FB, Can IO, Er A, Bozdag M, et al. Sex estimation in a contemporary Turkish population based on CT scans of the calcaneus. Forensic Sci Int 2017;279:310.e1–6. https://doi.org/10.1016/j.forsciint.2017.07.038
17. Wink AE. Pubic symphyseal age estimation from three-dimensional reconstructions of pelvic CT scans of live individuals. J Forensic Sci 2014;59(3):696–702. https://doi.org/10.1111/1556-4029.12369
18. Blaszkowska M, Flavel A, Franklin D. Validation of the İşcan method in clinical MSCT scans specific to an Australian population. Int J Legal Med 2019;133(6):1903–13. https://doi.org/10.1007/s00414-018-01992-0
19. Verna E, Piercecchi-Marti MD, Chaumoitre K, Bartoli C, Leonetti G, Adalian P. Discrete traits of the sternum and ribs: a useful contribution to identification in forensic anthropology and medicine. J Forensic Sci 2013;58(3):571–7. https://doi.org/10.1111/1556-4029.12111
20. Verna E, Piercecchi-Marti MD, Chaumoitre K, Adalian P. Relevance of discrete traits in forensic anthropology: from the first cervical vertebra to the pelvic girdle. Forensic Sci Int 2015;253:134.e1–7. https://doi.org/10.1016/j.forsciint.2015.05.005
21. Lopes PM, Moreira CR, Perrella A, Antunes JL, Cavalcanti MG. 3-D volume rendering maxillofacial analysis of angular measurements by multislice CT. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 2008;105(2):224–30. https://doi.org/10.1016/j.tripleo.2007.08.036
22. Franklin D, Cardini A, Flavel A, Kuliukas A, Marks MK, Hart R, et al. Concordance of traditional osteometric and volume-rendered MSCT

interlandmark cranial measurements. Int J Legal Med 2013;127(2):505–20. https://doi.org/10.1007/s00414-012-0772-9

23. Stull KE, Tise ML, Ali Z, Fowler DR. Accuracy and reliability of measurements obtained from computed tomography 3D volume rendered images. Forensic Sci Int 2014;238:133–40. https://doi.org/10.1016/j.forsciint.2014.03.005
24. Colman KL, Dobbe JGG, Stull KE, Ruijter JM, Oostra RJ, van Rijn RR, et al. The geometrical precision of virtual bone models derived from clinical computed tomography data for forensic anthropology. Int J Legal Med 2017;131(4):1155–63. https://doi.org/10.1007/s00414-017-1548-z
25. Colman KL, de Boer HH, Dobbe JGG, Liberton NPTJ, Stull KE, van Eijnatten M, et al. Virtual forensic anthropology: the accuracy of osteometric analysis of 3D bone models derived from clinical computed tomography (CT) scans. Forensic Sci Int 2019;304:109963. https://doi.org/10.1016/j.forsciint.2019.109963
26. Christensen AM, Passalacqua NV, Bartelink EJ. Forensic anthropology: current methods and practice. San Diego, CA: Academic Press, 2014;301–39.
27. Sekharan PC. Identification of skull from its suture pattern. For Sci Int 1985;27(3):205–14. https://doi.org/10.1016/0379-0738(85)90156-2
28. Sekharan PC. Personal identification from skull suture pattern. J Can Soc Forensic Sci 1989;22(1):27–34. https://doi.org/10.1080/00085030.1989.10757416
29. Kahana T, Hiss J. Positive identification by means of trabecular bone pattern comparison. J Forensic Sci 1994;39(5):1325–30. https://doi.org/10.1520/JFS13720J
30. Valenzuela A. Radiographic comparison of the lumbar spine for positive identification of human remains. A case report. Am J Forensic Med Pathol 1997;18(2):215–7. https://doi.org/10.1097/00000433-199706000-00024
31. Kahana T, Hiss J, Smith P. Quantitative assessment of trabecular bone pattern identification. J Forensic Sci 1998;43(6):1144–7. https://doi.org/10.1520/JFS14377J
32. Mann RW. Use of bone trabeculae to establish positive identification. Forensic Sci Int 1998;98(1–2):91–9. https://doi.org/10.1016/s0379-0738(98)00138-8
33. Smith DR, Limbird KG, Hoffman JM. Identification of human skeletal remains by comparison of bony details of the cranium using computerized tomographic (CT) scans. J Forensic Sci 2002;47(5):937–9. https://doi.org/10.1520/JFS15499J
34. Rogers TL, Allard TT. Expert testimony and positive identification of human remains through cranial suture patterns. J Forensic Sci 2004;49(2):203–7. https://doi.org/10.1520/JFS2003095
35. Christensen AM. Testing the reliability of frontal sinuses in positive identification. J Forensic Sci 2005;50(1):18–22. https://doi.org/10.1520/JFS2004145
36. Pfaefli M, Vock P, Dirnhofer R, Braun M, Bolliger SA, Thali MJ. Postmortem radiological CT identification based on classical ante-mortem X-ray examinations. Forensic Sci Int 2007;171(2–3):111–7. https://doi.org/10.1016/j.forsciint.2006.10.009
37. Smith VA, Christensen AM, Myers SW. The reliability of visually comparing small frontal sinuses. J Forensic Sci 2010;55(6):1413–5. https://doi.org/10.1111/j.1556-4029.2010.01493.x
38. Stephan CN, Winburn AP, Christensen AF, Tyrrell AJ. Skeletal identification by radiographic comparison: blind tests of a morphoscopic method using antemortem chest radiographs. J Forensic Sci 2011;56(2):320–32. https://doi.org/10.1111/j.1556-4029.2010.01673.x
39. Quatrehomme G, Biglia E, Padovani B, du Jardin P, Alunni V. Positive identification by X-rays bone trabeculae comparison. Forensic Sci Int 2014;245:e11–4. https://doi.org/10.1016/j.forsciint.2014.09.019
40. Mowafey B, Van de Casteele E, Youssef JM, Zaher AR, Omar H, Politis C, et al. Can mandibular lingual canals be used as a forensic fingerprint? J Forensic Odontostomatol 2015;33(2):26–35
41. De Angelis D, Gibelli D, Palazzo E, Sconfienza L, Obertova Z, Cattaneo C. Skeletal idiopathic osteosclerosis helps to perform personal identification of unknown decedents: a novel contribution from anatomical variants through CT scan. Sci Justice 2016;56(4):260–3. https://doi.org/10.1016/j.scijus.2016.03.003
42. Cappella A, Gibelli D, Cellina M, Mazzarelli D, Oliva AG, De Angelis D, et al. Three-dimensional analysis of sphenoid sinus uniqueness for assessing personal identification: a novel method based on 3D–3D superimposition. Int J Legal Med 2019;133(6):1895–901. https://doi.org/10.1007/s00414-019-02139-5
43. Jayaprakash PT. Skull sutures: radiographic contour of Wormian bone as an individualising epigenetic marker. J Can Soc Forensic Sci 1997;30(2):39–47. https://doi.org/10.1080/00085030.1997.10757085
44. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J-C, Pujol S, et al. 3D slicer as an image computing platform for the quantitative imaging network. Magn Reson Imaging 2012;30(9):1323–41. https://doi.org/10.1016/j.mri.2012.05.001
45. 3D Slicer. www.slicer.org (accessed July 17, 2020).
46. Ortner DJ. Identification of pathological conditions in human skeletal remains, 2nd edn. San Diego, CA: Academic Press, 2003;102–5.
47. Hauser G, De Stefano GF. Epigenetic variants of the human skull. Stuttgart, Germany: Schweizerbart, 1989;50–226.
48. Naderi S, Korman E, Citak G, Güvençer M, Arman C, Senoğlu M, et al. Morphometric analysis of human occipital condyle. Clin Neurol Neurosurg 2005;107(3):191–9. https://doi.org/10.1016/j.clineuro.2004.07.014
49. Scott GR, Irish JD. Human tooth crown and root morphology. New York, NY: Cambridge University Press, 2017;89–94.
50. Pilloud MA, Maier C, Scott GR, Edgar HJH. Molar crenulation trait definition and variation in modern human populations. Homo 2018;69(3):77–85. https://doi.org/10.1016/j.jchb.2018.06.001
51. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Measur 1960;20(1):37–46. https://doi.org/10.1177/001316446002000104
52. VassarStats. http://vassarstats.net (accessed July 17, 2020).
53. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–74. https://doi.org/10.2307/2529310

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14520
Available online at: onlinelibrary.wiley.com

TECHNICAL NOTE

ANTHROPOLOGY

Victoria M. Dominguez,[1,2,3,4,†] *Ph.D.; Angela L. Harden*,[4,†] *M.A.; Matthew Wascher*,[5] *M.A.; and Amanda M. Agnew*,[4] *Ph.D.*

Rib Variation at Multiple Locations and Implications for Histological Age Estimation*

ABSTRACT: Existing histological age estimation methods using the rib were developed mainly from the midshaft; however, in forensic practice, uncertainty of sampling location often arises due to fragmented or previously sampled ribs. The potential for error increases when sampling location is uncertain and utilizing a section beyond the midshaft (either anterior or posterior) may result in erroneous age estimates. Additionally, there is debate within the field regarding the minimum number of sections needed for accurate age estimation. The aim of this research is to determine the importance of the midshaft distinction for age-at-death assessment and the necessity of analyzing serial sections by evaluating histological variables at sampling locations along the length of the rib. Three seriated histological sections at three sampling locations (anterior, midshaft, and posterior) were obtained from sixth ribs of ten postmortem human subjects. Cortical area (Ct.Ar) and osteon population density (OPD) were collected from each section (n = 90). Significant differences were determined in Ct.Ar between sampling locations, demonstrating the variation present along the length of the rib. A comparison of OPD at sampling locations revealed significant differences, suggesting that sampling site is critical to accurate age estimates. When sampling location is uncertain, a more anterior section should be taken. Analysis of serial sections within locations revealed no significant differences in OPD or Ct.Ar, supporting the practice of collecting data from one section for age estimation. While an age estimate can be achieved through the analysis of one section, best practice suggests reading two sections to capture intraindividual variation.

KEYWORDS: age estimation, rib histomorphometry, osteon population density, OPD, cortical area, skeletal aging, forensic anthropology

Bone histology methods for estimating age-at-death generate reliable data that contribute to the biological profile both independently, as well as integrated with more traditional gross morphological approaches. The ability of these techniques to accurately predict age depends in large part on being able to correctly identify and section sampling locations, which can be difficult in forensic cases with fragmented or previously sectioned skeletal material. When sampling location is uncertain, the potential for error increases, but much of this uncertainty can be relieved by developing greater knowledge of microstructural variability within the skeleton.

Among the most commonly used elements in histological age analysis are the ribs, specifically mid-thoracic ribs when present. Mid-thoracic ribs are similarly cyclically loaded in respiration in all humans and are more biomechanically constrained than elements of the appendicular skeleton, providing a natural control for load variation between people (1). Additionally, the smaller cross-sectional area of the rib relative to other long bones (e.g., the femur) allows analysis of complete cross sections, reducing the likelihood of sampling error (2). Perhaps the best known method of rib histomorphometric aging is that developed by Stout (3) and subsequently revised a number of times by Stout et al. over the years (4–6). This method utilizes the midshaft region of the sixth rib, examining a complete cross section and quantifying osteon population density (OPD) through linear regression to generate age estimates. Despite the utility of this method, among its limitations is the fact that it is not always possible to reliably determine which rib is the sixth rib or what precisely constitutes the midshaft, particularly in the case of damaged or otherwise fragmented remains.

Previous work has addressed these questions at least in part. In an exploratory study, Crowder and Rosella (7) examined the accuracy of using the middle third of nonsixth, mid-thoracic ribs for estimating age and validated the assumption that the middle third of ribs 3–8 could be used when the sixth rib cannot be reliably identified. In addition to rib level, spatial variation in bone microstructure is accounted for by consistent sampling at the midshaft. However, how precise the midshaft designation needs to be remains uncertain. Stout's (4) original method drew its samples from the "middle third" of the sixth rib, which in an average adult can easily be a region several centimeters in length and thus prone to great microstructural variability considering

[1] Department of Anthropology, Lehman College, City University of New York, Bronx, NY, 10468.

[2] Department of Anthropology, The Graduate Center, City University of New York, New York, NY, 10016.

[3] New York Consortium of Evolutionary Primatology, New York, NY.

[4] Skeletal Biology Research Laboratory, Injury Biomechanics Research Center, The Ohio State University, Columbus, OH, 43210.

[5] Department of Statistics, The Ohio State University, Columbus, OH, 43210.

Corresponding author: Victoria M. Dominguez, Ph.D. E-mail: victoria.dominguez@lehman.cuny.edu

*Presented in part at the 70th Annual Scientific Meeting of the American Academy of Forensic Sciences, February 18-23, 2018, in Seattle, WA.

†Authors contributed equally.

[Correction added 11 August 2020. The Highlights section was removed.]

Received 21 May 2020; and in revised form 30 June 2020; accepted 6 July 2020.

that undecalcified bone sections are typically less than 130 microns in thickness. In other words, a section taken at the anterior-midshaft may or may not produce dramatically different results from those taken at the posterior-midshaft. Though significant variability along the length of the rib has been observed in cross-sectional geometry (8), microstructural differences, particularly in remodeling, have been poorly documented along the rib shaft and merit further attention.

There is one more important consideration when sampling: How much bone must be examined to account for incoherence? Incoherence refers to the random variation between serial sections within an individual sample, variation that is typically greater in smaller cross sections (such as the rib) than in larger cross sections (such as the femur) (2). Frost (9) proposed a minimum of 50 mm^2 of cortex be measured to account for this variance, which in the case of some individuals, such as small, osteoporotic females might require as many as five complete cross sections to be analyzed. Stout and Paine's (4) recommendation is to sample two entire cross sections, regardless of bone size, to account for this intraskeletal variance. This latter recommendation is echoed in Crowder and Rosella (7), though they acknowledge that while useful for examining remodeling variability, evaluating multiple sections may not improve age estimates. In practice, due to feasibility constraints, particularly the time-consuming nature of these analyses, sometimes only a single cross section, regardless of size, may be available for estimating age-at-death and the reliability of such estimates must be examined.

The goals of this study are twofold. First, we assess variability in commonly applied histological variables (number of intact and fragmentary osteons and OPD) and cortical area at three sampling locations along the rib shaft (i.e., anterior, midshaft, and posterior). Second, a set of serial sections from each location is assessed for incoherence variation. Our aim is to develop and provide the information necessary to develop specific best practice recommendations for sampling fragmented or damaged midthoracic ribs for use in histological aging techniques.

Materials and Methods

A single left or right sixth human rib specimen was ethically obtained from ten (5 female, 77–92 years old, mean = 85.6 years, SD = 6.01 years; 5 male, 76–90 years old, mean = 83.8 years, SD = 4.91 years) postmortem human subjects through The Ohio State Whole Body Donation Program. All bone samples were derived from fresh cadaveric material, and soft tissue was removed through a low-temperature maceration process. Two-centimeter transverse sections were taken from each rib at three specific measured sampling locations: posterior (25%), midshaft (50%), and anterior (75%) (Fig. 1). Each undecalcified section was embedded in methyl methacrylate and sectioned on a diamond wire saw (Delaware Diamond Knives, DE) at a thickness of ~70 μm. Three seriated sections were obtained from each rib at the three previously identified sampling locations, resulting in a total of nine sections from each individual (90 sections in total; Fig. 1). All thin sections were then mounted with Eukitt mounting medium and cover-slipped on glass slides. Prepared cross sections were photographed at 100× magnification on an Olympus VS120 slide scanner in brightfield and polarized light (Fig. 1). Four histological and three cross-sectional variables were quantified and analyzed for this study (Table 1).

Cross-sectional areas, specifically total area (Tt.Ar) and endosteal area (Es.Ar), were collected via semi-automated methods (10) by one qualified observer (VMD) using ImageJ software (11). Serial sections were assessed independently from one

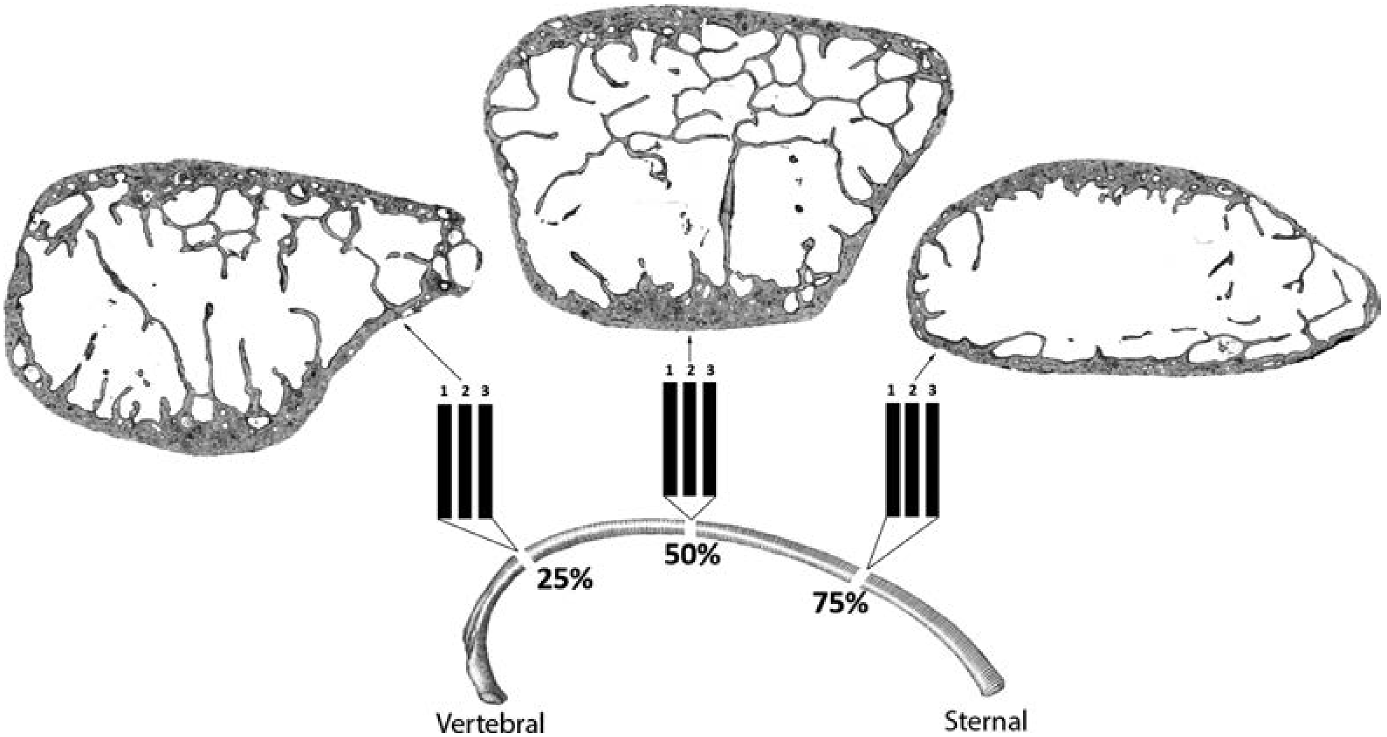


FIG. 1—*Sampling locations along the length of the rib, schematic of seriated sections, and exemplar cross-sectional images for each location (posterior = 25%, midshaft = 50%, anterior = 75%). Cutaneous for all images is oriented to the top and pleural to the bottom. [Color figure can be viewed at wileyonlinelibrary.com]*

TABLE 1—*Histological variables.*

	Abbreviation and Units	Definition and Formula	
Histological Variables			
Intact Osteon Number	On.N	The number of intact osteons*	
Fragmentary Osteon Number	Fg.On.N	The number of fragmentary osteons†	
Surface Area	Sa.Ar	Actual area of bone evaluated per microstructures per section	
Osteon Population Density	OPD	Sum of intact and fragmentary osteons visible per unit area	$\frac{\mathrm{On.N+Fg.On.N}}{\mathrm{Sa.Ar}}$
Cross-Sectional Variable			
Total Area	Tt.Ar (mm^2)	Area below the periosteum, including medullary cavity	
Endosteal Area	Es.Ar (mm^2)	Area below the endosteum, including trabecular area and medullary cavity	
Cortical Area	Ct.Ar (mm^2)	All cortical bone area including pores	Tr.Ar – Es.Ar

*Intact osteon: A secondary osteon with an intact Haversian canal bounded by a reversal line (17)

†Fragmentary osteon: A secondary osteon with a partially visible Haversian canal that has been breached either by a neighboring osteon or a resorptive bay or a secondary osteon with no remnants of a Haversian canal present (17).

another when defining cortical borders. The number of intact osteons (On.N), number of fragmentary osteons (Fg.On.N), and surface area (Sa.Ar) were manually counted live from an Olympus BX-63 microscope, fitted with a standard Merz counting reticule, using brightfield and polarized light at 200× magnification, following the methods of Stout and Paine (1992). OPD was calculated as shown in Table 1. Normal theory-based analysis of variance (ANOVA) *F* tests for OPD and Kruskal–Wallis tests for Ct.Ar (which was not normally distributed) were performed to examine the relationships between section location (anterior, middle, or posterior) within the entire sample, males only, females only, and between sexes. Repeated-measures ANOVA tests were performed to examine the relationships between serial sections for OPD and Ct.Ar. In order to have an overall type I error rate of 0.05, a Bonferroni correction for multiple comparisons was made (0.05/35 ≈0.00143) and the adjusted significance level (0.001) was utilized.

Results and Discussion

Descriptive statistics for OPD and Ct.Ar are presented in Table 2. For the entire sample, there was a statistically significant difference in OPD between sampling locations as determined by a one-way ANOVA ($F(2,87) = 9.082$, $p < 0.001$; Table 3). Post hoc comparisons using the Tukey HSD test indicated statistically significant differences between the posterior and anterior locations ($p < 0.001$) and the posterior and midshaft locations ($p < 0.001$). OPD between the midshaft and anterior locations ($p = 0.473$) showed no significant differences. A Kruskal–Wallis test determined a statistically significant difference in Ct.Ar between sampling locations within the entire sample ($\chi^2(2) = 15.743$, $p < 0.001$). A Dunn's all-pairs post hoc test demonstrated that the significant difference was between the posterior and anterior locations ($p < 0.001$). No significant differences were determined for Ct.Ar between the midshaft and the anterior or posterior locations ($p = 0.073$). These results reiterate the importance of sampling location for histological age estimation. Furthermore, if sampling location is uncertain, a more anterior section should be taken, as the histomorphology of the anterior rib is more consistent with that of the midshaft. Sampling at the midshaft or anterior location should result in similar OPD data, thus generating similar age estimates. Additionally, these data support previously described histomorphological methods utilizing the anterior (or sternal end) rib for age estimation (12,13) when the midshaft is unavailable.

When examining OPD at sampling locations in males or females, statistically significant differences were found between only anterior and posterior sampling locations for males ($F(2,42) = 14.376$, $p < 0.001$) and no statistically significant differences were determined for females ($F(2,42) = 2.8113$, $p = 0.071$). Additionally, when comparing OPD between males and females at the sampling locations, no statistically significant difference was determined at any of the sampling locations ($p \geq 0.030$ for all locations). These results support previous studies (4,14) that state accurate age estimates from the rib do not require sex-specific equations and thus known sex.

When examining sampling locations in males or females for Ct.Ar, no statistically significant difference was determined between sampling locations for males ($\chi^2(2) = 7.888$, $p = 0.019$) but statistically significant differences were found for females ($\chi^2(2) = 22.023$, $p < 0.001$). Contrary to the OPD results, significant differences existed in Ct.Ar at all three sampling locations between males and females (anterior, $p < 0.001$; midshaft, $p = 0.001$; posterior, $p < 0.001$). Elevated cortical loss in females versus males in an elderly sample such as this one is expected as a result of the rapid bone loss experienced by females after the onset of menopause (15). While significant differences in Ct.Ar between males and females were determined,

TABLE 2—*OPD and Ct.Ar (minimum, maximum, and mean) for total sample, males, and females.*

	Total Sample			Males			Females		
	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
OPD									
Anterior	19.43	38.79	28.79	26.07	38.79	30.98	19.43	33.53	26.72
Midshaft	12.23	40.06	27.82	24.23	35.84	28.86	12.23	40.06	27.04
Posterior	10.54	32.86	23.15	20.98	27.78	24.56	10.54	32.86	21.88
Ct.Ar (mm^2)									
Anterior	11.66	33.38	17.32	14.73	33.38	21.27	11.66	15.45	13.11
Midshaft	12.27	35.27	19.35	15.77	35.27	22.94	12.27	19.50	15.56
Posterior	15.00	51.45	23.55	22.26	51.45	29.63	15.00	21.00	17.25

TABLE 3—*OPD and Ct.Ar statistical analysis results.*

	Total Sample	Males	Females
OPD			
Anterior-midshaft	$p = 0.473^{\dagger}$	$\boldsymbol{p < 0.001}$*	$p = 0.071$*
Midshaft-posterior	$\boldsymbol{p < 0.001}^{\dagger}$		
Posterior-anterior	$\boldsymbol{p < 0.001}^{\dagger}$		
Ct.Ar			
Anterior-midshaft	$p = 0.073^{\dagger}$	$p = 0.019^{\ddagger}$	$\boldsymbol{p < 0.001}^{\ddagger}$
Midshaft-posterior	$p = 0.073^{\dagger}$		
Posterior-anterior	$\boldsymbol{p < 0.001}^{\dagger}$		

*ANOVA.

†Tukey Post Hoc.

‡Kruskal–Wallis.

§Dunn's Post Hoc.

Bolded text indicates significant values.

these differences did not affect OPD, which is designed to account for cross-sectional variability due to allometry, and are likely due to size differences between sexes.

Repeated-measures ANOVA tests determined no significant differences between serial sections for OPD or Ct.Ar ($p \geq 0.046$). This indicates that adjacent sections likely generate comparable age estimates and support the practice of collecting data from one section. However, while the OPD from a single section is not significantly different from that derived from multiple sections, best practice suggests examining at least two sections to measure enough bone to assess intraindividual variation.

This study has a few limitations that are important to note. First, though evenly distributed by sex, the sample size is small, including only ten individuals. This is partly due to sample availability, as complete ribs are necessary to accurately measure each location, and partly due to the labor-intensive nature of histological studies, demonstrated by the 90 slides generated for analysis here requiring several hours for preparation and analysis each. Second, the sample age is skewed, with all included samples falling into the elderly demographic. Current methods of histological age estimation do not perform well once remodeling has reached the osteon asymptote (16), usually in the fifth decade of life in the ribs, and as such age estimates were not generated for comparisons in this study. However, age estimates rely on the assumption of a consistent rate of remodeling throughout life; therefore, the OPD differences observed throughout the rib in this study should hold true for a younger sample. Future work should examine a larger, sex-matched sample and a wider age distribution to confirm these results.

One of the major advantages to histological aging methods is their utility even in highly fragmented remains. In the ribs, fragmentation can often be extensive though a trained osteologist can typically ascertain the general region to which a shard belongs (i.e., anterior, middle, posterior). The overall findings and recommendations presented here are intended to reduce uncertainty about proper sampling in such cases of fragmentation and to aid practitioners in generating age estimates from limited materials. This study indicates that when the precise midshaft cannot be determined, anterior-midshaft should be used rather than posterior-midshaft to generate comparable OPD values for generating age. In addition, while multiple sections are always ideal to better capture micromorphological variation, OPD can typically be calculated from the analysis of single cross section.

Acknowledgments

The authors would like to thank the anatomical donors for their generous gifts. The authors thank all the members of the Skeletal Biology Research Laboratory and the Injury Biomechanics Research Center for preparation and imaging of the slides used in this project.

References

1. Tommerup LJ, Raab DM, Crenshaw TD, Smith EL. Does weight-bearing exercise affect non-weight-bearing bone? J Bone Miner Res 1993;8(9):1053–8. https://doi.org/10.1002/jbmr.5650080905.
2. Wu K, Schubeck K, Frost H, Villanueva A. Haversian bone formation rates determined by a new method in a mastodon, and in human diabetes mellitus and osteoporosis. Calcif Tissue Res 1970;6(3):204–19. https://doi.org/10.1007/BF02196201.
3. Stout SD. The use of bone histomorphometry in skeletal identification: the case of Francisco Pizarro. J Forensic Sci 1986;31(1):296–300.
4. Stout SD, Paine RR. Brief communication: histological age estimation using rib and clavicle. Am J Phys Anthropol 1992;87(1):111–5. https://doi.org/10.1002/ajpa.1330870110.
5. Stout SD, Porro MA, Perotti B. Brief communication: a test and correction of the clavicle method of Stout and Paine for histological age estimation of skeletal remains. Am J Phys Anthropol 1996;100(1):139–42. https://doi.org/10.1002/(SICI)1096-8644(199605)100:1<139:AID-AJPA12>3.0.CO;2-1.
6. Cho H, Stout SD, Madsen RW, Streeter MA. Population-specific histological age-estimating method: a model for known African-American and European-American skeletal remains. J Forensic Sci 2002;47(1):12–8.
7. Crowder C, Rosella L. Assessment of intra- and intercostal variation in rib histomorphometry: its impact on evidentiary examination. J Forensic Sci 2007;52(2):271–6. https://doi.org/10.1111/j.1556-4029.2007.00388.x.
8. Murach MM, Kang YS, Goldman SD, Schafman MA, Schlecht SH, Moorhouse K, et al. Rib geometry explains variation in dynamic structural response: potential implications for frontal impact fracture risk. Ann Biomed Eng 2017;45(9):2159–73. https://doi.org/10.1007/s10439-017-1850-4.
9. Frost HM. Tetracycline-based histological analysis of bone remodeling. Calcif Tissue Res 1969;3(3):211–37.
10. Dominguez VM, Agnew AM. The use of ROI overlays and a semi-automated method for measuring cortical area in ImageJ for histological analysis. Am J Phys Anthropol 2019;168(2):378–82. https://doi.org/10.1002/ajpa.23747.
11. Rasband WS. ImageJ. Bethesda, MD: U.S. National Institutes of Health, 1997–2016.
12. Stout SD, Dietze WH, Iscan MY, Loth SR. Estimation of age at death using cortical histomorphometry of the sternal end of the fourth rib. J Forensic Sci 1994;39(3):778–84.
13. Kim Y-S, Kim D-I, Park D-K, Lee J-H, Chung N-E, Lee W-T, et al. Assessment of histomorphological features of the sternal end of the fourth rib for age estimation in Koreans. J Forensic Sci 2007;52(6):1237–42. https://doi.org/10.1111/j.1556-4029.2007.00566.x.
14. Stout SD, Paine RR. Bone remodeling rates: a test of an algorithm for estimating missing osteons. Am J Phys Anthropol 1994;93(1):123–9. https://doi.org/10.1002/ajpa.1330930109.
15. Clarke B, Khosla S. Female reproductive system and bone. Arch Biochem Biophys 2010;503(1):118–28. https://doi.org/10.1016/j.abb.2010.07.006.
16. Frost HM. Secondary osteon population densities: an algorithm for estimating the missing osteons. Am J Phys Anthropol 1987;30(S8):239–54.
17. Heinrich J, Crowder C, Pinto DC. Proposal and validation of definitions for intact and fragmented osteons. Am J Phys Anthropol 2012;147(S54):163.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14541
Available online at: onlinelibrary.wiley.com

TECHNICAL NOTE

ANTHROPOLOGY

Stefania Tritella,[1] *M.D.; Zuzana Obertová,*[2] *Ph.D.; Luca Maria Sconfienza,*[1,2] *M.D., Ph.D.; Federica Collini,*[3] *M.D.; Enrica Cristini,*[4] *M.D.; Alberto Amadasi,*[3] *M.D.; Barbara Ciprandi,*[3] *M.D.; Riccardo Spairani,*[1] *M.D.; Domenico Albano,*[1] *M.D.; Alessia Viero,*[3,5] *M.D.; Annalisa Cappella,*[2] *Ph.D.; Paolo Cammilli,*[3] *M.D.; Francesco Sardanelli,*[1,2] *M.D., Ph.D.; and Cristina Cattaneo,*[3] *M.D., Ph.D.*

Multi-Rater Agreement Using the Adapted Fracture Healing Scale (AFHS) for the Assessment of Tubular Bones on Conventional Radiographs: Preliminary Study*

ABSTRACT: Better understanding of the timing of fracture healing may help in cases of interpersonal violence but also of personal identification. The intra- and inter-rater agreement for the adapted fracture healing scale (AFHS) assessing the post-traumatic time interval on radiographs were tested. This is a preliminary study, providing essential information on method reliability for upcoming studies using the AFHS. Five raters (two radiologists, a forensic pathologist, an orthopedist, and an anthropologist) were presented with a test in three parts consisting of 85 radiographs (from 30 adults) of fractures of tubular bones in different stages of healing purposefully selected from more than 1500 radiographs. The raters were firstly asked to assess 15 features describing fracture healing as present, absent, or not assessable. Thereafter, the raters were asked to choose from the AFHS a single-stage best representing the observed healing pattern. The intra- and inter-rater agreement were assessed using single-rating, absolute agreement, two-way mixed-effects intra-class correlation (ICC) coefficients. The intra-rater ICC of radiologist 1 ranged from 0.80 to 0.94. The radiologists' inter-rater ICC ranged from 0.68 to 0.74, while it ranged from −0.01 to 0.90 for the other raters. The good to excellent ICC among the radiologists and forensic anthropologist provides good foundation for the use of the AFHS in forensic cases of trauma dating. The poor to good results for the other physicians indicate that using the AFHS requires training in skeletal anatomy and radiology.

KEYWORDS: forensic anthropology, fracture, healing assessment, radiology, reliability, inter-rater agreement

The Fracture Dating Study has originated from the need to gather data allowing for comprehensive assessment and interpretation of cases of interpersonal violence, such as child abuse, and institutional torture in the forensic context, with focus on the post-traumatic time interval (PTI). PTI is relevant since the magistrates and judges often query how much time passed between the time when a bone was fractured and either the time of death or the time of examination in survivors. In addition, the assessment of PTI may also be of importance in cases of unidentified bodies. Radiological imaging is crucial in these investigations, particularly since for the living victims it is the only method available for fracture assessment. For postmortem trauma analysis, histological assessment of the post-traumatic time interval is a valuable alongside radiographs (1,2).

Biological literature provides descriptions of the individual stages of bone healing (3–5), but the timing of their appearance is not well known. However, before any association of healing stages to the temporal component can be made, it is important to know whether raters are able to recognize and assign the individual healing stages on radiological images. In regard to fracture dating, the radiological and forensic literature offers various types of assessment scales for bone healing, some of which are clinically driven focusing on features associated with late healing stages (6), while others describe the whole spectrum of the healing process (1,2,7–13).

Several of these studies lack the essential information on the reliability of the given scales (7–10), while some report inter-rater agreement for the individual healing stages as variations of the κ statistic (11–13) or intra-class correlation (ICC) coefficients (2).

[1]Department of Radiology, IRCCS Policlinico San Donato, Piazza Edmondo Malan 2, San Donato Milanese, Italy.

[2]Department of Biomedical and Health Sciences, University of Milan, Via Luigi Mangiagalli 37, Milan, Italy.

[3]Laboratory of Forensic Anthropology and Odontology (LABANOF), Section of Legal Medicine, University of Milan, Via Luigi Mangiagalli 37, Milan, Italy.

[4]Department of Orthopaedics and Traumatology, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Via Francesco Sforza 28, Milan, Italy.

[5]Department of Legal Medicine and Toxicology, University Hospital of Padova, Via Falloppio 50, Padova, Italy.

Corresponding author: Zuzana Obertová, Ph.D. E-mail: obertovazuzana@yahoo.co.nz

*All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Received 6 May 2020; and in revised form 22 June 2020; accepted 21 July 2020.

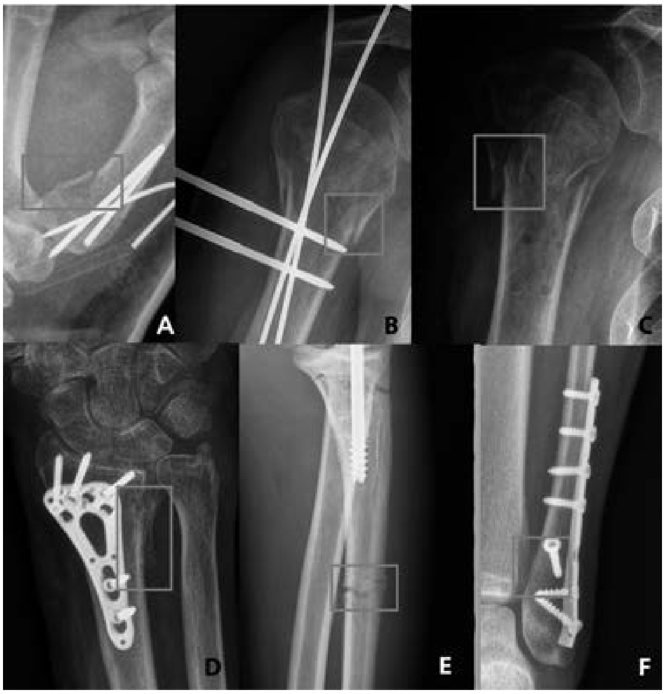


FIG. 1—*Examples of the stages in the adapted fracture healing scale (AFHS): (A) absorption; (B) periosteal reaction; (C) sclerosis; (D) callus; (E) bridging; (F) remodeling. [Color figure can be viewed at wileyonlinelibrary.com]*

The assessment of the reliability of a method is an essential initial step for introducing and subsequently applying any approach where human judgment is required (particularly one intended for forensic purposes); therefore, the aim of this paper is to assess the intra- and inter-rater agreement for the use of a purposefully adapted fracture healing scale (AFHS). The adaption process for the fracture healing scale was driven by the need to distinguish the early signs of healing (without the need of histological assessment) without including too many stages for a clear and concise differentiation and definition of the stages (Fig. 1).

Materials and Methods

The Fracture Dating Study is a retrospective observational study using series of clinical digital conventional radiographs of fractures of patients, who were admitted to the Emergency Department of the IRCCS Policlinico San Donato, between January 1, 2013, and June 30, 2015, and had at least one follow-up imaging with known date recorded at the same institution. The Study was approved by the relevant Ethics Committees (patients' informed consent was waived). The data collected within the Study include radiographs of fractures of a known date, and demographic and clinical variables, which may affect fracture healing. The main outcome of the study is the fracture "age" in relation to the healing stage, taking into account the effect of demographic and clinical variables on the temporal progression of fracture healing.

Data were collected from the hospital system database. An initial search for the term "fracture" returned 680 individual patients admitted during the given period. After the application of the exclusion criteria (no follow-up, initial access different from the hospital's Emergency Department), the final number of patients decreased to 467 totaling more than 1,500 radiographs. The vast majority of the patients was aged 18 years and older.

TABLE 1—*The adapted fracture healing scale (AFHS).*

Stage (short designation)	Description
0 (no healing)	No healing features; sharp contours of the fracture margin
1 (absorption)	Absorption of cortical bone at the fracture margin; blunted appearance of the margin
2 (periosteal reaction)	Start of periosteal reaction = linear elevation and calcification of periosteum in the vicinity or adjacent to the fracture site
3 (sclerosis)	Increased bone density at the fracture margin
4 (callus)	Appearance of callus; from the fluffy appearance of early new bone formation to well-demarcated callus, which may have margins as dense as the cortical bone
5 (bridging)	Fracture gap bridged by cortical bone to any extent (<50%, >50%)
6 (remodeling)	Bone returning to its original shape; from firm bony union, smoothening of contours to complete bone remodeling = bone at fracture site returned to its original shape

The AFHS (Table 1) used for this study was compiled from several sources (2,9,11). The initial evaluation using the predefined stages of the scale was undertaken by radiologist 1. A test was developed to assess the reliability of the AFHS. The test involved 85 radiographs (from 30 adult patients) of fractures of tubular bones (humerus, radius, ulna, metacarpal bones, femur, tibia, fibula, metatarsal bones) in different stages of healing, also including cases of clinical fixation. The radiographs were purposefully selected from the total pool of images to represent the whole range of healing stages. In addition, ten images were selected to represent nonideal settings (consisting of only one projection, or the fracture being obscured by cast) to resemble real-life scenarios. The option "not assessable" was included for all test questions. The healing process may not be consistent throughout all bone surfaces/margins, so the raters were asked to note the most advanced feature(s).

Part 1 of the test included 15 cases comprising the initial image from the emergency room (which was not evaluated) and two follow-up images from different points in time, both subject to evaluation (30 images to assess, 45 images in total). There were 15 features describing fracture healing, including no healing features, seven features assessing the fracture margin/fracture line/fracture gap, and six features addressing new bone formation adjacent to the fracture site and callus formation, and remodeling (Table 2). Each of these features needed to be evaluated as present, absent, or not assessable. Part 2 included 20 cases consisting of a single radiograph, which needed to be evaluated in the same manner as the series of radiographs in Part 1. For each of the images of Parts 1 and 2, the detailed features noted by the raters were subsequently translated into one of the stages of the AFHS using the principle of the most advanced feature. These were then used for the calculation of intra- and inter-rater agreement. The raters were informed that the order of the healing features was not necessarily related to time. Part 3 included 20 cases of a single radiograph, for which the raters selected only one-stage best representing the healing using the AFHS. Parts 2 and 3 involved only a single radiograph per individual to simulate real-life scenarios, where often only a single radiograph is assessed in forensic cases. For postmortem trauma analysis

TABLE 2—*Features describing fracture healing used in Parts 1 and 2 of the test.*

Healing features
No healing
Fracture margin/fracture line/fracture gap
Absorption of cortical bone at the fracture margin (blunted appearance)
Sclerosis (increased bone density) at the fracture margin
Bridging of < 50% of the fracture gap
Bridging of> 50% of the fracture gap
Start of blurring of the fracture line
Advanced blurring of the fracture line
No evidence of fracture line
New bone formation adjacent to the fracture site/callus formation
Start of periosteal and/or endosteal reaction
Fluffy appearance of early new bone around the fracture site
Periosteal and/or endosteal callus less dense than the cortex
Periosteal and/or endosteal callus of similar density to the cortex
Periosteal and/or endosteal callus more dense than the cortex
Periosteal and/or endosteal callus firmly attached/indistinguishable from adjacent bone tissue
Other
Remodeling (the fracture site returning to its original appearance)

including fracture dating, it is rare that suitable antemortem radiographs are available to help assess the progression of healing. Similarly, for living victims (for instance, suspected torture) it is also rare that imaging other than the queried radiograph taken at the time of victim examination is available for analysis.

Five raters (two radiologists, a forensic anthropologist, a forensic pathologist (legal medicine specialist), and an orthopedist) were presented with detailed written instructions on how to perform the test, which they undertook at their leisure and with no time restrictions. The imaging conditions (i.e., magnification options and software) were comparable among raters. All raters had approximately ten years of professional experience working with radiographs. The raters were chosen in relation to the forensic relevance of the study, including forensic anthropologists and forensic pathologists, but also clinicians who routinely work with radiological images of trauma.

The initial evaluation by radiologist 1 was considered as the optimal assessment, that is, other assessments were compared against it in the ICC calculation. Since there is no golden standard in the assessment of radiographs, the experience of radiologist 1 was set as the standard. Since the initial assessment was not blinded with respect to the time variable due to the evaluation being part of the data collection, the ICC was also calculated based on the blinded second observation by radiologist 1 to assess any potential bias. Moreover, the ICC was computed with the assessments of radiologist 2 as comparative values against other raters (except for radiologist 1) to estimate any potential discrepancies in the evaluation pattern of radiologist 1 with respect to the other raters.

The intra-rater agreement for radiologist 1 was tested on the combined sample of 70 images. Radiologist 1 undertook the second round of observations more than six months after the initial evaluation. The outcome of the intra-rater agreement for radiologist 1 may also be considered as an indication of the test–retest reliability. In addition, there were 15 images, which were repeated in two parts of the test (either in Parts 1 and 2, or Parts 2 and 3, or Parts 1 and 3) to test the intra-rater agreement for each of the raters or the repeatability of the results. Moreover, the values may also represent a measure of internal consistency.

Except for testing the reliability of the AFHS as such, we have also tested whether using multiple images in known order would improve the reliability of the results by comparing the ICC derived from Part 1 to those of Parts 2 and 3 (the assumption was that it would), and whether creating the final stage based on numerous detailed descriptions would improve reliability in comparison with assigning a single stage from a short list of stages by comparing the ICC of Parts 2 and 3 (the assumption was that it would not, since more options would result in more variability and complicate the decision process).

The statistical analysis was performed using Stata 12 (StataCorp, 2011, College Station, TX, USA). The intra- and inter-rater agreement as the measures of reliability were assessed using single-rating, absolute agreement, two-way mixed-effects intra-class correlation coefficients. In general, ICC is high when there is little variation among raters considering the individual stages (or feature descriptions) given to each radiograph.

The ICC values of intra- and inter-rater were categorized following the recommendations of Cicchetti (14): Less than 0.40 was categorized as poor agreement, between 0.40 and 0.59 as fair, between 0.60 and 0.74 as good, and between 0.75 and 1.00 as excellent agreement.

Results

The ICC coefficients for intra- and inter-rater agreement (with 95% CI) are presented in Table 3. The intra-rater ICC of radiologist 1 was classified as excellent, ranging from 0.80 to 0.94. The inter-rater ICC for the radiologists was classified as good, ranging from 0.68 to 0.74. The inter-rater ICC values for the other raters ranged from −0.01 to 0.90 for the three different parts of the test.

One case was purposefully selected to represent the scenario, where none of the features could be assessed due to the presence of a cast obscuring the fracture. All raters assessed this case as "not assessable." This case was not included in the ICC calculations.

The results of the inter-rater agreement for the combined sample derived from comparing the blinded round by radiologist 1 to other raters were as follows: 0.70 (95% CI 0.48–0.83) to radiologist 2, 0.86 (95% CI 0.75–0.93) to forensic anthropologist, 0.51 (95% CI 0.32–0.70) to forensic pathologist, and 0.22 (95% CI −0.03–0.44) to the orthopedist, while the ICC coefficient based on the comparison between radiologist 2 and forensic anthropologist was 0.83 (95% CI 0.69–0.94), between radiologist 2 and forensic pathologist was 0.50 (95% CI 0.32–0.65), and between radiologist 2 and the orthopedist was 0.30 (95% CI 0.08–0.50).

The ICC values representing intra-rater agreement computed on 15 cases repeated in two different parts of the test were 0.93 (95% CI 0.72–0.99) for radiologist 1, 0.98 (95% CI 0.93–1.00) for radiologist 2, 0.97 (95% CI 0.82–1.00) for forensic anthropologist, 0.80 (95% CI 0.29–0.95) for forensic pathologist, and 0.57 (95% CI, 0.06–0.82; 12 cases) for the orthopedist.

Discussion

The aim of this study was to assess the reliability of the AFHS, which has been developed to facilitate the forensic assessment of post-traumatic time interval in both deceased and living individuals. Measuring intra- and inter-rater agreement is essential for determining the reliability of a scientific method. When using a nominal scale, human raters will usually not be in complete agreement. In general, some of the variation in ratings may be due to chance, some due to differences in training or experience among raters but some may be purely due to the scale failing to measure the intended characteristic.

TABLE 3—*Intra- and inter-rater intra-class correlation (ICC) coefficient with 95% confidence interval (CI) for the three parts of the test separately and combined.*

Test (number of cases)	Intra-rater	Inter-rater Radiologists	Inter-rater Forensic Anthropologist	Inter-rater Forensic Pathologist	Inter-rater Orthopedist
	ICC (95% CI)				
Part 1 (30)	0.94 (0.83–0.98)	0.69 (0.38–0.86)	0.90 (0.78–0.96)	0.58 (0.25–0.75)	−0.01 (−0.35 to 0.28)
Part 2 (20)	0.81 (0.44–0.94)	0.68 (0.34–0.85)	0.78 (0.33–0.94)	0.49 (−0.01 to 0.73)	0.24 (−0.22 to 0.64)
Part 3 (20)	0.80 (0.48–0.90)	0.74 (0.40–0.89)	0.80 (0.42–0.94)	0.42 (−0.20 to 0.74)	0.65 (0.39–0.81)
Combined (70)	0.87 (0.78–0.93)	0.70 (0.54–0.80)	0.84 (0.73–0.91)	0.52 (0.28–0.67)	0.24 (0.04–0.41)

In this study, the inter-rater agreement between the radiologists and forensic anthropologist ranged from good to excellent, while the agreement between the radiologists and the other physicians (forensic pathologist and orthopedist) ranged from good to poor. The excellent agreement between the assessments of the radiologists and the forensic anthropologist may be due to the experience of the anthropologist working with conventional radiographs while assessing skeletal development and trauma. The forensic anthropologist was also involved in the development of the AFHS and was therefore well-aware of the specific definitions of the bone changes and stages of healing.

The intra-rater agreement for both radiologists, forensic anthropologist, and forensic pathologist was classified as excellent (confirming the reliability of the AFHS), while the intra-rater agreement of the orthopedist was fair. To cite the participating orthopedist, "Orthopedists use radiographs to visualize patients' complaints, but are not used to assess radiographs without any clinical information."

In previous publications, the inter-rater agreement was reported for a variety of stages with more or less detailed descriptions (2,11–13). de Boer et al. (2) reported the ICC values for 13 features observed during the radiological assessment of fractures. Prosser et al. (11) reported free-marginal multi-rater κ values separately for radiographs showing cases with and without cast, the cast being detrimental for evaluation. Two of the publications specifically mentioned that the raters were radiologists (11,13), while in the other two the profession of the raters is unclear (2,12).

As expected, using a number of different subcategories for a feature such as callus leads to low agreement between raters: For example, in one case, even though the raters agreed on callus being present, the subcategory given ranged from the early-stage "fluffy" callus, through callus being less dense, equally dense, or more dense than the cortical bone up to callus being firmly attached to the cortex (results not shown). When the state of the blurring of the fracture line was reported, a feature, which has been previously used to define certain stages in various fracture healing scales (8,10,12), it was found that, for example, the description "advanced blurring of the fracture line" was almost equally noted for each of the stages of the AFHS (results not shown). This finding suggests that it either posed great difficulty for the raters to assess when the blurring of the fracture line is actually advanced or that advanced blurring of the fracture line occurred at all stages of the healing process (which is unlikely for the early stages). Similarly, Halliday et al. (12) reported difficulties in the assessment of the definition of callus as opposed to presence/absence of callus as well as for the definition of fracture line. The reported κ values for the endosteal callus ranged from poor to good for three pair of observers, who were likely radiologists.

We assumed that using multiple images in known order would improve the reliability of the results, that is, the ICC coefficients will be greater for Part 1 than for Parts 2 and 3 of the test. This was the case for radiologist 1, forensic anthropologist, and forensic pathologist. We also hypothesized that in comparison with assigning a single stage from a small number of clearly distinguished healing stages as was the case in Part 3 of the test, the assignment of the final stage based on detailed descriptions as was the case in Part 2 would reduce the reliability due to the fact that using a multitude of different features and recognizing minute differences between the individual descriptions may complicate the decision process. Our assumption was confirmed by the ICC values for radiologist 2 and especially the orthopedist being higher in Part 3, even though radiologist 1 and forensic anthropologist had similar values in both parts. In summary, it seemed that using the AFHS with a small number of clearly differentiated stages and when possible evaluating a series of images showing the progress of fracture healing in known temporal order benefited the assessment process. Although it may be argued that including a smaller number of stages will result in broad fracture dating estimates, including a greater number of stages has been shown to result in an overall low reliability of the method due to the difficulty to distinguish among the multiple stages.

The agreement between the radiologists was poor when the absorption of bone at the fracture margin was considered (results not shown). This early stage of healing is important from the forensic point of view and is often addressed by forensic pathologists and anthropologists but is rarely if ever assessed by radiologists within the clinical context. Therefore, it is likely that the observations of this stage by radiologist 1 were guided by her awareness of the study's purpose, while radiologist 2 mostly assessed this stage as "no healing" possibly due to the need of using magnification features to correctly assess this stage. In contrast, forensic anthropologist and forensic pathologist showed good agreement with radiologist 1. Similarly, Boer et al. (2) reported poor to fair inter-rater agreement for early stages of fracture healing (smoothening of the lesion margin and absorption of cortical bone adjacent to lesion) when assessed on radiographs in comparison with assessments of histological specimens. Assuming that the technical parameters, such as the image resolution of digital radiographs, were sufficient for the assessment of the initial fracture healing stages, as shown by the good agreement among radiologist 1 and the forensic anthropologist and forensic pathologist, the assessment accuracy may be improved by training of raters less experienced in assessing the early stages of fracture healing.

The limitations of the study may include the fact that the results from the two parts of the test were based on less than 30 individuals (the recommended minimal number for the calculation of ICC), which resulted in broad CIs of the estimates. However, the overall number of individuals and radiographs (85) was sufficient and the different parts of the study provided valuable insight into the different aspects of the rating. Expanding the sample size would prove difficult, since the time required for the completion of the test needed to be restricted due to the workload of the raters.

This study focused solely on the healing patterns in tubular bones of adults. The healing of cranial bones shows a different pattern, lacking the typical callus formation stage. Using the initial, nonblinded assessment as the optimum may be considered questionable, but the excellent intra-observer agreement and almost identical ICC of the inter-rater agreement based on both the nonblinded and blinded assessment by radiologist 1 proved that the choice was appropriate. Moreover, there was little difference in the ICC coefficients between the radiologist and the other physicians when the assessment of radiologist 2 was used as the optimal staging, confirming that the assessment pattern of radiologist 1 did not show any major discrepancies when compared to that of radiologist 2.

The test was undertaken without specific training regarding the AFHS, only using detailed written instructions, to avoid introducing bias by trainers since it has been assumed that the descriptions of the stages are sufficient for practitioners who have got experience with the assessment of fractures on radiographs. Notably, the AFHS has been introduced mainly for forensic purpose. The stages used in the AFHS cover the whole spectrum of fracture healing. Although radiologists are used to assess and report on some of the stages, including callus formation, bridging, or remodeling, the results have shown that a specific training with focus on the less routinely used stages of the AFHS may be needed even for experienced radiologists.

The training in radiological assessment varies between countries, as, for example, in the United States radiology is part of a forensic anthropologist's training. Therefore, the findings would not necessarily apply to forensic anthropologists and other practitioners in the United States and other countries with different training background.

Conclusion

In conclusion, we tested the reliability of using the AFHS intended for the forensic assessment of the post-traumatic time interval on radiographs. The good to excellent ICC values achieved among the radiologists and forensic anthropologist provide good foundation for the use of the AFHS. The poor to good results for the other practitioners indicate that using the AFHS requires advanced training in skeletal anatomy and radiology. Nevertheless, a collaboration of experts from different fields is crucial for a comprehensive assessment of forensic cases of abuse and torture.

References

1. Maat GJR. Case study 5.3: dating of fractures in human dry bone tissue. In: Kimmerle EH, Baraybar JP, editors. Skeletal trauma: identification of injuries resulting from human rights abuse and armed conflict. Boca Raton, FL: CRC Press, 2008;245–54.
2. de Boer HH, van der Merwe AE, Hammer S, Steyn M, Maat GJR. Assessing post-traumatic time interval in human dry bone. Int J Osteoarchaeol 2015;25:98–109.
3. Marsell R, Einhorn TA. The biology of fracture healing. Injury 2011;42 (6):551–5. https://doi.org/10.1016/j.injury.2011.03.031
4. Schindeler A, McDonald MM, Bokko P, Little DG. Bone remodeling during fracture repair: the cellular picture. Semin Cell Dev Biol 2008;19 (5):459–66. https://doi.org/10.1016/j.semcdb.2008.07.004
5. van Rijn RR, Sieswerda-Hoogendoorn T. Imaging child abuse: the bare bones. Eur J Radiol 2012;171(2):215–24. https://doi.org/10.1007/s00431-011-1499-1
6. Axelrad TW, Einhorn TA. Use of clinical assessment tools in the evaluation of fracture healing. Injury 2011;42(3):301–5. https://doi.org/10.1016/j.injury.2010.11.043
7. Barbian LT, Sledzik PS. Healing following cranial trauma. J Forensic Sci 2008;53(2):263–8. https://doi.org/10.1111/j.1556-4029.2007.00651.x
8. Islam O, Soboleski D, Symons S, Davidson LK, Ashworth MA, Babyn P. Development and duration of radiographic signs of bone healing in children. Am J Roentgenol 2000;175(1):75–8. https://doi.org/10.2214/ajr.175.1.1750075
9. Malone CA, Sauer NJ, Fenton TW. A radiographic assessment of pediatric fracture healing and time since injury. J Forensic Sci 2011;56 (5):1123–30. https://doi.org/10.1111/j.1556-4029.2011.01820.x
10. Sanchez TR, Nguyen H, Palacios W, Doherty M, Coulter K. Retrospective evaluation and dating of non-accidental rib fractures in infants. Clin Radiol 2013;68(8):467–71. https://doi.org/10.1016/j.crad.2013.03.017
11. Prosser I, Lawson Z, Evans A, Harrison S, Morris S, Maguire S, et al. A timetable for the radiologic features of fracture healing in young children. Am J Roentgenol 2012;198(5):1014–20. https://doi.org/10.2214/AJR.11.6734
12. Halliday KE, Broderick NJ, Somers JM, Hawkes R. Dating fractures in infants. Clin Radiol 2011;66(11):1049–54. https://doi.org/10.1016/j.crad.2011.06.001
13. Walters MM, Forbes PW, Buonomo C, Kleinman PK. Healing patterns of clavicular birth injuries as a guide to fracture dating in cases of possible infant abuse. Pediatr Radiol 2014;44(10):1224–9. https://doi.org/10.1007/s00247-014-2995-z
14. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 1994;6(4):284–90.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14515
Available online at: onlinelibrary.wiley.com

TECHNICAL NOTE

CRIMINALISTICS

Minji Lee,[1] M.S.; Ju Yeon Jung,[1] M.S.; Sungsoo Choi,[1] M.S.; Ilung Seol,[2] Ph.D.; Seohyun Moon,[1] Ph.D.; and In Kwan Hwang,[1] Ph.D.

Single Nucleotide Polymorphism Assay for Genetic Identification of *Lophophora williamsii**

ABSTRACT: *Lophophora* is a member of the Cactaceae family, which contains two species: *Lophophora williamsii* and *L. diffusa*. *Lophophora williamsii* is an illegal plant containing mescaline, a hallucinogenic alkaloid. In this study, a novel method based on a single nucleotide polymorphism (SNP) assay was developed for identifying *L. williamsii*; this assay reliably detects SNPs within chloroplast DNA (*rbcL*, *matK*, and *trnL-trnF* IGS) and was validated for identifying *Lophophora* and *L. williamsii* simultaneously. The chloroplast DNA sequences from four *L. williamsii* and three *L. diffusa* plants were obtained and compared using DNA sequence data from approximately 300 other Cactaceae species available in GenBank. From this sequence data, a total of seven SNPs were determined to be suitable for identifying *L. williamsii*. A multiplex assay was constructed using the ABI PRISM® SNaPshot™ Multiplex Kit (Applied Biosystems, Forster City, CA) to analyze species-specific SNPs. Using this multiplex assay, we clearly distinguished the *Lophophora* among 19 species in the Cactaceae family. Additionally, *L. williamsii* was distinguished from *L. diffusa*. These results suggest that the newly developed assay may help resolve crimes related to illegal distribution and use. This multiplex assay will be useful for the genetic identification of *L. williamsii* and can complement conventional methods of detecting mescaline.

KEYWORDS: Cactaceae, *Lophophora williamsii*, mescaline, single-base extension, single nucleotide polymorphism, SNaPshot™ Multiplex Kit

The genus *Lophophora* belongs to the family Cactaceae, order Caryophyllales, and includes two species: *Lophophora williamsii* and *L. diffusa*. *Lophophora* are native to central Mexico and southern Texas in North America (1). This turquoise cactus has a round shape and lacks thorns. The flowers of *L. williamsii* are pink, whereas those of *L. diffusa* are white. *Lophophora williamsii*, also known as Peyote, is characterized by the presence of mescaline, a psychotropic alkaloid (2,3).

Both mescaline and *L. williamsii* were classified as "Schedule 1 Drugs" by the "Controlled Substances Act"; these were outlawed in 1970 in the United States and banned internationally by the 1971 "Convention on Psychotropic Substances" (4). In Korea, it is illegal to possess psychotropic drugs, including mescaline. *Lophophora williamsii* is designated as a psychotropic drug under Article 2 of the "Narcotics Control Act" (5). Therefore, possession of *L. williamsii* is illegal. Species identification of *L. williamsii* is performed by determining the mescaline content through component analysis. However, the mescaline content differs in different parts of the plant (crowns, stems, roots, etc.). Moreover, older plants contain higher levels of mescaline (6). In addition, the mescaline content may vary depending on the time of collection, individual differences, and cultivation environment. *Lophophora diffusa*, which does not contain mescaline, is similar in appearance to *L. williamsii*. Thus, *L. williamsii* or *L. diffusa* may be incorrectly identified when evaluating morphological features alone (7). In addition, there is concern that *L. williamsii* can be used for grinding, brewing, and powdered for capsules (8).

Plant species identification is usually based on morphological and anatomical characteristics. However, when these methods are not suitable, DNA barcoding methods based on sequencing may be substituted (9,10). *Lophophora williamsii* identification using sequencing chloroplast DNA region such as *rbcL* and *trnL-trnF* IGS, which contains 400–900 base pairs (bp), has also been reported (11,12). A previous study reported the application of loop-mediated isothermal amplification (LAMP) using six specific primers, including two loop primers, to evaluate *L. williamsii* and *L. diffusa* DNA by measuring turbidity due to the formation of magnesium pyrophosphate in real-time monitoring of LAMP (7). The LAMP method facilitates fast amplification speed and high specificity, although the experimental process is complex and interpreting the results requires technical skills (13). In this study, we developed a new identification method using SNPs to minimize identification errors and simplify interpretation of the results. SNPs are a form of genetic variation in intraspecies and interspecies; it has been primarily used in forensic investigations for identifying endangered wildlife (14–17). The results of SNP assays can be confirmed by polymer chain reaction (PCR) and capillary electrophoresis, which are standard procedures in forensic genetic laboratories.

[1] Forensic DNA Division, National Forensic Service, Wonju, 26460, Korea.

[2] Forensic Toxicology Division, National Forensic Service, Wonju, 26460, Korea.

Corresponding author: In Kwan Hwang, Ph.D. E-mail: inkai@korea.kr

*This work was supported by National Forensic Service (NFS2019DNA02), Ministry of the Interior and Safety, Republic of Korea.

Received 9 April 2020; and in revised form 5 June 2020; accepted 25 June 2020.

In this study, we identified specific SNPs for *L. williamsii* in chloroplast DNA (*rbcL*, *matK*, and *trnL-trnF* IGS regions), which are commonly used as DNA barcodes. We developed a new assay for specifically identifying *Lophophora* and *L. williamsii* simultaneously. Moreover, this method was applied to other samples from Cactaceae to verify the validation of *L. williamsii*.

Materials and Methods

Sample Collection

The details of all samples are shown in Table 1. Three samples of *L. williamsii* were submitted by the National Forensic Service for forensic analysis, and three samples of *L. diffusa* were purchased via a website (http://www.xplant.co.kr). Tissues of 19 species belonging to the 10 genera including *L. williamsii* were obtained from Munich Institute (Germany) and Korea National Arboretum. Tissues were not dried and were stored at −20°C before extracting their genomic DNA. Samples were analyzed by liquid chromatography-quadrupole time-of-flight mass spectrometry (SCIEX, Concord, ON, Canada), and mescaline was confirmed to be detectable.

DNA Extraction from Plant Tissue

Approximately 100 mg of plant material was collected from each of the 26 plants and crushed for about 3 min with a bead beater. DNA was isolated using a DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. The extracted DNA was confirmed by agarose gel electrophoresis.

Identification of Lophophora- *and* L. williamsii*-specific SNPs*

The *rbcL* (~500 bp), *matK* (~850 bp), and *trnL-trnF* IGS (~1,000 bp) sequences of *Lophophora* from NCBI GenBank were searched to extract base sequences from 100 Cactaceae with high similarity in each region (10,11). These sequences were aligned with four *L. williamsii* and three *L. diffusa* DNA sequences using MEGA software (Molecular Evolution Genetics Analysis 5.0 software, Pennsylvania State University, State College, PA; www.megasoftware.net). Based on the aligned sequences, comparisons were made between the *Lophophora* and other genera at the genus level, and between *L. williamsii* and *L. diffusa* in the *Lophophora* at the species level.

TABLE 1—*List of cactuses used in this study.*

Voucher No.	Species	Source	Mescaline
L.w-1	*Lophophora williamsii*	Arboretum	D*
L.w-2,3	*Lophophora williamsii*	Evidence	D
L.w-4	*Lophophora williamsii*	Evidence	ND†
L.d-1,2,3	*Lophophora diffusa*	Internet	ND
C1	*Astrophytum myriostigma*	Arboretum	NT‡
C2	*Astrophytum ornatum*	Arboretum	NT
C3	*Copiapoa bridgesii*	Arboretum	NT
C4	*Copiapoa haseltoniana*	Arboretum	NT
C5	*Coryphantha difficilis*	Arboretum	NT
C6	*Echinocereus arizonicus*	Arboretum	NT
C7	*Echinopsis multiplex*	Arboretum	NT
C8	*Ferocactus flavovirens*	Arboretum	NT
C9	*Ferocactus latispinus*	Arboretum	NT
C10	*Gymnocalycium comarapeuse*	Arboretum	NT
C11	*Gymnocalycium mihanovichii*	Arboretum	NT
C12	*Gymnocalycium tilcarense*	Arboretum	NT
C13	*Gymnocalycium zegarrae*	Arboretum	NT
C14	*Gymnocalycium hossei*	Arboretum	NT
C15	*Mammillaria decipiens*	Arboretum	NT
C16	*Mammillaria polythele*	Arboretum	NT
C17	*Matucana madisoniorum*	Arboretum	NT
C18	*Thelocactus ehrenbergii*	Arboretum	NT
C19	*Thelocactus rinconensis*	Arboretum	NT

C1–C19 are cactuses of different genera in the Cactaceae family.

*Detection.

†Not detectable.

‡Not tested.

SNaPshot Multiplex PCR Assay

After identifying *Lophophora*- and *L. williamsii*-specific SNPs, multiplex PCR primers were prepared with three primer pairs (rbcL_FR, matK_FR, and trnLF_FR) to amplify chloroplast DNA containing these SNPs (Table 2A).

The PCR amplification products were 243, 258, and 339 bp, and the Tm values were adjusted to ensure amplification by multiplex PCR. PCR was performed by mixing 3 units of AmpliTaq Gold DNA polymerase (Thermo Fisher Scientific, Waltham, MA), 1 µL of Gold ST*R 10X buffer (Promega, Madison, WI), 2–5 ng of template DNA, and distilled water to a volume of 10 µL in an ABI 9700 system (Thermo Fisher Scientific). After 11 min of predenaturation at 95°C, denaturation at 94°C for 20 sec, annealing at 57°C for 1 min, and extension at 72°C for 30 sec were performed for 35 cycles. Final extension was carried out at 72°C for 7 min.

The amplification products were purified by adding 5 µL of product to 2 µL of ExoSAP-IT® (Thermo Fisher Scientific) which were reacted for 45 min at 37°C, followed by 15 min at 80°C.

One microliter of purified PCR amplification products was subjected to mini-sequencing using an ABI PRISM® SNaPshot™ Multiplex Kit (Thermo Fisher Scientific). Single-base extension (SBE) multiplex primer information obtained by mini-sequencing for SNP site identification is shown in Table 2B. Each SBE primer had a poly T-tail attached to the $5'$ end with more than a 5-nucleotide difference between the amplification products for easy genotype determination of the SNPs in the electropherogram.

The mini-sequencing reaction product was purified by adding 10 µL of the amplification product to 1 µL of shrimp alkaline phosphatase (Thermo Fisher Scientific) which were reacted for 45 min at 37°C, followed by treatment for 15 min at 80°C. Capillary electrolysis was performed using 1 µL of the purified amplification product and 0.2 µL of GeneScan™ 120 LIZ® Size Standard (Thermo Fisher Scientific), 10 µL Hi-Di™ Formamide (Thermo Fisher Scientific), and Applied Biosystems 3500xL Genetic Analyzer (Thermo Fisher Scientific). The results were confirmed using GeneMapper™ ID-X software version 1.4 (Thermo Fisher Scientific).

Results and Discussion

Lophophora- *and* L. williamsii*-specific SNPs*

Based on the nucleotide sequences of seven samples of *Lophophora* and NCBI GenBank data for *rbcL*, *matK*, and *trnL-trnF* IGS of Cactaceae, four SNPs specific to *Lophophora* (rbcL235, matK1497, matK1502R, and trnLF285R) were selected, and three SNPs (rbcL304, trnLF385R, and trnLF476) that distinguished *L. williamsii* from *L. diffusa* were selected (Table 3).

SNPs rbcL235, rbcL304, and matK1497 are in protein-coding regions but are synonymous and do not cause amino acid

TABLE 2—*Primers used for the SNaPshot multiplex PCR assay. (A) Multiplex PCR primer. (B) SBE multiplex PCR primer.*

(A)

Primer	Primer Sequence (5′–3′)	Primer Length	Primer Concentration (μM)	Product Size (bp)
rbcL_F	TCGTTACAAAGGACGATGCTAC	22	0.05	243
rbcL_R	CTCTCAACTTGGATACCGTGA	21	0.05	
matK_F	TGCCTTATCTTATGGCCTTTCA	22	0.05	258
matK_R	CATTTGACTCCGTACCACTG	20	0.05	
trnLF_F	GTGCAGAGACTCAAAGGAAGT	21	1	339
trnLF_R	CAACTTGGAATCGATTCATAACCC	24	1	

(B)

Primer	Primer Sequence (5′–3′)	Primer Length	Primer Concentration (μM)	Identification Level
rbcL235	t CGTTCCTGGAAAAGACAATCAATA	25	0.1	Genus
rbcL304	(t)$_{33}$TTCTGTTACTAATATGTTTACTTCCAT	61	0.6	Species
matK1497	(t)$_{13}$GAAAATCAATTCTGGCTTCAAA	35	0.3	Genus
matK1502R	(t)$_{7}$TCCATTTATGCATCAGAAGAGAT	30	0.3	Genus
trnLF285R	(t)$_{18}$TCACTACACGTATATGCTTTAC	41	1	Genus
trnLF385R	(t)$_{26}$CTTATTATATAAAGTAGAATTCAGATTAT	55	1	Species
trnLF476	(t)$_{39}$TCTAAATAGAAATTTTAGAATATGAA	66	1.5	Species

SBE, single-base extension.

TABLE 3—Lophophora *and* L. williamsii *species-specific SNPs found by alignment of sequences of four* L. williamsii *and three* L. diffusa *compared with rbcL, matK, and trnL-trnF IGS sequences of genetically similar Cactaceae (comparison group) in Genbank.*

Gene	*rbcL*		*matK*		*trnL-trnF* IGS		
SNP site	235	304	1497	1502	285	385	476
L. williamsii	C	T	G	A	A	G	G
L. diffusa	C	C	G	A	A	T	A
Comparison group	T	C,T	G,A	C	G	G,A	A

mutations. In contrast, matK1502R is a nonsynonymous SNP that is a transversion mutation in *Lophophora* that carries an adenine (A) and translates into lysine (K); in other genera, it carries a cytosine (C) and translates into threonine (Thr). In addition to this, SNP, trnLF285R, and trnLF385R are designed with reverse primers for easy SBE multiplex PCR amplification. Although *trnL-trnF* IGS is a noncoding region, it is important in *L. williamsii* identification, as it has been previously reported to be related to morphology or the mescaline content (18).

SNaPshot Multiplex Assay

Multiple amplifications with primer sets rbcL_FR, matK_FR, and trnLF_FR designed to amplify the chloroplast regions containing all SNPs were successfully performed in seven *Lophophora* (Fig. 1). Subsequently, when the PCR products were purified and then mini-sequenced as templates, only four *L. williamsii*, including the positive control and three *L. diffusa*, showed the same SNP profiles (Fig. 2A,B). Notably, *L. williamsii* was confirmed as guanine (blue) at the trnLF476 site, but no results were obtained for *L. diffusa*. To determine the cause of this result, mono-PCR was further performed for the trnLF476 site; however, additional peaks were not confirmed, suggesting that PCR amplification was not possible for this sample because of differences in the primer binding sequences. Sequencing results showed that the sequence of the primer binding sites differed by one base in *L. diffusa* (*L. diffusa*: 5′-. . .T**T**AA-3′, *L. williamsii*: 5′-. . .T**G**AA-3′). Thus, amplification during PCR may have been insufficient.

Specificity of SNP Test

The specificity of the SNP assay developed in this study was demonstrated using 19 Cactaceae species. To interpret the results, only samples for which the seven SNP profiles expected according to the final capillary electrophoresis results were considered *L. williamsii*. The first PCR amplification products of C2, C8, C16, C18, and C19 were similar to those of *Lophophora*. In other species, some bands were deleted or differed from the expected size (data not shown). Mini-sequencing confirmed species-specific amplification differences. Although the peaks of seven sites were not confirmed in all species, the expected bases were confirmed by analyzing the sequencing results (Fig. 2C). The reason for the difference in size of rbcL235 and matK1502R on the electropherogram is that shorter fragments are more strongly affected by the fluorescent dye

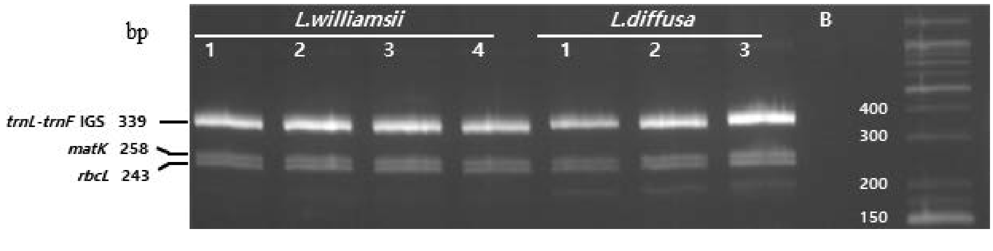


FIG. 1—*Analysis by 4% agarose gel of first multiplex PCR amplicon in* Lophophora.

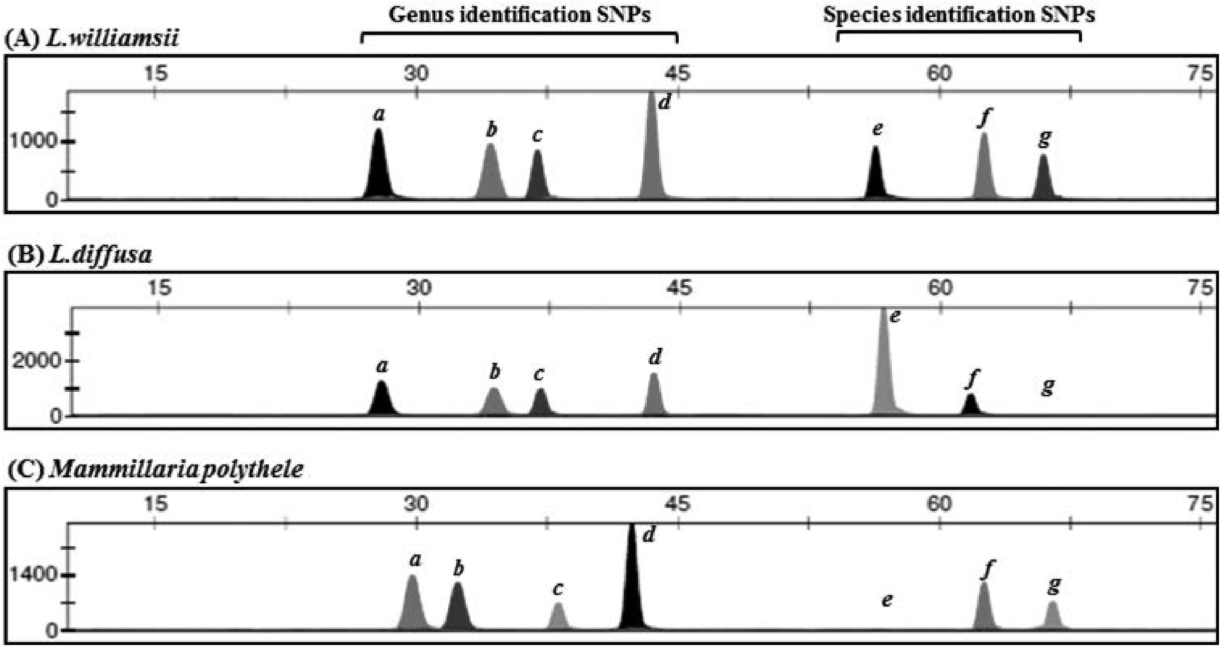


FIG. 2—*Electropherograms showing SNP typing of* L. williamsii *and* L. diffusa*. (A) L. williamsii, (B) L. diffusa, and (C) Mammillaria polythele (C16, one of the five most similar samples from the first PCR). a, rbcL235; b, matK1502R; c, matk1497; d, trnLF285R; e, trnLF385R; f, rbcL304; g, trnLF476. The "a–d" peaks identify the genus level and the "e–g" peaks identify the species level. [Color figure can be viewed at wileyonlinelibrary.com]*

(19,20). As a result, none of the 19 species produced the SNP profile of *L. williamsii*.

In this study, we developed an SNP method for identifying *L. williamsii*; this method is more rapid and convenient than conventional mescaline detection and sequencing methods. The specific SNP profile of *L. williamsii* was compared with that of Cactaceae (high genetic similarity as a result of NCBI search, 19 similar appearance species to *L. williamsii*) using the SNaPshot™ multiplex kit. The assay analyzed short DNA fragments up to 340 bp, allowing for analysis of older or degraded samples. This simple process enables multiple sample analysis in a short time. This assay can be used in forensic analysis to distinguish and identify *L. williamsii* from other Cactaceae such as *L. diffusa*, which are very similar in morphology.

References

1. IUCN. The IUCN red list of threatened species. 2014. www.iucnredlist.org (accessed May 25, 2020).
2. The vaults of EROWID. Visionary cactus guide. https://www.erowid.org/plants/cacti/cacti_guide/cacti_guide_lophopho.shtml (accessed May 25, 2020).
3. LOPHOPHORA. *Lophophora williamsii* v. *caespitosa* graft – 5th anniversary. http://lophophora.blogspot.com/2009/11/lophophora-williamsii-v-caespitosa.html (accessed May 25, 2020).
4. Anderson EF. Peyote: the divine cactus, 2nd rev edn. Tucson, AZ: University of Arizona Press, 1996;230–3.
5. Korea Ministry of Government Legislation. Narcotics Control Act. https://www.moleg.go.kr/index.es?sid=a1 (accessed May 25, 2020).
6. Klein MT, Kalam M, Trout K, Fowler N, Terry M. Mescaline concentrations in three principal tissues of *Lophophora williamsii* (Cactaceae): implications for sustainable harvesting practices. Haseltonia 2015;20:34–42. https//doi.org/10.2985/026.020.0107.
7. Sasaki Y, Fujimoto T, Aragane M, Yasuda I, Nagumo S. Rapid and sensitive detection of *Lophophora williamsii* by loop-mediated isothermal amplification. Biol Pharm Bull 2009;32(5):887–91. https//doi.org/10.1248/bpb.32.887.
8. The Recovery Village. Peyote addiction and abuse. https://www.therecoveryvillage.com/peyote-addiction/#gref (accessed May 25, 2020).
9. Ferri G, Alu M, Corradini B, Licata M, Beduschi G. Species identification through DNA "barcodes". Genet Test Mol Biomarkers 2009;13(3):421–6. https//doi.org/10.1089/gtmb.2008.0144.
10. Hollingsworth PM, Graham SW, Little DP. Choosing and using a plant DNA barcode. PLoS One 2011;6(5):e19254. https//doi.org/10.1371/journal.pone.0019254.
11. Ng AE, Sandoval E, Murphy TM. Identification and Individualization of Lophophora using DNA Analysis of the trn L/trn F Region and rbc L Gene. J Forensic Sci 2016;61(Suppl 1):S226–9. https//doi.org/10.1111/1556-4029.12936.
12. Lee MJ, Choi SS, Kim DH, Jung JY, Kim JY, Moon SH, et al. Case report: species identification of *Lophophora williamsii* (Psychoactive cactus) using botanical DNA barcodes. J Sci Crim Investig 2020;14(1):71–7. https//doi.org/10.20297/jsci.2019.14.1.71.
13. Karami A, Gill P, Motamedi MHK, Saghafinia M. A review of the current isothermal amplification techniques: applications, advantages and disadvantages. J Global Infect Dis 2011;3(3):293–302. https//doi.org/10.4103/0974-777X.83538.
14. Rafalski JA. Novel genetic mapping tools in plants: SNPs and LD-based approaches. Plant Sci 2002;162(3):329–33. https//doi.org/10.1016/S0168-9452(01)00587-8.
15. Lee JCI, Hsieh HM, Huang LH, Kuo YC, Wu JH, Chin SC, et al. Ivory identification by DNA profiling of cytochrome b gene. Int J Legal Med 2009;123(2):117–21. https//doi.org/10.1007/s00414-008-0264-0.
16. Verma SK, Singh L. Novel universal primers establish identity of an enormous number of animal species for forensic application. Mol Ecol Notes 2003;3(1):28–31. https//doi.org/10.1046/j.1471-8286.2003.00340.x.
17. Ogden R, McGough HN, Cowan RS, Chua L, Groves M, McEwing R. SNP-based method for the genetic identification of ramin Gonystylus spp. timber and products: applied research meeting CITES enforcement needs. Endanger Species Res 2008;9(3):255–61. https//doi.org/10.3354/esr00141.
18. Aragane M, Sasaki Y, Nakajima JI, Fukumori N, Yoshizawa M, Suzuki Y, et al. Peyote identification on the basis of differences in morphology, mescaline content, and trnL/trnF sequence between *Lophophora williamsii* and *L. diffusa*. J Nat Med 2011;65(1):103–10. https//doi.org/10.1007/s11418-010-0469-7.
19. Rotherham D, Harbison SA. Differentiation of drug and non-drug Cannabis using a single nucleotide polymorphism (SNP) assay. Forensic Sci Int 2011;207(1–3):193–7. https//doi.org/10.1016/j.forsciint.2010.10.006.
20. Kitpipit T, Tobe SS, Kitchener AC, Gill P, Linacre A. The development and validation of a single SNaPshot multiplex for tiger species and subspecies identification—implications for forensic purposes. Forensic Sci Int Genet 2012;6(2):250–7. https//doi.org/10.1016/j.fsigen.2011.06.001.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14494
Available online at: onlinelibrary.wiley.com

TECHNICAL NOTE

CRIMINALISTICS

Maria Fernanda M. Ribeiro,[1] M.D.; Fátima Bento,[2] Ph.D.; Antônio J. Ipólito,[3] M.S.; and Marcelo F. de Oliveira,[1] Ph.D.

Development of a Pencil Drawn Paper-based Analytical Device to Detect Lysergic Acid Diethylamide (LSD)*'†

ABSTRACT: The need for agile and proper identification of drugs of abuse has encouraged the scientific community to improve and to develop new methodologies. The drug lysergic acid diethylamide (LSD) is still widely used due to its hallucinogenic effects. The use of voltammetric methods to analyze narcotics has increased in recent years, and the possibility of miniaturizing the electrochemical equipment allows these methods to be applied outside the laboratory; for example, in crime scenes. In addition to portability, the search for affordable and sustainable materials for use in electroanalytical research has grown in recent decades. In this context, employing paper substrate, graphite pencil, and silver paint to construct paper-based electrodes is a great alternative. Here, a paper-based device comprising three electrodes was drawn on 300 g/m^2 watercolor paper with 8B pencils, and its efficiency was compared to the efficiency of a commercially available screen-printed carbon electrode. Square wave voltammetry was used for LSD analysis in aqueous medium containing 0.05 mol/L $LiClO_4$. The limits of detection and quantification were 0.38 and 1.27 μmol/L, respectively. Both electrodes exhibited a similar voltammetric response, which was also confirmed during analysis of a seized LSD sample, with recovery of less than 10%. The seized samples were previously analyzed by GCMS technique, employing the full scan spectra against the software spectral library. The electrode selectivity was also tested against 3,4-methylenedioxymethamphetamine (MDMA) and methamphetamine. It was possible to differentiate these compounds from LSD, indicating that the developed paper-based device has potential application in forensic chemistry analyses.

KEYWORDS: forensic chemistry, electrochemistry, LSD, paper-based electrodes, screen-printed electrode, voltammetry

In recent decades, the need for technology miniaturization and for portable compact devices containing the indispensable tools for a specific task has increased enormously. In the field of forensic analysis, the situation is no different. The possibility of conducting analysis at the crime scene not only speeds up the identification of samples, but it also helps to elucidate what may have happened. In addition to portability, nowadays there is great demand for low-cost environmentally friendly devices (1,2).

Although the chromatographic equipment used by the forensic scientists ensures correct identification of seized samples, it does not fit these three latter requirements. The devices are large and expensive and require the use of organic solvents during analysis, which generates considerable amounts of residues due to the extensive time each analysis lasts. Another disadvantage is the need for a larger amount of sample and specific preparation to suit the specifications of these analytical method (3,4).

For a rapid screening of samples, prior their analysis by chromatographic methods, electroanalytical methods are particularly adequate as they are accessible, portable, and environmental friendly (5). Miniaturized potentiostats can be found in the size of a smartphone (6,7). The price of these devices has also become attractive, especially because they can be used for simultaneous analysis at the same time. Further advantages of electroanalytical techniques include the possibility of performing assays in aqueous solutions, reducing generation of residues, reducing analysis time, and the ability to analyze small volumes, in the order of μL, which is an important consideration when only small amounts of sample are available (3–9).

The currently employed compact set of electrodes (working, reference, and auxiliary electrodes) in a single device also meet these requirements. As they are produced in smaller dimensions, a less volume of sample solution is needed, and disposable user-friendly electrodes are accessible at low cost. In recent years, screen-printed electrodes have emerged and became quite popular. Most of these electrodes use ceramic or polymers as substrate, even for a wearable device (1–3,10,11). Due to environmental concerns, the use of paper substrate for these

[1] Universidade de São Paulo, USP, Avenida Bandeirantes, Ribeirão Preto, SP 3900, Brazil.

[2] Centro de Química, Universidade do Minho, Campus de Gualtar, Braga, 4710 – 057, Portugal.

[3] Superintendência Polícia Técnica Científica, SPTC, Rua São Sebastião, Ribeirão Preto, SP 1339, Brazil.

Corresponding author: Marcelo F. de Oliveira, Ph.D. E-mail: marcelex@usp.br

*Presented at the XXII SIBEE Meeting, September 3, 2019, in São Paulo, Brazil.

†Financial support provided by the Polícia Científica do Estado de São Paulo for the partnership, and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP - Process 2016/23825-3), Fundação para a Ciência e Tecnologia, FCT, Portugal, (UID/QUI/00686/2016 and UID/QUI/00686/2019), and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Process Capes Pro Forenses 25/2014).

Received 23 Mar. 2020; and in revised form 28 May 2020; accepted 29 May 2020.

electrodes is attracting more researchers, as it is a biodegradable material (5,12–15).

Paper electrodes, also called electrochemical paper-based analytical devices (ePADs) or paper-based electrochemical devices (PEDs), have been successfully applied in pharmaceutical, biological, food, and environmental analysis (9,16–24). Also, in forensic analysis they have been employed to analyze analgesics and sedatives in whiskeys (25). Besides the diversity of electrodes geometry and arrangements that are possible, different materials may be used ranging from metals to carbon inks or even graphite from a simple drawing pencil, which is easily applied on paper (5,23–28).

The paper electrodes drawn with pencils are noteworthy due to the easy access to these materials and quick execution, since you only need to draw them. But this is also a disadvantage, the reproducibility of the drawing and the voltammetric response is a difficulty, but it has been circumvented by the use of stainless steel (5,26) and polyester (25) molds to draw the electrodes, and even by the use of pencil attached to a printer (29,30).

The use of electrochemistry in forensic chemistry is highlighted for the identification and quantification of drugs of abuse, including cocaine, Δ9-tetrahydrocannabinol (Δ9-THC), 3,4-methylenedioxymethamphetamine (MDMA), N-benzyl-substituted phenethylamines (NBOMe), and lysergic acid diethylamide (LSD) (31–43). LSD was commonly used in the 1960s, but it remains the recurrent hallucinogen among young people. It is usually found and consumed in the form of bottlers. Its correct identification is of particular interest to the forensic scientists given that drug diversity has increased (41–44).

Because a substantial amount of drugs is seized on a monthly basis and the demand for faster and less costly analyses is growing, this work aims to demonstrate how quickly, simply, and cheaply an electrode can be made with paper, pencil, and silver with an easier to paint design, increasing the reproducibility of the electrode, for LSD detection and quantification in seized samples.

Materials and Methods

Apparatus

A potentiostat/galvanostat (model PGSTAT128N form Metrohm) was used for voltammetric determinations. Comparative measurements were conducted between the paper-based electrodes presented in this work, and the SPCE from DropSens (model DR-110) to verify the LSD response. For comparison, results are expressed as current density. The active area of the electrodes was evaluated in potassium hexacyanoferrate solutions.

A gas chromatograph (Shimadzu GC-2010 Plus) coupled with a mass spectrometer (Shimadzu GCMS-QP2010 SE) was used to analyze seized samples for comparative purpose. Data were acquired and analyzed by using the GCMS solution version 4.41 software (Shimadzu Corporation). The analyte was separated on an amine column (OPTIMA 35MS; 30 m × 0.25 mm i.d.; 0.25-μm film thickness; Macherey-Nagel, Germany). The oven temperature was programmed to start at 90°C, to increase to 300°C at 30°C/min, to remain at 300°C for 2 min, to increase to 320°C at 10°C/min, to remain at 320°C for 2 min, to increase to 340°C at 10°C/min, and to remain at 340°C for 2 min. The carrier gas was helium (purity 99%) at a flow rate of 4.0 mL/min. The injection volume was 1.0 μL, and the temperatures of the injection port, ion source, and interface were 250, 250, and 280°C, respectively. A full scan mode in the *m/z* range from 50 to 450 was performed to acquire MS for the qualitative analyses. The seized samples were also identified by matching the full scan spectra against the instrument spectral library (SWGDRUG MS library version 3.3). For the quantitative analysis, the SIM scan mode was performed with *m/z* 323, 221, and 181.

Scanning electron microscopy (SEM) was performed on a Carl Zeiss microscope (model EVO 50). The paper-based electrode was coated with a thin Au layer with Bal-Tec Sputter Coater SEM Sample Prep (model SCD 050) and analyzed at an accelerating voltage of 20 kV in the low vacuum mode.

Standard Solutions and Samples

Acetonitrile, methanol, potassium hexacyanoferrate ($K_3[Fe(CN)_6]$), sulfuric acid (H_2SO_4), and lithium chloride (LiCl) were obtained from Merck (Germany); sodium chloride (NaCl), potassium chloride (KCl), sodium perchlorate ($NaClO_4$), and potassium perchlorate ($KClO_4$) were acquired from Vetec (Brazil); and perchloric acid ($HClO_4$) and lithium perchlorate ($LiClO_4$) were purchased from Sigma-Aldrich.

The $K_3[Fe(CN)_6]$ (1.0×10^{-3} mol/L) was prepared in 0.1 mol/L KCl and kept in a freezer while it was not being used for analysis. The same procedure was followed for the 0.1 mol/L H_2SO_4 solution. All the supporting electrolyte solutions were prepared on the same day of the analysis.

A standard commercial solution of lysergic acid diethylamide (LSD - 1.0 mg/mL, Cerilliant®) in acetonitrile was diluted to a concentration of 3.09×10^{-4} mol/L. The lysergic acid amide standard (LSA - 0.5 mg, Toronto Research Chemicals, Canada) was also diluted to 1.87×10^{-3} mol/L with methanol. To study the interfering substances that are usually apprehended with LSD, MDMA (1.0 mg/mL, Cerilliant®) and methamphetamine (0.1 mg/mL, LGC, U.K.) were employed.

Seized LSD samples were obtained through cooperation between this research group and the laboratory of toxicological analysis of the Institute of Criminalistics in Ribeirão Preto, state of São Paulo, Brazil. The drug was extracted from the blotter with three 2.0-mL aliquots of water/methanol (1:1 v/v) followed by ultrasonication for 6 min, which gave a total of 6.0 mL that was filtered through a 45-μm filter. For the chromatography analysis, the water/methanol was evaporated, and the sample was resuspended in 1.0 mL methanol.

Paper-based Electrodes

The electrode shapes and cell configuration were based on the commercial screen-printed electrode; they are arranged, from left to right, as counter, working, and reference electrodes. The paper electrode arrangement comprised a circular working electrode with 3-mm diameter and rectangular counter and reference electrodes, while the commercial screen-printed electrode has a circular shape for all electrodes and a 4-mm diameter working electrode (Fig. 1).

Four types of paper sheets were tested, with different weight and roughness. The thinnest and smoothest was sulfite paper with 75 g/m^2, and the thickest and roughest was watercolor paper with 300 g/m^2, all from Canson®.

Simultaneously, the graphite pencil used to paint the electrodes was also varied from 2B to 8B (Faber-Castell). The nomenclature given to the pencils is based on the amount of graphite that is present in the composition of the mine. The mixture is mainly composed of graphite, clay, and wax, and the pencil compositions are depicted in Table 1.

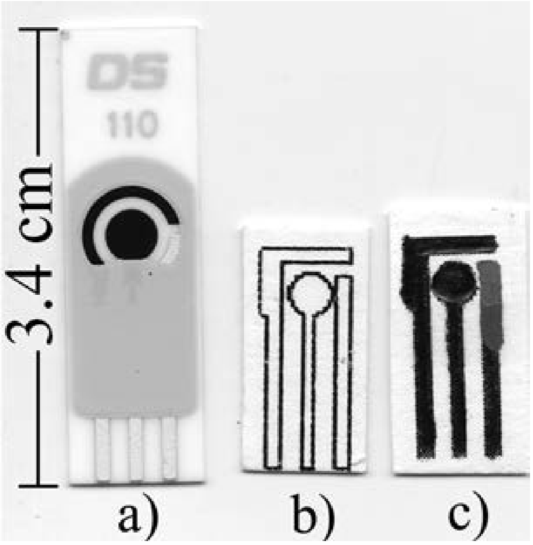


FIG. 1—*Comparison between the paper-based electrodes and the screen-printed electrodes: (a) commercial screen-printed carbon electrode; (b) paper electrode contour; (c) final configuration of the paper electrode painted with graphite pencil and silver ink. [Color figure can be viewed at wileyonlinelibrary.com]*

The paper-based electrodes contour was printed by a laser printer, and then, the drawing was filled in by hand with pencil, until a homogeneous graphite layer was obtained. Using a brush, a silver ink (Sigma-Aldrich) layer was deposited to form the reference electrode. The electrode was then left to dry at room temperature.

Once the silver paint had dried completely, a hydrophobic barrier was painted around the reference and counter electrodes and contacts with transparent nail polish (Colorama) purchased at the supermarket, so that the drop of sample solution covered the set of three electrodes and did not spread away. After the nail polish dried, the electrode was cut with scissors to fit the potentiostat connector and was read to use.

ANOVA Test

A one-way ANOVA analysis was performed by Excel (Microsoft Office 2010) to compare results from the home-made paper electrodes with the commercial screen-printed electrodes, with a confidence interval of 95%. The standard deviation for repeatability (S_r) was calculated by the square root of the mean square within groups (MS_w). Equation 1 was used for the reproducibility (S_R), where MS_b is the mean square between groups, and *n* is the number of replicas.

TABLE 1—*Pencil composition in percentage of graphite, clay, and wax (45)*

Pencil	Graphite (%)	Clay (%)	Wax (%)
2B	74	20	5
3B	76	18	5
4B	79	15	5
5B	82	12	5
6B	84	10	5
8B	90	4	5

$$S_R = \sqrt{MS_w + \frac{MS_b - MS_w}{n}} \quad (1)$$

Results and Discussion

Paper-based Electrodes Electrochemical Characterization

Unlike other works, which used stainless steel (5,26), adhesive paper label (23), and polyester (25) molds to draw the electrodes, it was decided to print the electrode contour on paper first. For painting, the electrode contours were followed, but were fully covered with graphite and silver paint, so they do not interfere with the voltammetric signal.

Different formats for the electrodes were tested, including a square format for the working electrode and larger areas for the counter and reference electrodes. The configuration that resulted in the best voltammetric response for potassium hexacyanoferrate was shown in the Fig. 1, with the circular working electrode and rectangular configurations for the auxiliary and reference electrodes.

The scanning electron microscopy (SEM) images reveal differences between the morphologies of the 300 g/m^2 paper surface before and after it was painted with 8B pencil (Fig. 2*a*). The left side of the image shows the cellulose fibers forming a disorganized arrangement. On the right side, it is possible to see the graphite layers covering and filling the space between the fibers, indicating the formation of the conductive material. The graphite layers are more visible at larger magnifications (Fig. 2*b,c*). Grooves in the circular direction formed by the pencil scratching can also be observed.

To test the more adequate type of paper/pencil combination for the construction of the paper-based electrodes, electrochemical response of potassium hexacyanoferrate was examined. The peaks shape (both anodic and cathodic) and peaks separation ($\Delta E_p = 0.79$ V) obtained in the graphite-paper electrodes indicated that they suffer from a severe ohmic distortion. This effect decreased substantially by applying a pretreatment to the working electrode consisting of 10 cycles of cyclic voltammetry in a 0.1 mol/L H_2SO_4 solution, from −1.5 to −0.5 V, at 50 mV/sec. After this activation, the peaks became more defined, the current was higher, and the ΔE_p decreased to 0.30 V.

The 300 g/m^2 watercolor paper combined with the 8B pencil (that has the highest percentage of graphite), and using the electrochemical activation procedure earlier described, it was possible to record a voltammetric signal that was most like the signal obtained with the commercial screen-printed electrode after the same electrochemical activation procedure (Fig. 3). The voltammograms acquired using the electrodes constructed using the different kinds of papers and pencils are available in the Figures S1 and S2, respectively.

The electroactive area was also calculated before and after the electrochemical activation of the working electrode in the acid solution. Voltammograms for ferricyanide were obtained at different scan rates, from 10 to 300 mV/sec. A linear relation was obtained by plotting the peak current from the anodic scan versus the square root of the scan rate, before the activation, $i(\mu A) = 0.22\, v^{1/2}$ (mV/sec) + 0.75, and after the activation $i(\mu A) = 0.55\, v^{1/2}$ (mV/sec) + 1.69.

From the Randle–Sevcik equation (Eq. 2), the electroactive area of the electrode was obtained from the angular coefficient,

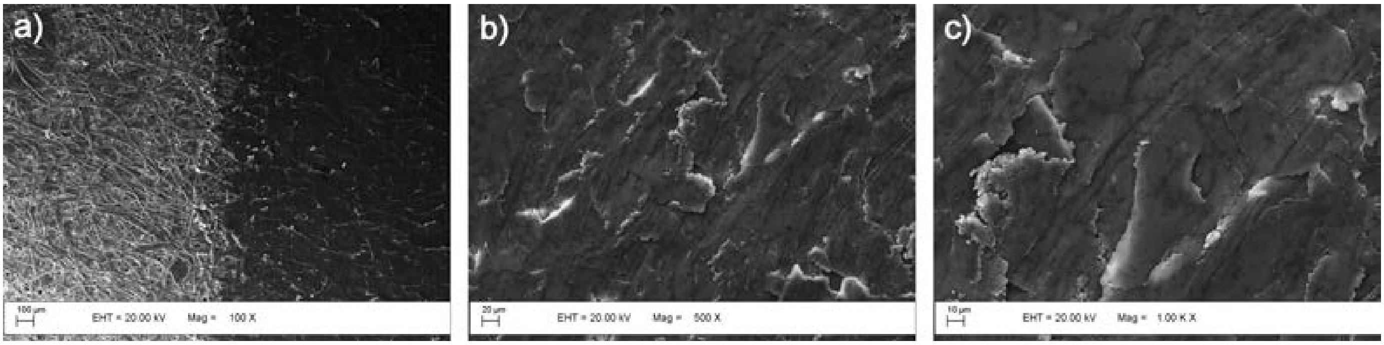


FIG. 2—*SEM images showing the morphology of (a) the border of the working electrode produced by drawing on a substrate of 300 g/m^2 paper with the pencil 8B under a magnification of 100×; (b) graphite layers on the surface of the working electrode under magnification of 500×; (c) graphite layers on the surface of the working electrode under magnification of 1000×.*

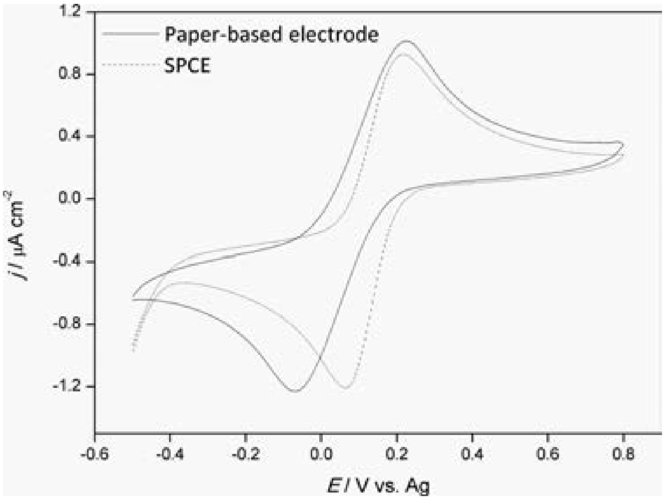


FIG. 3—*Cyclic voltammograms obtained from 1.0 × 10^{-3} mol/L $K_3[Fe(CN)_6]$ in 0.1 mol/L KCl after 10 activation cycles in 0.10 mol/L H_2SO_4, from − 1.5 to − 0.5 V at 50 mV/sec using: (——) the paper-based electrodes (using the 8B pencil and the 300 g/m^2 paper) and (-----) the commercial screen-printed carbon electrode.*

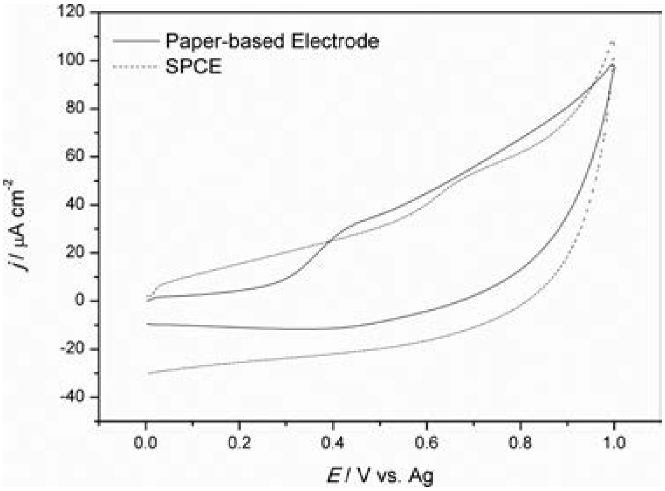


FIG. 4—*Cyclic voltammograms obtained from a 6.18 µmol/L LSD in 0.05 mol/L $LiClO_4$ supporting electrolyte solution at 100 mV/sec using: (——) the paper-based electrodes (using the 8B pencil and the 300 g/m^2 paper) and (-----) the commercial screen-printed carbon electrode.*

using the known values of the diffusion coefficient ($D^{1/2} = 2.75\times10^{-3}$ cm/sec$^{1/2}$), the number of electrons involved in the reaction (n = 1), and the analyte concentration, (C = 1.0 mmol/L).

$$i = 2.69 \times 10^5 a D^{1/2} n^{1/2} \nu^{1/2} C \qquad (2)$$

The active area of the electrode increased considerably from 0.031 cm^2 before activation to 0.075 cm^2 after activation, attaining a value close to the electrode geometrical area (0.071 cm^2).

The ability to produce devices with similar characteristics, with respect to geometry and dimensions of the electrode assembly and the morphology of the surface of the working electrode, was tested by means of analysis of variance, ANOVA. The reproducibility of results obtained from 10 electrodes was compared with repeatability associated to five consecutive repetitions of each measurement. The variations were tested for two parameters: ΔE_p and i_{oxi}/i_{red} (Tables S1 and S2, respectively). For the peaks separation, the standard deviation for the repeatability and the reproducibility was 1.30 and 3.04%, respectively; for the currents ratio, these values were slightly higher, 2.04 and 6.85%, respectively. These values are adequate for an analytical application and quite remarkable considering that the electrodes are hand-made.

LSD Analysis

The optimized paper-based electrodes, which construction and test details were provided previously, were applied for LSD detection and quantification. For comparison purposes, the paper-based electrodes results are compared to the screen-printed carbon electrodes (SPCE) results by the same voltammetric methodologies.

In both cyclic and square wave voltammetric techniques, different supporting electrolytes were tested. The best current signals at lower potentials were achieved using 0.05 mol/L $LiClO_4$ (Fig. 4). The oxidation peak potential was about 0.25 V lower for the paper electrodes (E_p = 0.43 V) as compared to the SPCE (E_p = 0.68 V). This difference can be due to the use of different pseudo-reference electrodes in the two assays. The Ag inks used by us may certainly have a different composition from that used by the manufacturer of the commercial SPCE.

In the voltammetry of LSD (Fig. 6), a single oxidative wave is visible in the oxidation range potentials. In an attempt to

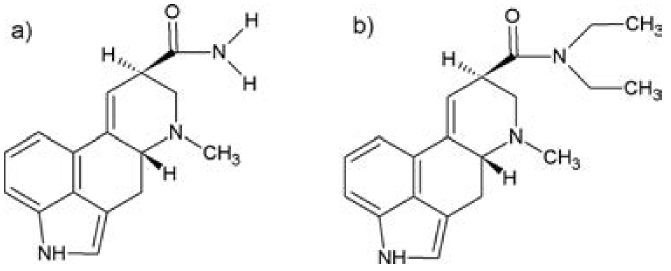


FIG. 5—*A comparison between the molecular structures of (a) LSA and (b) d-LSD.*

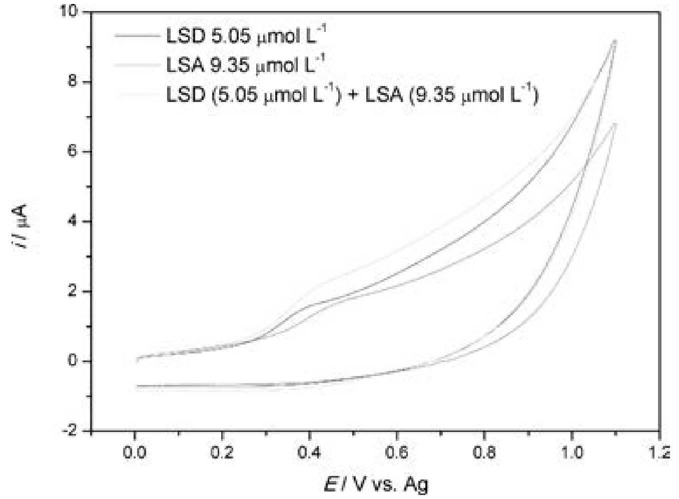


FIG. 6—*Cyclic voltammograms obtained from 5.05 μmol/L LSD (——), 9.35 μmol/L LSA () and a mixture of both species (——) using the paper-based electrode (using 8B pencil and the 300 g/m^2 paper) at 100 mV/sec. [Color figure can be viewed at wileyonlinelibrary.com]*

identify the oxidative reaction that was involved in the process, another psychoactive drug whose structure resembles the LSD was analyzed: lysergic acid amide. This substance occurs in the seeds of Morning Glory species and is commonly used in ritual ceremonies because it has almost the same hallucinogenic properties as LSD (see Fig. 5) (46,47).

The voltammetric resemblance between the drugs (Fig. 6) indicated that the reaction probably occurred in the indole group, as described in the literature (43). Nevertheless, the voltammetric response does not show a clear evidence for the presence of the second oxidation peak, which ruled out the second part of the dimerization. Although the voltammetric response of these drugs occurs at potentials, 0.38 V for LSD and 0.46 V for LSA, in a solution containing both drugs a single peak emerges at 0.42 V. This fact may result from the proximity of the two peak potentials, leading to a single signal that rise at close to the potential where the LSD response starts.

The peak current from LSD obtained at different scan rates follows a linear relation with the square root of the scan rate, i (μA) = 0.033 $v^{1/2}$ (mV/sec) − 0.17, $r^2 = 0.991$, which is typical of an electrode reaction controlled by diffusion. After performing 10 consecutive cycles, the peak current shows a different behavior that is consistent to a mixed diffusion-adsorption mechanism. This statement is corroborated by the log i *versus* log v plot, as a linear trend was obtained with a slope between 0.5 and 1.0 (log i(μA) = 0.83 log v − 8.43, $r^2 = 0.995$) (38)

The quantification of LSD was performed by square wave voltammetry and cyclic voltammetry (although this technique is less used for quantification as it usually does not attain the performance parameters of the pulsed techniques).

All the voltammetric parameters including scan rate, potential range, potential step, frequency, amplitude, and potential and time for preconcentration were optimized (Table 2). The pH was also varied from 2 to 10. Although the lower peak potential was attained for pH 10, the voltammetric signal suffers from lower reproducibility. The optimal conditions were settled for pH 6.

Voltammetric data lead to linear trends between the oxidation peak areas and LSD concentration, giving rise to four analytical curves, by combining the two electrochemical techniques with the two electrodes. Table 3 depicts all the analytical parameters for comparison of the four analytical curves. To ensure that no LSD adsorption blocked the analytical signal, the curve for the square wave voltammetry using the paper-based electrode was obtained by employing one electrode for each solution (Fig. 7).

The LoD and LoQ values were obtained, with 3 replicates for each value, on the basis of the relations 3SD/m and 10SD/m, respectively, where m is the amperometric sensitivity of the curve. By cyclic voltammetry, results from the two electrodes are comparable, whereas by square wave voltammetry there is a clear difference between results from the two sets of electrodes. This improvement may have resulted from the changing of the paper-based electrodes between the analysis of each solution, avoiding LSD adsorption on the electrode surface and the analytical signal block. The continuous recording of voltammograms using only one SPCE may have led to partial blocking of the electrode surface, resulting in higher values of LoD and LoQ.

Interference Analysis

As shown before, the present methodologies did not allow differentiation between LSD and LSA as both voltammograms occur at similar potentials. This fact does not translate into the actual problem as they have different distribution channels, LSD

TABLE 2—*Cyclic and square wave voltammetric parameters used for LSD analysis*

	SPCE		Paper-Based Electrode	
	Cyclic Voltammetry	Square Wave Voltammetry	Cyclic Voltammetry	Square Wave Voltammetry
Potential Range (V)	0.0–1.0	0.0–1.0	0.0–1.0	0.0–1.1
Scan Rate (mV/sec)	400	350	250	350
Step Potential (V)	0.003	0.014	0.03	0.014
Frequency (Hz)	–	25	–	25
Amplitude (V)	–	0.035	–	0.035
Preconcentration Time (sec)	90	150	120	150
Preconcentration potential (V)	0.0	0.0	0.0	0.0

TABLE 3—*Analytical parameters for LSD quantification and comparison between the SPCE and the paper-based systems*

	SPCE		Paper-Based Electrodes	
	Cyclic Voltammetry	Square Wave Voltammetry	Cyclic Voltammetry	Square Wave Voltammetry
Equation	$y = 2.98 \times 10^{-2}x + 1.27 \times 10^{-8}$	$y = 5.77 \times 10^{-2}x + 1.02 \times 10^{-8}$	$y = 3.33 \times 10^{-3}x - 2.49 \times 10^{-10}$	$y = 1.72 \times 10^{-2}x + 5.31 \times 10^{-9}$
R^2	0.996	0.997	0.996	0.996
SD	5.52 10^{-9}	1.23 10^{-8}	5.83 10^{-10}	2.19 10^{-9}
LoD (μmol/L)	0.55 (±0.10)	0.63 (±0.02)	0.52 (±0.08)	0.38 (±0.02)
LoQ (μmol/L)	1.85 (±0.25)	2.13 (±0.04)	1.74 (±0.20)	1.27 (±0.07)

In parentheses are standard deviations calculated based on $n = 3$.

is a synthetic drug that is produced in laboratory and can be found in blotters, microdots, and gels. LSA can be extracted from seeds, but the seeds are usually consumed without any kind of extraction (44,46,47). Thus, these two drugs are unlikely to be found together.

The police commonly apprehend the drugs MDMA and methamphetamine with LSD. Therefore, the ability of voltammetry to identify LSD and distinguish it from MDMA or methamphetamine is relevant. The voltammetric response of solutions containing methamphetamine and MDMA separately and from a mixture of LSD with MDMA is shown in Fig. 8. While for methamphetamine, there was no voltammetric response under the employed condition (not even for mM concentrations); MDMA exhibits a peak at about 0.96 V (assigned in Fig. 8 by 2). This peak is very well separated from the LSD (assigned in Fig. 8 by 1) and therefore does not interfere with its detection or quantification.

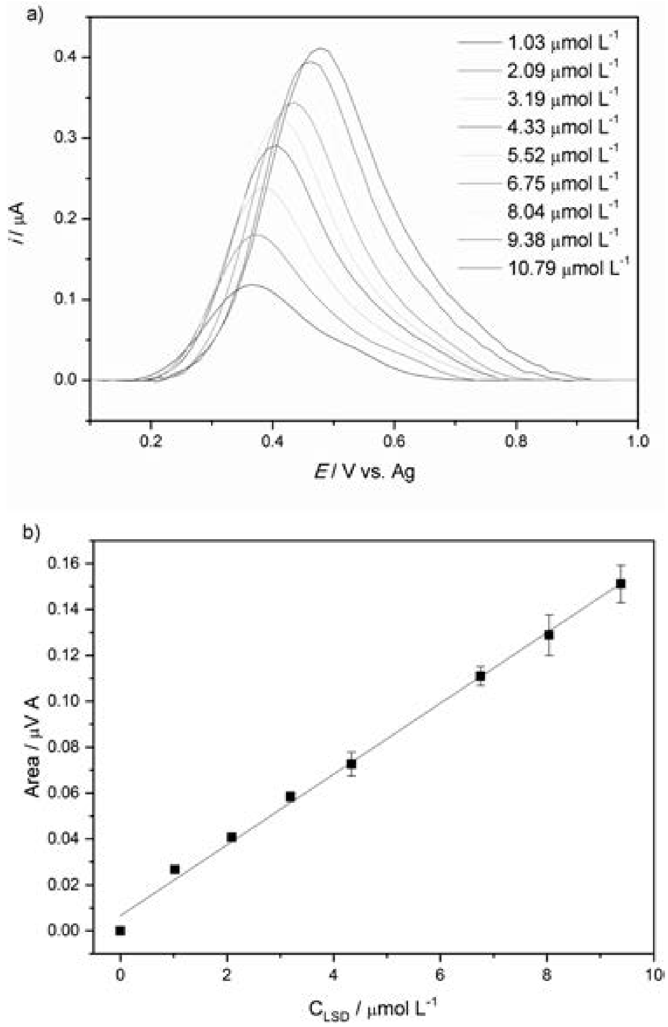


FIG. 7—*Voltammetric behavior of the paper-based electrode (using 8B pencil and the 300 g/m^2 paper) in response to different LSD concentrations; (a) voltammograms obtained by square wave voltammetric modality with* A = *0.035 V,* f = *25 Hz, and preconcentration of 0.0 V for 150 sec; (b) The analytical curve was obtained for LSD species employing one new electrode for each concentration. [Color figure can be viewed at wileyonlinelibrary.com]*

Sample Analysis

Two samples (A and B) that could contain LSD were provided by the Scientific Police of Ribeirão Preto and were analyzed by the developed methodologies and by GC-MS, for comparison.

Sample A was analyzed by both techniques, and no LSD was detected. The presence of LSD was detected in the sample B and was identified on the basis of the GC-MS library with 92% of compatibility.

After identifying the presence of LSD in sample B, its quantification was performed by both techniques (with SIM scan mode and monitoring ion 221 for the gas chromatography). Since only two samples of blotters were obtained, there was not enough data to provide a correct sample blank, but as the two samples had a very similar chromatographic profile, taking into account the baseline, sample A was considered as blank of the sample B, minimizing the matrix effect.

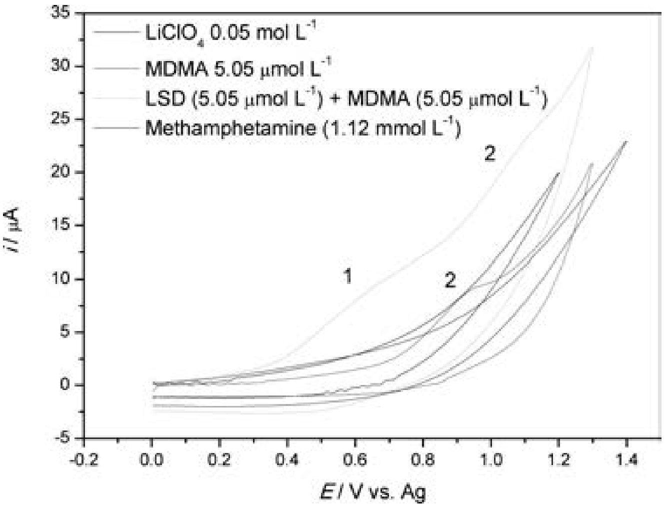


FIG. 8—*Voltammetric response to LSD, MDMA, and methamphetamine using the paper-based electrodes (using 8B pencil and the 300 g/m^2 paper), at 250 mV/sec. "1" designates the voltammetric peak current of LSD species, and "2" designates the voltammetric peak current of MDMA species. [Color figure can be viewed at wileyonlinelibrary.com]*

Then, the response obtained in the analysis of sample B was interpolated with the analytical curve obtained by cyclic and square wave voltammetry presented in Table 3, and the analytical curve provided by gas chromatography, which relates the peak area (y) to the concentration of the LSD standard solution (x), according to the Equation 3.

$$y = 2.75 \times 10^{8}x - 2.17 \times 10^{-2} \qquad (3)$$

The GC result for LSD quantification was 1.34×10^{-5} mol/L, and using the SPCE, the LSD concentration values obtained in sample B were 1.40×10^{-5} mol/L, by cyclic voltammetry, and 1.31×10^{-5} mol/L, by square wave voltammetry. Using the paper-based electrodes, these values were 1.45×10^{-5} mol/L and 1.35×10^{-5} mol/L, respectively. All the results obtained using the voltammetric techniques are in agreement between them and with the chromatographic data, as the differences between them remained below 10%. These results demonstrate that the paper-based device described in this paper is adequate for LSD detection and quantification and could be used for the rapid screening of samples *in loco*.

Conclusions

The paper-based devices build and optimized in this work are very promising tools for the electrochemical analysis. Besides being produced using accessible low-cost materials, its construction is simple and inexpensive, as it does not require any sophisticated equipment. The paper-based electrodes gave surprisingly good results, in terms of reproducibility, demonstrating that the hand-painting procedure does not introduce important variations in the size or in the deposited graphite layer. They are easy to handle, and eco-friendly (paper substrate and graphite electrodes). The paper-based electrodes and the developed methodologies of operation demonstrated to be adequate for LSD detection and quantification as they provided results similar to those obtained for a commercial screen-printed electrode and to chromatographic analysis.

This device can thus be considered a valid alternative for forensic analysis that is cheaper and more sustainable.

Acknowledgments

The authors thank Dr. Cynthia Maria de Campos Prado Manso for her assistance in revising and editing the text.

References

1. Renedo OD, Alonso-Lomillo MA, Martínez MJA. Recent developments in the field of screen-printed electrodes and their related applications. Talanta 2007;73(2):202–19. https://doi.org/10.1016/j.talanta.2007.03.050.
2. Araujo WR, Cardoso TMG, da Rocha RG, Santana MHP, Muñoz RAA, Richter EM, et al. Portable analytical platforms for forensic chemistry: a review. Anal Chim Acta 2018;1034(30):1–21. https://doi.org/10.1016/j.aca.2018.06.014.
3. Oliveira LP, Rocha DP, Araujo WR, Munoz RAA, Paixão TRLC, Salles MO. Forensics in hand: new trends in forensic devices (2013–2017). Anal Methods 2018;10(43):5135–63. https://doi.org/10.1039/c8ay01389f.
4. Slepchenko GB, Gindullina TM, Nekhoroshev SVJ. Capabilities of the electrochemical methods in the determination of narcotic and psychotropic drugs in forensic chemistry materials. Anal Chem 2017;72(7):703–9. https://doi.org/10.3390/bios6030045.
5. Bernalte E, Foster CW, Brownson DAC, Mosna M, Smith GC, Banks CE. Pencil it in: exploring the feasibility of hand-drawn pencil electrochemical sensors and their direct comparison to screen-printed electrodes. Biosensors 2016;6(3):45–64. https://doi.org/10.1134/S1061934817070127.
6. Liu J, Geng Z, Fan Z, Liu J, Chen H. Point-of-care testing based on smartphone: the current state-of-the-art (2017–2018). Biosens Bioelectron 2019;132:17–37. https://doi.org/10.1016/j.bios.2019.01.068.
7. Ji D, Liu L, Li S, Chen C, Lu Y, Wu J, et al. Smartphone-based cyclic voltammetry system with graphene modified screen printed electrodes for glucose detection. Biosens Bioelectron 2017;98:449–56. https://doi.org/10.1016/j.bios.2017.07.027.
8. Shaw L, Dennany L. Applications of electrochemical sensors: forensic drug analysis. Curr Opin Electrochem 2017;3(1):23–8. https://doi.org/10.1016/ j.coelec.2017.05.001.
9. Malon RSP, Heng LY, Córcoles EP. Recent developments in microfluidic paper-, cloth-, and thread-based electrochemical devices for analytical chemistry. Rev Anal Chem 2017;36(4):20160018. https://doi.org/10.1515/revac-2016-0018.
10. Hart JP, Crew A, Crouch E, Honeychurch KC, Pemberton RM. Some recent designs and developments of screen-printed carbon electrochemical sensors/biosensors for biomedical, environmental, and industrial analyses. Anal Lett 2004;37(5):789–830. https://doi.org/10.1081/AL-120030682.
11. Kim J, Kumar R, Bandodkar AJ, Wang J. Advanced materials for printed wearable electrochemical devices: a review. Adv Electron Mater 2007;3:1600260. https://doi.org/10.1002/aelm.201600260.
12. Tobjörk D, Österbacka R. Paper electronics. Adv Mater 2011;23(17):1935–61. https://doi.org/10.1002/adma.201004692.
13. Mahadeva SK, Walus K, Stoeber B. Paper as a platform for sensing applications and other devices: a review. ACS Appl Maters Inter 2015;7(16):8345–62. https://doi.org/10.1021/acsami.5b00373.
14. Kanaparthi S, Badhulika S. Solvent-free fabrication of paper based all-carbon disposable multifunctional sensors and passive electronic circuits. RSC Adv 2016;6(98):95574–83. https://doi.org/10.1039/c6ra21457f.
15. Nantaphol S, Kava AA, Channonr B, Kondo T, Siangproh W, Chailapakul O, et al. Janus electrochemistry: Simultaneous electrochemical detection at multiple working conditions in a paper-based analytical device. Anal Chim Acta 2019;1056:88–95. https://doi.org/10.1016/j.aca.2019.01.026.
16. Kaneta T, Alahmad W, Varanusupakul P. Microfluidic paper-based analytical devices with instrument-free detection and miniaturized portable detectors. Appl Spectrosc 2019;54(2):117–41. https://doi.org/10.1080/05704928.2018.1457045.
17. Adkins J, Boehle K, Henry C. Electrochemical paper-based microfluidic devices. Electrophoresis 2015;36(16):1811–24. https://doi.org/10.1002/elps.201500084.
18. Cate DM, Adkins JA, Mettakoonpitak J, Henry CS. Recent developments in paper-based microfluidic devices. Anal Chem 2015;87(1):19–41. https://doi.org/10.1021/ac503968p.
19. Garcia PT, Gabriel EFM, Pessôa GS, Júnior JCS, Filho PCF, Guidugli RBF, et al. Paper-based microfluidic devices on the crime scene: a simple tool for rapid estimation of post-mortem interval using vitreous humour. Anal Chim Acta 2017;974:69–74. https://doi.org/10.1016/j.aca.2017.04.040.
20. Santhiago M, Wydallis JB, Kubota LT, Henry CS. Construction and electrochemical characterization of microelectrodes for improved sensitivity in paper-based analytical devices. Anal Chem 2013;85(10):5233–9. https://doi.org/10.1021/ac400728y.
21. Santhiago M, Henry CS, Kubota LT. Low cost, simple three dimensional electrochemical paper-based analytical device for determination of p-nitrophenol. Electrochim Acta 2014;130:771–7. https://doi.org/10.1016/j.electacta.2014.03.109.
22. Orzari LO, Freitas RC, Andreotti IAA, Gatti A, Janegitz BC. A novel disposable self-adhesive inked paper device for electrochemical sensing of dopamine and serotonin neurotransmitters and biosensing of glucose. Biosens Bioelectron 2019;138:111310. https://doi.org/10.1016/j.bios.2019.05.015.
23. Orzari LO, Freitas RC, Andreotti IAA, Bergamini MF, Junior LHM, Janegitz BC. Disposable electrode obtained by pencil drawing on corrugated fiberboard substrate. Sensor Actuat B 2018;264:20–6. https://doi.org/10.1016/j.snb.2018.02.162.
24. Oliveira VXG, Dias AA, Carvalho LL, Cardoso TMG, Colmati F, Coltro WKT. Determination of ascorbic acid in commercial tablets using pencil drawn electrochemical paper-based analytical devices. Anal Sci 2018;34(1):91–5. https://doi.org/10.2116/analsci.34.91.
25. Dias AA, Cardoso TMG, Chagas CLS, Oliveira VXG, Munoz RAA, Henry CS, et al. Detection of analgesics and sedation drugs in whiskey using electrochemical paper-based analytical devices. Electroanalysis 2018;30(10):2250–7. https://doi.org/10.1002/elan.201800308.

26. Foster CW, Brownson DAC, Souza APR, Bernalte E, Iniesta J, Bertotti M, et al. Pencil it in: pencil drawn electrochemical sensing platforms. Analyst 2016;141(13):4055–64. https://doi.org/10.1039/c6an00402d.
27. Araujo WR, Paixão TRLC. Fabrication of disposable electrochemical devices using silver ink and office paper. Analyst 2014;139(11):2742–7. https://doi.org/10.1039/c4an00097h.
28. Dornelas KL, Dossi N, Piccin E. A simple method for patterning poly (dimethylsiloxane) barriers in paper using contact-printing with low-cost rubber stamps. Anal Chim Acta 2015;858:82–90. https://doi.org/10.1016/j.aca.2014.11.025.
29. Dossi N, Petrazzi S, Toniolo R, Tubaro F, Terzi F, Piccin E, et al. Digitally controlled procedure for assembling fully drawn paper-based electroanalytical platforms. Anal Chem 2017;89(19):10454–60. https://doi.org/10.1021/acs.analchem.7b02521.
30. Ataide VN, Mendes LF, Gama LILM, de Araujob WR, Paixão TRLC. Electrochemical paper-based analytical devices: ten years of development. Anal Methods 2020;12(8):1030–54. https://doi.org/10.1039/c9ay02350j.
31. Santhiago M, Strauss M, Pereira MP, Chagas AS, Bufon CCB. Direct drawing method of graphite onto paper for high-performance flexible electrochemical sensors. ACS Appl Mater Inter 2017;9(13):11959–66. https://doi.org/10.1021/acsami.6b15646.
32. Ribeiro MFM, Júnior JWC, Dockal ER, Mccord BR, Oliveira MF. Voltammetric determination of cocaine using carbon screen printed electrodes chemically modified with uranyl Schiff base films. Electroanalysis 2016;28(2):320–6. https://doi.org/10.1002/elan.201500372.
33. Oliva PHB, Katayama JMT, Oiye EN, Ferreira B, Ribeiro MFM, Ipólito AJ, et al. Determination of cocaine by square wave voltammetry with carbon paste electrodes. Braz J Forensic Sci Med Law Bioethics 2019;8(3):149–64. https://doi.org/10.17063/bjfs8(3)y2019149.
34. Silva TG, Araujo WR, Muñoz RAA, Richter EM, Santana MHP, Coltro WKT, et al. Simple and sensitive paper-based device coupling electrochemical sample pretreatment and colorimetric detection. Anal Chem 2016;88(10):5145–51. https://doi.org/10.1021/acs.analchem.6b00072.
35. Wanklyn C, Burton D, Enston E, Bartlett CA, Taylor S, Raniczkowska A, et al. Disposable screen printed sensor for the electrochemical detection of delta-9-tetrahydrocannabinol in undiluted saliva. Chem Cent J 2016;10:1–11. https://doi.org/10.1186/s13065-016-0148-1.
36. Balbino MA, Oiye EN, Ribeiro MFM, Júnior JWC, Eleotério IC, Ipólito AJ, et al. Use of screen-printed electrodes for quantification of cocaine and Δ9-THC: adaptions to portable systems for forensic purposes. J Solid State Electr 2016;20(9):2435–43. https://doi.org/10.1007/s10008-016-3145-3.
37. Tadini MC, Balbino MA, Eletotério IC, Oliveira LS, Dias LG, Jean-François Demets G, et al. Developing electrodes chemically modified with cucurbit[6]uril to detect 3,4-methylenedioxymethamphetamine (MDMA) by voltammetry. Electrochim Acta 2014;121:188–93. https://doi.org/10.1016/j.electacta.2013.12.107.
38. Couto RAS, Costa SS, Junior BM, Pacheco JG, Fernandes E, Carvalho F, et al. Electrochemical sensing of ecstasy with electropolymerized molecularly imprinted poly(o-phenylenediamine) polymer on the surface of disposable screen-printed carbon electrodes. Sensors Actuat B Chem 2019;290:378–86. https://doi.org/10.1016/j.snb.2019.03.138.
39. Oiye EM, Katayama JMT, Ribeiro MFM, Oliveira MF. Electrochemical analysis of 25H-NBOMe by square wave voltammetry. Forensic Chem 2017;5:86–90. https://doi.org/10.1016/j.forc.2017.07.001.
40. Elbardisy HM, Foster CW, Marron J, Mewis RE, Sutcliffe OB, Belal TS, et al. Quick test for determination of n-bombs (phenethylamine derivatives, NBOMe) using high-performance liquid chromatography: a comparison between photodiode array and amperometric detection. ACS Omega 2019;4(11):14439–50. https://doi.org/10.1021/acsomega.9b01366.
41. Ribeiro MFM, Oiye EN, Katayama JMT, Ipólito AJ, Oliveira MF. Simple and fast analysis of LSD by cyclic voltammetry in aqueous medium. ECS Trans 2017;80(10):1259–68. https://doi.org/10.1149/08010.1259ecst.
42. Oiye EN, Ipólito AJ, Oliveira MF. Quantification of LSD in seized samples using one chromatographic methodology for diode array detection and electrochemical detection. Forensic Sci Criminol 2017;2(3):1–7. https://doi.org/10.15761/FSC.1000116.
43. Merli D, Zamboni D, Protti S, Pesavento M, Profumo A. Electrochemistry and analytical determination of Lysergic Acid Diethylamide (LSD) via adsorptive stripping voltammetry. Talanta 2014;130:456–61. https://doi.org/10.1016/j.talanta.2014.07.037.
44. Passie T, Halpern JH, Stichtenoth DO, Emrich HM, Hintzen A. The pharmacology of lysergic acid diethylamide: a review. CNS Neurosci Ther 2008;14(4):295–314. https://doi.org/10.1111/j.1755-5949.2008.00059.x.
45. Sousa MC, Buchanan JW. Observational models of graphite pencil materials. Comput Graph 2000;19(1):22–49. https://doi.org/10.1111/1467-8659.00386.
46. Mercurio I, Melai P, Capano D, Ceraso G, Carlini L, Bacci M. GC/MS analysis of morning glory seeds freely in commerce: can they be considered "herbal highs"? Egypt J Forensic Sci 2017;7(16):1–6. https://doi.org/10.1186/s41935-017-0016-8.
47. Paulke A, Kremer C, Wunder C, Achenbach J, Djahanschiri B, Elias A, et al. Argyreia nervosa (Burm. f.): receptor profiling of lysergic acid amide and other potential psychedelic LSD-like compounds by computational and binding assay approaches. J Ethnopharmacol 2013;148(2):492–7. https://doi.org/10.1016/j.jep.2013.04.044.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

FIG. S1. Voltammetric response obtained by using the paper-based electrodes made of different types of papers and painted with the 8B pencil from a solution 1,0 10^{-3} mol L^{-1} $K_3[Fe(CN)_6]$ in 0.1 mol L^{-1} KCl, after 10 activation cycles in 0.1 mol L^{-1} H_2SO_4, from −1.5 to −0.5 V at 50 mV s^{-1}.

FIG. S2. Voltammetric response using the paper-based electrodes made of different types of pencils painted on the 300 g m^{--2} watercolor paper from a solution 1,0 10^{-3} mol L^{-1} $K_3[Fe(CN)_6]$ in 0.1 mol L^{-1} KCl, after 10 activation cycles in 0.1 mol L^{-1} H_2SO_4, from −1.5 to −0.5 V at 50 mV s^{-1}.

Table S1. ANOVA results for the ΔE_p data obtained from the voltammetric response of a 1,0 10^{-3} mol L^{-1} $K_3[Fe(CN)_6]$ solution.

Table S2. ANOVA results for the i_{oxi}/i_{red} data obtained from the voltammetric response of a 1,0 10^{-3} mol L^{-1} $K_3[Fe(CN)_6]$ solution.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14544
Available online at: onlinelibrary.wiley.com

TECHNICAL NOTE

CRIMINALISTICS

He Zhang,[1] M.S.; Luoxi Liu,[1] Ph.D.; Yaping Luo,[2] Ph.D.; and Ran Chang,[3] M.S.

Determining Shoe Length from Partial Shoeprints

ABSTRACT: The length of a shoe has significant value in reflecting characteristics of the owner, and thereby, it can help in tracking suspects in criminal cases. However, the shoeprints left at a crime scene are often incomplete, resulting in difficulties in assessing shoe length. To find a way to estimate the shoe length from a partial shoeprint, 109 shoes with different sizes and general patterns were collected, and their prints were lifted using magnetic powder. Four feature points were defined on a shoeprint, and the longest distance between the feature points was defined as shoe length. Using linear, quadratic, and cubic regression analyses, a total of 15 equations were obtained between the shoe length and the other distances between the feature points. Out of these, the five most accurate equations were selected as the optimal equation. The verified test, including another 18 pairs of shoes, showed an average error of equations between 0.591 cm and 0.732 cm. The equations were also applied in two practical cases, resulting in good accuracy. The study demonstrates that shoe length could be determined from partial shoeprints through the proven equations.

KEYWORDS: footwear examination, shoeprint identification, partial shoeprint, shoe length, feature point, regression equation

Footwear impression, which reflects the characteristics of the outsole, is a forensic evidential item that appears at crime scenes and has played a significant role in solving criminal cases. Shoe length, meaning the distance from tiptoe to heel, is one of the class characteristics of the outsole. Shoe length can be used as one of the important class characteristics of imprint or impression evidence to convict suspects in the court, as different people have different shoe sizes. In the field of forensic science, many studies on shoe length have been accomplished, including on the relationship between footwear impression length and state of motion (1), the effects of shoe size on foot volumetric (2,3), length dispersion of shoes labeled with the same size (4), and the correlation between shoe-length fit, and diabetic peripheral neuropathy (5).

As a basic and vital characteristic, shoe length can easily be determined from a complete shoeprint. However, shoeprints found at crime scenes are usually incomplete, in some cases being distorted or obscured by movement or overstepping. Simultaneously, under various circumstances, partial shoeprints of the same shoe might seem different in a footwear impression. They can be affected by motion and substrate surface. Thus, it is difficult for investigators to obtain complete information regarding the outsole and shoe length.

Therefore, it is significant to ascertain the shoe length from partial shoeprints, and it is indispensable to develop a standard to study samples for unified analysis. In this study, we defined four common and easily identifiable points as feature points of partial shoeprints, studied quantitative relationships between feature lines (defined by connecting each feature point) and shoe length, and evaluated the equations we obtained. This study aimed to help investigators and footwear examiners determine the shoe length from partial shoeprints.

Materials and Methods

Shoes

A total of 109 pairs of shoes with different general patterns were collected from eight provinces of China through mail or directly asking for them in residential areas. The shoes (size 5–14) were worn by 106 people (43 males and 63 females). The owners lived in cities or rural regions, and their ages ranged from 16 to 70 years.

Exemplar Outsole Impressions

Exemplar outsoles were obtained by dusting the outsoles with magnetic fingerprint powder and covering them with adhesive film. A qualified sample can be obtained after the following five steps: clean, dust, adhere, tear off, and cover (6).

Clean: Outsoles were cleaned up by wet towel, which afterward was inverted to dry to confirm the neatness of the outsole and to exclude shoes with gum or tar so that the black magnetic powder would easily hold on.

Dust: Outsoles were dusted twice with black magnetic powder (Beijing Bulant Police Equipment Co., Ltd, 30 mL), using a magnetic powder brush (Hangzhou Silverarrow Co., Ltd, 150 mm long, 27 mm in brush tip diameter) to ensure complete coverage of the magnetic powder; the excess powder was removed by gently tapping the shoe three to four times.

[1]School of Forensic Science, People's Public Security University of China, Beijing, 100038, China.
[2]Graduate School, People's Public Security University of China, Beijing, 100038, China.
[3]Beijing Forensic Science Institute, Beijing, 100038, China.
Corresponding author: Yaping Luo, Ph.D. E-mail: yaping_luo@126.com

Received 15 Mar. 2020; and in revised form 23 May 2020; accepted 24 July 2020.

Adhere: The shoes were fixed on a shoe holder, and the outsoles were carefully covered by a piece of self-adhesive film (AD-Fix, white back, trimmed to the size of 15 × 33 cm) from tiptoe to heel. A gentle and full-scale press was applied to make the film sufficiently contact the black magnetic powder.

Tear off: The film was slowly torn off from heel to tiptoe.

Cover: The detached film was covered by a transparent PVC sheet (Elife, A4, 12.5c).

Following these steps, the samples were scanned (HP LaserJet Professional M1136 MFP, A4, 1200 dpi) and later saved on a computer. Dust impression is shown in Fig. 1.

Selection of Feature Points

Based on Yin et al. (7), four feature points were defined and named, respectively, as tiptoe, lateral metatarsal, heel, and medial metatarsal.

- Tiptoe: most prominent point of the shoe tip;
- Heel: most prominent point of the shoe heel;
- Lateral metatarsal: most prominent point of the lateral contour edge of metatarsal;
- Medial metatarsal: most prominent point of the medial contour edge of metatarsal.

Measurements of Feature Lines

Six lines were established after connecting feature points in pairs. The distance from tiptoe to heel was defined as shoe length. The distance from tiptoe to lateral metatarsal was labeled as L_1; from lateral metatarsal to heel as L_2; from heel to medial metatarsal as L_3; from medial metatarsal to tiptoe as L_4; and from medial metatarsal to lateral metatarsal as L_5 (Fig. 1). FIJI 1.0 software was used for the measurements of the six feature lines on each dust impression (Fig. 2). Theoretically, the six lines on the left shoeprint should be of the same length with the corresponding lines of the right shoeprint. However, there were subtle deviations caused by the selection of feature points. In order to minimize these deviations, mean values of measurements on left and right shoeprints, which were relatively reliable, were employed for analysis.

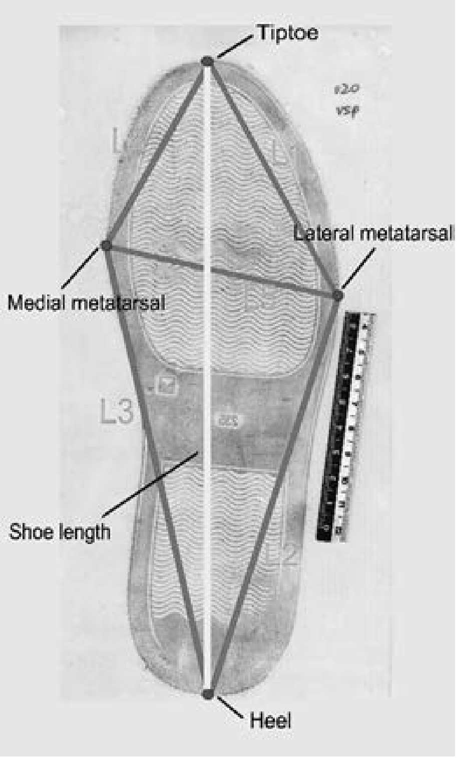


FIG. 1—*Feature points (indicated by red spots), feature lines (indicated by blue lines) and the shoe length (indicated by yellow line). [Color figure can be viewed at wileyonlinelibrary.com]*

Regression Analysis

In this study, three kinds of regression analysis (linear, quadratic, and cubic regression) were used to study the correlation between the shoe length and the other five feature lines. Their equations were established by the Minitab 17 software. The distributions of the five variables (L_1, L_2, ... L_5) and shoe length were analyzed using the IBM SPSS Statistics 23.0 software.

Selection of Optimal Equation

The optimal regression equation was selected by comparing R-sq and R-sq(adj) – also called R^2 and R^2_{adj} – which are the fit index criteria in the equation (Formula 1–2). R-sq represents the multiple correlation coefficient, and R-sq(adj) is the adjusted multiple correlation coefficient. SS_E is the explained sum of squares for error, and SS_T represents the total sum of squares, also called total variation. (Formula 3–4). The items n and p are the total number of variables and the total number of terms in the regression equation, respectively. R-sq and R-sq (adj) can predict how well the regression model fits, and it ranges from 0% to 100%. The bigger the value, the better the model fits. The regression equation would be more accurate if R-sq became higher, and the difference between R-sq and R-sq (adj) lower (8). Therefore, the optimal equation should be determined by comprehensively assessing these two values above.

Verified Test

A verified test was conducted to evaluate the accuracy of the optimal equations. Nine volunteers were asked to walk normally on the floor and on sand with two pairs of their shoes (a pair each of leather shoes and trainers or casual shoes). All the shoes (size 6–12) that volunteers wore had completely different outsole patterns (Fig. 3). Afterward, their shoeprints were captured with a camera (Nikon D7000); all shoeprint samples were clear and complete (Fig. 4). Using these samples, we determined the feature points and measured the length of the five feature lines on the substrates using a ruler. Following measurements and statistics, data analysis was performed. The accuracy of the equations was evaluated by comparing their errors.

$$R-\text{sq} = 1-\frac{\text{SS}_\text{E}}{\text{SS}_\text{T}}$$

Formula 1 – Calculation formula of R-sq

$$R-\text{sq(adj)} = 1-\frac{\text{SS}_\text{E}/(n-p)}{\text{SS}_\text{T}/(n-1)}$$

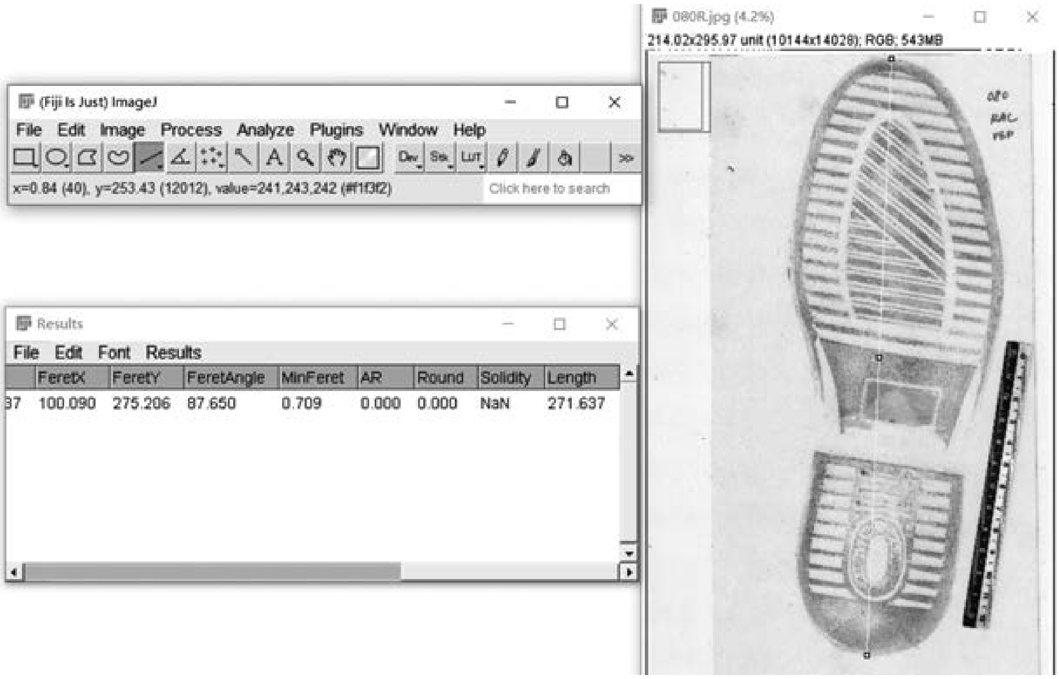
Formula 2 – *Calculation formula of R-sq(adj)*

$$\text{SS}_\text{E} = \sum_{i=1}^{n}(y_i - y_i)^2$$

Formula 3 – *Calculation formula of SSE*

$$\text{SS}_\text{T} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Formula 4 – *Calculation formula of SST*



x=0.84 (40), y=253.43 (12012), value=241,243,242 (#f1f3f2)

214.02x295.97 unit (10144x14028); RGB; 543MB

	FeretX	FeretY	FeretAngle	MinFeret	AR	Round	Solidity	Length
37	100.090	275.206	87.650	0.709	0.000	0.000	NaN	271.637

FIG. 2—*FIJI 1.0 software was used for distance measurements. [Color figure can be viewed at wileyonlinelibrary.com]*

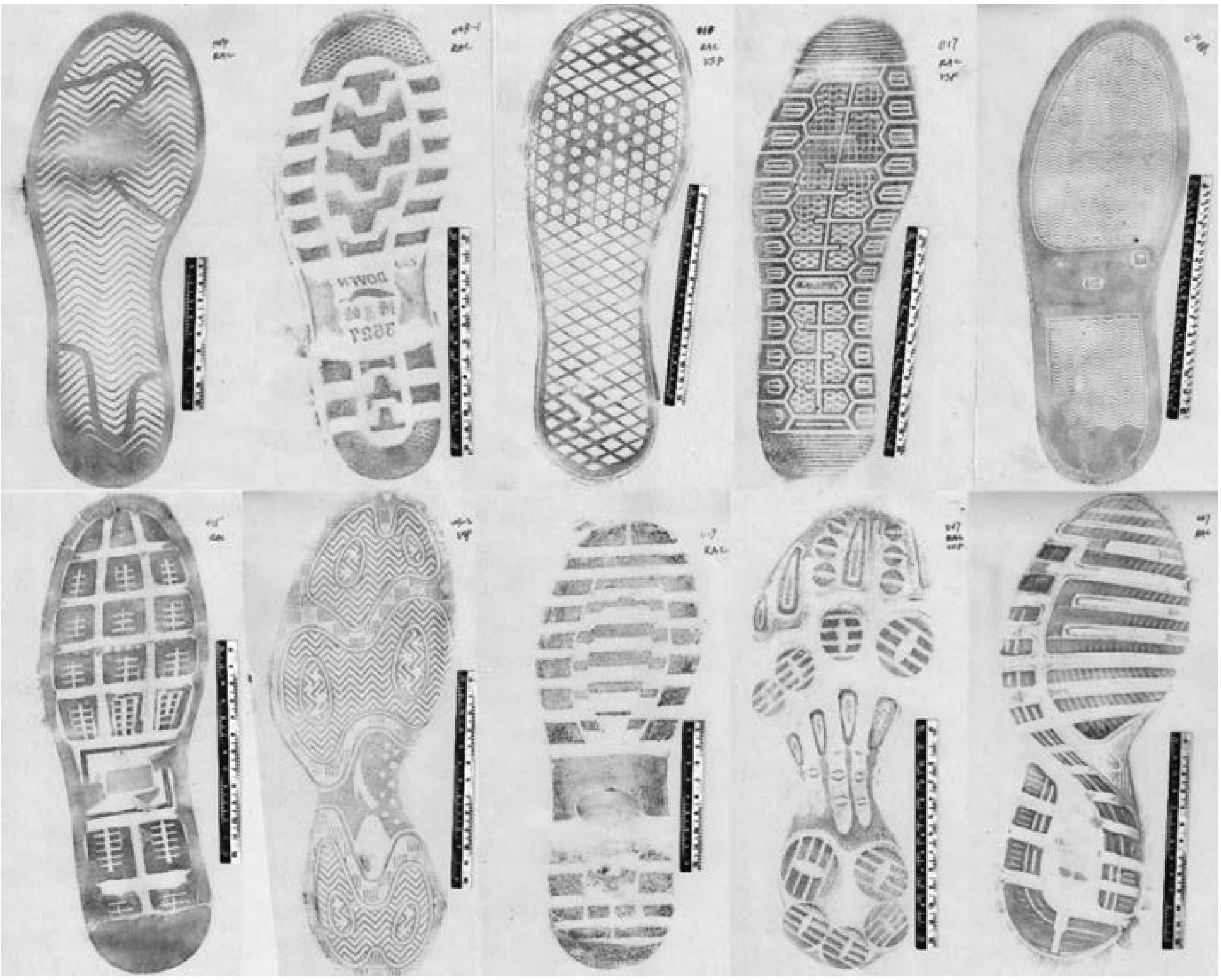


FIG. 3—*Part of different outsole patterns. [Color figure can be viewed at wileyonlinelibrary.com]*

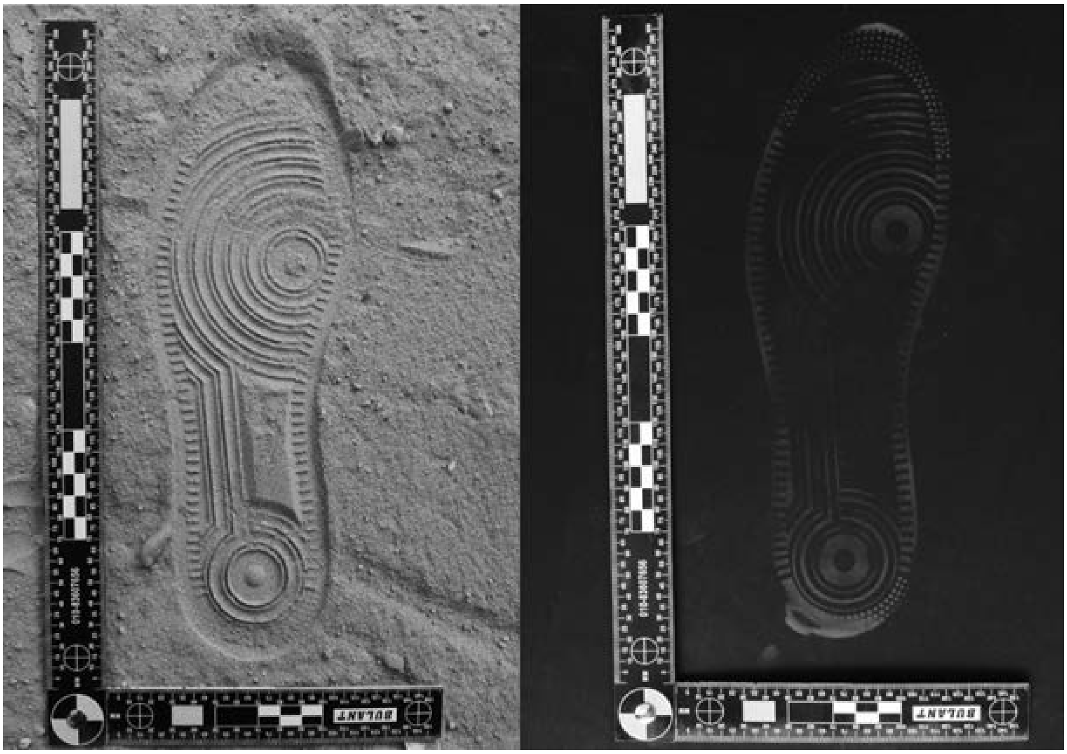


FIG. 4—*Shoeprints of the same shoe on the sand and floor. [Color figure can be viewed at wileyonlinelibrary.com]*

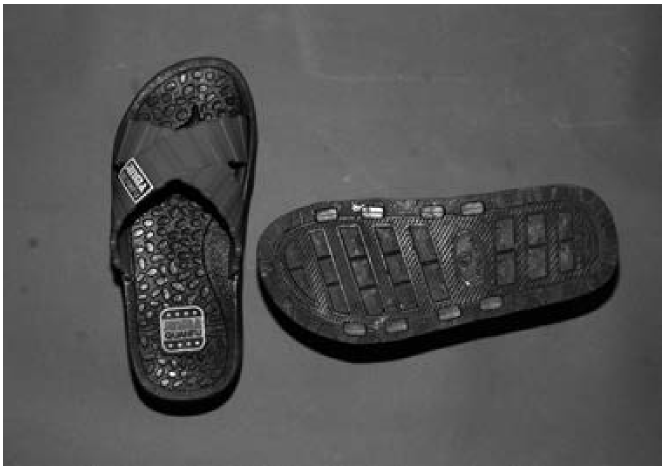


FIG. 5—*A pair of black slippers found at the crime scene. [Color figure can be viewed at wileyonlinelibrary.com]*

Cases

In this part of the study, in order to test their reliability and accuracy with partial shoeprints, the equations were applied in two real cases.

Case 1

In May 2015, two bodies were found in a room in Beijing. Investigators found a pair of black slippers and a partial blood shoeprint on the floor at the crime scene. By identifying footwear impression patterns, footwear examiners concluded that this partial blood shoeprint was left by the right black slipper (Figs 5 and 6).

Case 2

In December 2018, there was a murder case in Beijing. Investigators found a partial blood shoeprint on the south floor at the crime scene. The police found a pair of gray trainers worn by the suspect, after capturing him. The footwear examiner concluded this partial blood shoeprint to had been left by a left gray trainer, after identifying footwear impression patterns (Figs 7 and 8).

Results

Primary Test

Some measurements of the five feature lines and shoe lengths are presented in Table S1, and details of the data we collected are shown in Figs 9 and 10.

We obtained their Pearson's r to study the correlation between feature lines and shoe length (Table 1). The results showed great correlation between them and indirectly proved the reliability of the test. Therefore, it was valid and significant to choose L_1–L_5 as variables in our regression analysis.

Three equations (linear, quadratic, and cubic regression) and their value of *R-sq* and *R*-sq (adj) were obtained (Table 2, Figs S1–S15). The distribution of scatters in the scatterplot was relatively serried and tended to fit a regression equation. Furthermore, the high value of *R*-sq and R-sq (adj) verified the pictures. Thus, the regression analysis was appropriate, valid, and scientific in this study.

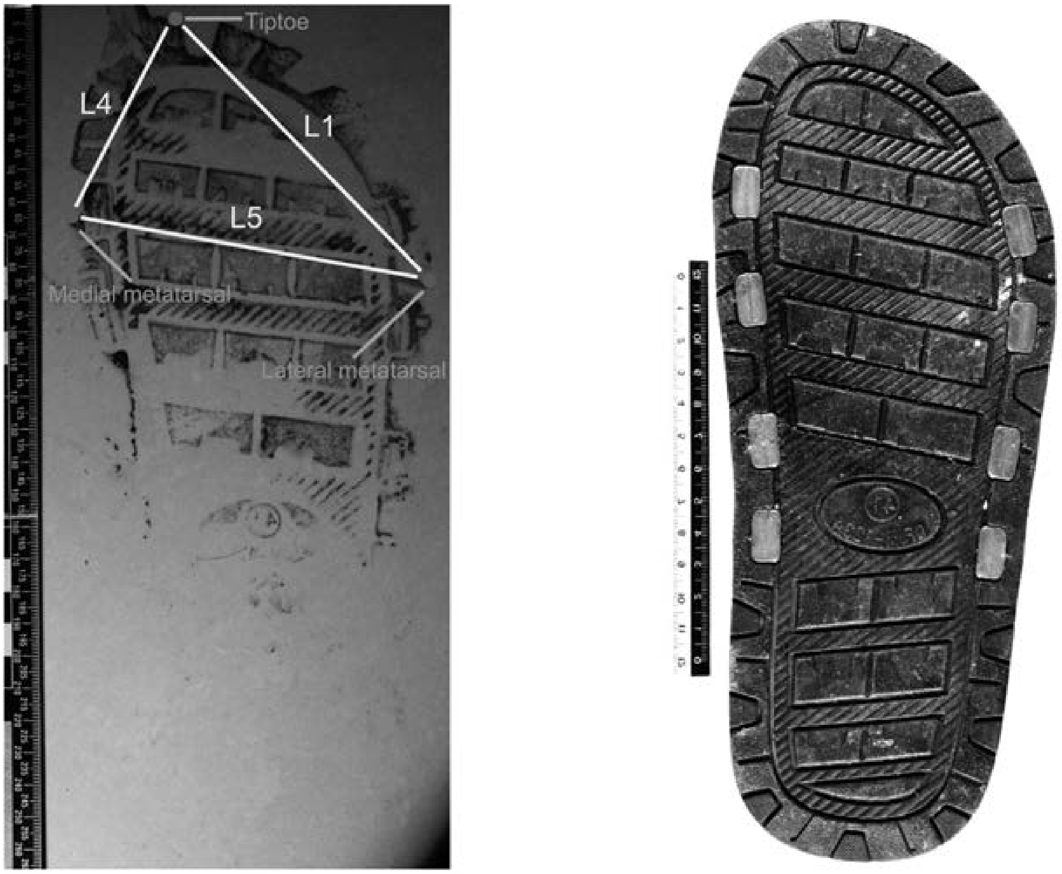


FIG. 6—*Partial blood shoeprint and outsole of left slipper. [Color figure can be viewed at wileyonlinelibrary.com]*

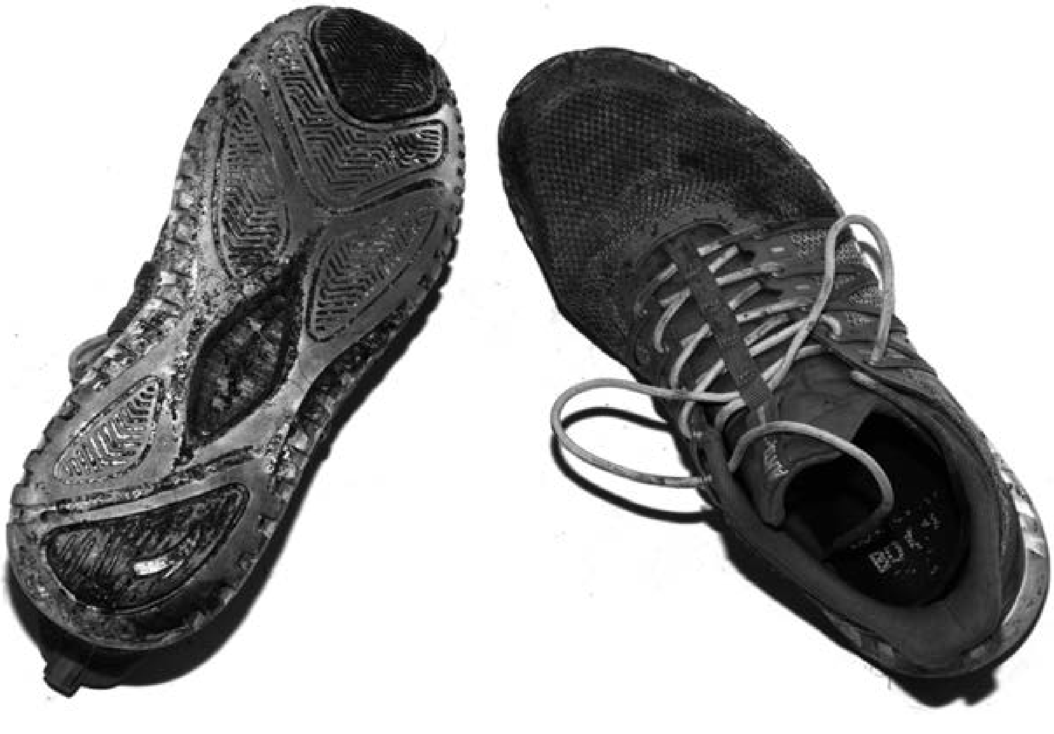


FIG. 7—*A pair of gray trainers worn by the suspect. [Color figure can be viewed at wileyonlinelibrary.com]*

The result showed that the fit indices in the three equations were different for *R*-sq and *R*-sq(adj). Further, all three equations had a certain correlation. The result analysis was objective and comprehensive, choosing the optimal equation through a joint analysis on *R*-sq and *R*-sq(adj) instead of only *R*-sq. Among them, the cubic regression equation was chosen to be the optimal

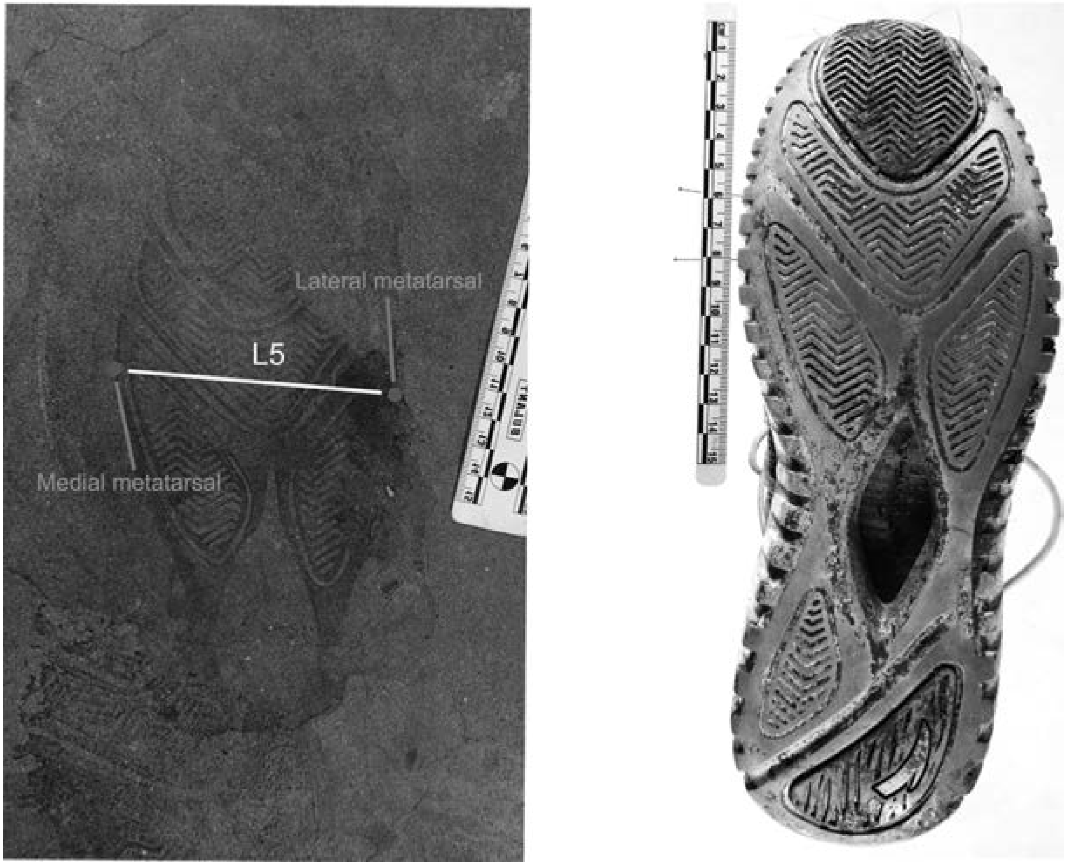


FIG. 8—*Partial blood shoeprint and outsole of left trainer. [Color figure can be viewed at wileyonlinelibrary.com]*

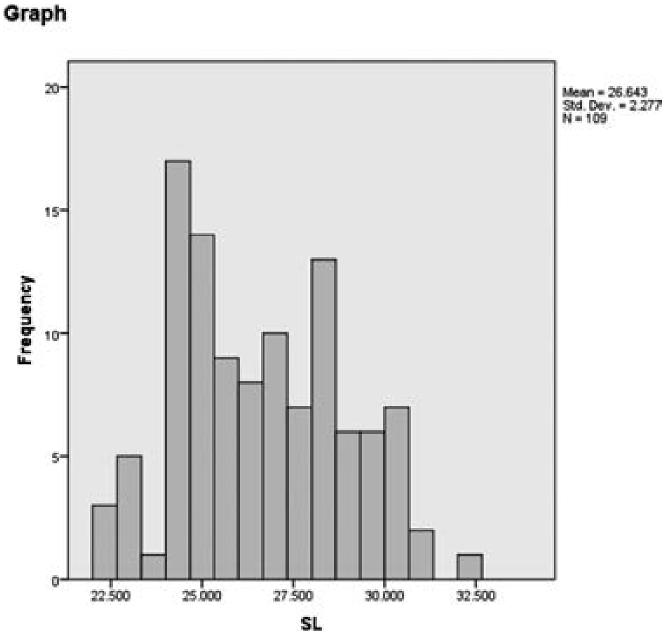


FIG. 9—*Histogram of collected shoe lengths. [Color figure can be viewed at wileyonlinelibrary.com]*

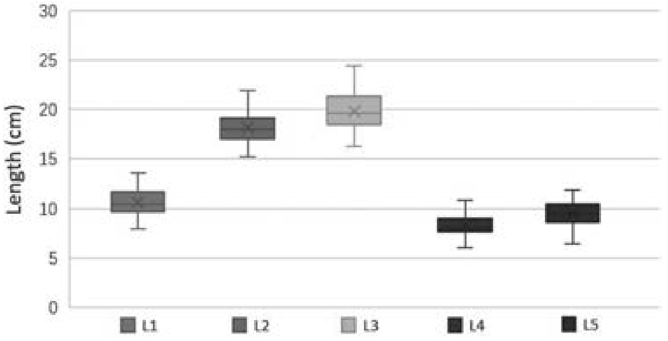


FIG. 10—*Box-plot of feature lines lengths. [Color figure can be viewed at wileyonlinelibrary.com]*

TABLE 1—*Pearson's r of feature lines and shoe length.*

	L_1	L_2	L_3	L_4	L_5
Pearson's *r*	0.908	0.920	0.938	0.863	0.911

equation, as it had a high *R*-sq value, and the difference between *R*-sq and *R*-sq(adj) was little.

Through analysis above, five regression equations were obtained and numbered as results:

$$\mathrm{SL} = 48.10 - 10.39L_1 + 1.218L_1^2 - 0.0404L_1^3 \tag{1}$$

$$\mathrm{SL} = 143.7 - 22.81L_2 + 1.363L_2^2 - 0.02545L_2^3 \tag{2}$$

$$\mathrm{SL} = 60.90 - 7.87L_3 + 0.4651L_3^2 - 0.007844L_3^3 \tag{3}$$

TABLE 2—*Lengths of feature lines and shoes (cm).*

	Regression equation	*R*-sq	*R*-sq(adj)
L_1			
Linear	SL = 9.505 + 1.600 L_1	82.50%	82.30%
Quadratic	$\mathrm{SL} = -1.918 + 3.726L_1 - 0.09756L_1^2$	83.10%	82.80%
Cubic	$\mathrm{SL} = 48.10 - 10.39L_1 + 1.218L_1^2 - 0.0404L_1^3$	83.30%	82.90%
L_2			
Linear	SL = 1.641 + 1.378 L_2	84.70%	84.50%
Quadratic	$\mathrm{SL} = -10.6 + 2.751L_2 - 0.03743L_2^2$	84.90%	84.60%
Cubic	$\mathrm{SL} = 143.7 - 22.81L_2 + 1.363L_2^2 - 0.02545L_2^3$	85.20%	84.80%
L_3			
Linear	SL = 1.412 + 1.271 L_3	88.00%	87.90%
Quadratic	$\mathrm{SL} = -1.604 + 1.573L_3 - 0.00759L_3^2$	88.00%	87.80%
Cubic	$\mathrm{SL} = 60.90 - 7.87L_3 + 0.4651L_3^2 - 0.007844L_3^3$	88.10%	87.80%
L_4			
Linear	SL = 10.34 + 1.967 L_4	74.50%	74.30%
Quadratic	$\mathrm{SL} = 3.732 + 3.558L_4 - 0.09441L_4^2$	74.80%	74.30%
Cubic	$\mathrm{SL} = 91.78 - 28.69L_4 + 3.796L_4^2 - 0.1545L_4^3$	76.00%	75.30%
L_5			
Linear	SL = 9.728 + 1.774 L_5	85.00%	82.90%
Quadratic	$\mathrm{SL} = 17.79 + 0.0359L_5 + 0.09277L_5^2$	83.50%	83.20%
Cubic	$\mathrm{SL} = 83.10 - 21.77L_5 + 2.483L_5^2 - 0.08609L_5^3$	84.60%	84.10%

TABLE 3—*Feature line lengths and equation error (cm).*

		Feature Lines (cm)						Error (cm)				
	Number	L1	L2	L3	L4	L5	SL	E1	E2	E3	E4	E5
Sand	1L	14.6	19.8	22.1	10.7	10.9	30.0	0.3	−1.1	−0.5	0.1	−0.7
	1R	14.5	19.8	22.0	10.8	10.9	30.0	0.4	−1.1	−0.7	0.1	−0.7
	2L	11.8	19.5	22.0	8.9	10.4	28.8	−0.1	−0.3	0.5	−0.6	−0.4
	...	...	...	...	...	...	...	...	...	...	...	...
	17R	8.7	16.5	17.7	7.7	8.3	24.0	−0.7	0.1	−0.2	1.3	0.2
	18L	9.1	17.7	18.2	7.7	9.0	24.8	−0.8	1.0	−0.3	0.6	0.7
	18R	9.2	17.6	18.1	7.6	8.9	24.8	−0.7	0.9	−0.5	0.4	0.6
Floor	1L	14.3	18.8	21.7	10.1	10.7	29.5	0.9	−2.0	−0.6	0.5	−0.6
	1R	14.2	18.9	21.6	10.2	10.8	29.5	0.9	−1.9	−0.7	0.6	−0.3
	2L	11.6	19.3	21.8	8.5	10.2	28.7	−0.3	−0.5	0.4	−1.5	−0.7
	...	...	...	...	...	...	...	...	...	...	...	...
	17R	8.6	15.9	17.5	7.3	8.2	23.8	−0.6	−0.5	−0.3	0.7	0.3
	18L	8.8	16.8	17.6	7.2	8.6	24.4	−0.8	0.2	−0.6	0.0	0.4
	18R	8.8	16.9	17.5	7.4	8.6	24.5	−1.0	0.1	−0.9	0.3	0.3

$$\mathrm{SL} = 91.78 - 28.69L_4 + 3.796L_4^2 - 0.1545L_4^3 \qquad (4)$$

$$\mathrm{SL} = 83.10 - 21.77L_5 + 2.483L_5^2 - 0.08609L_5^3 \qquad (5)$$

Verified Test

We determined the difference between the measured shoe lengths and shoe lengths calculated through Equations 1–5, separately (Table 3).

The equation errors (gap between measured and calculated shoe lengths) are shown in Fig. 11. As it can be seen in Table 4, Equation 3 is the most optimal equation to determine the shoe length, with an average error of 0.591 cm, while Equation 4 is the most unreliable one, with an average error of 0.732 cm. Equations 1, 2, 5 had a similar medium accuracy in estimating shoe length. In addition, Table 4 shows that the errors generated by the shoeprints left on sand were smaller than the ones from the floor prints. The results would be more accurate if the equations were used for trainer prints or casual shoe prints. Tables 1, 2, 4 show that the average error of the equations was inversely proportional to both *R-sq* and Pearson's *r* values.

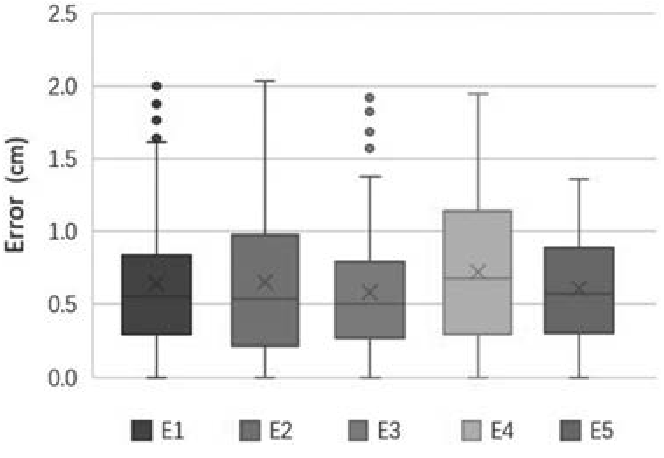


FIG. 11—*Box-plot of equations error. [Color figure can be viewed at wileyonlinelibrary.com]*

TABLE 4—*Average error of the equations.*

	E1	E2	E3	E4	E5
Sand	0.636	0.610	0.581	0.740	0.571
Floor	0.663	0.718	0.601	0.723	0.667
leather shoes	0.802	0.790	0.658	0.958	0.612
Trainers/Casual shoes	0.513	0.551	0.532	0.530	0.626
Total	0.649	0.664	0.591	0.732	0.619

TABLE 5—*Application of equations in the cases.*

	Feature line	Length	Calculated shoe length	Shoe length	Error
	L1	10.377	26.296	26.008	0.288
Case 1	L4	7.733	25.474	26.008	0.534
	L5	11.208	29.803	26.008	3.795
Case 2	L5	9.928	27.462	27.773	0.311

Cases

We measured all feature lines available on the partial shoeprints from two cases, namely three feature lines (L_1, L_4, L_5) for Case 1 and only one (L_5) for Case 2. The shoe lengths were calculated using the corresponding equations (Table 5). The test showed errors within a reasonable confidence interval in three out of four results. This proved the equations had a certain reliability and accuracy in determining the shoe length from partial shoeprints. Simultaneously, however, the test registered one error of 3.795 cm, far beyond the reasonable error range. The peculiar shape of some specific shoes was one of the factors causing this result. Large errors will occur if the equation is applied to some uncommon or special shoe types.

In these two cases, the first thing we did was locating the shoeprint. Once found, we tried to clarify the blood shoeprint using 3, 3',5,5'-tetramethylbenzidine. Finally, eligible feature points were obtained on the processed shoeprints and the estimated shoe length was calculated through the stated equations. To summarize, shoe length can be estimated through the equations we calculated, if at least two from any of the four feature points can be obtained.

Limitations

There are some limitations to our study. Some shoeprints, such as athletic cleats with highly pronounced design elements, had no actual contour of the outsoles. Thus, due to an inevitable error that occurred when applying the above-mentioned method, real lengths of feature lines as well as shoe lengths could hardly be obtained (Fig. 12). In this research, only three kinds of regression equations were used to study the correction of feature lines and shoe lengths, namely linear, quadratic, and cubic regression equations. Multiple regression equations above cubic were never used. In our test, we did not study equations based on shoe types. In addition, the relatively small sample size confined the accuracy of the results to some extent.

Conclusion

This work developed a method to determine shoe length from partial shoeprints, using the following equations:

$$\text{Equation 1: Shoe length} = 48.10 - 10.39L_1 + 1.218L_1^2 - 0.0404L_1^3$$

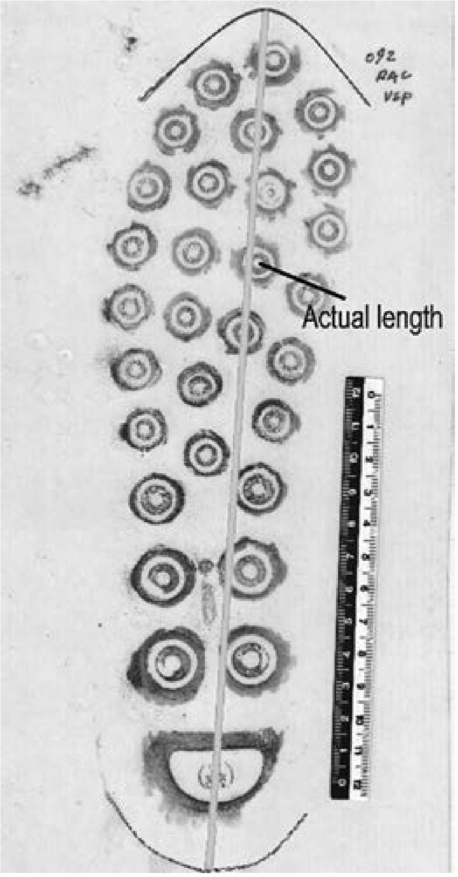


FIG. 12—*An outsole with a special general pattern that can be seen from its impression. [Color figure can be viewed at wileyonlinelibrary.com]*

$$\text{Equation 2: Shoe length} = 143.7 - 22.81L_2 + 1.363L_2^2 - 0.02545L_2^3$$

$$\text{Equation 3: Shoe length} = 60.90 - 7.87L_3 + 0.4651L_3^2 - 0.007844L_3^3$$

$$\text{Equation 4: Shoe length} = 91.78 - 28.69L_4 + 3.796L_4^2 - 0.1545L_4^3$$

$$\text{Equation 5: Shoe length} = 83.10 - 21.77L_5 + 2.483L_5^2 - 0.08609L_5^3$$

To some extent, the above equations have a certain ability to determine the length of a shoe from partial shoeprints. The equations would be more suitable for shoeprints left in the sand (compared with the floor). Furthermore, they would be more accurate if they were applied for trainer prints or casual shoe prints (compared with leather shoes). These equations can be used to obtain reliable physical measurements from partial footwear impressions, which may assist in the investigation or used in a laboratory analysis when suspect shoes are determined.

References

1. Neves FB, Arnold GP, Nasir S, Wang W, MacDonald C, Christie L, et al. Establishing state of motion through two-dimensional foot and shoe print analysis: a pilot study. Forensic Sci Int 2018;284:176–83. https://doi.org/10.1016/j.forsciint.2018.01.008
2. Mcwhorter JW, Wallmann H, Landers M, Altenburger B, LaPorta-Krum L, Altenburger P. The effects of walking, running, and shoe size on foot volumetrics. Phys Ther Sport 2003;4(2):87–92. https://doi.org/10.1016/S1466-853X(03)00031-2

3. Kazuro BABA. Foot measurement for shoe construction with reference to the relationship between foot length, foot breadth and ball girth. J Hum Ergol (Tokyo) 1975;3(2):149–56. https://doi.org/10.11183/jhe1972.3.149
4. Ales J, Saso D. Length dispersion of shoes labelled with the same size in the UK shoe-size system. Footwear Sci 2013;5(Supp 1):S39–S41. https://doi.org/10.1080/19424280.2013.799543
5. Mcinnes AD, Hashmi F, Farndon LJ, Church A, Haley M, Sanger DM, et al. Comparison of shoe-length fit between people with and without diabetic peripheral neuropathy: a case–control study. J Foot Ankle Res 2012;5(1):9. https://doi.org/10.1186/1757-1146-5-9
6. Liu L, Wu J, Luo Y, Lin S. Reproducibility of artificial cut on heel area of rubber outsole. J Forensic Sci 2020;65(1):229–37. https://doi.org/10.1111/1556-4029.14148
7. Yin A, Chen J, Du P, Hu J. 根据穿鞋残缺足迹进行身高分析系数修正的研究 [Study on the correction of height analysis coefficient based on the incomplete shoeprints]. Police Technol 2015;6:29–31. https://doi.org/10.3969/j.issn.1009-9875.2015.06.010
8. Fletcher J. Multiple linear regression. Measurement 2009;338:b167. https://doi.org/10.1136/bmj.b167

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. L_1 Linear regression equation.
Fig. S2. L_1 Quadratic regression equation.
Fig. S3. L_1 Cubic regression equation.
Fig. S4. L_2 Linear regression equation.
Fig. S5. L_2 Quadratic regression equation.
Fig. S6. L_2 Cubic regression equation.
Fig. S7. L_3 Linear regression equation.
Fig. S8. L_3 Quadratic regression equation.
Fig. S9. L_3 Cubic regression equation.
Fig. S10. L_4 Linear regression equation.
Fig. S11. L_4 Quadratic regression equation.
Fig. S12. L_4 Cubic regression equation.
Fig. S13. L_5 Linear regression equation.
Fig. S14. L_5 Quadratic regression equation.
Fig. S15. L_5 Cubic regression equation.
Table S1. Lengths of Feature Lines and shoe length(cm).

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14540
Available online at: onlinelibrary.wiley.com

TECHNICAL NOTE

CRIMINALISTICS

Tiffany Chan,[1] B.A.; Guy Robinson,[2,3] Ph.D.; Jonathan Liu,[4]; Marin Kurti,[5] Ph.D.; Yi He,[1] Ph.D.; and Klaus vonLampe,[6] Ph.D.

Identifying Counterfeit Cigarettes Using Environmental Pollen Analysis: An Improved Procedure*

ABSTRACT: Traditional pollen preparation techniques provide clear residues for pollen identification; however, such methods are time-consuming, requiring repeated centrifugation, heating, and digestion with high-concentration hazardous chemicals. Tobacco leaves can effectively trap environmental pollen due to hairy surface and terpene-rich exudates. A new tobacco sample processing method was developed by using different extraction chemistry with surfactant. Marlboro Gold cigarettes were employed as model samples for method development. Parameters critical for pollen extraction, which include number of cigarette sticks used, extraction solution, and extraction temperature, were optimized. By using 1% dishwashing detergent to treat three cigarettes at room temperature, the improved method was able to recover sufficient pollen for microscopic analysis in three repeated centrifuge-washing steps and omit hazardous chemicals involved in traditional methods. We focused on the pollen of common ragweed (*Ambrosia artemisiifolia*), a plant native to North America, as an indicator to differentiate genuine and counterfeit U.S. brand cigarettes. Results from analyzing randomly purchased genuine (authenticated by forensic examination) and known counterfeit Marlboro Gold provided by law enforcement revealed that a significant amount (39%) of *Ambrosia* were consistently present in all genuine samples, while counterfeit contained none or only trace count. Similar results were found in other counterfeit U.S. brand cigarettes (all seized in the U.S.) involved in this study as well. Lack of *Ambrosia* in cigarette strongly indicates the product was not originated in the United States.

KEYWORDS: forensic palynology, pollen, *Ambrosia*, counterfeit, cigarette

Globally, the illegal cigarette trade represents 11.6% of the cigarette market, resulting in government revenue loss of $40.5 billion USD (1). An understudied aspect of this illicit trade is counterfeit cigarettes that are clandestinely produced without the consent of the trademark holder. Counterfeit cigarettes also pose significant harm to human health because they have elevated levels of toxic heavy metals (2–4). Reliable scientific methods for identification of counterfeit cigarettes, especially with the capability to trace back the geographical origin of production and potential trade route, are highly needed in the criminalistics community for investigation purposes.

Elemental isotope analysis combined with chemometrics data treatment using strategies such as principal component analysis (PCA) is a widely accepted method for verification of the geographical origin of a sample of interest. It has been used for tracing the origin of high-value meat (5) produce (6,7), rice (8), dairy product (9), wine (10), olive oil (11) etc. Stable isotopes were employed for forensic science studies with applications to both natural and manufactured products. For example, the technique was used for the recognition of marijuana trade patterns over a large geographical area in the continental United States (12).

Despite superb sensitivity and effectiveness, elemental analysis, especially stable isotope analysis, requires highly sophisticated and expensive instruments such as inductively coupled plasma—mass spectrometry (13,14), which is not accessible to most laboratories, and the operation and maintenance cost is significant. As an attractive alternative, pollen analysis provides a very affordable choice for revealing a sample's geographic origin.

Forensic palynology is the science of analyzing pollen and spores to solve criminal and civil cases. Pollen is a useful marker due to its size, resistance to biological and chemical degradation, and abundance in the environment. At the microscopic level, pollen grains have distinct features making them identifiable often to the level of plant genus (15). The aerodynamic properties of pollen grains and the vast amounts released by many plants result in an airborne "pollen rain" which settles onto soil, dust, mud, and other surfaces such as tobacco leaves. A

[1] Department of Sciences, John Jay College of Criminal Justice, The City University of New York, 524 W59th Street, New York, NY, 10019.

[2] Department of Natural Sciences, Fordham University, 113 West 60th Street, New York, NY, 10023.

[3] Plant Research Laboratory, New York Botanical Garden, 2900 Southern Boulevard, Bronx, NY, 10458.

[4] Department of Chemistry, Harvey Mudd College, 301 Platt Blvd, Claremont, CA, 91711.

[5] Department of Sociology, Anthropology, Criminology and Social Work, Eastern Connecticut State University, 83 Windham Street, Willimantic, CT, 06226.

[6] Department of Police and Security Management, Berlin School of Economics and Law, Campus Lichtenberg, Alt-Friedrichsfelde 60, Berlin, 10315, Germany.

Corresponding author: Yi He, Ph.D. E-mail: yhe@jjay.cuny.edu

*This research was supported by PSC-CUNY, Grant 62408-00 50.

Received 29 April 2020; and in revised form 11 June 2020, 16 July 2020; accepted 20 July 2020.

sample taken from these surfaces can reflect the particular combination of plants growing in a geographic region (15), as well as the time of year the sample was collected.

Tobacco leaves can effectively trap environmental pollen because of its hairy surface structure (16). Previous studies (16-19) demonstrated that tobacco leaves contain sufficient pollen grains for forensic work and they are relatively easy to recover. In addition, self-contamination seems not a problem due to the standard agronomic practices of topping and suckering to remove tobacco buds before flowering to encourage plant growth.

Several notable studies regarding establishing tobacco origin from pollen analysis have been reported in recent years (17-19). Although there are different views on whether particular pollen taxa can identify the region of tobacco production, there is agreement on the value of forensic pollen in tobacco research.

Donaldson et al. (17) investigated pollen content of two samples representing U.S. and Chinese brands. Their findings indicated that "palynology has the potential to constrain geographical source(s) of tobacco, particularly if regionally localized species can be recognized among the pollen." This article shows the potential value of pollen analysis in identification of counterfeit cigarettes. Williams et al. concluded that pollen signature can distinguish broad geographic areas (19). They analyzed a tobacco sample from Brazil in an effort to identify signature taxa from the state of Minas Gerais. Their work evaluated the role of honey additives to tobacco pollen profile too.

Bryant et al. (18) investigated tobacco pollen in an archaeological project attempting to determine potential sources of the tobacco imported and used in the Netherlands during the lifetime of Rembrandt. They also analyzed modern brands of commercial pipe tobacco. They suggested it is questionable to use pollen profile to identify a specific location of production or shipment of pipe tobaccos because most pipe tobaccos are blends of tobacco grown in different geographical regions during different years.

Collectively, above-mentioned work not only demonstrated the value of application of palynology to match tobacco samples to broad geographical regions and in fighting illicit trade of counterfeit cigarettes, but also highlighted the precautions that should be taken in case study and data interpretation.

Forensic palynology is a valuable yet underutilized technique. Bryant et al. (20) indicated that the reasons for limited attempt to use pollen evidence in either criminal or civil cases included "a lack of available information about the technique, a very limited number of specialists trained to do forensic pollen work, and an almost total absence of academic centers able to train needed specialists for forensic facilities able, or willing, to fund research in this area." (20).

To answer this call, in this study, we developed a simple, fast, and user-friendly method to recover pollen grains retained on the tobacco leaves for forensic identification. The goal of this study is to ease the operation of sample preparation and lower the entry barrier for palynology's forensic application. This work is an interdisciplinary effort integrating the strength and expertise in the fields of analytical chemistry, palynology, and counterfeit tobacco research. This article thus focuses on demonstration of the feasibility of the improved technique and provides a convenient alternative for palynologists who are interested in studying tobacco pollen in general, rather than examining extensive collections of samples.

Standard pollen preparation techniques, typically including acetolysis, provide clear residues in which the cellular content of pollen grains is removed to enhance their external features for precise identification (21). However, such methods are time-consuming, requiring repeated centrifugation, heating, and digestion with high-concentration hazardous chemicals. Acetolysis as indicated by Williams et al. (19) is "a process that can be dangerous if not executed properly." Since we focus on the best recovery of total pollen grains, we propose a simpler and safer protocol that produces residues more akin to airborne samples used for daily pollen counts. Key extraction parameters, including the number of cigarettes to use per sample, washing method, and extraction temperature, were investigated and optimized to recover a more substantial amount of trapped pollen.

We used Marlboro brand cigarettes as model sample in analytical method development, since it is one of the most popular and also the most counterfeited cigarette brand in the world (22). We paid special attention to *Ambrosia* pollen for authentication since the *Ambrosia artemisiifolia* is native to the regions of North America, where genuine tobacco destined for the domestic market is harvested. To validate the feasibility of the method, we applied the procedure to analyzing a collection of genuine and counterfeit cigarettes brands commonly seen in the United States, such as Marlboro Gold, Marlboro Red, and Newport. All counterfeit cigarettes were seized in the United States by the U.S. Bureau of Alcohol, Tobacco, Firearms and Explosives (ATF).

Materials and Methods

Cigarette Samples

Genuine Marlboro Gold (formerly known as Marlboro Light) cigarette packs were purchased at asking price from licensed tobacco retailers in New York City in 2014–2015. The authenticity of these packs was verified through forensic examination of the tax stamp fixed at the bottom of the package and other packaging characteristics (23). The known counterfeit cigarettes were provided by ATF from different seizures in the United States.

Cigarette Sample Selection

Our first experiment centered on optimizing the number of cigarette sticks used for analysis. Twelve (12) cigarettes were randomly selected from each pack and separated into groups of 3, 4, and 5 cigarettes, and then, the samples were processed.

Pollen extraction parameters were optimized by using three cigarette sticks, which were randomly selected from two genuine packs with two cigarettes from the first pack and one cigarette from the second pack.

Optimization of Pollen Extraction Parameters

In the experiments of optimizing extraction solution, three types of extraction solutions were prepared as follows: (i) surfactant solutions with 1% Aquet lab detergent, (ii) 1% dishwashing detergent (*Palmolive*, *Colgate-Palmolive*), and (iii) ethanol water (1:1 ratio) mixture. Three extraction temperatures were investigated for optimum performance: (i) The solution was either chilled in an ice bath (0°C), (ii) heated in a temperature-controlled water bath until equilibrium was established (40°C), or (iii) kept at room temperature (23°C). Extraction using traditional 95% denatured ethanol was performed for comparison.

Preparation of Phenosafranine Stain

Phenosafranine stain (*Sigma Aldrich*) stock solution (2 mg/mL) was prepared in deionized water. 0.01 g of phenosafranine stain was measured and added to 5 mL of deionized water. Twenty-five mL of glycerol gelatin (*Fisher*) was measured in an Erlenmeyer flask and heated in a microwave for approximately 40 sec. 2.5 mL of the phenosafranine stain stock solution was added to the heated glycerol gelatin and mixed thoroughly to result in a shade of red color, which was used to stain pollen grains.

Pollen Extraction

Pollen extraction method was developed using genuine Marlboro Gold cigarettes. Three group sizes (three, four, and five cigarette sticks) were used for analysis. After removing the wrapping paper and filter from the cigarette, the tobacco leaves were weighed on an analytical balance and transferred to a 50-mL centrifuge tube.

Thirty-five mL of extraction solution was added to the centrifuge tube to extract pollen grains. After vortexing (*Vortex-Genie 2*, *Scientific Industries)* 30 sec, each sample was filtered through a 250-micron mesh sieve and the extraction solution was collected into a 250-mL beaker. The centrifuge tube was rinsed with deionized water, which was washed through the residue in the mesh, into the beaker. The combined solution was returned to the 50-mL tube and centrifuged at 1500 rpm for 10 min. After decanting the supernatant, deionized water was added to bring each tube to 35 mL, followed by vortex and centrifugation. The above washing procedure was repeated twice.

After the final decantation of the washing solution, one *Lycopodium* tablet was placed into each tube (24). Each tablet contains 20,848 *Lycopodium* spores (+/-1546; batch, No. 1031, Department of Quaternary Geology, University of Lund, Sweden). *Lycopodium* counts are used to assess the relative richness of pollen within each sample (25). This was followed by addition of 3 mL of deionized water, and 4 mL of 1 M hydrochloric acid to dissolve the *Lycopodium* tablet, which serves as an internal standard for quantifying the number of pollen grains present in cigarettes. When the tablet was completely dissolved, each tube was vortexed and centrifuged under the same condition again. After decanting the supernatant, 7 mL of deionized water was added for a final wash and a solid pollen pellet was obtained.

Pollen Analysis

Approximately 1 mL of glycerol was placed into each sample tube and mixed thoroughly with pollen pellet to form a homogenized solution. A total of 15 μL solution was collected and transferred to a microscope slide by pipetting 5 μL of solution from the top, middle, and bottom section of the centrifuge tube respectively. One drop of heated glycerine jelly mounting medium with phenosafranine stain was added and mixed with the residue on the slide. A glass coverslip was added and the slide placed on a hot plate at a moderate setting for a few seconds, so that the residue mixed with the mounting medium spreads uniformly to the outer edges of the coverslip. After cooling down, the slide was ready for pollen analysis using a compound microscope at 400× magnification (*Olympus*, *CH series*). Both *Lycopodium* spores and pollen grains were counted. In order to ensure 95% confidence of identifying *Ambrosia artemisiifolia* if present in trace amounts (approximately 2%), approximately 150 pollen grains were counted for each sample analyzed.

Results and Discussion

We were first interested in identifying whether there were distinct pollen types found in our samples. Preliminary results found several pollen types, including *Ambrosia*, *Pinus*, Amaranthaceae, Poaceae, Apiaceae, *Plantago*, *Artemisia*, and high-spined Asteraceae, in cigarette samples. Table 1 compares a typical pollen distribution between a genuine Marlboro Gold and a counterfeit Marlboro Gold from China. A significant amount (42.9%) of pollen extracted from genuine samples was *Ambrosia*, while no *Ambrosia* was observed in counterfeit extract. The major pollen species found in the counterfeit samples was Poaceae, accounting for 34%. This result indicates *Ambrosia* is a viable marker for identifying genuine Marlboro Gold produced in North America. Common ragweed (*Ambrosia artemisiifolia*) is native to North America (26) and releases its pollen during late summer to early fall, when tobacco is typically harvested. Large numbers of *Ambrosia* pollen observed in the genuine sample, therefore, are very indicative of the region where tobacco plants were grown. On the contrary, lack of *Ambrosia* strongly suggests the cigarettes were not U.S. originated. The high counts of Poaceae are characteristic for products produced in China. These results are in line with previous findings from Donaldson and Stephens (17) in their pollen analysis of Hongtashan cigarettes produced in China. It is worth noting that our pollen analysis focuses on distinguishing whether *Ambrosia* pollen is present or not in a cigarette sample rather than identifying pollen which is classified as "Other."

Based on the preliminary results, we decided to experiment with different pollen recovery parameters to increase pollen counts and *Ambrosia* recovery. These parameters include the number of cigarettes used for extraction, extraction solution, and extraction temperature.

Effect of Number of Cigarettes Used in Each Sample

Pollen extraction using three, four, and five cigarettes was investigated. We found there are no statistically significant differences in yields of pollen (F = 0.51, df = 2, 21, p = 0.605) and *Ambrosia* (F = 0.55, df = 2,21, p = 0.583) count among three, four, and five cigarettes (Fig. 1). The reason is 50-mL centrifuge tubes (a commonly used size) were employed. The extraction and washing were less efficient when the tube was filled with more tobacco leaves as in using four or five cigarettes. Since there were no statistically significant differences, a

TABLE 1—*Preliminary pollen counts and percentage for a genuine and a counterfeit Marlboro Gold.*

Pollen Type	Genuine (%)	Pollen Concentration[a,b] in Genuine Marlboro Gold	Counterfeit (%)	Pollen Concentration[a,b] in Counterfeit Marlboro Gold
Ambrosia	66 (42.9%)	5639	0	0
Apiaceae	0	0	0	0
Plantago	0	0	5 (5.0%)	563
Poaceae	17 (11.0%)	1453	34 (34.0%)	3832
Pinus	6 (3.9%)	513	3 (3.0%)	338
Artemisia	0	0	4 (4.0%)	451
Amaranthaceae	8 (5.2%)	683	1 (1.0%)	113
Other	55 (35.7%)	4699	53 (53.0%)	5973
Lycopodium	244		185	
Total Pollen Count	154		100	

"Other" consists of pollen unable to be identified due to unclear visualization of its physical characteristics resulting from damage to the pollen and/or concealment from debris, as well as pollen types not commonly seen in American branded cigarettes.

[a,b] Concentration of pollen grains was calculated for 3 cigarettes per sample for each pollen type.

sample size of 3 cigarettes was sufficient for analysis with the procedure developed in this work.

Effect of Extraction Solution on Pollen Recovery

We examined the effects of pollen extraction using the traditional method (17), ethanol, and several user-friendly solutions, including 1% Aquet lab detergent solution, 1% dishwashing detergent solution (*Palmolive*, *Colgate-Palmolive*), and 1:1 ethanol:water mixture. Tobacco plants consist of dense coverage of trichomes, which result in the ability of its fine hairs to trap particles (27). In addition, terpene-rich exudates secreted by trichomes efficiently retain pollen on tobacco leaves (27). Detergent is a surfactant. Using detergent or methanol may change the surface tension so that pollen particles can be easily washed away from the leaf surface.

Figure 2 summarizes counts of pollen and *Ambrosia* grains extracted by using different extraction solutions. There were no statistically significant differences in pollen (F = 4.62, df = 3,4, p = 0.086) and *Ambrosia* recovery (F = 5.20, df = 3,4, p = 0.072). On average, the 1% dishwashing detergent produced an average yield of 285 pollen grains, of which 102 grains were *Ambrosia*. The 1% dishwashing detergent, which is a more readily available and user-friendly solution, was chosen as the best extraction solution for following experiments.

Effect of Extraction Temperature

Effect of temperature on extraction was investigated since temperature may affect the surface tension (28). Extractions were performed at 0, 23, and 40°C. Figure 3 shows the total pollen and *Ambrosia* obtained at different temperatures. Overall, we found no statistically significant differences across extraction temperatures for pollen (F = 0.78, df = 2,3, p = 0.538) and ambrosia recovery (F = 1.39, df = 2,3, p = 0.374). Room temperature (23°C),

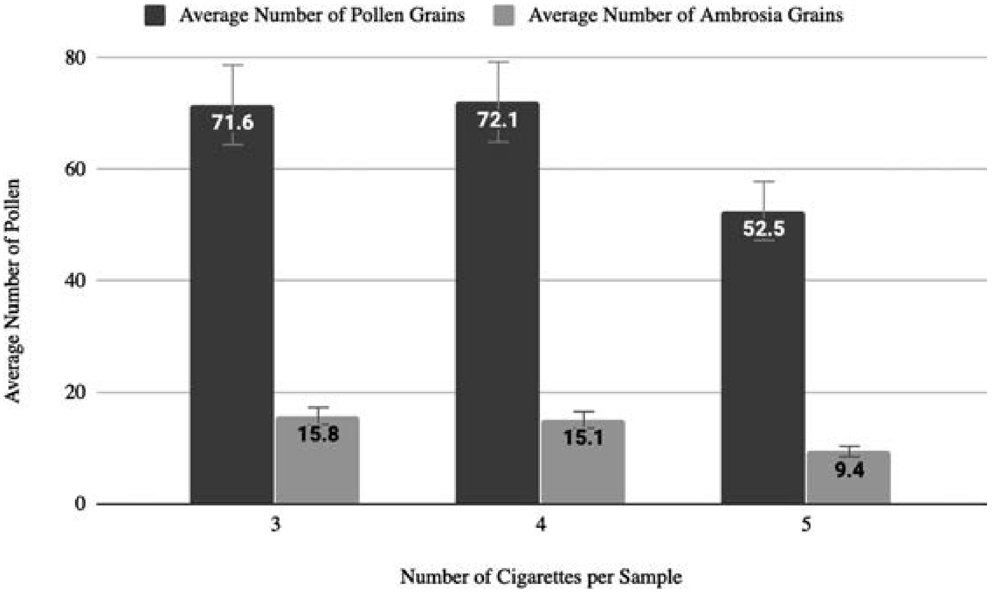


FIG. 1—*Comparison of pollen recovery using different numbers of cigarettes.*

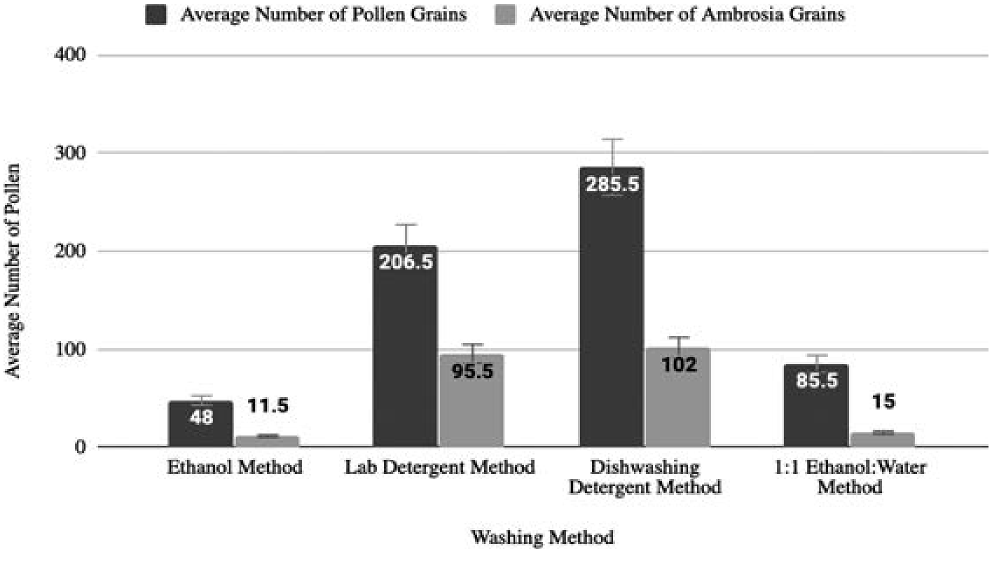


FIG. 2—*Comparison of pollen recovery by using different extraction solutions.*

therefore, was selected as the optimal method since it did not require additional preparation or instrumentation.

Based on experimental results, the optimized extraction parameters were using three cigarette sticks and processing the samples at room temperature (23°C) with 1% dishwashing detergent. Compared with the traditional method employed by different studies (17–19), this work uses less tobacco filler, fewer extraction steps, kitchen dishwashing detergent instead of harmful chemicals, and operates at room temperature. Table 2 compares the sample processing methods used in various studies.

Method Application

This method was applied to analyze ten genuine and ten counterfeit Marlboro Gold packs. Table 3 summarizes the average pollen counts from these samples. We found statistically

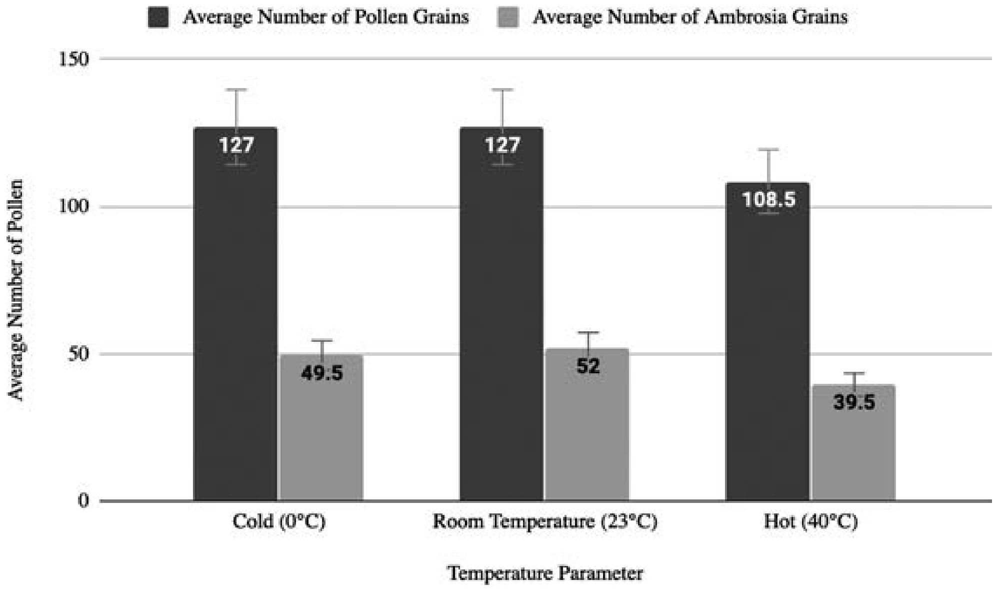


FIG. 3—*Comparison of pollen recovery at different extraction temperatures.*

TABLE 2—*Comparison of sample processing methods used in various studies.*

	Donaldson (17)	Williams (19)	Bryant (18)	This study
Sample size	5 cigarette sticks	15.04 g tobacco	2 g pipe tobacco	3 cigarette sticks
Acetolysis	Yes	Yes	Yes	No
Reagent	No detailed chemical regent information was provided, but the method is referred as standard pollen preparation involving NaOH, sieving, and acetolysis.	KOH, glacial acetic acid; acetic anhydride: sulfuric acid (8:1) Zinc bromide solution; ethanol	KOH, HCl, glacial acetic acid; acetic anhydride: sulfuric acid (9:1) ethanol (ETOH)	Dishwashing detergent; 1 M HCl for dissolving *Lycopodium* tablet
Processing temperature	N.A.	Hot water bath (99°C) for 20 min for acetolysis; Hot plate (35°C) evaporation of ethanol	Heating block at 80°C for 10 min for acetolysis; Heating block for evaporation of ethanol	Room temperature

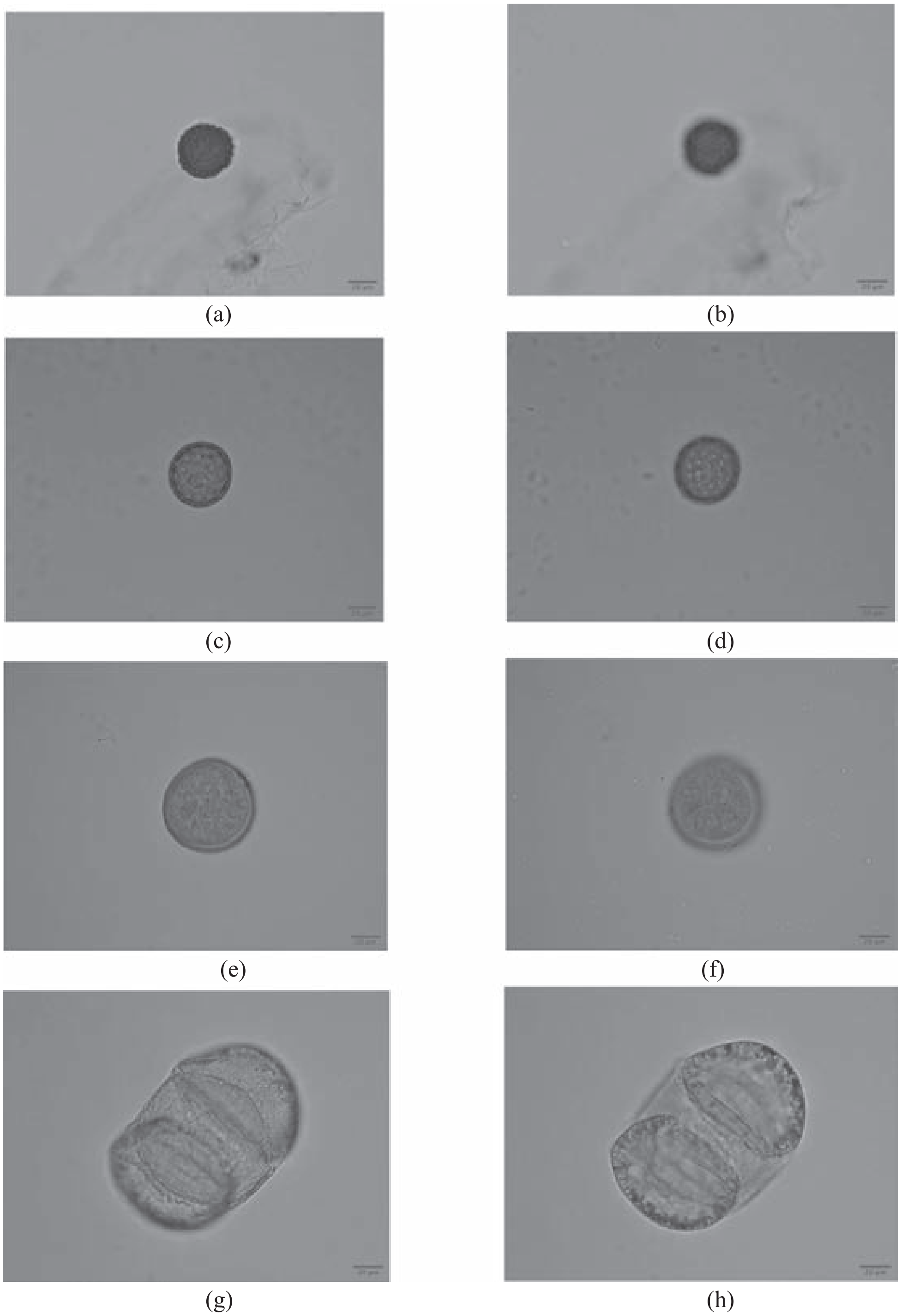
TABLE 3—*Average pollen counts and percentage for genuine and counterfeit Marlboro Gold.*

	Genuine Marlboro Gold (n = 10)		Counterfeit Marlboro Gold (n = 10)		
Pollen Type	Pollen Counts (%)	Pollen Concentration[a,b]	Pollen Counts (%)	Pollen Concentration[a,b]	Test of Equality
Ambrosia	56.3 (39%)	4636	0.6 (1.1%)	93	p<.001
Apiaceae	0.4 (0.3%)	31	0.4 (0.7%)	62	p = 1.00
Plantago	1.3 (0.9%)	101	1.5 (2.7%)	233	p = 0.79
Poaceae	18.3 (12.7%)	1418	13.2 (23.6%)	2054	p = 0.22
Pinus	7.4 (5.1%)	574	1.2 (2.1%)	187	p < 0.01
Artemisia	1.1 (0.7%)	85	1.7 (3.1%)	264	p = 0.407
Amaranthaceae	9.6 (6.6%)	744	1.9 (3.4%)	295	p<.001
High-Spined Aster	0.3 (0.2%)	23	1.1 (1.9%)	171	p = 0.09
Other	49.1 (34.1%)	3805	34.3 (61.3%)	5336	p = 0.06
Lycopodium	269		134		
Total Pollen Count	144		56		

"Other" consists of pollen unable to be identified due to unclear visualization of its physical characteristics resulting from damage to the pollen and/or concealment from debris, as well as pollen types not commonly seen in American branded cigarettes.

[a,b]Concentration of pollen grains was calculated for 3 cigarettes per sample for each pollen type.

significant differences across several types of pollen between genuine and counterfeit. For example, we found a significantly higher count of *Ambrosia* in genuine cigarettes (p<.001). On average, 39% of the pollen in genuine Marlboro cigarettes was *Ambrosia*, when compared to only 1.1% in counterfeit cigarettes. Similarly, significantly higher counts of *Pinus* and Amaranthaceae were found in genuine cigarettes. Figure 4 shows the pictures of *Ambrosia*, Amaranthaceae, Poaceae, and *Pinus*. Figure 5 is the picture from a counterfeit cigarette sample. Total pollen extracted from genuine samples (144 grains) is much higher than that from counterfeit (56 grains). No *Ambrosia* was found in seven out of 10 counterfeit samples, and between one and four *Ambrosia* grains were found in the rest of the three counterfeit samples. Since the counterfeits were seized in the



(a) (b)

(c) (d)

(e) (f)

(g) (h)

FIG. 4—*Pictures of Ambrosia artemisiifolia, Amaranthaceae, Poaceae, and Pinus captured on Nikon Eclipse E400. The diameter of pollen grains shown is approximately 20 micrometers. The figure depicts: (a) Ambrosia artemisiifolia in peripheral view, (b) Ambrosia artemisiifolia in near view, (c) Amaranthaceae in peripheral view, (d) Amaranthaceae in near view, (e) Poaceae in peripheral view, (f) Poaceae in near view, (g) Pinus in peripheral view, and (h) Pinus in near view. [Color figure can be viewed at wileyonlinelibrary.com]*

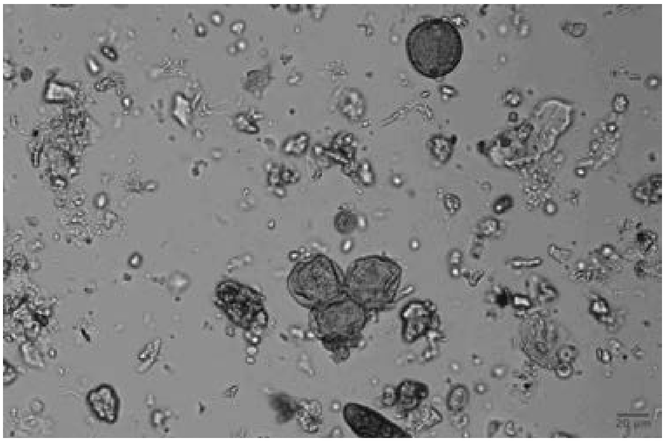


FIG. 5—*Counterfeit cigarette sample: Picture of a cluster of three crumpled Poaceae pollen grains along with a pollen grain identified as "Other" captured on Nikon Eclipse E400. Presence of Curvularia and Myxomycete mold spores, along with debris resulting from extraction procedures were apparent in the sample. [Color figure can be viewed at wileyonlinelibrary.com]*

United States, the negligible amount of *Ambrosia* may come from the environment during sample transportation and storage. Thus, it is important to note that the trace amounts of *Ambrosia* identified in counterfeit cigarette samples do not suggest genuine authenticity of the tobacco product in question.

We applied the method to analyzing approximately 45 additional counterfeit cigarettes seized in the United States in different regions of the country by various law enforcement agencies, which included other brand names such as Marlboro Light, Marlboro Red, and Newport. Consistent results were obtained across all samples examined. No *Ambrosia* was observed in the majority of samples, and only trace was found in the rest. Because abundant information was associated with the counterfeit analysis, detailed study of the pollen profile in the counterfeits will be summarized in a separate report.

In addition to the presence of pollen, foreign materials such as charcoal, clay, silt, insect appendages, mold spores, and *Alternaria* (spores of a plant pathogen) were also found in the extract after the sample was chemically processed; however, it did not interfere with identification of pollen species with research interests. Although acetolysis could be used to efficiently dissolve debris and organic material present in the samples so as to more easily stain and visualize pollen grains (29), we chose not to include this procedure since the method we developed uses less time and yields sufficient resolution. Furthermore, foreign material may provide additional environmental information regarding manufacture (30). For example, in the case of forensic palynology analysis of falsified antimalarial drugs, information such as whether the source of manufacture was close to road or rice or maize fields was obtained through analyzing foreign particles like charcoal and insect exoskeletons (30).

Conclusion

A simple, fast, and more user-friendly tobacco sample processing method was developed in this work. Through changing extraction chemistry by using surfactant instead of strong base such as KOH or NaOH, the improved procedure yields pollen samples sufficiently refined to detect a distinct plant type abundant in the tobacco growing regions of eastern North America, yet rare or absent elsewhere tobacco is grown commercially. As Williams et al. (19) point out to discover where cigarette tobacco is grown contributes to an understanding of the illicit trade networks, even if the information is not used in a court of law. In this case, authentic samples contained a significant amount of ragweed (*Ambrosia*) pollen. Counterfeit U.S. brand cigarette samples (seized in the U.S. by law enforcement) examined in this study either lacked ragweed or only had trace amounts. Lack of *Ambrosia* therefore serves as a strong indicator that the cigarette is not originated from the United States. The future application of forensic palynology to tobacco analysis certainly calls for more information collected by researchers among different continents and efforts building a sufficient tobacco pollen profile database.

Acknowledgments

We thank the U. S. Bureau of Alcohol, Tobacco, Firearms and Explosives (ATF) for providing counterfeit cigarette samples.

References

1. Joossens L, Merriman D, Ross H, Raw M. How eliminating the global illicit cigarette trade would increase tax revenue and save lives. Paris, France: International Union Against Tuberculosis and Lung Disease, 2009.
2. Stephens WE, Calder A, Newton J. Source and health implications of high toxic metal concentrations in illicit tobacco products. Environ Sci Technol 2005;39(2):479–88. https://doi.org/10.1021/es049038s.
3. Pappas RS, Polzin GM, Watson CH, Ashley DL. Cadmium, lead, and thallium in smoke particulate from counterfeit cigarettes compared to authentic US brands. Food Chem Toxicol 2007;45:202–9. https://doi.org/10.1016/j.fct.2006.08.001.
4. He Y, von Lampe K, Wood L, Kurti M. Investigation of lead and cadmium in counterfeit cigarettes seized in the United States. Food Chem Toxicol 2015;81:40–5. https://doi.org/10.1016/j.fct.2015.04.006.
5. Nie J, Shao SZ, Xia W, Liu Z, Yu CC, Li R, et al. Stable isotopes verify geographical origin of yak meat from Qinghai-Tibet plateau. Meat Sci 2020;165:108113. https://doi.org/10.1016/j.meatsci.2020.108113.
6. Choi SH, Bong YS, Park JH, Lee KS. Geographical origin identification of garlic cultivated in Korea using isotopic and multi-elemental analyses. Food Control 2020;111:107064. https://doi.org/10.1016/j.foodcont.2019.107064.
7. Aguzzoni A, Bassi M, Pignotti E, Robatscher P, Scandellari F, Tirler W, et al. Sr isotope composition of Golden Delicious apples in Northern Italy reflects the soil Sr-87/Sr-86 ratio of the cultivation area. J Sci Food Agric 2020;100(9):3666–74. https://doi.org/10.1002/jsfa.10399.
8. Wang JS, Chen TJ, Zhang WX, Zhao Y, Yang SM, Chen AL. Tracing the geographical origin of rice by stable isotopic analyses combined with chemometrics. Food Chem 2020;313:126093. https://doi.org/10.1016/j.foodchem.2019.126093.
9. Zhao SS, Zhao Y, Rogers KM, Chen G, Chen AL, Yang SM. Application of multi-element (C, N, H, O) stable isotope ratio analysis for the traceability of milk samples from China. Food Chem 2020;310:125826. https://doi.org/10.1016/j.foodchem.2019.125826.
10. Epova EN, Berail S, Seby F, Barre JPG, Vacchina V, Medina B, et al. Potential of lead elemental and isotopic signatures for authenticity and geographical origin of Bordeaux wines. Food Chem 2020;303:125277. https://doi.org/10.1016/j.foodchem.2019.125277.
11. Bontempo L, Paolini M, Franceschi P, Ziller L, Garcia-Gonzalez DL, Camin F. Characterisation and attempted differentiation of European and extra-European olive oils using stable isotope ratio analysis. Food Chem 2019;276:782–9. https://doi.org/10.1016/j.foodchem.2018.10.077.
12. Cerling TE, Barnette JE, Bowen GJ, Chesson LA, Ehleringer JR, Remien CH, et al. Forensic stable isotope biogeochemistry. Annu Rev Earth Planet Sci 2016;44(1):175–206. https://doi.org/10.1146/annurev-earth-060115-012303.
13. Techer I, Medini S, Janin M, Arregui M. Impact of agricultural practice on the Sr Isotopic composition of food products: application to discriminate the geographic origin of olives and olive oil. Appl Geochem 2017;82:1–14. https://doi.org/10.1016/j.apgeochem.2017.05.010.
14. Mihaljevic M, Ettler V, Sebek O, Strnad L, Chrastny V. Lead isotopic signatures of wine and vineyard soils – tracers of lead origin. J

Geochem Explor 2006;88(1–3):130–3. https://doi.org/10.1016/j.scitotenv.2015.07.133.

15. Bryant VM. Pollen and spore evidence in forensics. In:Jamieson A, Moenssens A, editor. Wiley encyclopedia of forensic science. Hoboken, NJ: Wiley-Blackwell, 2014;1–3. https://doi.org/10.1002/9780470061589.fsa085.pub2.
16. Poethig RS, Sussex IM. The developmental morphology and growth dynamics of the tobacco leaf. Planta 1985;165(2):158–69.
17. Donaldson MP, Stephens WE. Environmental pollen trapped by tobacco leaf as indicators of the provenance of counterfeit cigarette products: a preliminary investigation and test of concept. J Forensic Sci 2010;53(3):738–41. https://doi.org/10.1111/j.1556-4029.2010.01319.x.
18. Bryant VM, Kampbell SM, Hall JL. Tobacco pollen: archaeological and forensic applications. Palynology 2012;36(2):208–23. https://doi.org/10.1080/01916122.2011.638099.
19. Williams S, Hubbard S, Reinhard KJ, Chaves SM. Establishing tobacco origin from pollen identification: an approach to resolving the debate. J Forensic Sci 2014;59(6):1642–9. https://doi.org/10.1111/1556-4029.12569.
20. Bryant VM, Jones GD. Forensic palynology: current status of a rarely used technique in the United States of America. Forensic Sci Int 2006;163(3):183–97. https://doi.org/10.1016/j.forsciint.2005.11.021.
21. Hesse M, Waha M. A new look at the acetolysis method. Plant Syst Evol 1989;163(3/4):147–52. https://doi.org/10.1007/bf00936510.
22. World Customs Organization. Customs and tobacco report. Brussels, Belgium: World Customs Organization, 2009.
23. Kurti M, He Y, Silver D, Giorgio M, von Lampe K, Macinko J, et al. Presence of counterfeit Marlboro gold packs in licensed retail stores in New York City: evidence from test purchases. Nicotine Tob Res 2018;21(8):1131–4. https://doi.org/10.1093/ntr/nty096.
24. Stockmarr J. Tablets with spores used in absolute pollen analysis. Pollen Spores 1971;13:615–21.
25. Slater SM, Wellman CH. A quantitative comparison of dispersed spore/pollen and plant megafossil assemblages from a Middle Jurassic plant bed from Yorkshire, UK. Paleobiology 2015;41(4):640–60. https://doi.org/10.1017/pab.2015.27.
26. Gentili R, Asero R, Caronni S, Guarino M, Montagnani C, Mistrello G, et al. Ambrosia artemisiifolia L. temperature-responsive traits influencing the prevalence and severity of pollinosis: a study in controlled conditions. BMC Plant Biol 2019;19(1):155. https://doi.org/10.1186/s12870-019-1762-6.
27. Wagner GJ. Secreting glandular trichomes: more than just hairs. Plant Physiol 1991;96(3):675–9. https://doi.org/10.1104/pp.96.3.675.
28. Palmer SJ. The effect of temperature on surface tension. Phys Educ 1976;11(2):119–20. https://doi.org/10.1088/0031-9120/11/2/009.
29. Jones GD. Pollen analyses for pollination research, acetolysis. J Pollinat Ecol 2014;13:203–17. https://doi.org/10.26786/1920-7603%282014%2919.
30. Mildenhall DC. The role of forensic palynology in sourcing the origin of falsified antimalarial pharmaceuticals. Palynology 2017;41:203–6. https://doi.org/10.1080/01916122.2016.1156587.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14493
Available online at: onlinelibrary.wiley.com

TECHNICAL NOTE

PATHOLOGY/BIOLOGY

Lorenzo Gitto,[1] M.D.; Laura Donato,[2] M.S.; Alessandro Di Luca,[3] M.D.; Stephanie M. Bryant,[1] M.D.; and Serenella Serinelli,[1] M.D., Ph.D.

The Application of Photogrammetry in the Autopsy Room: A Basic, Practical Workflow*

ABSTRACT: Photogrammetry is a technique that uses two-dimensional photographs taken from different angles and positions to determine three-dimensional coordinates and distances. Becoming familiar with the photography technique for photogrammetry purposes is the first step to obtaining high-quality results. Ten human cadavers were studied to develop this protocol. Appropriate equipment settings, measurements, and suitable ambient conditions were determined. Finally, the protocol was tested on one cadaver wherein a full postmortem examination was conducted, allowing accurate 3D modeling and measurements of the human body. This straightforward, step-by-step workflow will help users become familiar with this technique. A thorough description of the necessary steps is reported, including equipment, environment requirements, body placement, how to take photographs, and the minimum suggested number of photographs. Numerous graphics show the protocol's main steps to help users understand and reproduce the entire process. Photogrammetry allows the permanent recording and storage of photographic evidence of conditions that existed at the time of autopsy and accurate measurements of the body. The 3D model can have a powerful effect in court, where the findings can be accurately depicted without elicitation of strong emotion that may influence the judge or jurors. The primary disadvantage of photogrammetry for forensic pathology is its time-consuming nature. However, the widespread use of the photogrammetry technique in postmortem rooms would allow in-depth testing to detect and fix potential pitfalls, making this technique more reproducible and verifiable, increasing its admissibility in courts.

KEYWORDS: forensic pathology, autopsy, photogrammetry, postmortem imaging, forensic photography, 3D-modeling

Photogrammetry is a technique that uses two-dimensional photographs taken from different angles and positions to determine three-dimensional coordinates and distances. The process involves overlapping photographs of an object or a landscape to create a highly accurate and realistic digital 3D model. In this technique, the camera moves in three-dimensional space to determine 3D coordinates (x, y, and z) of a subject and a specific software detects the common areas between photographs producing a 3D model with measurement estimation ability.

The interest in using photogrammetry as a postmortem imaging tool has increased in recent years (1). Forensic photography, an essential part of forensic autopsy, allows the recording of the initial appearance of a body and of the autopsy findings as a means of creating a permanent record for legal purposes. However, even high-quality photographs may not allow for a detailed review of an injury or an organ. The primary limitation of photographs lies in the two-dimensional representation of a 3D human body. Moreover, if a specific finding or measurement is missed during postmortem examination, a review of data may no longer be possible once the autopsy is over. The 3D documentation of a subject addresses these issues allowing for digital analysis and visual rendering of data. To increase the application of photogrammetry in morgues for the purposes of forensic pathology, we propose a simple and reproducible step-by-step protocol for photogrammetry.

Materials and Methods

Subjects

Preliminary tests were performed on ten human cadavers donated for anatomical purposes to the Anatomical Gift Program of the State University of New York, Upstate Medical University. The test results were used to determine the appropriate equipment settings, to estimate measurements, and to set suitable ambient conditions. Once the final protocol was developed, it was tested on one cadaver wherein a full postmortem examination was conducted (external and internal examinations). The subject was a white female aged 76 years, 153 cm in length, and 37 kg in weight.

Equipment Used for the Tests

All tests were performed using the following equipment:

- Cameras. Nikon® (Tokyo, Japan) D5600 DSLR camera 24.2 MP with AF-P DX NIKKOR 18–55 mm f/3.5–5.6 G Lens; Apple (Cupertino, CA, USA) iPhone® XS plus dual camera 12 MP with built-in optical image stabilization (one

[1]Department of Pathology, State University of New York, Upstate Medical University, 750 E Adams St, Syracuse, NY 13202.

[2]Forensic Anthropologist, Independent Researcher, Via tripolitana, 195, Rome, 00199, Italy.

[3]Section of Legal Medicine, Department of Public Health, Catholic University “Sacro Cuore” of Rome, Largo Francesco Vito, 1, Rome, 00168, Italy.

Corresponding author: Lorenzo Gitto, M.D. E-mail: gittol@upstate.edu

*Presented at the 72nd Annual Scientific Meeting of the American Academy of Forensic Sciences, February 17–22, 2020, in Anaheim, CA.

Received 14 Mar. 2020; and in revised form 27 May 2020; accepted 1 June 2020.

wide-angle lens with f/1.8 aperture and one 2 × telephoto lens with f/2.4 aperture).

- Computer. Laptop: Intel® (Santa Clara, CA, USA) Core™ i5-7200 CPU @ 2.50 GHz 2.71 GHz, 12 GB RAM, OS Microsoft® Windows 10 – ×64.
- Software. Photogrammetry: Zephyr Lite (3D Flow®, Udine, Italy); digital sculpting tool: ZBrush (Pixologic®, Los Angeles, CA, USA); and digital clinometer: Rotating Sphere Clinometer (© Nils Calandar 2016, available on the Apple AppStore).
- Others: personal protective equipment, standard autopsy tools (blades, scissors, forceps, and saw), and ruler.

Photogrammetry Protocol Workflow

Equipment

There are no strict requirements in terms of the equipment to be used. Standard tools can be used, as follows: an inexpensive or expensive camera (smartphone camera may also be used), a computer (desktop or laptop), and conventional autopsy tools. However, the definition of the final 3D rendering will depend mostly on the quality of camera used. In the same way, a higher performing computer allows the software to more quickly complete photogrammetric elaboration. A mechanical or digital clinometer or a protractor can be used to set the inclination of the camera (see below for further information).

Environment

- An adequate source of light is essential. The light should be able to illuminate the entire surface of the body (Fig. 1A). When necessary, more light sources should be added to obtain the best results (Fig. 1B). The use of camera flashes should be avoided because of irregular light patterns and shadows that can be produced.
- The room background should be as clean as possible to allow the software to recognize and delete digital noise from the final 3D model. When necessary, monochromatic bed screens or hospital sheets should be used when taking photographs.

Placement of the Body

- The body must be placed on a monochromatic surface. A standard, clean autopsy table is acceptable, even those with stainless steel grid plates. Stained, old, or rusted tables should not be used or must be covered with clean white hospital sheets.
- The body must be placed first in a supine position and then in prone position (Fig. 2):
  - *Anterior* aspect in the supine position.
    - The head must be facing upward.
    - The arms must be placed parallel to the body with the palms of the hands facing downward and in contact with the table surface. Since a certain degree of rigor mortis in the upper extremities is often observed, which keeps the hands in the prone position, it can be difficult to place the body in the anatomical position with the dorsal aspects of the hands in contact with the table. Thus, placing the palms of the hands downward will reduce the need for body manipulation and potential distortions in the final 3D model.
    - The legs should be positioned with the heels close together and the toes extended outwards.
  - *Posterior* aspect in the prone position.
    - The head must be facing downward.
    - The arms must be placed along the body with the palms of the hands facing upward.

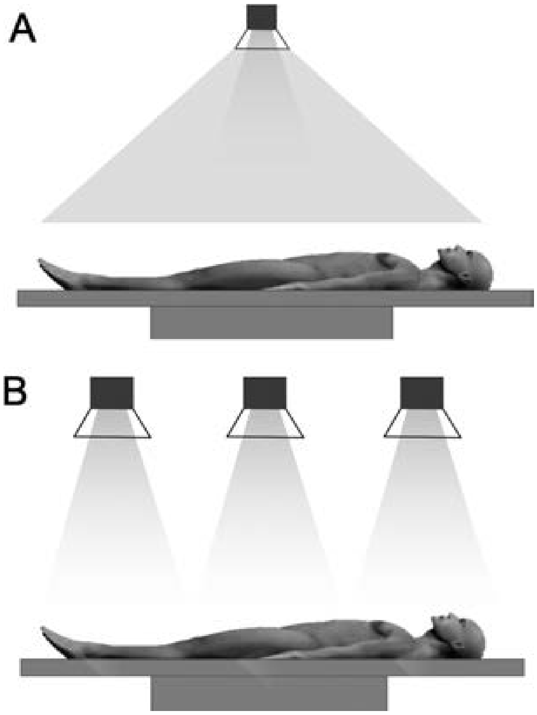


FIG. 1—*(A) Single source of light: the light should be able to illuminate the whole surface of the body. (B) Multiple sources of light: if a single source of light is not enough, more lights can be used. [Color figure can be viewed at wileyonlinelibrary.com]*

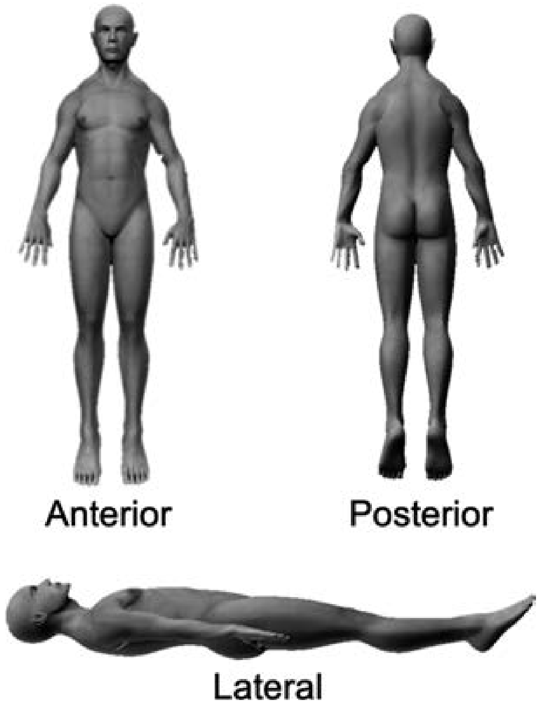


FIG. 2—*Placement of the body. The body must be placed supine and subsequently prone. The figure shows the anterior, posterior, and lateral aspects of the body. [Color figure can be viewed at wileyonlinelibrary.com]*

- The legs should be positioned with the heels close together and the toes extending outwards.

How to Take Photographs for Photogrammetry

The steps below will ensure uniform intervals and angles between photographs when moving the camera around the body allowing the software to effectively recognize the overlapping areas.

- Basic photography skills are required. Photographs can be taken either by the forensic photographer, the forensic pathologist, the autopsy technician, or other assistants.
- The length of the body should be divided into sections.

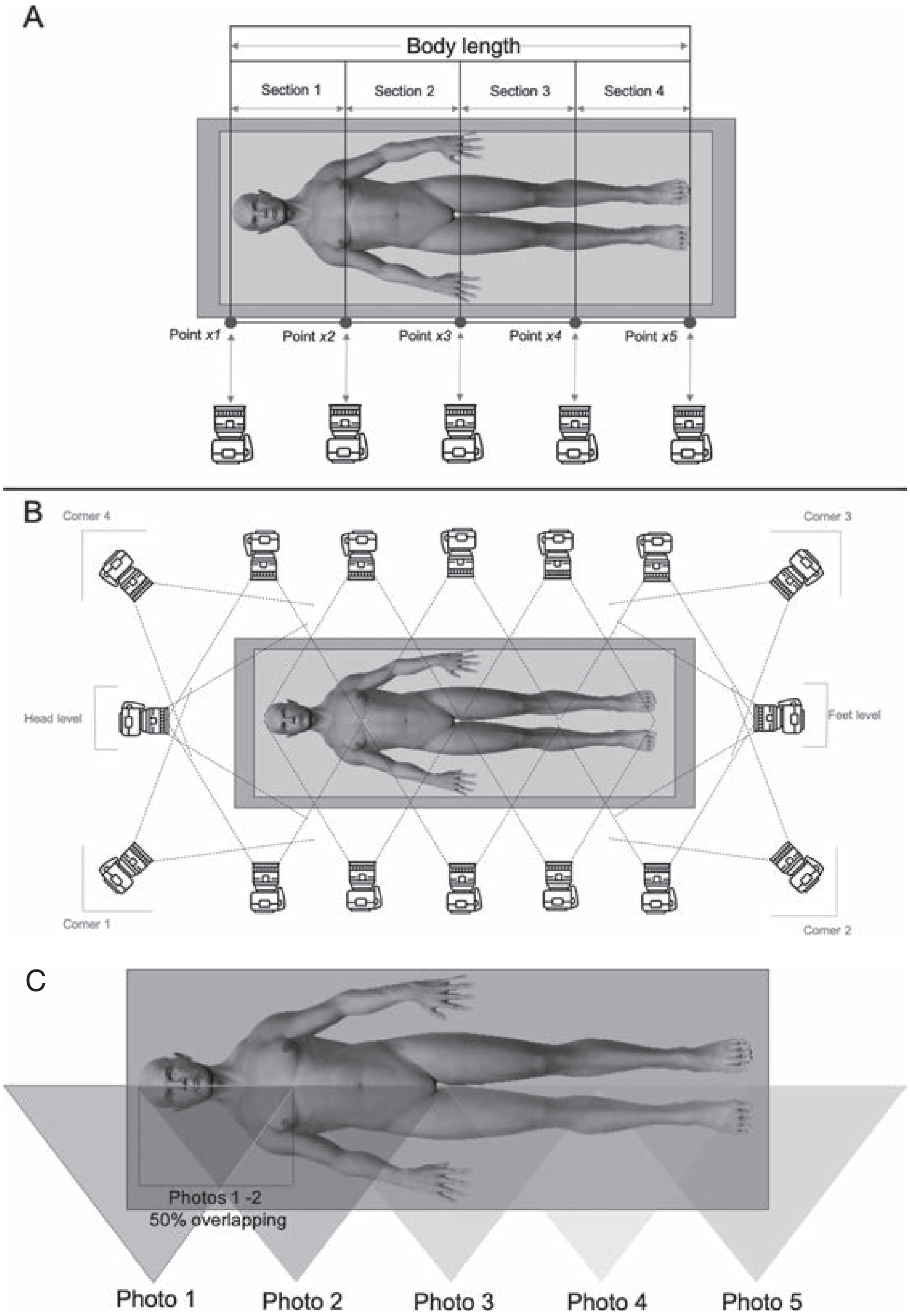


FIG. 3—*(A) In this example, the body is divided into four sections. Considering a body that is 160 cm in length, each section will measure 40 cm. The first section will be delimited by two virtual points placed at 0 cm (*x1*—proximal) and 40 cm (*x2*—distal); the second section will be delimited by two virtual points placed at 40 cm (*x2*—proximal, corresponding to the distal point of the previous section) and 80 cm (*x3*—distal); and so on. In total, five points (red circles) are obtained by the division of the body into four sections, and pictures will be taken for each of these points. (B) Additional photographs from the head level, the feet level, and the corners are required. (C) Each photograph must overlap by at least 50% adjacent photographs: the colored triangles represent the camera fields to show the overlapping pictures. [Color figure can be viewed at wileyonlinelibrary.com]*

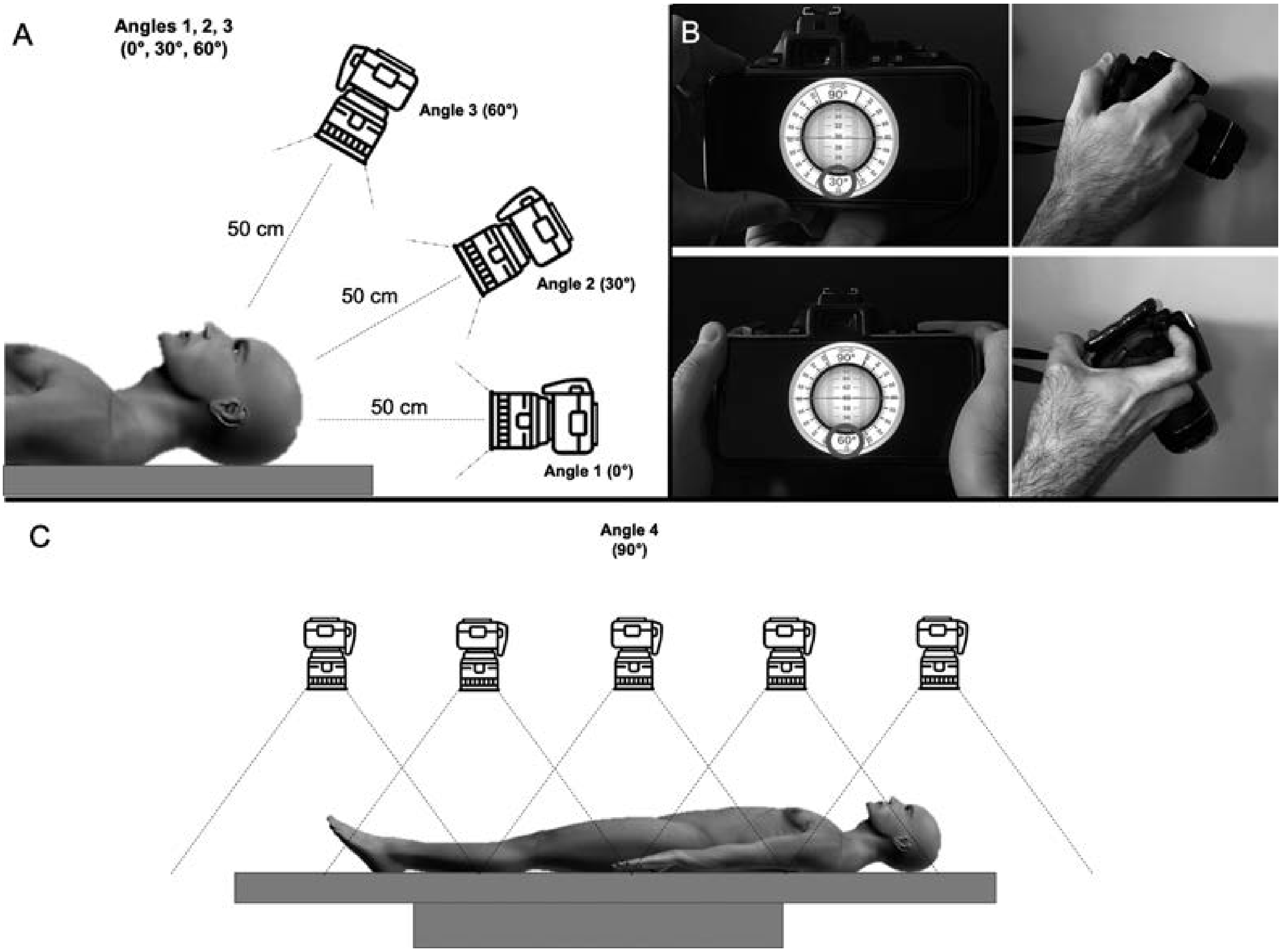


FIG. 4—*Photographs should be taken at different angles (inclination of the camera). (A) Angles 1 (0°), 2 (30°), and 3 (60°) are shown. (B) While taking photographs, a clinometer or a protractor bond to the camera can be used to determine the inclination of the camera. The present protocol was developed using a smartphone application (digital clinometer). In the picture, the smartphone was leaning against the back of the camera while running the digital clinometer to set the right photograph angle (in the red circles, the degrees of inclination are shown). (C) Angle 4 (90°) is shown. [Color figure can be viewed at wileyonlinelibrary.com]*

  - The user decides the number of sections in which the body will be divided: more sections will require more photographs and lead to higher quality rendering. The length of the body is divided by the chosen number of sections so that all the sections are of the same size. Each section is delineated by two “virtual points” (proximal and distal), from which photos will be taken. The total number of points will equal the number of the sections plus one. The first section will be delineated by two points at *x*1 (0 cm proximal) and *x*2 (distal); the second section will be delineated by two points at *x*2 (proximal, the same distal point of the previous section) and *x*3 (distal); and so on. For example, if the body is divided into four sections, five points are obtained, and photographs are taken at each point (Fig. 3A).
  - Additional photographs are taken from the head level, the feet level, and the four corners of the table (Fig. 3B).
  - Each photograph must overlap the adjacent photographs by 50%. (Fig 3C).
- Two essential variables will determine the position of the camera in space (including the height from the floor) while taking each photograph:
  - The distance from the camera lens to the virtual points.
    - Photographs should be taken at the same distance from each virtual point. An adequate distance is 50 cm from each point, although a minimal variability (±10 cm) is permitted due to physiologic body surface irregularities.
  - The inclination of the camera (angles).
    - Photographs should be taken at different angles: angle 1 (0°), angle 2 (30°), angle 3 (60°), and angle 4 (90°). Minimal variability (±5°) is permitted. While holding the camera at the same distance from the points, the correct inclination of the device can be determined using a clinometer (digital or mechanical) or a protractor bound to the camera (Fig. 4A–C).
- A ruler should be placed close to the body in one photograph for the software to recognize the size of that area and estimate other body measurements.
- Default settings with automatic camera focus are recommended since they ensure that images remain sharp. No camera zoom should be used.
- The steps mentioned above can be applied to single organs or specific parts of the body.

Minimum Number of Photographs

For satisfactory results, we suggest taking at least 106 photographs of the whole body if divided into four sections (five virtual points): 53 for the anterior aspect and 53 for the posterior aspect. For both aspects of the body, photographs should be taken as follows: 48 photographs taken from angles 1, 2, and 3 at each point on the left and right of the body, at the head level, at the feet level, and at the four corners, and five photographs from angle 4 (moving the camera from head to toe). See Figs 3 and 4 for further details.

Photogrammetric Elaboration

Detailed description of the steps necessary to properly use the software to obtain the final 3D rendering is beyond the scope of the present protocol. Numerous software programs exist and it is not possible to report a single step-by-step guide that will be applicable to each program. The following information is intended to summarize the main steps necessary to obtain a final 3D model of the body using generic software.

Two software programs are required to obtain 3D models using this protocol: photogrammetry software and a 3D digital sculpture tool software. Once the photographs are taken, the images are loaded into a photogrammetry software that processes the pictures and recognizes common areas between them. A point cloud is produced by the software, which forms the basis for the creation of the three-dimensional matrix. At the end of this process, two separate 3D meshes are created (one each for the anterior and posterior aspects of the body). Once the processing is done, the project is saved and exported in the adequate format (in our case, *.obj* filetype).

A digital sculpture tool software that allows modeling, texturing, and painting of the 3D meshes is then used. The file created by the photogrammetry software is imported into the sculpture software. Once loaded, the software will recognize the areas in common between the anterior and posterior meshes of the body (depicted by the white areas) and automatically merges the meshes to produce a single 3D reconstruction of the entire body. At the end of the process, the 3D model is obtained, and it can be rotated 360°.

Results

The present photogrammetry protocol allowed us to create high-definition 3D models of the body: two-dimensional views of the anterior, posterior, and lateral aspects of the body are shown in Fig. 5, while the corresponding 3D models are available as Video Clips S1, S2 and S3. The 3D rendered models of the body obtained with the protocol were very accurate. Figure 6 shows a comparison between two-dimensional autopsy photographs (on the left) and 3D graphic models of the same areas (on the right).

The accuracy of photogrammetry in estimating body measurements was also evaluated. Measurements recorded during the postmortem examination were compared with those estimated by the software on the digital 3D models. A ruler placed close to the right hand of the body in one photograph allowed the software to create a virtual ruler. The measurements obtained digitally from the 3D rendering were highly accurate and comparable to actual measurements. For example, the donor body had a circular chemotherapy port implanted under the right chest skin, measuring 2.0 cm in diameter. During the photogrammetric elaboration, the virtual ruler was digitally moved close to the chemo port, showing comparable results with the actual measurements (Fig. 7).

A virtual probe demonstrating a bullet pathway was then added to the 3D model in a simulated case of a gunshot wound. The software allowed us to rotate the body 360 degrees, showing the direction of fire (white arrow in Fig. 8) with high accuracy.

Specific autopsy dissections and organs were also documented and elaborated with the photogrammetry. The 3D model of a leg dissection to demonstrate deep venous thrombosis in case of pulmonary embolism is shown in Fig. 9. En-block dissections of the neck and chest internal organs were performed. We deliberately took a low number of photos with a low-resolution camera setting to show the difference in quality between this 3D model (Fig. 10) and the high-definition models reported earlier in the text.

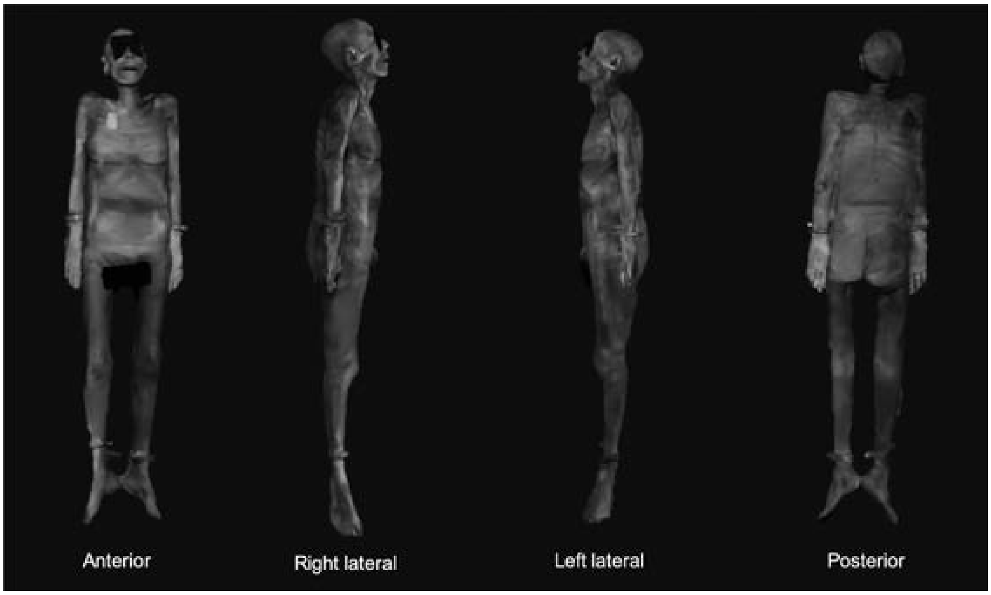


FIG. 5—*Two-dimensional views of the anterior, posterior, and lateral aspects of the body in the final 3D model. [Color figure can be viewed at wileyonlinelibrary.com]*

Discussion and Conclusions

The present protocol allowed accurate 3D modeling and measurements of the human body using photogrammetry. Due to its accuracy in elaborating detailed 3D models of human bodies, photogrammetry may become an essential tool in forensic pathology. This technique is simple, cost-effective, and its acceptance in the scientific community is increasing (2–6).

Photogrammetry has several significant advantages. Firstly, it is an accessible, simple, and relatively inexpensive technique that does not require professional skills or unusual equipment. Next, it allows the permanent recording and storage of photographic evidence, in three dimensions, of conditions that existed at the time of autopsy. Accurate measurements can then be accessed long after the autopsy is concluded. Virtual probes, rulers, or other objects can easily be added to 3D rendered models to highlight specific findings. The 3D models can then be rotated 360 degrees with the virtual device in place. The presentation of concrete and detailed visual 3D models of a body or specific injury can have a powerful effect in court. When called for deposition, the forensic pathologist may be asked to describe a particular autopsy finding using body diagrams or autopsy photographs. Unfortunately, body diagrams are not very accurate, and photographs can show only a two-dimensional view of the body, which may be inadequate in some scenarios (e.g., gunshot wounds). The 3D model of the body can be easily moved, rotated, or zoomed to highlight specific findings that may be difficult to describe using two-dimensional images. In addition, actual photographs may lead to emotional reactions in court: jurors or judges might react with sympathy or with disgust, fear, or

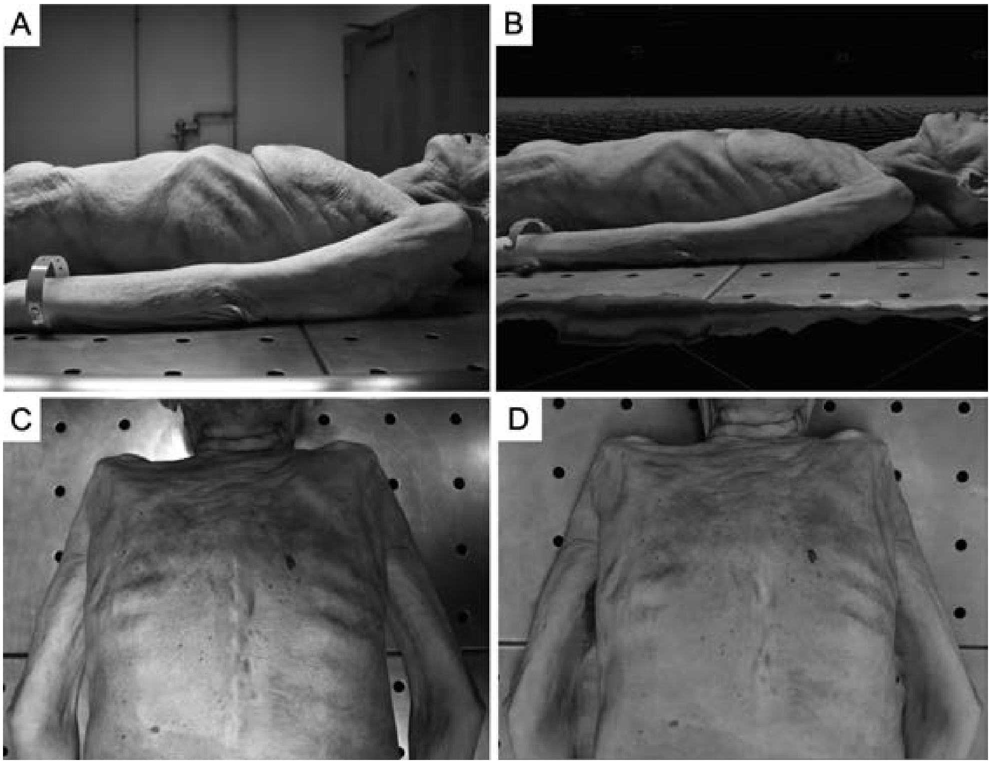


FIG. 6—*Comparison between autopsy photographs and digital images. (A, B) View of the left lateral aspect of the body: autopsy photographs (left) and 3D graphic reconstruction (right). The multiple blue spots represent the location and the number of photos taken during the external examination. (C, D) View of the upper posterior aspect of the body: autopsy photographs (left) and 3D graphic reconstruction (right). [Color figure can be viewed at wileyonlinelibrary.c om]*

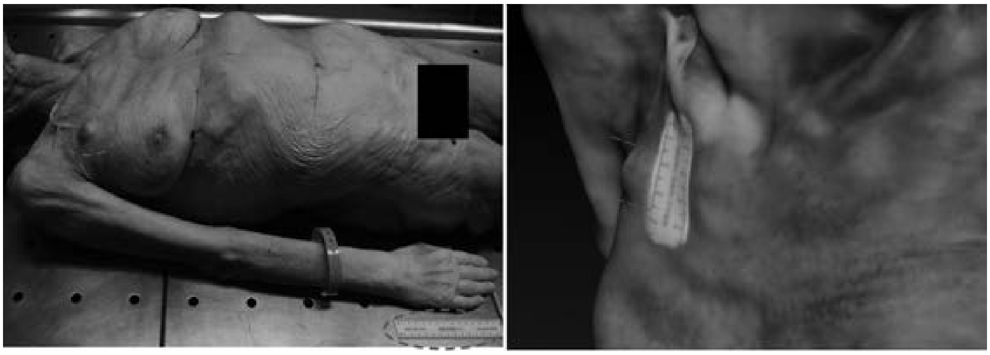


FIG. 7—*Measurements estimated from digital images. On the left, an autopsy photograph shows the chemo-port device on the right upper chest and the ruler close to the right hand (red dashed circles). On the right, the estimation of the size of the device using a virtual ruler on the 3D model (left anterior–lateral view). [Color figure can be viewed at wileyonlinelibrary.com]*

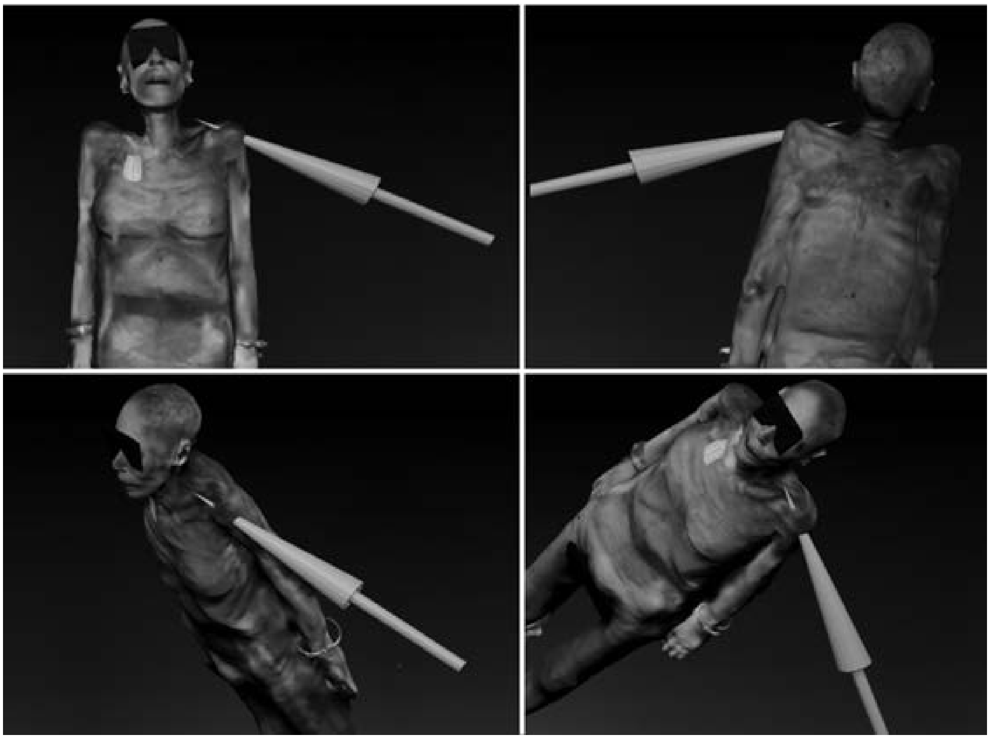


FIG. 8—*Direction of fire using a virtual probe (white arrow) on the 3D model in a simulated case of a gunshot wound. [Color figure can be viewed at wileyonlinelibrary.com]*

anger, potentially influencing their decisions. The 3D model can be manipulated in such a way that the injury is accurately depicted without elicitation of such strong emotion. There is also increasing interest in 3D modeling of the internal organs (7), in order to accurately document the original appearance of an organ surface and any specific anomalies.

The primary disadvantage of photogrammetry for forensic pathology is its time-consuming nature. When the operator

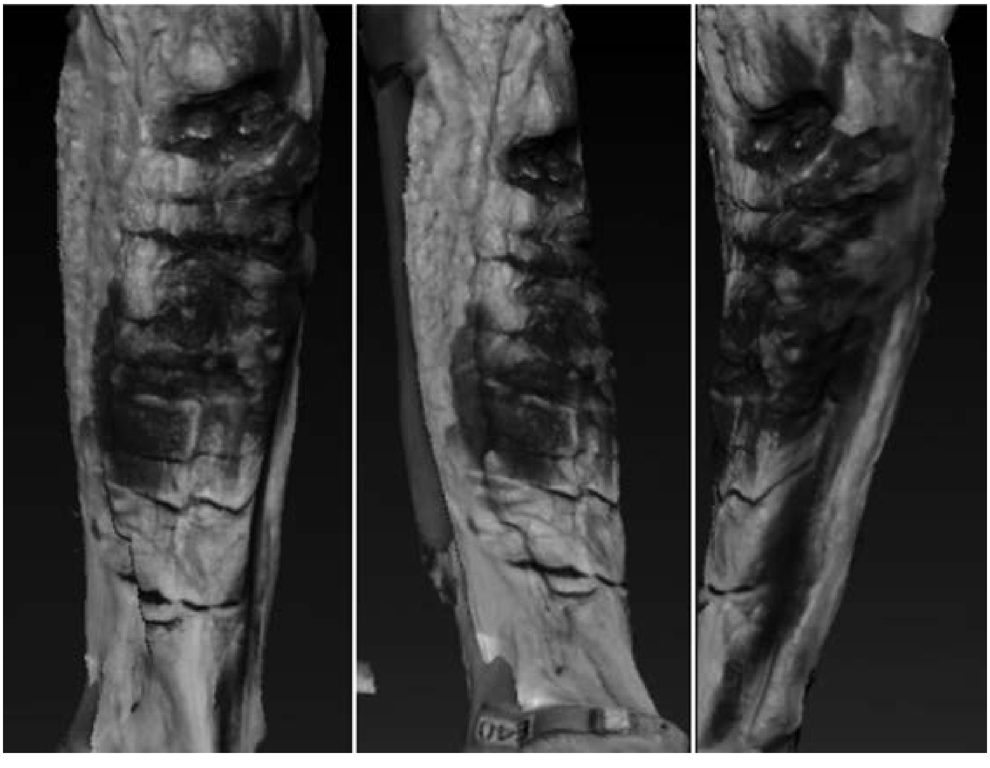


FIG. 9—*3D modeling of autopsy dissection technique: leg dissection for determination of deep venous thrombosis. [Color figure can be viewed at wileyonlinelibrary.com]*

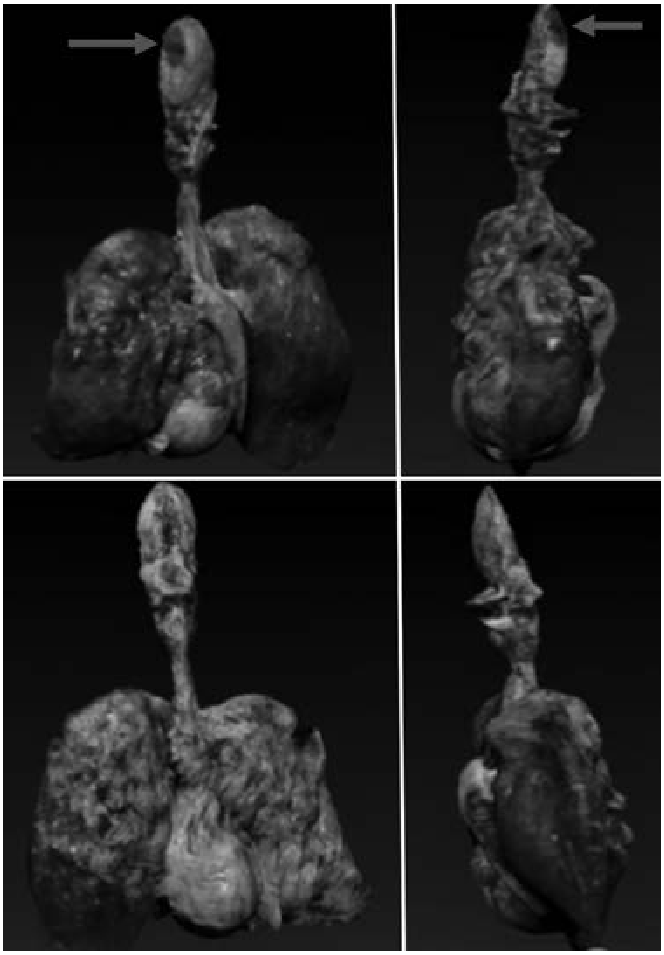


FIG. 10—*Low-definition 3D modeling of internal organs: neck and chest organs eviscerated en-block. An area of hemorrhage can be seen on the left anterior tongue (red arrow). [Color figure can be viewed at wileyonlinelibrary.com]*

becomes familiar with the device and the workflow, it still requires at least 20–40 min to take the necessary photographs of the body. The required time will change depending on the skill of the user, the desired quality of the 3D model, and the size of the area of interest (focusing on a specific anatomical area requires less time than performing the procedure on the whole body). The average smartphone or digital camera user should not have any difficulty in mastering the technique essential to performing a satisfactory photogrammetry process. Beginners may need time and practice to become familiar with their device and the angles required to take suitable photographs. A quick and straightforward way to practice is taking photographs of everyday objects (a bottle, a fruit, etc.). This allows identification and correction of user weaknesses and provides practice using the photogrammetry software for small-scale 3D rendering and taking digital measurements. Some of the authors of this paper have never performed the photogrammetry process before, and they became familiar with it after one or two autopsies. Other possible difficulties that may be faced are adequate sources of light and a clean environment. Both are mandatory to obtain good results and are easily remedied with additional lighting and clean linens.

Although not directly related to the postmortem examination process, additional time is also needed by the software to process the images. The time required can vary based on the software itself and on the computer performance. There are numerous open-source and commercial photogrammetry and 3D modeling software available online for the major operating systems (Windows, MacOS, Linux) (8,9). Finally, as with every digital data, they are susceptible to hacks and loss due to hardware damage or viruses. For these reasons, we suggest performing the photogrammetry protocol only in relevant cases.

The presented protocol describes a simple way to take photographs of the human body for photogrammetry purposes. This workflow allows users to take pictures in a standardized and reproducible way, regardless of the software used to elaborate the digital images. Widespread use of the photogrammetry technique in postmortem rooms would allow in-depth testing to detect and fix potential pitfalls, making this technique more reproducible and verifiable. Validating the application of photogrammetry will increase its admissibility in court.

For those interested in becoming familiar with the theory of the photogrammetry knowledge base and technology, numerous resources are available, including dedicated books and online information. (10,11)

References

1. Grabherr S, Egger C, Vilarino R, Campana L, Jotterand M, Dedouit F. Modern post-mortem imaging: an update on recent developments. Forensic Sci Res 2017;2(2):52–64.
2. Urbanová P, Hejna P, Jurda M. Testing photogrammetry-based techniques for three-dimensional surface documentation in forensic pathology. Forensic Sci Int 2015;250:77–86.
3. Flies MJ, Larsen PK, Lynnerup N, Villa C. Forensic 3D documentation of skin injuries using photogrammetry: photographs vs video and manual vs automatic measurements. Int J Legal Med 2019;133(3):963–71.
4. Sheppard K, Cassella JP, Fieldhouse S. A comparative study of photogrammetric methods using panoramic photography in a forensic context. Forensic Sci Int 2017;273:29–38.
5. Slot L, Larsen PK, Lynnerup N. Photogrammetric documentation of regions of interest at autopsy—a pilot study. J Forensic Sci 2014;59 (1):226–30.
6. Leipner A, Obertová Z, Wermuth M, Thali M, Ottiker T, Sieberth T. 3D mug shot-3D head models from photogrammetry for forensic identification. Forensic Sci Int 2019;300:6–12.
7. Turchini J, Buckland ME, Gill AJ, Battye S. Three-dimensional pathology specimen modeling using "structure-from-motion" photogrammetry: a powerful new tool for surgical pathology. Arch Pathol Lab Med 2018;142(11):1415–20.
8. All3dp. 2020 best photogrammetry software. 2020. https://all3dp.com/1/best-photogrammetry-software/ (accessed May 23, 2020).
9. All3dp. 2020 best 3D modeling software/ 3D design software. 2020. https://all3dp.com/1/best-free-3d-modeling-software-3d-cad-3d-design-software/ (accessed May 23, 2020).
10. Linder W, editor. Digital photogrammetry. A practical course, 4th edn. Berlin, Germany: Springer-Verlag, 2016.
11. 3DSCANEXPERT. The beginners guide to 3d scanning & photogrammetry on a budget. 2020. https://3dscanexpert.com/beginners-guide-3d-scanning-photogrammetry/ (accessed May 27, 2020).

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Video Clip S1. 3D model of the posterior aspect of the body (01:12).

For photogrammetric purposes, it is mandatory to follow specific criteria while taking photos. As explained in the paper, the more photos are taken, the higher will be the definition of the final 3D model. This high-quality 3D model is the result of more than 50 photos of the posterior aspect of the body. Photos were taken at different inclinations to obtain an accurate and reliable 3D elaboration of the body. Different inclinations of the camera need to be considered, ranging from angle 1 (0 degrees) to angle 4 (90 degrees). For each angle, multiple photos should be taken, moving the camera around the body, and keeping the same distance from the points of interest. Here, you can see a ruler place in proximity to the body. It is necessary to allow the software to recognize the size of the digital area surrounding the ruler. In this way, a reliable virtual estimation of the measurements of body areas can be made on the 3D model. The virtual camera can be easily moved around the 3D model as well as stopped on areas of interest. While moving the virtual camera, it is also possible to zoom in and zoom out on specific body areas to allow customized views of the body.

Video Clip S2. 3D model of the anterior aspect of the body (00:36).

In this video, you can see the 3D model of the anterior aspect of the body. The multiple blue marks on the background represent the number and the location of the camera. The user can move the virtual camera all around the body, highlining specific areas of interest. Here, a zoom to the chemo port device is shown. It is possible to hide selected body areas. In this case, the eyes and the genitals were deleted. The high-quality 3D model allows seeing even small details, such as the scant pink lividity on the left lateral aspect of the body and the skin wrinkles on the left breast.

Video Clip S3. Final 3D model of the body (00:53).

This video shows the final 3D model obtained by merging the anterior and posterior meshes. The software recognizes the area in common between the meshes, elaborating a three-dimensional model of the body. The 3D model can be easily moved and zoomed, and special effects can be added. Here, you can see the dissection of the right leg to demonstrate deep venous thrombosis in a case of pulmonary embolism. The specific body area can be moved to show the desired details. Moreover, the digital model can be easily manipulated to add virtual probes. In this example, a white arrow was added to depict the direction of fire in a simulated case of a gunshot wound. Rotating the body in every direction provides views of the area of interest from different perspectives. In this way, users can review, as well as display specific findings that are hard to describe using two-dimensional images.

Audio comments to the videos have been added using the text-to-speech software “Amazon Polly” (Seattle, Washington, United States, © 2008 - 2020, Amazon Web Services).

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14497
Available online at: onlinelibrary.wiley.com

TECHNICAL NOTE

PATHOLOGY/BIOLOGY

Andrea Porzionato,[1] M.D., Ph.D.; Diego Guidolin,[1] Ph.D.; Aron Emmi,[1] Ph.D.; Rafael Boscolo-Berto,[1] M.D.; Gloria Sarasin,[1] Ph.D.; Anna Rambaldo,[1] Ph.D.; Veronica Macchi,[1] M.D., Ph.D.; and Raffaele De Caro,[1] M.D.

High-quality Digital 3D Reconstruction of Microscopic Findings in Forensic Pathology: The Terminal Pathway of a Heart Stab Wound*

ABSTRACT: High-quality digital three-dimensional (3D) reconstructions of microscopic findings have been used in anatomical and histopathologic research, but their use in forensic pathology may also be of interest. This paper presents an application of these methods to better characterize the pathway of a stab wound of the anterior surface of the heart in a case of suicide. A portion of the heart wall including the stab wound was serially sectioned for microscopic analysis along the full extent of the wound. Histologic sections were digitally acquired, and a 3D reconstruction was created with ImageJ software for 3D computer graphics. This showed a full-thickness wound path extending to the endocardial surface of the left ventricle, curvilinear in appearance. After correction for shrinkage, 3D reconstruction allowed estimation of the dimensions of the myocardial injury and comparison of the appearance of the wound with the suspected knife used. The curvilinear appearance was considered to reflect injury during myocardial contraction. Complete microscopic sectioning and 3D reconstruction may allow virtual sectioning through various orientations and also provide useful forensic information for selected injuries.

KEYWORDS: 3D reconstruction, microscopic findings, forensic pathology, stab wound, histopathology, virtual cut

In the field of forensic sciences, three-dimensional (3D) visual representation of medical and forensic findings obtained with various methods is a constantly evolving aspect of the discipline. Three-dimensional volume rendering techniques through computed tomography (CT) or magnetic resonance imaging (MRI) have been widely employed for morphological evaluation of various aspects of pathology (1,2). Photogrammetry (3,4) and structured light and laser surface scanning (5) can also be employed for 3D reconstructions. These approaches have been applied not only to crime scene investigations but also to analysis of lesions of soft tissues and bone (6,7). Further development in the forensic involvement of 3D image-acquiring technology has been integrated with information from CT and MRI, through merging of surface 3D images with deep visceral images by CT and MR: This approach may be useful in providing a first complete analysis of lesions before autopsy (6,7).

Three-dimensional reconstructions of microscopic sections allow us not only to visualize 3D microscopic aspects, but also to evaluate additional morphometric parameters. Although the importance of 3D reconstructions of microscopic findings from serial sections has been extensively stressed in various fields of morphological research and histopathology (8–11), its feasibility and usefulness in *forensic histopathology* has not yet been fully evaluated. The present work addresses the methodological implications of 3D microscopic reconstruction in the forensic setting through practical application involving the terminal heart pathway of a thoracic stab wound.

Materials and Methods

Case History and Autopsy Findings

A man was found dead in the bathroom of his home, with many blood stains on his clothes and on the floor. A blood-covered knife was lying on the floor. External examination revealed multiple incised wounds on the wrists, three stab wounds at the level of the neck, and a single stab wound in the chest (Fig. 1*A*). At autopsy, injuries to the neck and wrists were found to be superficial, whereas the thoracic stab wound had penetrated the chest wall and pericardium (Fig. 1*B*). Blood was present in the pericardial space (about 100 mL) and left pleural cavity (about 300 mL); the right pleural cavity was free of blood. Examination of the heart showed a stab wound 8 mm long in the anterior surface of the left ventricle (Fig. 1*C*). Heart sectioning and inspection of the corresponding internal aspect of the left ventricle did not show an evident macroscopic injury, partly due to the irregularity of the trabeculae

[1]Section of Human Anatomy, Department of Neuroscience, University of Padova, Via Gabelli 65, Padova, 35127, Italy.

Corresponding author: Aron Emmi, PhD. E-mail:aronemmi@hotmail.it

*All procedures performed in the study were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Received 10 Feb. 2020; and in revised form 29 May 2020; accepted 5 June 2020.

carneae, although a full-thickness wound was clearly suggested by hemopericardium (Fig. 1*D*). The stab wound in the heart wall was paraffin-embedded and subjected to complete sectioning for microscopic analysis along the whole length of the wound.

Methods of Image Acquisition and 3D Reconstruction

Tissue Preparation

Following fixation of the whole heart in 4% paraformaldehyde buffered in PBS for 10 days, a rectangular full-thickness 4 × 3 × 1.8 cm tissue block of the cardiac wall was sampled, containing the 8-mm-long wound found during autopsy (Fig. 1*E, F*). The tissue was dehydrated in a series of progressive alcohol solutions, cleared in xylene, and embedded in paraffin.

Scaled photographs were taken of the cut surface before and after tissue processing, in order to evaluate the amount of shrinkage induced by further processing, as reported by Hyde et al. (12) and Schneider and Ochs (13), who assumed uniform shrinkage across the various tissue components (12,14). The width, length, and area of the cut surface prior to processing were divided by the measurements on the stained sections, providing reliable shrinkage values which were then used to correct the size of the 3D models and related morphometric data.

Serial sections 10 µm thick were obtained from the tissue sample, each including the full thickness of the cardiac wall, on

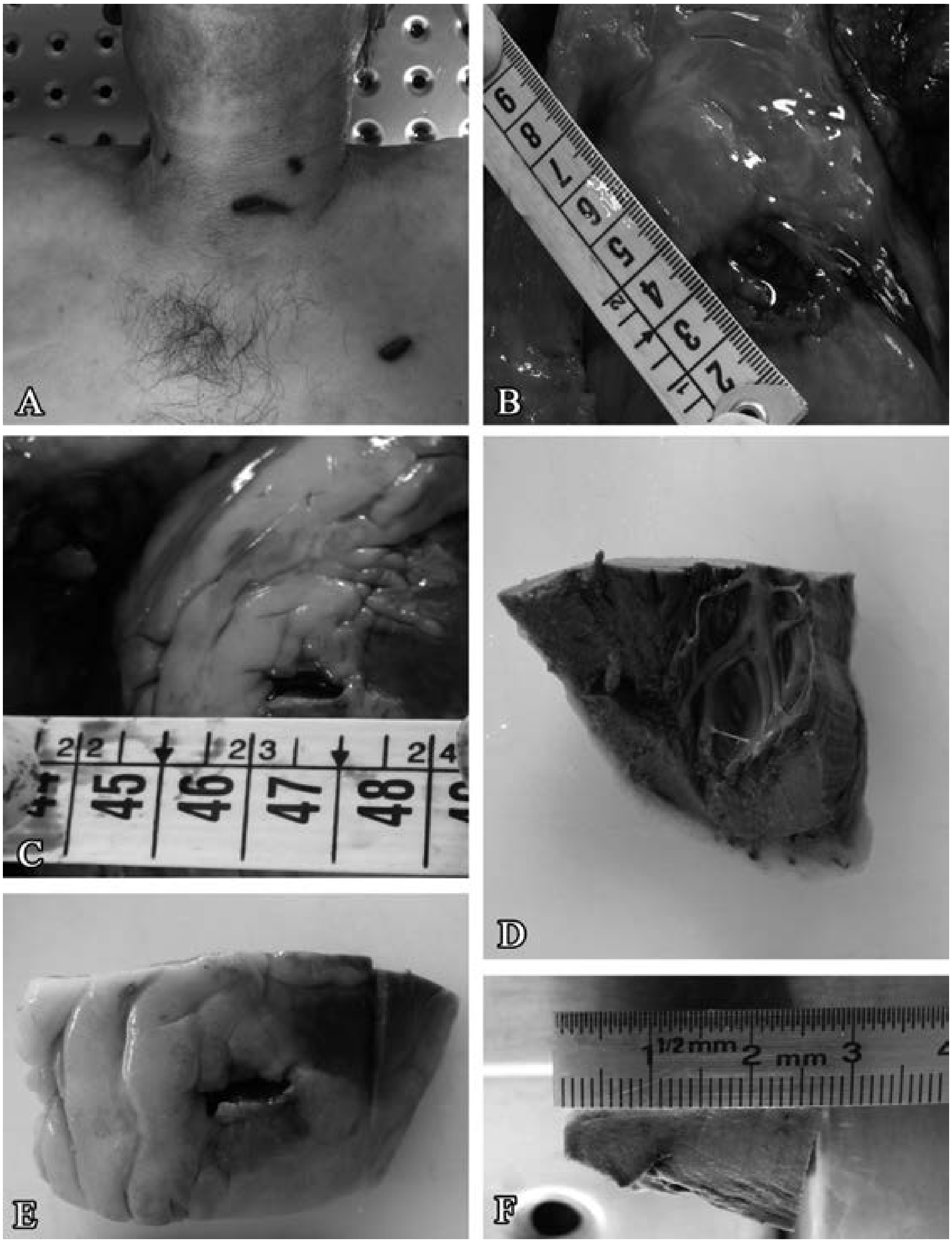


FIG. 1—*(A) External view of three stab wounds of neck and a single stab wound of chest. (B) Stab wound of anterior pericardial sac. (C) Stab wound, anterior left ventricle. (D, E) Formalin-fixed tissue block of the left ventricle prior to sectioning, containing full-thickness stab wound and showing internal (D) and external (E) surfaces. (F) Orientation of tissue block for sectioning, inferior left ventricle toward left, endocardial surface facing downward. [Color figure can be viewed at wileyonlinelibrary.com]*

a calibrated rotary microtome (RM 2235, Leica Microsystems, Wetzlar, Germany). Disposable metal microtome blades (Bio-Optica, Milan, Italy) were used at a cutting angle of about 5°. A set of 118 consecutive sections regularly sampling the whole tissue block at 200-μm intervals was then selected. The sections were dewaxed in xylene and rehydrated through a series of alcohol solutions of 99%, 95%, 70%, and 50%, ending with dH_2O. Sections were soaked for 8 min in hematoxylin and washed in flowing water for 5 min. Subsequently, the sections were left in eosin for 1 min, washed in distilled water, dehydrated in progressive alcohol solutions, cleared in xylene, and cover-slipped with a mounting medium.

Image Acquisition and Segmentation

Color images of the consecutive sections were acquired on a digital camera (Optikam HDMI, Optika srl, Italy) attached to a stereo microscope (SZM-2, Optika srl, Italy) operating with transmitted light at a first magnification of 0.67. Before acquisition, images of consecutive sections were manually aligned under the microscope, to compare live color images of the section examined with the contours of the previous section in an overlay display. Images were then acquired, corrected for shading, and stored as 24-bit RGB TIFF files (size 1280 × 720 pixels) (Fig. 2*A–D*). ImageJ software ([15]; freely available at http://rsb.info.nih.gov/ij/) was used, and specific routines were designed by the authors to facilitate the acquisition procedure (i.e., generation of the overlay of the previous section, image acquisition of the current section, shading correction, and storage of acquired images).

Before the images were used to create 3D models, an automated registration procedure was carried out for more accurate alignment of the images. “Registration” is the process of finding the spatial transformation which maps points from one image to corresponding points in another image (16). In the approach followed here, pixel intensity values were considered as important images, as they can facilitate the inclusion of the entire data contents, and automatic recording was performed with the “StackReg” plug-in for ImageJ of Philippe Thévenaz ([17]; http://bigwww.epfl.ch/thevenaz/stackreg/). This procedure maximizes the cross-correlation between consecutive images with the Marquardt–Levenberg algorithm for nonlinear least-squares optimization (18). The geometric deformation model applied was restricted to rigid-body motion (rotation and translation) which is usually used in applications involving only intra-subject registration (18).

The images from the aligned set were then segmented to identify the most important tissue components (Fig. 2*E,F*). The myocardium was identified by a conventional thresholding procedure; the knife wound and blood vessels were interactively traced. ImageJ software was always used to perform these processing steps.

Three-dimensional Reconstruction

Three-dimensional surface rendering was performed with the Visualization Toolkit (VTK, version 7.0), an open-source library

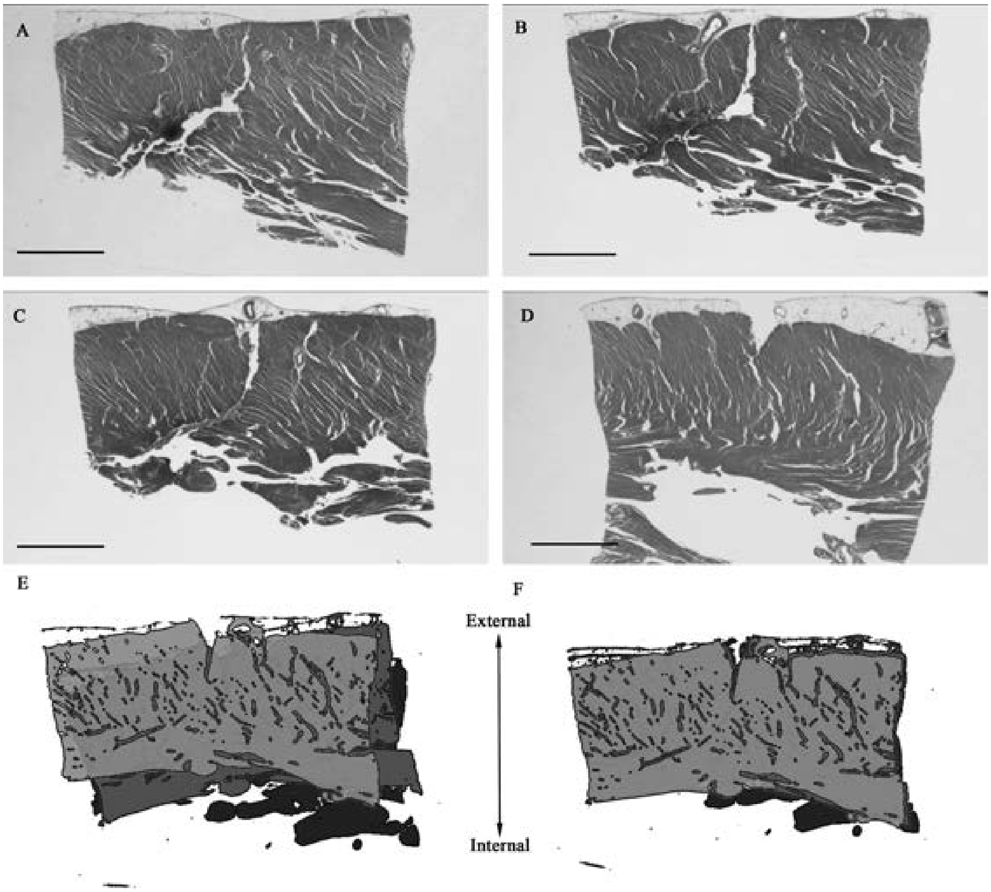


FIG. 2—*(A–D) Serial sections stained in hematoxylin–eosin. Scale bar = 7.5 mm. (E) Superimposed sections following thresholding. (F) Recording and alignment of thresholded sections. [Color figure can be viewed at wileyonlinelibrary.com]*

developed by Kitware Inc. (New York, freely available at http://www.vtk.org). VTK-based procedures were built with the VisTrails 2.2.3 tool (http://www.vistrails.org; see Silva et al. [19]). Briefly, for each selected tissue component, a volume of 118 binary images was read from TIFF files, with (relative) pixel spacing set to 1:1:4 to maintain correct scaling in all directions; the VTK "ContourFilter" function was applied to extract iso-surfaces encompassing the 3D structure as a combination of vertices and polygons connecting them. The resulting 3D surface was then smoothed with the VTK "WindowedSincPolyDataFilter" function to refine the geometric mesh (20). The resulting 3D object was finally stored as a vtk file. The whole 3D structure of the tissue block was then visualized with Paraview software (version 5.0.1, Kitware Inc.) which can render multiple 3D objects as a single scene with emulation of light and camera standpoint (Fig. 3).

Results and Discussion

Preliminary examination of the anatomo-microscopic sections showed a lesion extending toward the internal surface of the left ventricular wall, limited in extension but clearly visible. The lesion, as shown by its profile within the sections, presented a curvilinear pattern in the myocardium (Fig. 2), consistent with myocardial contraction at the moment of injury (rectilinear pattern) and following acquiring of a curvilinear shape due to differential muscle relaxation of the internal and external components of the myocardial wall. Thus, this lesion characteristic also confirmed the vitality of the lesion.

In 3D reconstruction of microscopic findings from embedded samples, preliminary evaluation of global shrinkage is essential in order to obtain reliable morphometric data. In this case, the global surface areas of the preprocessed and postprocessed cut surfaces were 353.1 mm^2 and 334.7 mm^2, respectively. A shrinkage value of 0.948 was estimated from these measurements. The 2D and 3D measurements on stained sections and 3D reconstructions were corrected on the basis of this shrinkage value.

Three-dimensional reconstructions of both tissues and lesion were first evaluated independently and then merged into a single scene (Fig. 3*A*). The lesion appeared as an approximately triangular structure with a linear base, two curvilinear lateral sides, one slightly concave, and the other slightly convex. The convex side curves progressively formed an obtuse angle, continued parallel to the base, and then merged with the concave side toward the apex. The distinct curvature of the whole structure was evident along its transverse axis.

Morphometric evaluation of the lesion within the tissue evidenced a − 41° angle downwards from the external to the internal surface and a 50° angle in medio-lateral direction (Fig. 3*C*). The base of the lesion was 7.9 mm thick; the convex side, forming the above-mentioned obtuse angle, was 30.2 mm (22 mm from the first segment and 8.2 mm from the second segment, which merged toward the apex); and the concave side was 18 mm (Fig. 3*D*). The lesion extended from the external to the internal surface of the left ventricular wall, with an approximate depth of 10.8 mm. On the internal surface, the lesion was masked by the trabeculae carneae, but nonetheless appeared visible through 3D rendering (Fig. 3*B*). In addition, the curvilinear pattern of the lesion within the myocardium (quantified as the ratio between Arch/Chord, of 1.122) further confirmed the findings on the microscopic sections (Fig. 3*E*).

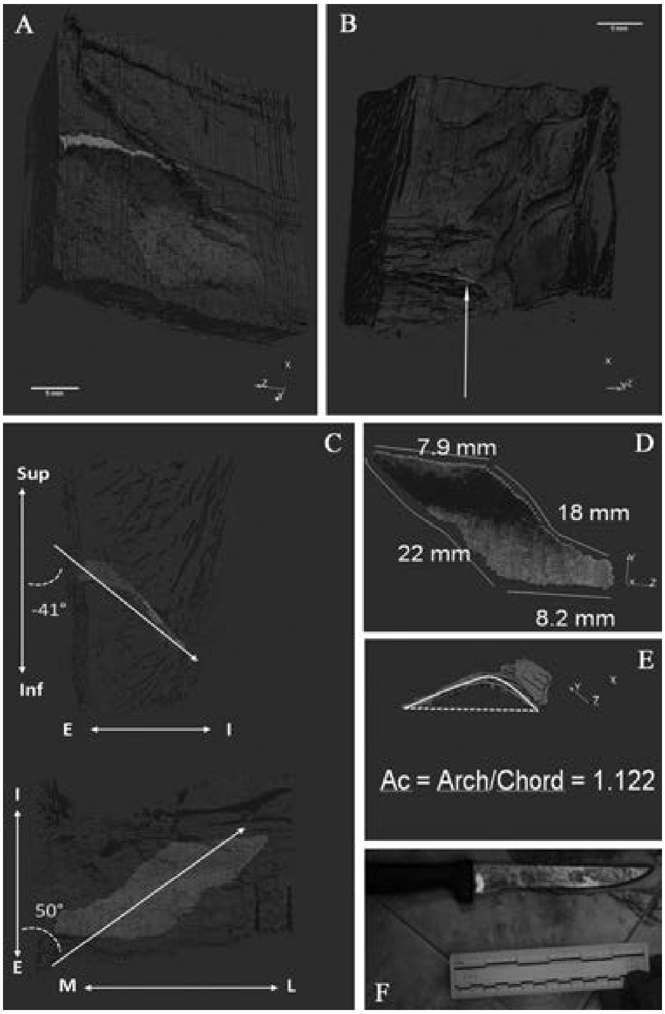


FIG. 3—*(A) 3D reconstruction of myocardium (brown, semi-transparent), lesion (pale blue), and blood vessels within tissue (red) rendered as a single scene within Paraview. Scale bar = 5 mm. (B) Internal surface of 3D reconstructed tissue block and lesion; yellow arrow points toward lesion (pale blue) at level of internal surface of ventricular wall, partially hidden by trabeculae. Scale bar = 5 mm. (C) Angulation, orientation, and trajectory of lesion (pale blue) within tissue block (brown). Sup, superior; Inf, inferior; E, external; I, internal. (D) Morphometrical values of margins of lesion. Convex margin was measured as an oblique segment of 22 mm and a transverse segment of 8.2 mm. (E) Curvilinear pattern of lesion, indicating that injury occurred during myocardial contraction. Continuous yellow line, Arch; dotted yellow line, chord. Ac, ratio between Arch and Chord. (F) Photograph of knife found at crime scene. [Color figure can be viewed at wileyonlinelibrary.com]*

Another feature achievable thanks to 3D reconstruction of the lesion and injured tissue is the possibility of rotating the model on various planes. The production of the so-called "Virtual Cut" is also possible, and allows us virtually to section the reconstructed 3D model in order to achieve perspectives of lesions which cannot be appreciated from single microscopic sections, as seen in Fig. 4. The virtual section of the 3D reconstruction also allows us to measure the lesion accurately—something not otherwise possible, due to the modalities of physical sections. In this case, however, it was possible to measure the width of the lesion along the main axis of the object causing it (7.9 mm)—a parameter which had not been clear in the microscopic sections (Fig. 4, right). In the case of stab wounds, virtual sections perpendicular to the pathway may be

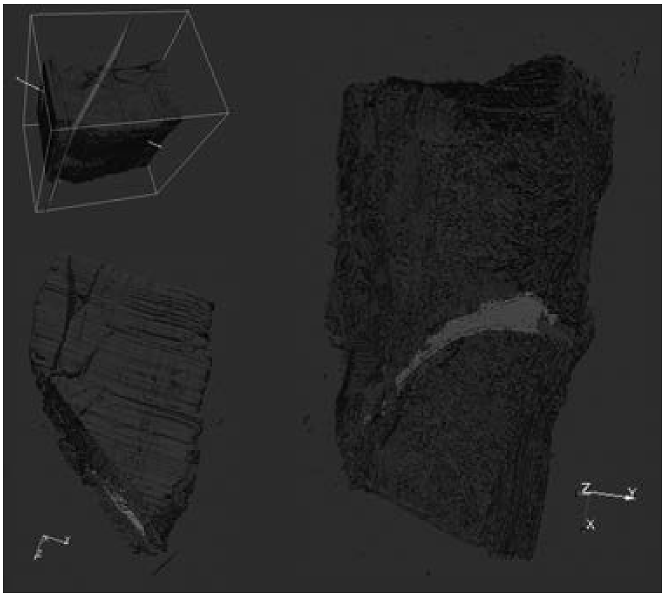


FIG. 4—*Virtual cut of 3D reconstruction. Top: Red line indicates level of section within block. Right: note width of lesion (pale blue) within tissue (brown) along main axis of object causing it. [Color figure can be viewed at wileyonlinelibrary.com]*

particularly useful in identifying the width of the blade when the trajectory is oblique.

In the present study, the heart sample was quite large, but adequate microscopic seriation is also possible in smaller samples. The main limitation of the technique is due to the need for exhaustive seriation of the sample, with consequent acquisition of many sections. This may not be easy to perform in the everyday work of a laboratory. However, 3D reconstruction of a totally sectioned sample should be carried out in specific cases, when sometimes important questions cannot easily be answered after macroscopic and "regular" microscopic analysis. Moreover, nowadays, image acquisition and 3D reconstruction are relatively economical, easy, and rapid to perform, involving simple open-source software.

In conclusion, this report presents an example of demonstrative application of 3D reconstruction in forensic histopathology. Although it was not strictly necessary for identification of the cause and dynamics of death, it did provide further useful information on the features of the shape of the pathway and evaluation of consistency with the suspected weapon. The size and shape of the wound appeared to be compatible with the distal part of the blade found at the crime scene (Fig. 3*A,B,F*).

In a general view, 3D reconstruction of microscopic findings may represent an additional scientific tool for ascertaining morphological/morphometric characteristics of specific lesions aiding identification of their cause. In this study, a myocardial lesion was analyzed, but 3D reconstructions of traumatic histopathologic findings may be of potential interest also for other tissues, such as skin, skeletal muscle, bone, visceral walls, and others. In addition, 3D reconstructions may include not only injury pathways but also histopathologic aspects of the surrounding tissues, such as hemorrhagic infiltration or immunohistochemical expression of vitality markers. Three-dimensional reconstructions can also be integrated with the surface 3D images of injuries obtained with optical digitizers: This may be particularly useful for lesions characterized by superficial and deep aspects of interest.

References

1. Oehmichen M, Gehl HB, Meissner C, Petersen D, Höche W, Gerling I, et al. Forensic pathological aspects of postmortem imaging of gunshot injury to the head: documentation and biometric data. Acta Neuropathol 2003;105(6):570–80. https://doi.org/10.1007/s00401-003-0683-4.
2. Biggs MJP, Morgan B, Rutty GN. Using freely-available 3D software to reconstruct traumatic bone injuries detected with post mortem computed tomography. Forensic Sci Med Pathol 2020;16(1):113–8. https://doi.org/10.1007/s12024-019-00205-3.
3. Sansoni G, Cattaneo C, Trebeschi M, Gibelli D, Porta D, Picozzi M. Feasibility of contactless 3D optical measurement for the analysis of bone and soft tissue lesions: new technologies and perspectives in forensic sciences. J Forensic Sci 2009;54(3):540–5. https://doi.org/10.1111/j.1556-4029.2009.01041.x.
4. Zancajo-Blazquez S, Gonzalez-Aguilera D, Gonzalez-Jorge H, Hernandez-Lopez D. An automatic image-based modelling method applied to forensic infography. PLoS One 2015;10(3):e0118719. https://doi.org/10.1371/journal.pone.0118719.
5. Buck U, Buße K, Campana L, Gummel F, Schyma C, Jackowski C. What happened before the run over? Morphometric 3D reconstruction. Forensic Sci Int 2020;306:110059. https://doi.org/10.1016/j.forsciint.2019.110059.
6. Thali MJ, Braun M, Dirnhofer R. Optical 3D surface digitizing in forensic medicine: 3D documentation of skin and bone injuries. Forensic Sci Int 2003;137(2–3):203–8. https://doi.org/10.1016/j.forsciint.2003.07.009.
7. Thali MJ, Braun M, Buck U, Aghayev E, Jackowski C, Vock P, et al. VIRTOPSY – scientific documentation, reconstruction and animation in forensics: individual and real 3D data-based geometric approach including optical body/object surface and radiological CT/MRI scanning. J Forensic Sci 2005;50(2):428–42. https://doi.org/10.1520/JFS2004290.
8. Handschuh S, Schwaha T, Metscher BD. Showing their true colors: a practical approach to volume rendering from serial sections. BMC Dev Biol 2010;10:41. https://doi.org/10.1186/1471-213X-10-41.
9. Macchi V, Porzionato A, Guidolin D, Parenti A, De Caro R. Morphogenesis of the posterior inferior cerebellar artery with three-dimensional reconstruction of the late embryonic vertebrobasilar system. Surg Radiol Anat 2005;27(1):56–60. https://doi.org/10.1007/s00276-004-0303-6.
10. Wollatz L, Johnston SJ, Lackie PM, Cox SJ. 3D Histopathology – a lung tissue segmentation workflow for microfocus X-ray-computed tomography scans. J Digit Imaging 2017;30(6):772–81. https://doi.org/10.1007/s10278-017-9966-5.
11. Porzionato A, Macchi V, Guidolin D, Parenti A, Ferrara SD, De Caro R. Histopathology of carotid body in heroin addiction. Possible chemosensitive impairment. Histopathology 2005;46(3):296–306. https://doi.org/10.1111/j.1365-2559.2005.02060.x.
12. Hyde DM, Tyler NK, Putney LF, Singh P, Gundersen HJ. Total number and mean size of alveoli in mammalian lung estimated using fractionator sampling and unbiased estimates of the Euler characteristic of alveolar openings. Anat Rec A Discov Mol Cell Evol Biol 2004;277(1):216–26. https://doi.org/10.1002/ar.a.20012.
13. Schneider JP, Ochs M. Alterations of mouse lung tissue dimensions during processing for morphometry: a comparison of methods. Am J Physiol Lung Cell Mol Physiol 2014;306(4):341–50. https://doi.org/10.1152/ajplung.00329.2013.
14. Dorph-Petersen KA, Nyengaard JR, Gundersen HJ. Tissue shrinkage and unbiased stereological estimation of particle number and size. J Microsc 2001;204(Pt 3):232–46. https://doi.org/10.1046/j.1365-2818.2001.00958.x.
15. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods 2012;9(7):671–5. https://doi.org/10.1038/nmeth.2089.
16. Ng L, Ibanez L. Medical image registration: concepts and implementation. In: Yoo TS, editor. Insight into images: principles and practice for segmentation, registration and image analysis. Wellesley, MA: AK Peters Ltd., 2004;239–306.
17. Thévenaz P, Ruttimann UE, Unser M. A pyramidal approach to subpixel registration based on intensity. IEEE Trans Image Process 1988;7(1):27–41. https://doi.org/10.1109/83.65084.
18. Marquardt DW. An algorithm for least-squares estimation of non-linear parameters. J Soc Indust Appl Math 1963;11(2):431–41.
19. Silva CT, Anderson E, Santos E, Freire J. Using VisTrails and Provenance for teaching scientific visualization. Comput Graph Forum 2011;30(1):75–84. https://doi.org/10.1111/j.1467-8659.2010.0183.x.
20. Frey PJ. About surface remeshing. In: Owen S, editor. Proceedings of 9th International Mesh Round Table; 2000 Oct 2–5; New Orleans, LA. Albuquerque, NM: Sandia National Laboratories, 2000;123–36.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14503
Available online at: onlinelibrary.wiley.com

TECHNICAL NOTE

PATHOLOGY/BIOLOGY

Lawrence Hill,[1] *M.Sc. (Med), M.B.A.; Allison E. Gilbert,*[2,3] *M.Sc. (Med); and Maureen Coetzee,*[2,3] *Ph.D., F.R.E.S.*

Modeling Temperature Variations Using Monte Carlo Simulation: Implications for Estimation of the Postmortem Interval Based on Insect Development Times*'†

ABSTRACT: The association between insect development and temperature is well established. Thermal summation using accumulated degree-day measures is commonly used. However, the time at which evidence is collected is important in these estimates. The aim of this study was to provide a simulated model of the effect of temperatures on six dipteran species commonly associated with cadavers, from the death scene to the refrigerator, and finally at the time of autopsy. Temperatures measurements were sampled over a 16-month period from the external environment (external to the mortuary), within the mortuary refrigerator, and within the mortuary autopsy suite. Monte Carlo simulation using accumulated degree-days (ADD) was used to estimate the variations based on the mean and standard deviation of the temperature measurements. It was found that there was a negative correlation between the base temperature of the fly species (lowest temperature at which the flies will survive) and developmental likelihood. Species with high base temperatures (*Chrysomya albiceps, Chrysomya chloropyga*, and *Musca domestica*) were less likely to continue development in refrigerators than species with lower base temperatures (*Lucilia sericata* and *Piophila casei*). The findings of this study highlight the importance of recording temperature measurements and the period of refrigeration on PMI estimation especially when continued development occurs in spite of a period of cooling of the insect evidence.

KEYWORDS: forensic entomology, modeling, Monte Carlo simulation, postmortem interval, accumulated degree-days, thermal summation

The positive association of insect development with temperature is well established (1,2). It is this association that allows for insect development to be used as a proxy in the estimation of the postmortem interval (PMI) of bodies where time of death is unknown (3,4). The estimation of time since insect colonization of a body provides a minimum PMI that can be used to infer the time of death (3).

One of the commonly applied methods in estimating the time since death is the use of accumulated degree-days (ADD) or hours (ADH) (5). This method utilizes the physiological energy budget required for development with an assumed linear regression which predicts development based on different temperatures over the developmental history of the larva (2,3). The average development per hour or day is added together to provide an accumulated effect of temperature and time, and is thus also known as the thermal summation model (2). This method therefore requires precise recordings of temperatures and time to improve the accuracy of the PMI estimation.

When the collection of insect evidence is only performed at the time of the autopsy (6,7), or even after the autopsy (8), there are a number of factors that must be considered. Firstly, the temperature of the death scene and the time that the body was removed are essential pieces of information (2). Secondly, information regarding the storage of the body prior to the autopsy (time and temperature of refrigeration) (1) and, lastly, when the body was removed from the refrigerator are additional data that must be recorded.

The period and temperature of cooling during refrigeration may affect the PMI estimation (2,6—10). Myskowiak and Doums (9) found that a period of cooling could alter the overall development of the dipteran (fly) larvae through behavioral inactivity as a response to harsh conditions, known as diapause. Diapause is essential to winter survival for many insect species and may occur during short-term mortuary refrigeration of dipteran larvae, especially when refrigerator temperatures are below the lower thermal threshold when development would be affected (11).

Huntington *et al*. (7) found that a significantly warmer microclimate was created when body bags were used, through bacterial activity producing additional heat. This meant that larvae in morgue refrigerators could remain active instead of entering diapause, which would skew results of the PMI because of

[1] Department of Forensic Science, School of Clinical Medicine, Faculty of Health Sciences, University of the Witwatersrand, Private Bag 3, WITS, Johannesburg, Gauteng, 2050, South Africa.

[2] Wits Research Institute for Malaria, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Private Bag 3, WITS, Johannesburg, Gauteng, 2050, South Africa.

[3] Vector Control Reference Laboratory, Centre for Emerging Zoonotic and Parasitic Diseases, National Institute for Communicable Diseases, Private bag X4, Sandringham, Johannesburg, Gauteng, 2131, South Africa.

Corresponding author: Lawrence Hill, M.Sc. (Med), M.B.A. E-mail: Lawrence.Hill@wits.ac.za

*Funding provided by the National Research Foundation (NRF) via a Scarce Skills Grant for Masters Research to LH, and AEG and MC were supported by a DST/NRF Research Chairs Initiative Grant.

†Opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the NRF.

Received 27 May 2020; and in revised form 15 June 2020; accepted 16 June 2020.

temperature differences. Thevan *et al.* (8) also found that within maggot masses, insect development did not cease with refrigeration. This suggests that refrigeration may still allow for some level of development given certain conditions (microclimate and maggot mass formation). Huntington *et al.* (7) found that although the fridge temperatures were set to 4°C they, in fact, fluctuated above 5°C, suggesting that even "set" temperatures could not be used as true representations of temperature. Depending on the fluctuations, the temperature could be higher than the lower threshold temperatures allowing some species to be active in spite of refrigeration.

The aim of this study was to provide a simulated model of the effect of temperatures, from the death scene to the refrigerator, and finally at the time of autopsy. The model simulated the estimated PMI using thermal summation for six dipteran species: *Calliphora vicina* Robineau-Desvoidy, 1830; *Chrysomya albiceps* (Wiedemann, 1819); *Chrysomya chloropyga* (Wiedemann 1818); *Lucilia sericata* (Meigen, 1826); *Musca domestica* Linnaeus, 1761; and *Piophila casei* (Linnaeus, 1758).

Methods and Materials

Site of Study

This study was conducted at the Johannesburg Forensic Pathology Medico-Legal Laboratories in Johannesburg, South Africa. The facility's primary function is to conduct medicolegal postmortem examinations in cases of unnatural deaths (including suicidal, homicidal, accidental, or deaths where circumstances are unknown or questionable).

Temperature Measurements

Temperature measurements were recorded over a 16-month period from January 2012 to June 2013 from the external environment, the mortuary refrigeration unit, and the autopsy suite. The temperature of the mortuary refrigerator was set to 4°C (±2°C). Data on the temperatures of all three locations were collected using Thermocron® iButtons® (Model: DS1922L). Two iButtons® were placed in each location to ensure that a more uniform measure of temperature was obtained for the site. In the refrigeration unit, one iButton® was placed inside near the door and the second iButton® was placed on the opposite wall furthest from the door. The refrigerator had a sliding door that had to be manually closed to maintain temperatures within the room. In the autopsy suite, iButtons® were placed on opposite walls, one near the door and the second on the opposite wall (an air conditioner was present in the room, but it was not functional throughout the duration of the study). For the external environment, iButtons® were placed outside the facility in the receiving area of the mortuary that received no direct sunlight. Additional temperature measurements were obtained from the South African Weather Services (12) to ensure that the readings obtained by the external iButton® were close to the SAWS ambient temperature (accounting for microclimatic variation).

PMI Estimation and Modeling Variation Using Monte Carlo simulation

The accumulated degree-day (ADD or ADH for Hours) or thermal summation method was used (3,11). The formula for ADD is as follows:

$$\mathrm{ADD} = \mathrm{T} \times (\Theta - \Theta_0)$$

where T is the developmental time, Θ is the ambient temperature, and Θ_0 is the species base temperature. The base temperature is the lowest possible temperature at which development can still occur in a specific insect species. When the ambient temperature is below the base temperature, the ADD becomes zero as development ceases. The base temperature is calculated using the species-specific development at a range of temperatures plotted against time. The thermal summation method assumes a straight-line relationship which is then extrapolated backward to the x-intercept, which is the base temperature estimate for that species (3).

For the purposes of this study, the PMI was estimated based on the following simulation. After removal from the death scene, on average a body will take an hour to be transported to the medicolegal forensic facility. The body will then remain refrigerated for 24 h before undergoing an autopsy. Finally, the postmortem examination (autopsy) should take no longer than four hours to be completed. The assumption is that at the end of this hypothetical scenario, the insect taxa needed to estimate PMI would be given to, or collected by, the forensic entomologist. The equation used in the simulation was therefore:

$$\begin{aligned}\text{Total ADD} &= \text{ADD during transportation}\\ &\quad + \text{ADD during refrigeration} + \text{ADD during autopsy}\end{aligned}$$

The scenario described above was used to perform a Monte Carlo simulation for the hypothetical time period, and to calculate the ADD estimation to determine the associated best and worst-case scenario estimations for six dipteran (flies) taxa (Table 1).

Monte Carlo simulation is a method of analyzing unknowns or variations in a system through a number of analytical assumptions based on an input model (13). The method has been used widely for simulating financial forecasts, but has also been suggested for estimations of PMI (14), and to describe variations in estimates (11). Monte Carlo simulation utilizes the mean and standard deviation of input variables to provide the unknown variability in a system. The data are then modeled to fit a normal distribution. The data were fitted to an output model, which was used to provide an initial iteration of the model and then replicated multiple times until a sufficient number of iterations had been performed.

Temperatures obtained from the three locations in this study were used as the input data for the model. From these temperatures, a mean and standard deviation was obtained for each. The

TABLE 1—*Dipteran taxa and their associated base temperatures as determined by previous research.*

Species	Base Temperature (°C)	Reference
Calliphora vicina	1.00	Donovan *et al.* (22)
Chrysomya albiceps	11.29	Salimi *et al.* (21)
Chrysomya chloropyga	11.44	Richards *et al.* (20)
Lucilia sericata	8.83	Roe and Higley (19)
Musca domestica	11.98	Wang *et al.* (18)
Piophila casei	7.29	Russo *et al.* (23)

For the purposes of the Monte Carlo simulation, an average base temperature was used to provide an estimation irrespective the of the larval instar stage.

mean and standard deviation was fitted to a normal distribution using the NORM.INV function in Microsoft Excel 2016 for the outdoor temperatures and the autopsy room temperatures. The NORM.INV function provides a normal distribution for a specified mean and standard deviation. The mean and standard deviations used for the outdoor temperatures and autopsy room temperatures were those previously recorded over the 16-month period using the iButton® data logger.

The refrigerator temperatures were fitted to a uniform distribution due to the minimum and maximum temperature set on the refrigeration unit. The model variables to describe the variability in the output were obtained using the RAND() function (Microsoft Excel 2016) to provide a random probability for the uniform distribution fit, in conjunction with the input mean and standard deviation of the refrigerator. The refrigerator uniform distribution was modeled using the function: minimum temperature + (RAND() × maximum temperature) to obtain a random value within the expected range of the refrigerator. The minimum and maximum temperatures were those recorded over the 16-month period from actual mortuary refrigerator temperatures.

The remaining model variables were those required to estimate PMI using thermal summation from insect development and the base temperature. From these input variables, a simulated output was obtained as the first iteration, which was repeated for 1000 iterations to obtain a comprehensive number of simulation replicates.

From the 1000 iterations, proportions of likelihood estimates were obtained for each species based on no development occurring (i.e., negligible larval development over the three day period) versus development occurring (i.e., development occurring between the time of removal from the scene and subsequent collection after autopsy).

Data Analysis

Monte Carlo simulation and likelihood of development measures were conducted using Microsoft Excel© 2006. SAS Enterprise Guide 9.1 (SAS institute, Cary, NC) was used for all statistical analyses. For the purposes of this study, a 95% level of significance ($\alpha = 0.05$) was used. Likelihood of development was calculated based on the 1000 iterations for each species. Iterations which were found to be zero, where the simulated temperature was below the base temperature of that species, were classified as "no development." Iterations greater than zero where temperatures were greater than the base temperature were classified as "continued development." The final likelihood of development was calculated as a proportion of each continued development and no development for the 1000 iterations simulated for each species. Likelihood of development measures was compared using Pearson's correlation using the base temperature of each species.

Results

The Monte Carlo simulation data obtained are summarized in Table 2 for each of the six dipteran species when insect samples are only collected after the completion of the autopsy. *Calliphora vicina* had the highest mean ADD and maximum ADD of all the species simulated. The minimum values were limited to zero to indicate development ceasing when temperatures were below the base temperatures indicated in Table 1.

The distributions of the ADD for all six species were found to be right skewed. The skewness obtained is an effect of limiting all the ADD minimum values to zero, in addition to the effects of the simulated uniform distribution of the refrigerator temperatures. The percentile values presented in Table 2 provide a more accurate representation of the actual distribution by focusing on the proportion of values rather than the actual value itself. The results presented in Table 2 indicate the expected variation in ADD for each species based on the modeled temperature and times, from the time the body is removed from the scene until the collection of insects following refrigeration and autopsy.

The effect of refrigeration alone was modeled separately to provide a better understanding of the effect on development of each species based on the variation in refrigerator temperature above and below the base temperatures of each species (Table 1). Table 3 provides a summary of the effect on development.

Development during refrigeration varied between the six species. Development likelihood values were found to be negatively correlated with the base temperatures of the species (Pearson Correlation = −0.99, $p = 0.0001$), where species with higher base temperatures (*Ch. albiceps, Ch. chloropyga* and *M. domestica*) were less likely to continue development than those with lower base temperatures (*L. sericata* and *P. casei*).

Discussion

This study provides a simulation of the uncertainty present when estimating the PMI using the thermal summation method from actual temperature variations recorded over 16 months at a South African mortuary. Prior to the autopsy, all bodies that are received at the mortuary are usually placed into refrigerators to prevent further decomposition before examination by the forensic medical practitioners. This has two obvious effects: Firstly, it slows the rate of decomposition ensuring a more productive autopsy, and secondly it slows the rate of development of any insects present on the remains (15).

A number of studies have analyzed the effect of refrigeration on maggot development (6,9) and survival following rapid cooling (16,17). This study expands on the potential effects of refrigeration through simulating the effect that refrigeration will have on PMI estimation and ADD measures. The simulated time period in this study provides a measure of uncertainty for ADD during transportation, storage, and finally the autopsy when insects are collected from the human remains. The ADD values for each species were relatively small, but the effect can result in an increased ADD estimation if thermal summation is used to estimate PMI. This would result in incorrect PMI estimations and potentially misleading time of death estimations.

The skewed distributions of this study were caused by the uniform distribution of refrigeration temperatures used in the simulation and the limiting of the ADD to zero. This further reiterates the warnings of Higley and Haskell (11), and Villet *et al.* (10) of the potential issues of estimating PMI when temperatures are close to the lower threshold of insect development.

Johl and Anderson (6) studied the effects of 24-h storage on the development *Calliphora vicina* where the simulated morgue refrigerator was set to 3°C over a 24-h period. The authors found variations in development of the larvae when reared after removal from the refrigerator, as a result of the different ages of the larvae. The findings of Johl and Anderson (6) suggest that the simulation of the present study may still not represent the complete variation that may occur when larval age is also considered.

TABLE 2—*Summary statistics of ADD for six forensically important Dipteran species from 1000 iterations of Monte Carlo simulations.*

						Percentiles*							95% Confidence Limits*	
Species	N	Mean*	Std Dev*	Min*	Max*	5th	10th	Lower Quartile	Median	Upper Quartile	90th	95th	Lower	Upper
Ca. vicina	1000	13.37	7.42	0.00	37.76	1.80	3.88	7.44	13.15	18.50	23.29	25.98	12.91	13.83
Ch. albiceps	1000	11.64	7.07	0.00	35.12	0.46	2.26	5.86	11.60	16.82	21.32	23.53	11.20	12.08
Ch. chloropyga	1000	11.45	6.79	0.00	30.10	0.91	2.64	5.75	11.34	16.66	20.38	22.81	11.03	11.87
L. sericata	1000	11.75	7.01	0.00	35.18	1.12	2.66	6.23	11.18	16.79	21.29	23.98	11.32	12.19
M. domestica	1000	11.61	6.85	0.00	31.61	1.09	2.59	6.12	11.40	16.66	20.81	23.31	11.19	12.04
P. casei	1000	12.21	7.08	0.00	36.71	1.42	2.81	6.58	12.05	17.26	21.66	24.64	11.77	12.64

Ca., *Calliphora*; *Ch.*, *Chrysomya*; *L.*, *Lucilia*; *M.*, *Musca*; Max, Maximum; Min, Minimum; *P.*, *Piophila*; SD, standard deviation.

*All measures are in ADD. Each percentile represents the ADD value at each corresponding point in the distribution (i.e., the 90th percentile represents the ADD value that is greater than 90% of the other ADD values in the distribution).

TABLE 3—*The simulated likelihood of development on six species based on a uniform distribution of a refrigerator set to 4°C, modeled using Monte Carlo simulation.*

Species	Likelihood no Development	Likelihood Continued Development
Calliphora vicina	24.00%	76.00%
Chrysomya albiceps	71.60%	28.40%
Chrysomya chloropyga	70.80%	29.20%
Lucilia sericata	58.10%	41.90%
Musca domestica	71.20%	28.80%
Piophila casei	56.80%	43.20%

The simulation model presented in this study highlights the importance of temperature and period of refrigeration on PMI estimation from specific fly species. The species used in this study include the most common taxa sampled from human cadavers and those which have been studied extensively for PMI estimations (8,18–23). Based on the findings of the present study, the temperature variations during refrigeration will result in continued development of fly larvae in a large proportion of cases. This highlights the importance for forensic entomologists to carefully consider the insect species present and the species-specific base temperatures especially following cooling of the body. Species with low base temperatures, such as *Ca. vicina* (22), *P. casei* (23) and *L. sericata* (19), are more likely to provide problems in PMI estimation using ADD following refrigeration as their continued development may skew the PMI estimation if refrigerator temperatures fluctuate too greatly. Walker *et al*. (24), recommends that the temperatures within refrigerators must be evaluated and if significant variations are present appropriate measures must be taken in the analysis of insects taken from human remains. This is an area for concern though, as it requires death-scene technicians and/or mortuary staff to ensure they are recording these temperatures and that the temperatures are not merely assumed to be that of the set temperature of the refrigerator. The findings of the present study agree with the suggestions of Higley and Haskell (11) that there is a need to clearly understand the variations present in PMI estimations.

Variations do exist in mortuary refrigerators, as found in this study, with temperatures ranging from −4.44°C to 22.32°C due to mechanical failures in the refrigerator. This failure in the refrigerator did increase the variability above and beyond that of a "normal" scenario but for the purposes of this study it presented a perfect "worst-case" scenario which could be used to test high levels of variation using Monte Carlo simulation. There is a need to understand what factors cause these variations or at least in ensuring that they are adequately recorded to aid in the accuracy of PMI estimations (24). Lack of adequate death-scene temperature information and data collection can also affect the estimation of the PMI (25). The results of the present study provide a good baseline of potential "worst-case" temperature variations using the mean and standard deviation to provide a range in the ADD estimation. While many authors have investigated the accuracy and precision of PMI estimation (10,26), the accuracy of the actual data collected is important. The results of this study suggest that when there is continued likelihood of development, those situations call for the use of careful consideration of ADD estimation depending on the temperatures to which the insect species have been exposed. The mean values obtained in the present study range between 11.45 and 13.37 ADD which highlight that ADD estimates may be underestimated if these factors are not considered. This variation may only result in underestimates in PMI estimations by several hours, or in worst-case scenarios days, depending on environmental temperatures and ADD measures at death scenes. Death-scene investigators and mortuary officials can aid in collecting these data. Failing this, there is a need to implement new measures such as data loggers on mortuary vehicles to monitor external temperatures and data loggers within body bags (when used) to monitor temperature conditions to which the body is exposed.

Conclusions

Insect development on bodies may be halted by refrigeration, but the effect this has on subsequent development and estimation of the PMI using thermal summation methods needs to be carefully considered (6,9). The findings of the present study suggest that, while more research to increase the accuracy of PMI estimations is needed, better management procedures of those working with the dead and the data they collect may be just as important. This study did not look at the effect that the number of bodies or human remains may have had on the temperature of the refrigerator as this effect was not considered until after the data were being analyzed. It is therefore suggested that future research observe the number of bodies present within the refrigerator and the effect this may have on temperature fluctuations. The period during which the refrigerator doors were open during movement of bodies into and out of the refrigerator was also unknown. Consideration also needs to be given to different insect species, presence of body bags, presence of maggot masses, stage of larval development when placed into the refrigerator, period within the refrigerator, and accuracy of refrigerator temperature gauges.

References

1. Amendt J, Campobasso CP, Gaudry E, Reiter C, LeBlanc HN, Hall MJ. Best practice in forensic entomology – standards and guidelines. Int J Legal Med 2007;121(2):90–104. https://doi.org/10.1007/s00414-006-0086-x.
2. Harvey ML, Gasz NE, Voss SC. Entomology-based methods for estimation of post-mortem interval. Res Rep Forensic Med Sci 2016;6:1–9. https://doi.org/10.2147/RRFMS.S68867.
3. Gennard DE. Forensic entomology: an introduction. West Sussex, U.K.: John Wiley & Sons Ltd, 2007;115–25.
4. Goff ML. Estimation of post-mortem interval using arthropod development and successional patterns. Forensic Sci Rev 1993;5(2):81–94.
5. Sharma R, Garg RK, Gaur JR. Various methods for the estimation of the post mortem interval from Calliphoridae: a review. Egypt J Forensic Sci 2015;5(1):1–12. https://doi.org/10.1016/j.ejfs.2013.04.002.
6. Johl HK, Anderson GS. Effects of refrigeration on development of the blow fly, *Calliphora vicina* (Diptera: Calliphoridae) and their relationship to time of death. J Entomol Soc Brit Columbia 1996;93:93–8.
7. Huntington TE, Higley LG, Baxendale FP. Maggot development during morgue storage and its effect on estimating the post-mortem interval. J Forensic Sci 2007;52(2):453–8. https://doi.org/10.1111/j.1556-4029.2007.00385.x.
8. Thevan K, Ahmad AH, Rawi CS, Singh BP. Growth of *Chrysomya megacephala* (Fabricus) maggots in a morgue cooler. J Forensic Sci 2010;55(6):1656–8. https://doi.org/10.1111/j.1556-4029.2010.01485.x.
9. Myskowiak JB, Doums C. Effects of refrigeration on the biometry and development of *Protophormia terraenovae* (Robineau–Desvoidy) (Diptera: Calliphoridae) and its consequences in estimating post-mortem interval in forensic investigations. Forensic Sci Int 2002;125:254–61. https://doi.org/10.1016/S0379-0738(02)00003-8.
10. Villet MH, Richards CS, Midgley JM. Contemporary precision, bias and accuracy of minimum post-mortem intervals estimated using development of carrion-feeding insects. In: Amendt J, Campobasso CP, Goff ML, Grassberger M, editors. Current concepts in forensic entomology. New York, NY: Springer, 2010;109–37.
11. Higley LG, Haskell NH. Insect development and forensic entomology. In: Byrd JH, Castner JL, editors. Forensic entomology: the utility of arthropods in legal investigations, 2nd edn. Boca Raton, FL: CRC Press, 2010;389–406.
12. South African Weather Service. Weather data January 2012 to July 2013. Pretoria, South Africa: South African Weather Service, 2013.
13. Fedra K. Environmental modeling under uncertainty: Monte Carlo simulation. Laxenburg, Austria: International Institute for Applied Systems Analysis, 1983;1–84.
14. Reibe S, Doetinchem PV, Madea B. A new simulation-based model for calculating post-mortem intervals using developmental data for *Lucilia sericata* (Dipt.: Calliphoridae). Parasitol Res 2010;107(1):9–16. https://doi.org/10.1007/s00436-010-1879-x
15. Goff ML. Early post-mortem changes and stages if decomposition in exposed cadavers. Exp Appl Acarol 2009;49(1–2):21–36. https://doi.org/10.1007/s10493-009-9284-9.
16. Coulson SJ, Bale JS. Characterization and limitations of the rapid cold hardening response in the house fly *Musca domestica* (Diptera: Muscidae). J Insect Physiol 1990;36(3):207–11. https://doi.org/10.1016/0022-1910(90)90124-X.
17. Chen C, Denlinger DL, Lee RE. Cold-shock injury and rapid cold hardening in the flesh fly *Sarcophaga crassipalis*. Physiol Zool 1987;60 (3):297–304. https://doi.org/10.1111/j.1365-3032.1996.tb00866.x.
18. Wang Y, Yang L, Zhang Y, Tao L, Wang J. Development of *Musca domestica* at constant temperatures and the first case report of its application for estimating the minimum postmortem interval. Forensic Sci Int 2018;285:172–80. https://doi.org/10.1016/j.forsciint.2018.02.004.
19. Roe A, Higley LG. Development modeling of *Lucilia sericata* (Diptera: Calliphoridae). PeerJ 2015;3:e803. https://doi.org/10.7717/peerj.803.
20. Richards CS, Crous KL, Villet MH. Models of development for blowfly sister species *Chrysomya chloropyga* and *Chrysomya putoria*. Med Vet Entomol 2009;23(1):56–61. https://doi.org/10.1111/j.1365-2915.2008.00767.x.
21. Salimi M, Rassi Y, Oshaghi M, Chatrabgoun O, Limoee M, Rafizadeh S. Temperature requirements for the growth of immature stages of blowflies species, *Chrysomya albiceps* and *Calliphora vicina*, (Diptera:Calliphoridae) under laboratory conditions. Egypt J Forensic Sci 2018;8 (1):28. https://doi.org/10.1186/s41935-018-0060-z
22. Donovan SE, Hall MJ, Turner BD, Moncrieff CB. Larval growth rates of the blowfly, *Calliphora vicina*, over a range of temperatures. Med Vet Entomol 2006;20(1):106–14. https://doi.org/10.1111/j.1365-2915.2006.00600.x.
23. Russo A, Cocuzza GE, Vasta MC, Simola M, Virone G. Life fertility tables of *Piophila casei* L. (Diptera: Piophilidae) reared at five different temperatures. Environ Entomol 2006;35(2):194–200. https://doi.org/10.1603/0046-225X-35.2.194.
24. Walker MW, Butler JM, Higdon HE, Boone WR. Temperature variations within and between incubators – a prospective, observational study. J Assist Reprod Genet 2013;30(12):1583–5. https://doi.org/10.1007/s10815-013-0104-0.
25. Hofer IM, Hart AJ, Martin-Vega D, Hall MJ. Optimising crime scene temperature collection for forensic entomology casework. Forensic Sci Int 2017;270:129–38. https://doi.org/10.1016/j.forsciint.2016.11.019.
26. Dabbs GR. Caution! All data are not created equal: the hazards of using National Weather Service data for calculating accumulated degree days. Forensic Sci Int 2010;202(1):e49–52. https://doi.org/10.1016/j.forsciint.2010.02.024.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14505
Available online at: onlinelibrary.wiley.com

CASE REPORT

CRIMINALISTICS

Nir S. Finkelstein,[1] *B.Sc., M.A.; Hila Rosengarten,*[1,†] *M.Sc.; and Ophir Levy,*[1,†] *Ph.D.*

Photographic Image Comparison of Vegetation

ABSTRACT: Images and videos are common types of evidence in crime scene investigations and laboratory analysis. Images may be taken by the suspect and/or by crime scene investigators and may serve as crucial elements in forensic laboratory analysis. Forensic photographic image comparison is the process of comparing one or more objects or persons in photographs/images when at least one image is known to be related to a crime. The forensic examiner usually compares the images in order to determine whether or not an association between the exhibits in the images can be made. This paper proposes an extension to the currently prevalent photographic image examination method. The extension introduces comparison of landscape and vegetation over time. It is revealed that similarities between images may still be found between the period of time the suspect photograph was taken and the period the crime scene investigator took the photograph from the same area. In this case report, two suspects to be involved in growing a marijuana field were arrested by the police. The forensic experts were asked to examine images taken by the crime scene investigators and to compare them to the images found in the suspects' phones. They then tried to determine whether the suspects could be linked to the specific locations. While applying techniques derived from morphological comparison methodologies, the plants at the scene provided significant additional information. A tree trunk, branches, and twigs on a hedge in the photographs revealed specific individual characteristics that led the examiner to reach a conclusive decision.

KEYWORDS: criminalistics, forensics, landscape, photographic, image comparison, image analysis, vegetation

Forensic photographic image (the terms "photograph" and "image" will hereinafter be used interchangeably) comparison is the process of comparing images of one or more objects or persons, when at least one image is known to be related to a crime (i.e., a *suspect image*). The comparison may be conducted on virtually any item, subject, or image (1). There are several methodologies for comparison. Among others, there are photogrammetric, superimposition, or morphological methodologies. In the morphological methodology, features of the object for comparison are identified, are classified, and then comparison can be made (2). The final result of the comparison should be an expert report regarding identification or elimination of the person/object (3). Common methods for identification include facial image comparison or object comparison. Whether faces or objects are to be compared, the method of comparison remains very similar: A reference image ("known" image) is compared to the suspect image. Regarding object comparison, the objects that are most commonly compared are clothing and vehicles. The photographic images are obtained by many types of digital cameras such as phone cameras, surveillance car cameras, and street cameras (4). Images of the same scene may be acquired at different times and possibly under different conditions. Therefore, reference recordings are made under controlled conditions that are similar to those of the actual recordings in order to avoid differences introduced by the recording device (5) and/or scene conditions.

The objective of photographic image comparison is to recognize the features, to find similar characteristic marks and components in the image, and to extract intended information for comparison and/or analysis (6). A major factor toward a successful examination and comparison is the quality of the images; the image must be of sufficient quality (7). If the suspect image is of insufficient quality to distinguish between characteristic marks, then it may be impossible to reach a conclusion. Another major factor is the potential of a given object to exhibit temporary or permanent changes in its characteristics, such as changes in shape, or changes over time (7). These properties make the comparison even more challenging.

The process of comparison is as follows: Stage 1 is examination of the suspect image as is and determination of the level of details in the image. In stage 2, artifacts and visible features (4) will be sought for, characterized, and classified into categories. In stage 3, a comparison is made to the reference image.

In general, the criteria for comparison are divided into two categories: class characteristics and individual characteristics:

A class characteristic—Measurable features of a specimen that indicate a restricted group source. The features result from design factors and therefore are determined prior to manufacture. An individual characteristic—An object or person that has a special mark produced by random imperfections or irregularities (8). Specifically, with regard to objects, these random imperfections or irregularities are produced incidentally to manufacture or are caused by use, wear, and damage. They are unique to that specific object and thus distinguish it from all other objects.

Traditionally, in "classic" photographic image comparison, forensic experts look for class and individual characteristics in objects that are part of an image. The number of such

[1] Toolmark and Materials Laboratory, Division of Identification and Forensic Science, Israel Police, Jerusalem, 91906, Israel.

Corresponding author: Nir S. Finkelstein, B.Sc., M.A. E-mail: nirf7@walla.com

[†] Authors contributed equally.

Received 2 April 2020; and in revised form 15 June 2020; accepted 16 June 2020.

characteristics necessary for identification is case-dependent. The value of each characteristic will depend on its quality, rarity, pattern, and construction, combined with the presence of random features (9). For example, a single piece of class characteristic evidence can rarely be used to make a conclusion. When multiple types of class characteristic evidence associate one suspect with the crime, the weight of that evidence becomes stronger (10). However, an individual characteristic, namely a unique mark, will always outweigh a class characteristic when comparing their value as forensic evidence.

Various types of physical evidence/marks can be found at almost any crime scene. These marks can be produced by a vast number of objects (9). Each of these pieces of physical evidence/marks is a valuable exhibit which is a source of information about spatial relationships between objects, suspects, and events.

In this current case report, no comparison was made between the classical objects, that is, clothing or vehicles, but by comparison of plants. Plants naturally grow and change over time, and even the rate of change may differ. The main object that was chosen to be compared was the tree in the middle of the images. In addition, the unique landscape was compared.

Trees are characterized by their long and thick trunks. The height of a dominant tree species and the shape of its crown determine the overall landscape of many terrestrial plant communities (11). Plants continuously face changing environmental conditions, and therefore, it is conferred upon them to adapt their morphology, physiology, and metabolism to survive (12). Numerous studies have shown that leaf morphology (13), and tree crown shape and height (11) are influenced by environmental factors such as temperature, species of trees in their surroundings, and light. All of these factors may influence the individual characteristics on the tree itself and on the surroundings, and can be used by the forensic expert. This case shows the efficacy of using vegetation as an object. The olive tree life cycle is governed by a development cycle bound to the seasons of the year, such as changes in the crown and in the combination of leaves and branches. Therefore, other tree parts that almost *do not* change over time will serve as unique characteristics of it. Here, because of the uniqueness of the tree trunk and the surrounding plants, they were chosen to constitute the main parameters for comparison.

Case Details

About 500 kg of marijuana plants were found in a field. Two suspects were caught near the crime scene. The suspects claimed that they had just come from elsewhere and denied any connection to the crime scene. Examination of the suspects' phones uncovered images of landscape that resembled the crime scene (Fig 1). The images contained only the marijuana plants without any facial images or tools. The Toolmark and Materials Laboratory team were given the task to compare images found in one of the suspects' phones (Fig 1) and those taken as reference by the crime scene officers (Fig 2). The laboratory team was tasked to conclude whether the images found in the phone and those taken by the crime scene officers relate to the same crime scene.

Suspect images were made use of to determine the lighting conditions and camera position. Very similar lighting and camera position were applied when taking images by the crime scene officers. Reference images of the most expressed similarity to the images found in the phone were chosen to continue with toward an in-depth examination.

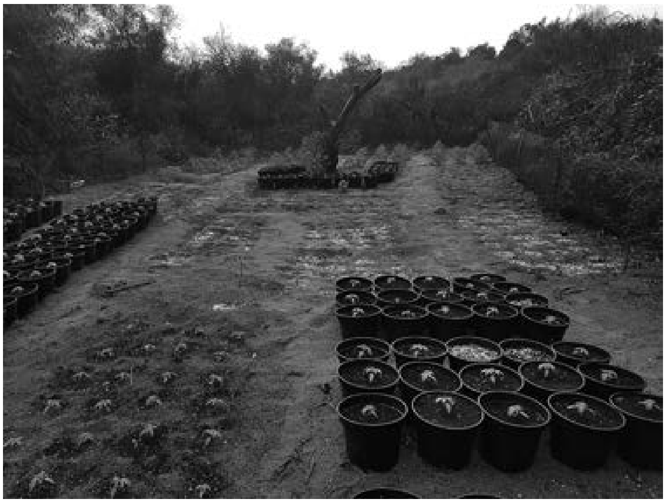


FIG. 1—*Scene from a suspect's phone [Color figure can be viewed at wileyonlinelibrary.com]*

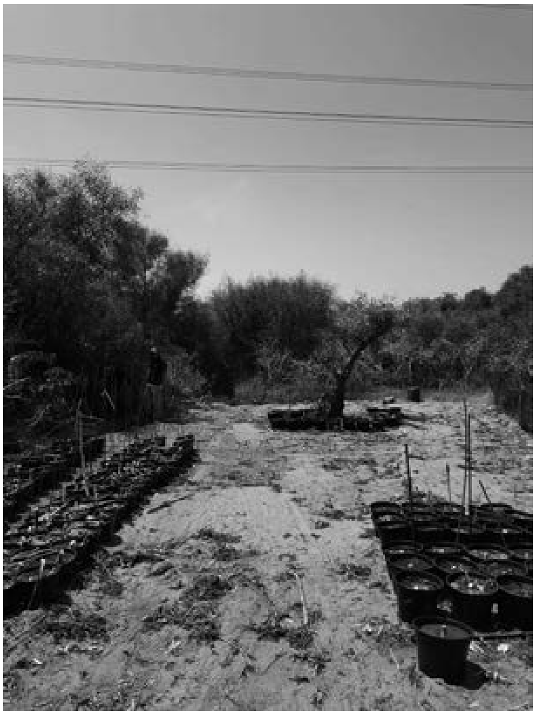


FIG. 2—*Crime scene photographs [Color figure can be viewed at wileyonlinelibrary.com]*

An olive tree in the middle of the image, surrounded by buckets and vegetation, was chosen to be the main objects for comparison. Examination of the suspect image features both class and individual characteristics. First, the buckets will be addressed. The number of buckets, shape, size, color, and the arrangement among them are all categories of marks that may be used for comparison. As for the number, shape, size, and color—these are class characteristic marks. Such marks cannot assist in deducing individual characteristics. So, at best, the buckets here serve as class characteristic evidence. Additionally, buckets are mobile objects. This fact further weakens the value of an already nonindividual mark evidence.

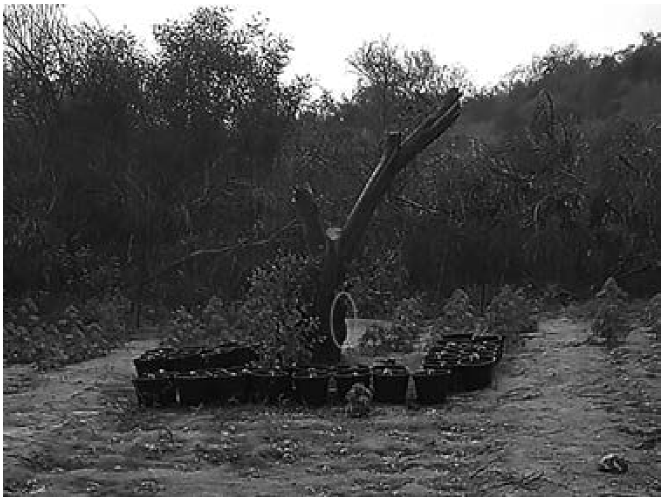


FIG. 3—*Magnification of Fig. 1. The unique shape of the trunk is indicated by a red circle [Color figure can be viewed at wileyonlinelibrary.com]*

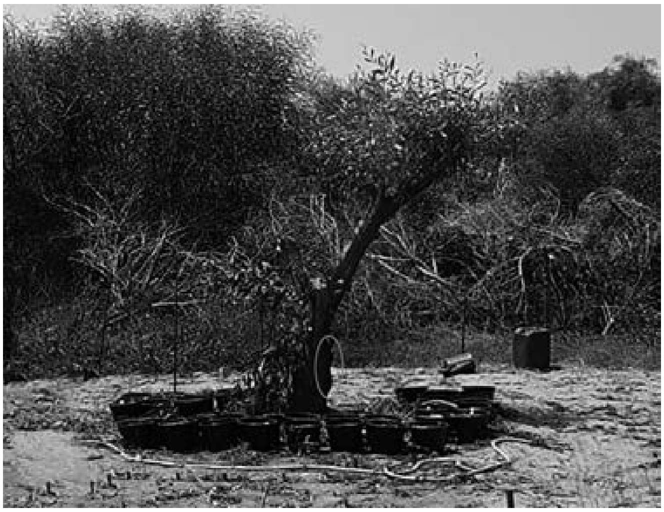


FIG. 4—*Magnification of Fig. 2. The unique shape of the trunk is indicated by a red circle [Color figure can be viewed at wileyonlinelibrary.com]*

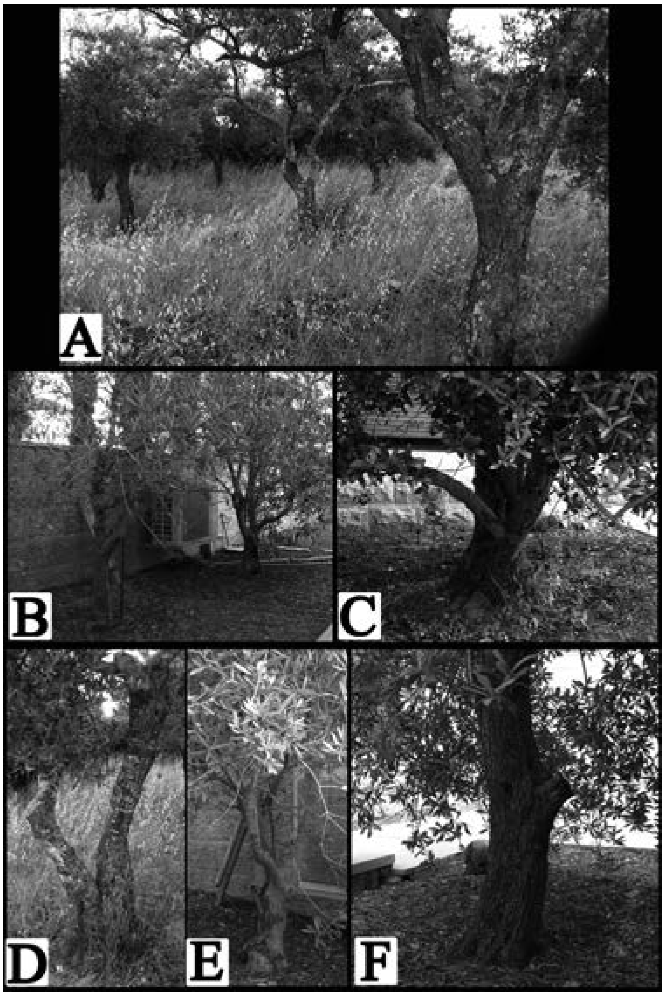


FIG. 5—*Eleven exemplars of olive tree trunks. Note that the photographs are not related to the crime scene; they are merely random photographs, taken in order to demonstrate the unusual shape and features of the tree trunk. (A and B) Inspection of several tree trunks in the same photograph demonstrates the unusual nature and variety of trunk shapes. (C–F) The unusual and individual features of single tree trunks may be observed [Color figure can be viewed at wileyonlinelibrary.com]*

In this particular case, in addition to buckets, more objects can be observed in the scene, such as fences and fence posts. As still object comparison in photographs is known to be routinely applied in analyzing crime scenes, this case report focuses on objects of vegetation. As opposed to buckets, objects of vegetation in this image may well serve as individual characteristics. The following discussion will focus on three objects of vegetation in the image, namely the single tree in upper-middle parts of Figures 1 and 2, branches of the hedge behind it, and a line of tree crown tops behind the hedge. Zooming in, an individual characteristic shape at the bottom of the single tree trunk can be observed (Figs 3 and 4). This shape resembles two humps or arches facing rightwards, like the right-hand side of the digit 3 (encircled in red, Figs 3 and 4). In order to demonstrate that the trunk shape feature is an individual characteristic in the photograph, random olive tree trunks were photographed. In Fig. 5, the unusual nature of the shapes, as well as their variety, can be observed. In zoom-in views, random features on single trunks can be seen.

In addition to the tree trunks, there are two more objects of vegetation that possess features of interest. The tree crown tops behind the hedge form a unique wave-like pattern. This pattern is formed by a number of trees (Fig. 6(*A*)). The branch and twigs on the hedge in the middle left-hand side of the images in Figs 6(B) and 7 feature another individual characteristic mark. A hedge branch that grows groundward features at least five more twigs that their general growing direction is left (encircled in red). The twigs are characterized by their position on the main branch, directionality, and approximate angles with respect to the main branch.

The unique match combinations lead to the conclusion that the suspect crime scene image and the crime scene image that was taken by the officers practically refer to the same location. The individual characteristics recognized by the forensic expert are consistent between the suspect images and the reference images. This conclusion links the suspects to the crime scene. The conclusion can be made use of, alongside, additional data from global positioning systems (GPS) and/or other phone

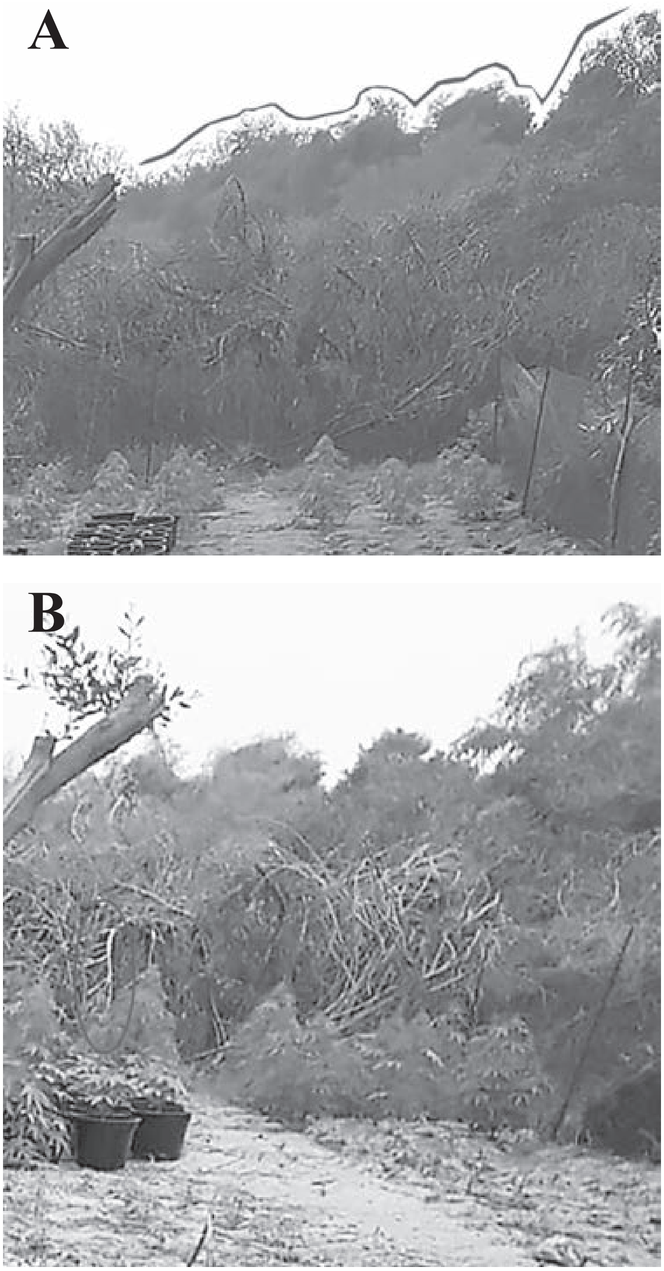


FIG. 6—*(A) Scene from a suspect's phone. The pattern that is formed by the tree top crowns is presented by a red curved line. (B) Scene from a suspect's phone. The unique branch and its twigs are indicated by a red circle [Color figure can be viewed at wileyonlinelibrary.com]*

tracking capabilities, if available, to link the suspects to the crime scene.

The performed image comparison in this case report showed the efficacy of using vegetation in photographic comparison. In this case, the comparison of objects other than vegetation-related was not sufficient to deduce identification. Growing plants and other vegetation objects eventually lead to a match. The fine details in the images shown here, especially the individual characteristics in the tree trunk, the tree crown tops, and the special hedge branch were sufficient to establish a conclusive conclusion.

Traditionally, it is common practice to perform photographic image comparisons on people's faces, still objects, or vehicle images. The comparison done here is based on other photographic features, ones that change in time. Identifying and isolating individual characteristics that *do not* change over time made this kind of comparison feasible. For forensic sciences, this method of comparison can further be utilized in hard-to-solve cases, which mainly contain vegetation.

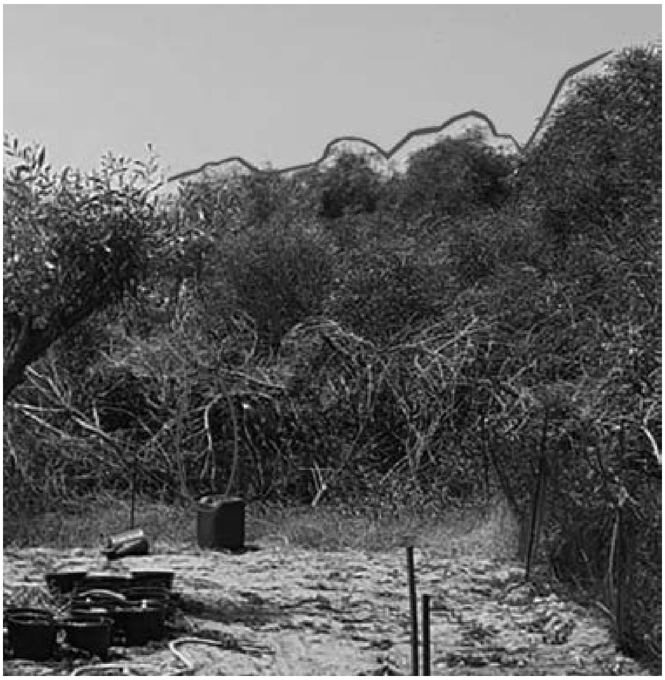


FIG. 7—*Crime scene photographs. The unique branch and its twigs are indicated by a red circle. The pattern that is formed by the tree top crowns is presented by a red curved line [Color figure can be viewed at wileyonline library.com]*

Acknowledgments

The authors would like to thank A. Chaikovsky, A. Silchenko, and A. Levi for their assistance with preparation of the artwork in this publication.

References

1. SWGDE – Scientific Working Group on Digital Evidence – SWGDE Current Documents. SWGDE technical overview for forensic image comparison. 2019. https://www.swgde.org/documents/Current%20Documents/SWGDE%20Technical%20Overview%20for%20Forensic%20Image%20Comparison (accessed March 26, 2020).
2. Facial Identification Scientific Working Group – FISWG Documents. Facial comparison overview and methodology guidelines, 2019. https://fiswg.org/fiswg_facial_comparison_overview_and_methodology_guidelines_V1.0_20191025.pdf (accessed March 26, 2020).
3. International Association for Identification – Forensic Photography & Imaging Certification. SWGIT guidelines for the forensic imaging practitioner, 2016. https://theiai.org/docs/SWGIT_Guidelines.pdf (accessed March 26, 2020).
4. Forensic image comparison and interpretation evidence: guidance for prosecutors and investigators. Issue: 2. National Crime Agency (NCA), Metropolitan Police, the Crown Prosecution Service (CPS), and Forensic Science Regulator, 2016. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/511168/Image_Comparison_and_Interpretation_Guidance_Issue_2.pdf (accessed June 16, 2020).
5. Verlome E, Mieremet A. Application of forensic image analysis in accident investigations. Forensic Sci Int 2017;278:137–47. https://doi.org/10.1016/j.fordviint.2017.06.039.
6. Hanji RB, Rajpurohit VS. Forensic image analysis – a frame work. Int J Forensic Comput Sci 2013;1:13–9. https://doi.org/10.5769/J201301002

7. Vorder Bruegge RW. Photographic identification of denim trousers from bank surveillance film. J Forensic Sci 1999;44(3):613–22. https://doi.org/10.1520/JFS14519J.
8. AFTE Standardization Committee. Glossary of the Association of firearm and tool mark examiners. 3rd ed. Chicago, IL: Available Business Printing, 1994;36,60.
9. Luong S, Roux C. Marks or impressions of manufactured items. In: Jamieson A, Moenssens A, editors. Wiley encyclopedia of forensic science. Hoboken, NJ: John Wiley & Sons, 2009.
10. Firearm Examiner Training. Class and individual characteristics. https://projects.nfstc.org/firearms/module06/fir_m06_t04_05.htm (accessed March 26, 2020).
11. Iwasa Y, Cohen D, Leon JA. Tree height and crown shape, as results of competitive games. J Theor Biol 1985;112(2):279–97. https://doi.org/10.1016/S0022-5193(85)80288-5.
12. Avin-Wittenberg T. Autophagy and its role in plant abiotic stress management. Plant, Cell Environ 2019;42(3):1045–53. https://doi.org/10.1111/pce.13404.
13. Li X, Li Y, Zhang Z, Li X. Influences of environmental factors on leaf morphology of Chinese jujubes. PLoS One 2015;10(5):e0127825. https://doi.org/10.1371/journal.pone.0127825.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14500
Available online at: onlinelibrary.wiley.com

CASE REPORT

PATHOLOGY/BIOLOGY

Luca Doro,[1] M.D.; Barbara Bonvicini,[1] M.D.; Elena Beccegato,[1] M.D.; and Claudio Terranova,[1] A.P.

Lying on the Road Before Being Run Over: Vehicular Manslaughter, Suicide, or Accident? Two Case Reports and Literature Review

ABSTRACT: We present two apparent hit-and-run cases where two women were run over. The vehicles involved were subsequently traced and their owners charged with manslaughter. Autopsy evidence, scientific investigation of the scene and circumstances of the deaths, technical inspection of the vehicles, and DNA analysis strongly suggested that both victims were lying on the road before the accident. Case 1 was a suicide. In Case 2, the victim had fallen to the ground following acute alcohol intoxication. Victimological analysis was pivotal in reconstructing the dynamics of the events. We suggest that a hit-and-run fatality should not be regarded as a manslaughter case until each piece of evidence has been carefully considered. We also propose an interdisciplinary method of reconstructing run over occurrences based on the following three steps: (i) identify whether there was a primary impact when the victim was in an upright position; (ii) identify victim drug/alcohol intoxication and/or presence of acute or chronic disease or injury, which may have contributed to the impact; and (iii) consider suicide intent.

KEYWORDS: forensic pathology, autopsy, pedestrian fatalities, run over, hit-and-run, road accident reconstruction, psychological autopsy

In Italy, 612 pedestrians were killed in 2018 (1). In that year, 63 hit-and-run fatalities of pedestrians were reported (2). The majority of traffic accidents are caused by factors related to the people and vehicles, and only a few can be considered the result of chance (3). Therefore, each piece of evidence (i.e., autopsy evidence, scientific investigation of the scene and circumstances of the death, technical inspection of the vehicles, and DNA analysis) should be carefully taken into account in the reconstruction of the event in order to determine the responsibilities of the parties involved in the causation of the juridical event.

The forensic pathologist is frequently asked to provide an objective interpretation of the sequence of events that led to a collision and subsequently to a pedestrian death. In hit-and-run cases, the technical reconstruction can be particularly difficult when there is a lack of scene findings and/or witnesses (4).

Only a few studies have investigated the reasons why a person could be lying on the ground before being run over by a vehicle (5,6). The reason can have important implications for the driver's juridical responsibility. In the instances where the reasons have been reported, they predominantly were one of the following: (i) a previous accident and subsequent fall to the ground, (ii) falling to the ground because of acute sickness or intoxication, or (iii) deliberately lying on the road to cause self-harm.

We describe two cases of persons lying on the road before being run over in apparent hit-and-run incidents. Case 1 was a suicide of a woman with psychiatric disorders. Case 2 was of a woman with acute alcohol intoxication. The two cases are analyzed in the context of the three reasons mentioned above.

Case Reports

Case 1

At late night, the dead body of a young woman was found in the middle of the road in a supine position. There were no available witnesses on the scene. The police found plastic fragments traceable to a specific make of vehicle. The speed limit on the road was 70 km/h, and there was no evidence of skid marks on the road. A blood trail of one meter in length was detected. The event was registered as a pedestrian hit-and-run case. The victim, 27 years old, was subsequently identified. She lived near the accident scene. She suffered from bulimia and had been discharged the week before from a psychiatric clinic after a suicide attempt through ingestion of medications.

In the postmortem external examination, it was observed that: (i) her clothes had several dirt stains, but there was no laceration of the trousers, (ii) there was a wide brush abrasion with laceration of the auricle on the right side of the head and an 8 cm wide laceration wound surrounded by a reddish-colored abrasion on the left side of the forehead, (iii) there were tire tread marks on the chest and abdomen, and (iv) there were several circular and linear scars on the anterior surface of the left forearm, consistent with previous self-injury. The internal examination revealed brain stem transection, several fractures (cranium, sternum, right clavicle, multiple ribs, cervical spine, and right sacroiliac joint), many visceral lacerations, hemothorax, and hemoperitoneum. No deep injury to the lower limbs was observed. Histology revealed intraparenchymal hemorrhages associated with internal lesions, and toxicological analysis detected therapeutic concentrations of

[1] Legal Medicine and Toxicology, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padova, via G. Falloppio n.50, Padova, 35121, Italy.

Corresponding author: Luca Doro, M.D. E-mail: luca_doro@yahoo.com

Received 29 Mar. 2020; and in revised form 26 May 2020, 9 June 2020; accepted 10 June 2020.

benzodiazepines, antipsychotics, and antidepressants, all of which had been prescribed to her at the time of her discharge from the psychiatric clinic. Death was caused by significant blood loss and brainstem laceration.

The intercurrent investigations led to a small black van whose owner was charged with manslaughter. The man stated that he did not realize he had hit the victim. The vehicle was inspected by a court-appointed engineer and a forensic pathologist. The inspection revealed damage on the right inferior plastic surface of the anterior bumper. In the same area, there was also a dust-free spot that matched the aforementioned forehead laceration in size and shape (Fig. 1), with DNA evidence matching to the woman. Tire marks on the corpse precisely reflected the tread pattern of the right rear wheel of the van. The reconstruction of the accident excluded the woman being in a standing position at the time of impact. It was assumed that the victim was lying in a supine position on the road and raised her head enough to be hit by the right inferior surface of the anterior bumper (where the dust-free area was detected). Subsequently, the right side of her head hit the ground and skidded across the road surface producing the wide abrasion observed in that area. She was then run over. Based on the experts' evaluation, the manslaughter charge against the driver was dropped. The manner of death was determined to be suicide.

Case 2

On a foggy night, a woman was driving down a country road with speed limit 50 km/h. She saw something similar to a bundle on the ground. As she was getting out of her vehicle to inspect the object, she witnessed another car driving in the opposite direction that ran over the bundle and drove away without stopping. Immediately afterward, the witness realized that it was an apparently lifeless human body, and she alerted the authorities. The police found the body in the middle of the road and collected a plastic grille from the scene. There were no apparent signs of dragging for a distance. The ensuing investigations led to the identification of both the victim and the vehicle owner. The victim was a 55-year-old woman who lived nearby. The vehicle owner, who was charged with manslaughter, stated that he did not realize he had hit the victim. The victim's neighbors reported that she had consumed copious quantities of alcohol that night before leaving her flat on foot.

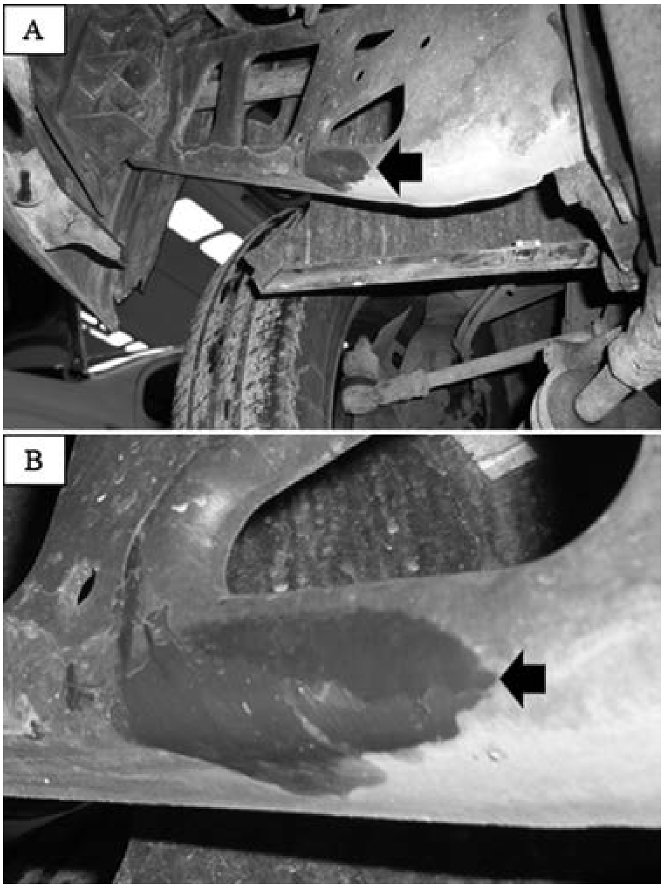


FIG. 1—*(A) Dust-free spot in the area of the anterior bumper (arrow). (B) Detail (arrow). [Color figure can be viewed at wileyonlinelibrary.com]*

External examination of the body principally revealed several dirt stains on the garments and a wide abrasion on the forehead and nasal regions. The internal examination found several fractures (sternum, multiple ribs, and pubic symphysis), many visceral lacerations, hemothorax, and hemoperitoneum. No significant deep injury was present on the lower limbs. The pathological findings were consistent with a crushing of the thorax. Histology revealed intraparenchymal hemorrhages associated with internal lesions, and toxicological analysis detected a high blood alcohol concentration (0.3 g/dL). Death was caused by significant blood loss secondary to polytrauma. Further evidence was provided by the technical inspection of the car, which showed that the lower front grille was missing. The reconstruction of the accident suggested that the victim was lying on the road in a supine position prior to the run over. She probably fell to the ground because of profound psychomotor impairment induced by acute alcohol intoxication. Based on the experts' evaluation, the manslaughter charge against the driver was dropped. The juridical event was considered an accident.

Discussion

We presented two hit-and-run cases with subsequent identification of the vehicles involved. From the perspective of reconstruction, the first question that needed to be answered was whether there had been a previous collision in the upright position before the victim was run over or the victim had been lying on the road since the beginning.

Both accidents occurred at night when visibility is low and traffic is less intense. In the autopsy, there was evidence of being run over (tire marks on the chest and abdomen in Case 1, and crush injuries to the thorax in Case 2) and absence of significant injuries on the lower limbs that would be indicative of the victim being hit in a standing position. Although avulsion pockets and extensive regions of "decollement" of the skin usually provide meaningful evidence of run over, we did not observe them (3). Because injuries to the lower limbs may not be evident if only an external examination is performed, it is recommended to perform the so-called peel-off procedure (exposure of soft tissues and musculature of the back and lower limbs) to reveal any deep bruises (bumper injuries) and/or fractures of the knee/diaphysis, such as bending fractures, which also support a hypothesis of impact on the side in the erect position (7). Karger et al. (2001) identified the following factors as being specific to a primary hit in a standing position: wedge-shaped fracture (Messerer fracture), typical glass fragment injuries, traumatic amputations, traces of car paint on the lower extremities, and abrasions of the shoe soles (5). However, because many of these findings do not occur regularly and can be a source of error if considered alone, a deeper analysis is advisable in uncertain cases (3). For example, Teresinìski et al. (2002) recommended cross sections of the knee epiphyses to reveal bone bruises as a marker of the limb load by body mass in upright hits (3,8). Bone bruises on sections of the greater trochanter and central fractures of the hip are also significant (9). The best results for reconstructive purposes are achieved by combining autopsy and postmortem imaging, as shown below. However, this combined method may not be possible because of limited economic resources (3,10,11).

In the presented cases, autopsy evidence was integrated with scene findings and a technical analysis of the traced vehicles. In Case 1, damage was concentrated on the right inferior surface of the anterior bumper, and in Case 2, only the lower front grille was missing from the vehicle. Thus, it was corroborated that the two victims were hit while in a lying position. Moreover, in Case 1, the positive DNA hit from the dust-free spot on the van (Fig. 1), which matched the size and shape of the laceration on the left side of the victim's forehead, indicated that the woman was lying in a supine position on the road and raised her head enough to be hit by the right inferior surface of the anterior bumper. In this regard, 3D technology led to a significant improvement in the quality and objectivity of the technical reconstruction of the accident. With photogrammetry, it is now possible to generate 3D models of bodies, clothes, and vehicles. The combination of optical surface scanning data with radiological data (postmortem multislice computed tomography and magnetic resonance imaging), along with the use of animation software, enables a better correlation between body injuries and vehicle damages (10). In recent years, the use of complex biomechanical models adopted from the vehicle safety improvement field, such as finite element models, has also been proposed for forensic reconstructive purposes (12).

Once it was ascertained that the primary impact with the victim occurred while she was lying on the ground, the second step pertained to possible acute sickness or intoxication of the woman in Case 2. A high blood alcohol concentration (0.3 g/dL) was detected. It was suspected that because of psychomotor impairment, the woman most likely fell to the ground and was subsequently run over. Alcohol and drugs are well-known risk factors for road traffic fatalities. Therapeutic medications have also been associated with an increased risk of pedestrians being involved in accidents. Because a high frequency of detection of psychoactive substances in pedestrian postmortem specimens has been reported, toxicological analysis should always be undertaken (13).

The fact that alcohol is also frequently involved in suicide cases leads us to discuss the third and last step with regard to Case 1. The young woman was at high risk for suicide—she was affected by an eating disorder (bulimia) and had a previously reported self-destruction attempt and a recent discharge from a psychiatric clinic. Additionally, there was autopsy evidence suggesting previous self-injury. All psychiatric diagnoses, including eating disorders, have been associated with suicide attempts and suicide with varying degrees of risk (14–16). Psychological autopsy studies have shown that more than 90% of people who commit suicide have a diagnosable psychopathological disorder (17,18). In particular, a critical period is the first few months after discharge from a psychiatric ward. The psychiatric diagnosis also includes a high probability of intercurrent psychopharmacological therapy and thus the potential eventuality of a psychomotor impairment leading to an accident. Toxicological analysis is therefore mandatory in these cases and should not be limited to alcohol and illegal drugs (19). In this specific case, toxicological evidence excluded overdoses and revealed therapeutic concentrations of medications. The overall evidence pointed to suicidal intent of the victim even if road traffic suicide is an uncommon method of self-destruction (around 2% of all road fatalities in Europe) (20,21). In fact, such events may be underestimated by up to 5% because they are often misinterpreted as accidents or remain as an open verdict (21–23). Among others, an important issue is that a positive alcohol and/or drug finding complicates the evaluation of the manner of death and the differentiation between suicide and crash under the influence of psychoactive substances. Alcohol is a general risk factor for road traffic fatalities, and its abuse is also a specific risk factor for suicide (21,24). To overcome these difficulties, Gauthier et al. (2015) suggested that suicide should be considered for each case, and a psychological autopsy conducted by an interdisciplinary team could help to clarify the manner of death by focusing on psychological/psychiatric aspects (21). More research is needed, particularly for pedestrian suicide, which is rare. Among a few other studies, Routley et al. (2003) found that a history of mental disease and alcohol abuse were the main factors involved in this peculiar method of self-destruction (20,21,23).

In summary, we highlighted that a hit-and-run fatality should not be regarded as a manslaughter case until each piece of evidence has been carefully considered. In both cases, the manslaughter charges against the two drivers were dropped. Victimological analysis was key in revealing the psychiatric background and acute alcohol intoxication in the presented fatalities. In cases of road users being run over, the proposed three-step reconstruction may help clarify the dynamics of the event, particularly when there is a lack of scene findings. The three steps can be summarized as follows: (i) identify whether there was a primary impact when the victim was in an upright position; (ii) identify victim drug/alcohol intoxication and/or presence of acute or chronic disease or injury, which may have contributed to the impact; and (iii) consider suicide intent.

References

1. ISTAT - Italian Institute of Statistics. Incidenti Stradali in Italia, comunicato stampa [Road traffic accidents in Italy, press release], 2018. https://www.istat.it/it/archivio/232366 (accessed February 3, 2020).
2. ASAPS - Association of Supporter and Friends of the Traffic Police. Pirateria stradale, report dell'osservatorio 2018 [Hit and run cases, 2018 report of the monitoring centre], 2019. https://www.asaps.it/ (accessed February 3, 2020).
3. Teresiński G, Madro R. Evidential value of injuries useful for reconstruction of the pedestrian-vehicle location at the moment of collision. Forensic Sci Int 2002;128(3):127–35. https://doi.org/10.1016/S0379-0738(02)00185-8.
4. MacLeod KE, Griswold JB, Arnold LS, Ragland DR. Factors associated with hit-and-run pedestrian fatalities and driver identification. Accid Anal Prev 2012;45:366–72. https://doi.org/10.1016/j.aap.2011.08.001.
5. Karger B, Teige K, Fuchs M, Brinkmann B. Was the pedestrian hit in an erect position before being run over? Forensic Sci Int 2001;119(2):217–20. https://doi.org/10.1016/S0379-0738(00)00430-8.
6. Brinkmann B, Schwarz G, Stichnoth E. Problems concerning pedestrians run over while in prone position] [Article in German. Arch Kriminol 1985;175(5–6):137–44.
7. Brinkmann B. Harmonisation of medico-legal autopsy rules. Int J Legal Med 1999;113(1):1–14. https://doi.org/10.1007/s004140050271.
8. Teresiński G, Madro R. Knee joint injuries as a reconstructive factors in car-to-pedestrian accidents. Forensic Sci Int 2001;124(1):74–82. https://doi.org/10.1016/S0379-0738(01)00569-2.
9. Teresiński G, Madro R. Pelvis and hip joint injuries as a reconstructive factors in car-to-pedestrian accidents. Forensic Sci Int 2001;124(1):68–73. https://doi.org/10.1016/S0379-0738(01)00567-9.
10. Buck U, Buße K, Campana L, Gummel F, Schyma C, Jackowski C. What happened before the run over? Morphometric 3D reconstruction.

Forensic Sci Int 2020;306:110059. https://doi.org/10.1016/j.forsciint.2019.110059.

11. Chatzaraki V, Thali MJ, Ampanozi G, Schweitzer W. Fatal road traffic vehicle collisions with pedestrian victims: forensic postmortem computed tomography and autopsy correlation. Am J Forensic Med Pathol 2018;39(2):130–40. https://doi.org/10.1097/PAF.0000000000000382.
12. Yu C, Wang F, Wang B, Li G, Li F. A computational biomechanics human body model coupling finite element and multibody segments for assessment of head/brain injuries in car-to-pedestrian collisions. Int J Environ Res Public Health 2020;17(2):492. https://doi.org/10.3390/ijerph17020492.
13. Thomas M, Riemann B, Jones J. Epidemiology of alcohol and drug screening among pedestrian fatalities in the United States, 2014–2016. Traffic Inj Prev 2019;20(6):557–62. https://doi.org/10.1080/15389588.2019.1622006.
14. Udo T, Bitley S, Grilo CM. Suicide attempts in US adults with lifetime DSM-5 eating disorders. BMC Med 2019;17(1):120. https://doi.org/10.1186/s12916-019-1352-3.
15. Terranova C, Sartore D. Suicide and psychiatrist's liability in Italian law cases. J Forensic Sci 2013;58(2):523–6. https://doi.org/10.1111/1556-4029.12039.
16. Kaplan HI, Sadock BJ. Emergenze psichiatriche, suicidio [Psychiatric emergencies, suicide]. In: Psichiatria. Manuale di scienze del comportamento e psichiatria clinica [Psychiatry. Bevaioral sciences and psychiatry manual]. 8th ed. Philadelphia, PA: Lippincott Williams & Wilkins, 1999;864–71.
17. Cavanagh JTO, Carson AJ, Sharpe M, Lawrie SM. Psychological autopsy studies of suicide: a systematic review. Psychol Med 2003;33(3):395–405. https://doi.org/10.1017/S0033291702006943.
18. Isometsä ET. Psychological autopsy studies – a review. Eur Psychiatry 2001;16(7):379–85. https://doi.org/10.1016/S0924-9338(01)00594-6.
19. Al-Abdallat IM, Al Ali R, Hudaib AA, Salameh GAM, Salameh RJM, Idhair AKF. The prevalence of alcohol and psychotropic drugs in fatalities of road-traffic accidents in Jordan during 2008–2014. J Forensic Leg Med 2016;39:130–4. https://doi.org/10.1016/j.jflm.2016.01.018.
20. Wyatt JP, Squires T, Collis S, Broadley R. Road traffic suicides. J Forensic Leg Med 2009;16(4):212–4. https://doi.org/10.1016/j.jflm.2008.12.003.
21. Gauthier S, Reisch T, Ajdacic-Gross V, Bartsch C. Road traffic suicide in Switzerland. Traffic Inj Prev 2015;16(8):768–72. https://doi.org/10.1080/15389588.2015.1021419.
22. Scott CL, Swartz E, Warburton K. The psychological autopsy: solving the mysteries of death. Psychiatr Clin North Am 2006;29(3):805–22. https://doi.org/10.1016/j.psc.2006.04.003.
23. Routley V, Staines C, Breman C, Haworth N, Ozanne-Smith J. Suicide and natural deaths in road traffic. Monash Univ Accid Res Cent Rep 2003;216(December). https://www.monash.edu/muarc/archive/our-publications/reports/muarc216 (accessed February 3, 2020).
24. Tsuang MT, Boor M, Fleming JA. Psychiatric aspects of traffic accidents. Am J Psychiatry 1985;142(5):538–46. https://doi.org/10.1176/ajp.142.5.538.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14506
Available online at: onlinelibrary.wiley.com

CASE REPORT

PATHOLOGY/BIOLOGY

Jas K. Rai,[1] B.Sc. (Hons); Jens Amendt,[2] Ph.D.; Victoria Bernhardt,[2] Ph.D.; Thierry Pasquerault,[3]; Anders Lindström,[4] Ph.D.; and M. Alejandra Perotti,[1] Ph.D.

Mites (Acari) as a Relevant Tool in Trace Evidence and Postmortem Analyses of Buried Corpses

ABSTRACT: This report interprets the presence of mite species in three clandestine graves in Europe, evaluating their potential use as trace evidence or markers. Grave 1 (Sweden): Two mite species *Rhizoglyphus robini* Claparède, 1869 and *Parasitus loricatus* (Wankel, 1861) were recovered from the surface of a body buried in a shallow grave in an area surrounded by trees, in close vicinity to house gardens. Grave 2 (Germany): Phoretic deutonymphs of *Gamasodes spiniger* (Trägårdh, 1910) were attached to an adult fly (Diptera: Sphaeroceridae) found within a shallow grave containing two human bodies covered in soil and dung. Grave 3 (France): *P. loricatus* were recovered from the soil around a body buried in a deep grave (80 cm under). In graves 1 and 3 both corpses were undergoing advanced decay and skeletization, the locations match with the subterranean habit of *P. loricatus*, highlighting the value of this species as a marker of graves or burials in soil and during late decomposition. *R. robini* is a soil mite that feeds on decayed roots and bulbs; this mite species confirms the location of the corpse within top soil, agreeing with a more specific type of superficial burial, a shallow grave. In case 2, the presence of both coprophiles, the mite *G. spiniger* and the carrier fly confirm association of remains with dung or animal feces. The three mite species are reported for the first time in human graves. There are no previous records of *R. robini* from Sweden.

KEYWORDS: trace evidence, burial, clandestine grave, soil mite, decomposition, marker of decomposition, corpse, Acari, Acaridae, Parasitidae

To conceal a murder (homicide), perpetrators often bury their victims. Such clandestine graves are typically shallow, use a mixture of plant materials and soil and are <50 cm in depth (1,2). VanLaerhoven and Anderson already stated 30 cm as the most common depth for clandestine burials (3). However, in rarer cases illegal graves may also be at much greater depths (4). As decomposition of the body progresses through the five most frequently recognized stages of cadaver decay: fresh, bloated, active, advanced, and dry/remains, it forms a rich source of organic material that is able to sustain a large community of arthropod scavengers (5,6). A number of early studies already showed that arthropods arrive at a carcass in a relatively predictive and successive pattern; different species are attracted to different stages of decay. Analysis of the composition of the arthropod community associated with each decomposition stage and the rate of decay can be used for estimation of the minimum postmortem interval (PMI min) (5–7) or as trace evidence. Of the great variety of animals accessing corpses in soil, insects such as Diptera and Coleoptera and minute arachnids such as Acari (mites) are often the most abundant and diverse (3,8).

The majority of PMI estimations of exposed corpses utilize necrophagous dipterans, frequently blow flies, as they can colonize a corpse within minutes after death and are therefore important markers of time (6). Estimating the PMI is crucial in every murder investigation. However, it is a challenging task because a decomposing body represents such a rapidly changing and ephemeral habitat. A major factor that can influence the decomposition rate and the succession, diversity, and abundance of decomposer arthropod communities in and around a cadaver is burial (3). Concealment of a carcass results in reduced insect activity which significantly decreases the rate of decay (3,9). Accordingly, the diversity of species, the ecological succession, and the colonizing time periods of major forensic insects are significantly altered or even prevented in a grave environment (2,3,10,11). In such circumstances, the acarological fauna (mites) may become useful as forensic indicator. Mites are a major part of the carrion fauna in outdoor decomposition, particularly those species sheltering in soil (8,12) but are often unnoticed or ignored because of their small size and difficulties in identification. Nevertheless, they are present through all stages of vertebrate decomposition and therefore have huge potential in interpreting a crime scene (13–21).

The vertical distribution of mites in soils (22) means that they can rapidly colonize a buried carcass at varying depths to feed on

[1] Ecology and Evolutionary Biology Section, School of Biological Sciences, Reading University, Whiteknights, Reading, RG6 6AS, U.K.

[2] Institut für Rechtsmedizin, Goethe University, Frankfurt am Main, 60596, Germany.

[3] Institut de Recherche Criminelle de la Gendarmerie Nationale Department Faune – Flore – Forensiques, Pontoise, 95000, France.

[4] Department of Microbiology, National Veterinary Institute SVA, Uppsala, 751 89, Sweden.

Corresponding author: M. Alejandra Perotti, Ph.D. E-mail: m.a.perotti@reading.ac.uk

Received 11 May 2020; and in revised form 13 June 2020; accepted 16 June 2020.

cadaverous tissue as well as predate on micro-organisms, insect larvae, micro-arthropods, and nematodes already inhabiting the carcass or the neighboring soil (8). Mites will also arrive at a buried carcass phoretically, carried by specific dipteran and coleopteran species that can access the corpse through cavities in the soil (23). Phoresy is the dispersal of one organism (the phoront) through the attachment to a host organism (24,25). This relationship is often transient and is displayed by many species of mites during ontogenesis to rapidly exploit ephemeral habitats, such as dung heaps and carrion (23,26). The host–phoront relationship between mites and insects is sometimes highly specific; where the choice of host is restricted to a single or a handful of species.

Therefore, a forensic acarologist can reconstruct the presence of the carrier species even in its absence, from analyzing the species of mites found at the crime scene (18). Mites may also be introduced on a carcass through material transfer on the victim or the perpetrator from an entirely different location and the habitat specificity of mites can be valuable as trace evidence (20,21,27). Jean Pierre Mégnin, the founder of Forensic Acarology, was the first to place mites along with insects and other arthropods throughout the 8 waves of arthropod colonization of exposed cadavers, where the 6th wave was composed entirely of mites (28). Mégnin listed mites as part of the 4 waves of arthropods associated with buried cadavers along with Diptera, Coleoptera, and Lepidoptera (28). In 1898, Motter reviewed bodies buried in coffins up to 150 cm in depth, mites were the most abundant arthropods, and *Uropoda depressa* (described by Mégnin) was the most common species (8,29). Recent analyses of buried carcasses have demonstrated that mites are plentiful in human graves though mites are unidentified or their role mainly unknown (8,14,22,30).

The main aim of this work is to document the mite species occurring in graves in three different biogeographical locations in countries in Europe: Sweden, Germany, and France, as well as to interpret the occurrence of certain species as markers of specific "burial" environments.

Materials and Methods

Description of Studied Graves

Grave 1

During construction work in Central Sweden, the remains of a male were discovered in a shallow grave (<50 cm) on 17 March 2015. The body was found in a small grove near an old manor surrounded by several houses and gardens. Homicide was suspected, and on 24 March 2015, the remains were autopsied. The corpse was partly skeletonized and the abdomen had a layer of adipocere. The internal organs were partly decomposed but relatively intact and the head was almost completely skeletonized. Ten individual mites were collected directly from the clothing during the autopsy and preserved in 70% ethanol. The sample containing the mite specimens were later sent to the Acarology laboratory, University of Reading (U.K.), for identification and interpretation of the acarological evidence. Insect fauna was also collected from the grave. It consisted of Phoridae adults, Piophilidae larvae, and Muscidae pupae. Several individuals of *Rhizophagus parallelocollis* (graveyard beetle—several millimeters in length) indicated a PMI of 10–24 months.

Grave 2

The bodies of 2 individuals were discovered on a horse ranch in a rural area in Germany in June 2014 (Fig. 1). The bodies were positioned horizontally adjacent to each other in a shallow grave of approximately 30 cm depth and was covered with horse manure and soil. The 2 individuals displayed evidence of gunshot wounds. A small adult Diptera was recovered from the samples taken from the grave soil carrying two phoretic mites grasped dorsally to the fly. The Dipteran was identified to the family Sphaeroceridae, *Spelobia* sp., because of the minute size (approximately 1–2 mm) and a characteristically thickened tarsomere of the posterior leg. All specimens were preserved in 70% ethanol. No more insect evidence was present. Mites were sent to the Acarology laboratory, University of Reading (U.K.), for identification and interpretation of the acarological evidence. Based on the clarified identity of the dead and the case reconstruction, a six-week PMI could be assumed.

Grave 3

A skeletonized body was exhumed from a deep soil grave (approximately 80 cm) in the west of France in 2004 (Fig. 2) (4). The soil was mainly clay based, and the body was covered by 40cm of farm quicklime. Entomological evidence was collected: Larvae, pupae, and empty puparia of Heleomyzidae (Diptera) (at 40–60 cm); pupae and empty puparia of scuttle fly *Megaselia* sp. (Diptera: Phoridae)—in hair, reaching 90 cm in depth; adults of unidentified lesser dung flies (Diptera: Sphaeroceridae) at 40–60 cm and adults of rove beetles (Coleoptera: Staphylinidae) *Aleochara* sp. The absence of Calliphoridae and Sarcophagidae

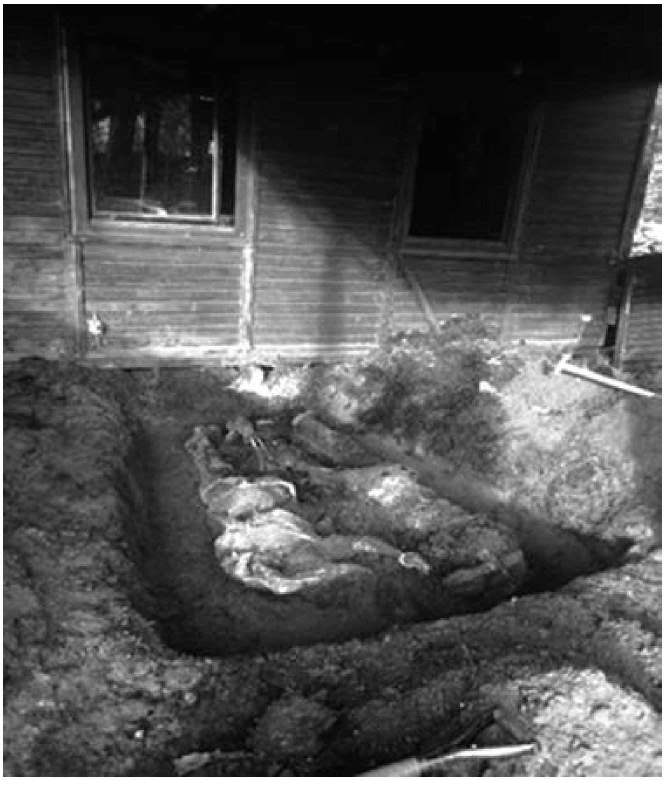


FIG. 1—*Grave site of remains of two individuals on discovery in June 2014 in Germany (grave depth approximately 30 cm). [Color figure can be viewed at wileyonlinelibrary.com]*

confirmed burial shortly after death. It was suspected that the victim was killed 6 months before the discovery of the body, during early autumn when the temperatures were likely to be favorable. A few mite specimens were also recovered from the grave, preserved in 70% ethanol, and were sent to the Acarology laboratory, University of Reading (UK), for identification and interpretation of the acarological fauna.

Identification of Mites

The clearing and mounting of mites was based on previously described methods (31). A Nikon Optiphot phase contrast light microscope was used for identification (objectives used were 10×, 40× and 100×). Images were captured with Motic Image Plus 3.0. Several taxonomical keys were used for the identification of mites. For case study 1, key for Astigmata species by Hughes (1976) was mainly used for identification to the genus and species level (32). A number of other keys and descriptions of Astigmata, Acaridae were also used (33–36). For identification of the Mesostigmata, Parasitidae, for cases 1, 2, and 3, a key to Mesostigmata families was first used to identify the mite to the family level (Parasitidae) (37) followed by a key to Parasitidae species (38).

Results and Discussion

Grave 1

Of the sample received, five individuals were identified as the bulb mite species *Rhizoglyphus robini* Claparède, 1869 (Fig. 3). All individuals of *R. robini* were in the hypopial stage, a heteromorphic deutonymph adapted to phoresy (Fig. 4*a*,*b*). The rest of the five specimens were identified as adults of the Parasitidae species *Parasitus loricatus* (Wankel, 1861) (Figs 5 and 6).

The hypopi of the genus *Rhizoglyphus* are similar in morphology to those of *Caloglyphus* (*Sancassania*) (Astigmata: Acaridae). Differences can be found in some morphologies such as minute pits evident on the dorsal surface of *R. robini*, the presence of shorter legs, and a transverse line separating the sternal and ventral shield. They also show similarities to *Acarus farris* hypopi (Astigmata: Acaridae).

FIG. 2—*Case study 3. Grave site of remains of an individual in France in 2008 (grave depth, approximately 80 cm). [Color figure can be viewed at wileyonlinelibrary.com]*

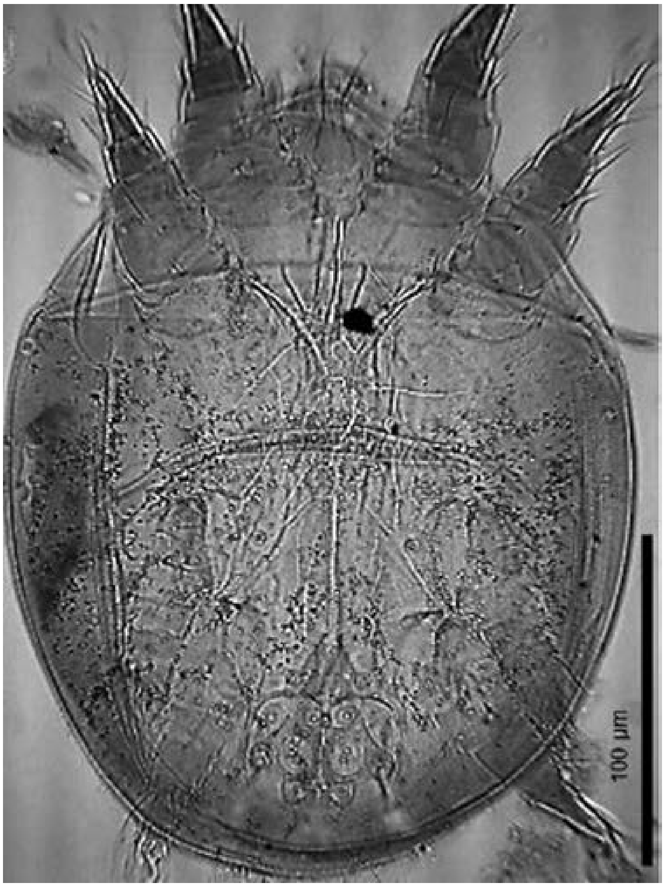


FIG. 3—*Hypopus of* Rhizoglyphus robini *(ventral).*

However, apodemes IV do not curve or run parallel for a short while as in *A. farris*, but rather meet at a point. Some diagnostic characteristics of *R. robini* are (i) the protrusion of the rostrum covers the entire gnathosoma, (ii) the apodemes do not reach the posterior edge, and (iii) the sucker plate, almost identical to the diagrammatic description shown by Fan and Zhang in 2004 (36), with 2 large central suckers with 6 smaller bordering suckers that are equal in size. In the contested specimens, vertical dorsal setae were not as distinguishable as expected, however, are expected to be relatively short in *R. robini*. Legs IV were slightly longer than expected and visible when viewed dorsally. The morphology of *R. robini* is closely related to *R. echinopus* (Astigmata: Acaridae). However, a number of diagnostic characters unique to *R. robini* rather than *R. echinopus* were identified (33,39–41). For example, these specimens show a gnathosoma entirely covered by the rostrum and not visible dorsally, agreeing with Radwan and co-authors (42).

This is the first report of *R. robini* in Sweden, although the species has a cosmopolitan distribution worldwide and is frequently reported in synanthropic habitats such as greenhouses and gardens in Europe (Table 1). Species of the family Acaridae are important pests of agricultural plants. Within the Acaridae family, bulb mites from the genus *Rhizoglyphus* typically attack bulbs, tubers, or corms of potato, carrot, onion, and garlic plants among other vegetables, as well as flower bulbs in greenhouses and fields (34–36,43). Among the broad variety of plants that *Rhizoglyphus* mites damage, they are most commonly associated with members of the Liliaceae family, one of the largest families of (garden) plants (35). The bulb mite undergoes 6 stages during

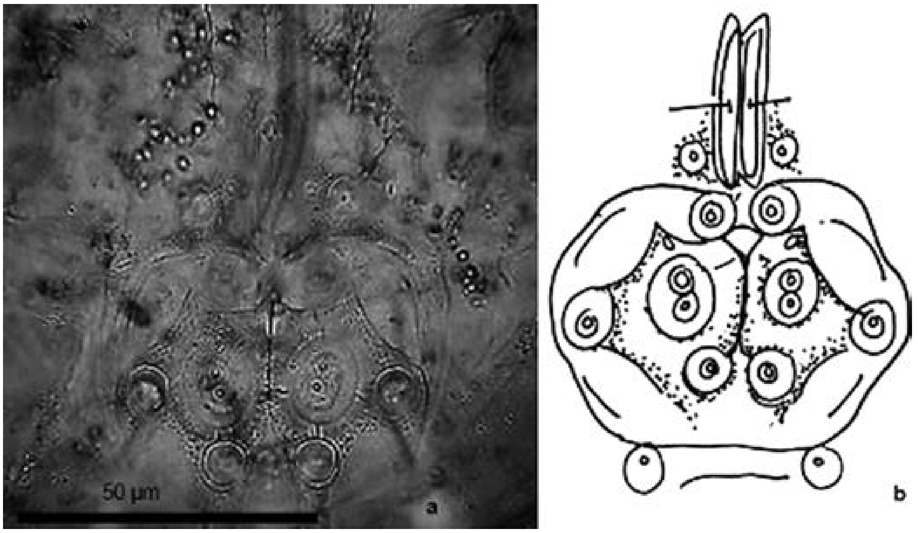


FIG. 4—*Ventral sucker plate of* Rhizoglyphus robini *hypopus of case specimens. b. Schematic drawing of the sucker plate of* R. robini*, adapted from previous work (32,36), showing a pair of large central suckers and 3 pairs of bordering suckers that are roughly equal in size.*

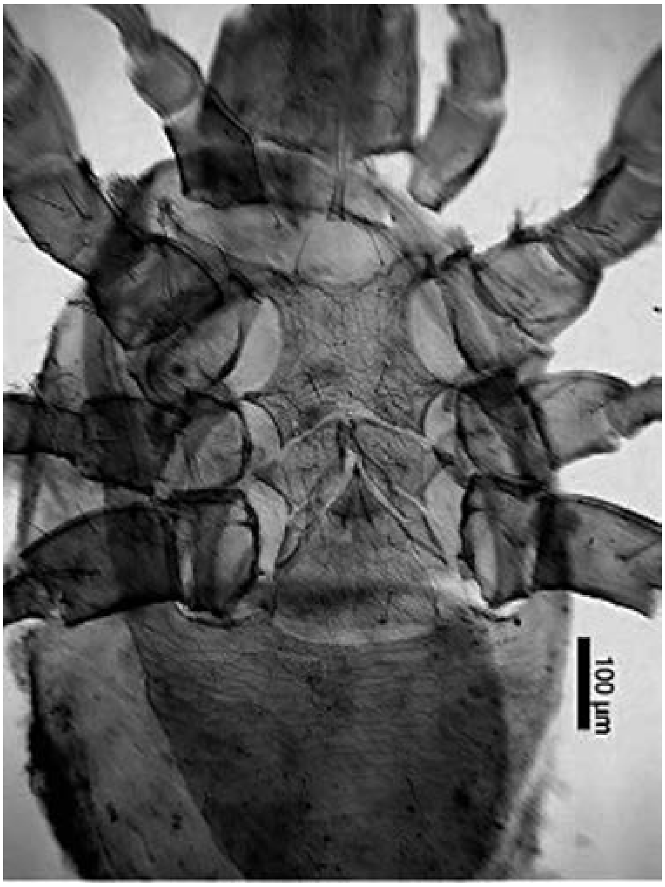


FIG. 5—Parasitus loricatus *female (ventral).*

its life cycle: egg, larva, protonymph, deutonymph, tritonymph, and adult (44).

Astigmata mites transform into the hypopi (nonfeeding, phoretic deutonymph) in response to deteriorating environmental conditions such as extremes of temperature and humidity and poor food quantity and quality (26). *Rhizoglyphus* hypopi are known to attach to several species of Diptera and Coleoptera and have been found phoretically associated with Scarab beetles, including *Osmoderma eremicola*, *Bothynus gibbosus*, and *Phyllophaga* spp., which are opportunistic colonizers of animal and human remains (45,46). The abundance of *Rhizoglyphus* hypopi found within populations in the field is generally low since most individuals will molt directly from a protonymph to a tritonymph if there is food available (35).

Only one past study has recovered *R. robini* from soil associated with decomposing surface animal remains, and no previous study has documented its occurrence in graves. Anderson and VanLaerhoven found *R. robini* in the soil beneath surface pig carcasses, along with Dipterans and Coleopterans, in a rural farming area of British Columbia (47). The life stage was not noted. Between one to 10 individuals were found in the soil when the pigs were undergoing the dry remains stage. Considering the location of the case, a rural area surrounded by some houses and gardens, *R. robini* places the origin of the corpse in the environment where it was found. *Rhizoglyphus robini* are considered to favor living plant matter such as the bulbs of common garden plants and ornamentals, to decomposing matter (34). However, the soil surrounding the body was devoid of such vegetation, and this had triggered the production of hypopi. The occurrence of *R. robini* is supported by the presence of a population of *R. parallelocollis* in the grave, which is a small (approximately 4 mm) root eating beetle that feeds on buried organic matter, commonly found in gardens and compost heaps as well as buried corpses (48).

Other 5 mite specimens of *P. loricatus* were recovered from the corpse, 3 females and 2 males. The females (Fig. 5) show the typical roughly triangular opisthonotal shield with the genital shield sharply pointed anteriorly, and the presence of the metasternal shield (38). The lack of diffusion between the genital and opisthogastric plates helped distinguished them from a closer species *P. fimetorum*. The males bear the specific diagnostic characters of the species, such as the leg apophyses (protrusions) on legs II (Fig. 6), a deeply bifid and V-shaped spur of femur II (Fig. 6*a*), and a clefted corniculli (Fig 6*b*) (38).

Parasitus loricatus is not restricted to isolated or secluded habitats and has been reported from a wide variety of biotopes

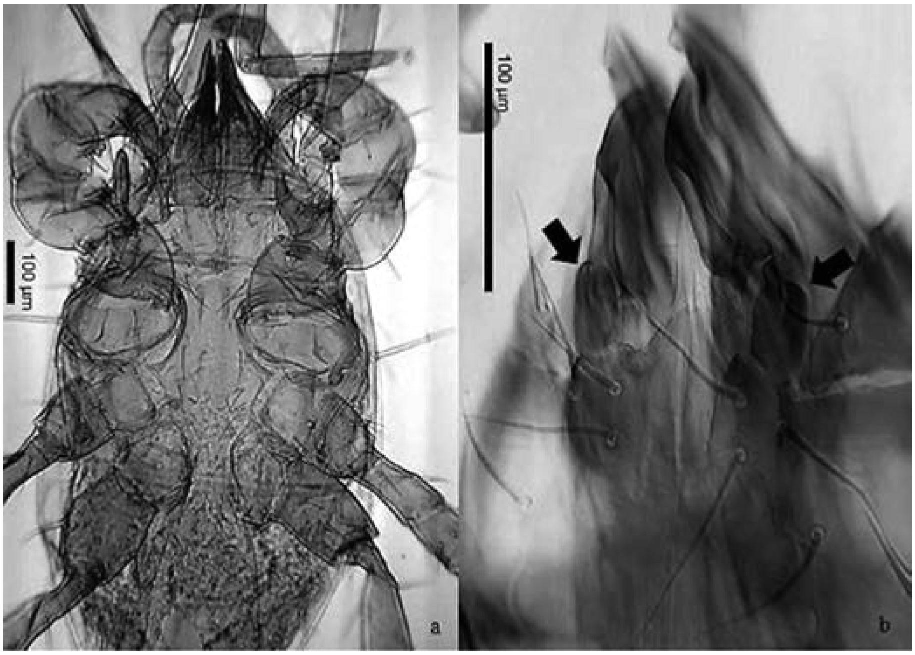


FIG. 6—*(a)* Parasitus loricatus *male (ventral); (b) Clefted corniculi of male* P. loricatus *(arrows).*

TABLE 1—*Occurrence of* Rhizoglyphus robini *in European countries: habitat type, host, and life stage.*

Location	Habitat	Host	Life Stage	Reference
Hungary	Unknown	Sour cherry tree	Unknown	(61)
Poland	Unknown	Bird Nests	Unknown	(62)
Poland	Garden	Onions	Ad	(42)
Canada	Rural Farm	Soil beneath pig carcass	Ad	(47)
Norway	Agriculture	Onion dust	Unknown	(63)
Holland	Lily fields	Lily plants	Ad	(64)
Poland	Dwelling	Dust samples	Unknown	(65)
Denmark	Beech woodland soil	Soil	Unknown	(66)
Denmark	Dwelling	Dust from a mattress (*Rhizoglyphus* sp.) (Unspecified)	Unknown	(67)
United	Kingdom	Unknown	*Freesia* sp.,	
	Narcissus sp.	Unknown	(68)	
Poland	Rye field	Rye	Ad	(69)
Greece	Unknown	*Dahlia* sp.	Unknown	(33)*
Holland	Unknown	*Amaryllis, Gladioulus sp., Iris* sp., *Lilium* sp.	Unknown	(33)*
United	Kingdom Unknown	House (70)	Stored	products
Italy		Bulbs	Unknown	(71)*
Austria		Bulbs	Unknown	(71)*

Ad, Adult.
*Review paper.

such as forest soil, nests of birds and mammals and semi-aquatic habitats such fish pond litter (Table 2). During analysis of the existent literature on *P. loricatus*, a common and major problem in acarology became apparent. The majority of reports that cite this species describe it as a eu-troglophile species; assuming its origin is from caves. However, the original publication by Wankel in 1861, written in Dutch, describes the species as a soil dwelling mite found in underground tunnels, often associated with micro nests of small mammals and arthropods (49). The species is found in subterranean habitats such as below-ground nests of rodents (50), justifying its occurrence in graves and on surface terrains such as compost, bird nests and in excavations like graves (this study). There is no past documentation of the association of *P. loricatus* with buried or surface cadavers and this is the first report of this species from a human grave. This species is frequently found in Europe, especially in Southern Sweden, Baltic Island of Gotland, and Norway and is often the most common species of caves (Table 2).

Grave 2

Two mites were found attached to the dorsal surface of a *Spelobia* fly (Sphaeroceridae) and were identified as *Gamasodes spiniger* (Trägårdh, 1910) deutonymphs (Figs 7 and 8) (38). *G. spiniger* deutonymphs are characterized by the presence of spurs on Leg II, on femur, genu, tibia, and tarsus, where femur and tibia bear one spur each (Fig. 7*b*). The femur spur is thumb shaped with a curved tip, the genu has a shorter more pointed spur, the tibia a rounded spur and the tarsus a short conical spur. Presternal shields are wide and elongated, and the sternal shield is characteristically outlined and partly punctate (i.e., bearing holes). The dorsal setae are mainly short where more than two pairs of dorsal setae are stouter and pilose; the opisthonotal shield (dorsal) bears 14 pairs of setae, where setae Z1, Z3, and J5 are stouter and pilose.

The sternal and opisthogastric setae are typically fine and slender. The specimens differ slightly from the description in Hyatt (1980) (38) in the shape of the sternal shield and lateral spines of the tectum, with dentate lateral margins. This species is a saprophile (associated with dead or decaying matter) and a coprophile (associated with dung); therefore, it is also frequently found in dung or manure (Table 3). The deutonymphs of *G. spiniger* are known to be phoretic with Coleoptera such

TABLE 2—*Occurrence of* Parasitus loricatus *in European countries, habitat type, and life stage.*

Location	Habitat	Life Stage	Reference
Poland	Caves	Unknown	(72)
Belgium	Subterranean cavities	Ad	(73)
Italy	Caves and subterranean cavities	Unknown	(74)
Slovakia	Fields, surrounded by farms. Subterranean nests of mound-building mouse	Ad, Dt	(50)
Romania	Mountain soil	Unknown	(75,76)
Western Slovakia	Nests of Anseriformes and Passeriformes	Unknown	(77)
South West Slovakia	Forest soil	Unknown	(78)
Poland	Fur of Voles	Dt	(79)
Slovakia	Soil and litter of fishponds and Mallard nests	Unknown	(80)
Northern Slovakia	Caves	Unknown	(81)
Slovakia	Bat dung and soil/sediment of caves	Unknown	(82)
Hungary	Cave	Unknown	(82)
Slovakia	Caves	Unknown	(83)
Sweden	Caves	Ad, Dt	(84)
United Kingdom	Yew	Unknown	(85)
Poland	Grassland	Unknown	(86)

Ad; Adult, Dt; Deutonymph.

as *Copris hispanus* (Coleoptera, Scarabaeidae) as well as Diptera (51), especially small specimens, for example, sciarid flies, which are well known pests of greenhouses (52) (Table 4). Many *Gamasodes* species are predators, existing as parasitic and free-living mites and practice phoretic activity for dispersal into bird nests and the nests of small mammals (38). *Gamasodes* species have also been found phoretically associated with several species of dung beetles (53). *G. spiniger* is a common soil dwelling species in European countries and is also frequently found inhabiting nests of mammals and birds (Table 3).

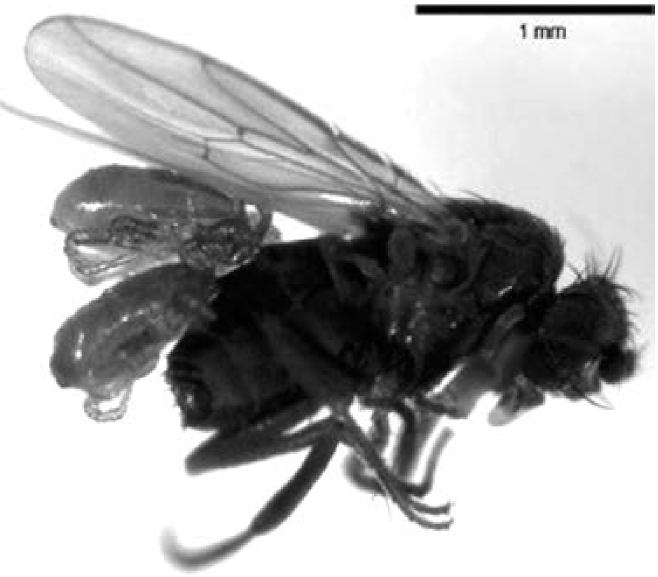


FIG. 8—*Two deutonymphs of* Gamasodes spiniger *attached to dipteran (*Spelobia *species).*

There are only three documented cases of *Gamasodes* species associated with animal carcasses. *Gamasodes spiniger* was collected from beneath exposed pig carcasses in a rural farming area during the very early fresh stage of decomposition, with no further occurrence of this species throughout the rest of decomposition (47). Mesostigmatid mites were found in high abundance in the soil directly associated with decaying surface rabbit carcasses in Malaysia, with Macrochelidae species occurring throughout decomposition and Parasitidae mites such as *Gamasodes* sp. (unidentified species) dominating in the late stages of decomposition (46).

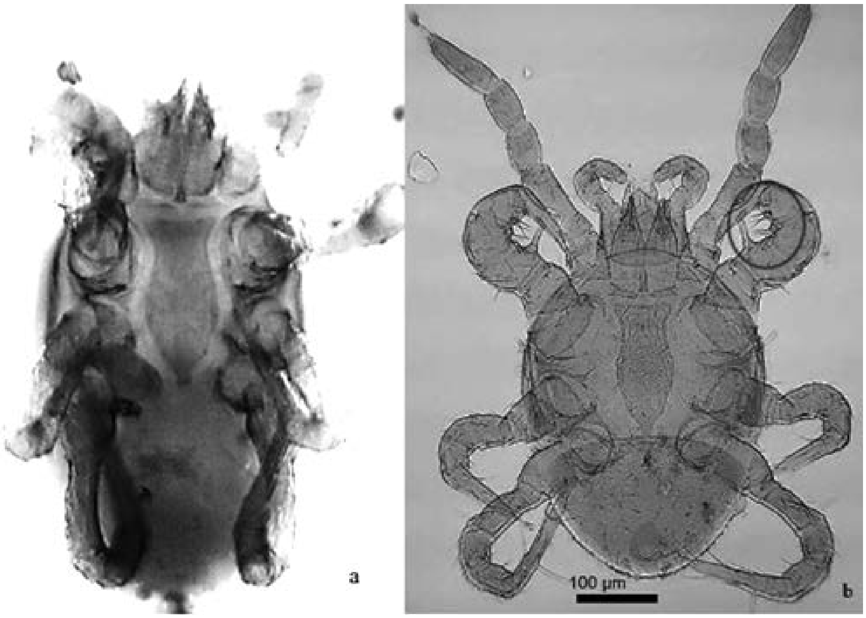


FIG. 7—Gamasodes spiniger *deutonymph (ventral). (a) Image of specimen from case study. (b) Example of* G. spiniger *(deutonymph), using another specimen not-related to this case, to show the diagnostic features such as the sternal shield and presternal shields which have a characteristic shape, and spurs on femur and genu on Leg 2 (circled).*

TABLE 3—*Occurrence of* Gamasodes spiniger *in European countries, habitat type, and the life stage.*

Location	Habitat	Host	Life Stage	Reference
Romania	Forest	Soil	Unknown	(75)
Belgium	Underground cavities		Dt	(73)
Slovakia	Fields, surrounded by farms	Subterranean nests of mound-building mouse	Ad, Dt	(50)
Poland	Forest	Nests of black stork	Ad, Dt	(87)
Slovakia		Bearded tit	Unknown	(88)
Italy		Bearded tit	Unknown	(88)
Austria		Bearded tit	Unknown	(88)
Southern Sweden	Garden Lawn	Fly (Sphaeroceridae)	Dt	(56)
Northern Ireland	Bramley apple orchard	Leaf and pitfall	Unknown	(89)
Eastern Germany	Sterile soil	Soil	Unknown	(90)
Latvia	Strawberry field	Leaf and pitfall	Unknown	(91)
Poland	Nests of white stork	Diptera	Dt	(56)
Poland and Czech Republic	Mountain range	Bank vole and common Vole	Unknown	(79)
Slovakia	Fishponds and Mallard nests	Soil and litter of fishponds and Mallard nests	Unknown	(80)
Slovakia	Farmland	Nests of Red-backed shrike	Ad, Dt	(92)
England and Wales	Grassland, garden	Yew	Unknown	(85)
United Kingdom	Grassland	Slurry	Unknown	(93)
United Kingdom	Farm	Poultry litter	Unknown	(94)

Ad; Adult, Dt; Deutonymph.

TABLE 4—*Literature citing deutonymphs of* Gamasodes spiniger *with associated arthropod host.*

Location	Habitat	Host	Reference
Southern Sweden	Garden Lawn	Sphaeroceridae (Diptera)	(56)
Poland	Nests of white stork	Diptera	(55)
Spain	Forest, on pig carrion bait	Flying insects	(54)
Czech republic	Unknown	Sphaeroceridae (in dung)	(52)
Israel	Forest	*Copris hispanus* (Coleoptera: Scarabaeidae)	(51)

Deutonymphs of *G. spiniger*, phoretic with flying insects, were recovered at irregular intervals colonizing unconcealed pig carrion baits placed on forest soil in North Spain (54). This however is the first report of *G. spiniger* from a human grave.

Phoresy of *G. spiniger* with dipterans has been previously documented but these handful studies have not always indicated the species of Diptera. For example, *G. spiniger* has been found in the nests of white storks in Poland; thought to have arrived through phoretic activity, attached to dipterans, however, the species were not identified (55). There are previous records of deutonymphs of *G. spiniger* associated with flies of Sphaeroceridae; and of unattached *G. spiniger* deutonymphs found along with Sphaeroceridae flies in manure; but the flies were never identified to species (52). In another study on phoront-host associations between mites and insects in a garden lawn in Southern Sweden, in 1998, a single *G. spiniger* mite was found attached with its chelicera to the abdomen of a Sphaeroceridae fly, at day 100 of a study. More so, a further 17 deutonymphs of *G. spiniger* were collected during the same study (56). However, none of the records identified the species of flies. This is the first confirmation of *G. spiniger* traveling on *Spelobia* species (Sphaeroceridae).

The association of *G. spiniger* with Sphaeroceridae is interesting from the forensic point of view. This is a family of Diptera with a global distribution, occurring in most terrestrial habitats, commonly known as lesser dung flies, which thrive in dung, but also feed on dead animal matter (57). Sphaeroceridae are generally less abundant on vertebrate carrion than Calliphoridae (blow flies). However, in cases where blow flies cannot access corpses, as in the case of burials and particularly if the environment of the grave contains animal dung, the smaller Sphaeroceridae are more adapted to detect and colonize such remains than Calliphoridae. In a study of buried and surface pigs in Michigan, larvae of Sphaeroceridae were recovered from pigs buried at 30 cm but not from pigs buried at 60 cm, 60 days after burial. In the same study, no Sphaeroceridae were found colonizing the surface pigs (2). The puparia of Sphaeroceridae were found in the lead coffin graves of Archbishop Greenfield, buried in 1315 (48) and in 1968, Payne found Sphaeroceridae colonizing pigs buried 50–100 cm in soil during bloating and active decay (58). Interestingly, the species *Spelobia luteilabris* has been previously reported among the dominating dipteran species in both open habitats and forests in Southern Germany feeding on various forms of carrion baits, and in exposed and buried up to 5 cm (57). *Spelobia* species have been collected from exposed pigs in a meadow undergoing late fresh stage during winter in Germany (59). The studied grave in the present work is located in Germany.

In this case, both the fly, *Spelobia* sp. and *G. spiniger* mites were likely attracted to the horse manure that was used to conceal the grave. The occurrence of *G. spiniger* during mid to late stage of decay of the corpses in this case study is not surprising as the species predate on other soil-inhabiting micro-arthropod decomposers of cadavers such as bacteria, fungi, nematodes, and other micro-arthropods. *Spelobia* sp. along with *G. spiniger* seemed to have arrived shortly before the bodies were discovered due to the recovery of a low number of specimens of both species. The simultaneous occurrence of the two species is indicative of late decomposition in graves in livestock-related environments or habitats.

Grave 3

Of the five mites recovered from this corpse, three were identified as *P. loricatus* (male, female, and deutonymph). The other two specimens were relatively fragmented, which prevented their preparation for identification; however, they still showed general similarities with the species. The presence of both adults and deutonymphs suggests at least a single life cycle within the grave, indicating that the decomposition process of the body might have occurred within the isolated grave; information was also complemented by the absence of Calliphoridae and Sarcophagidae flies. This is the first report documenting the occurrence of *P. loricatus* in a human grave of approximately 80 cm depth, which is considered a deep grave (4). Small size arthropods mainly occupy upper horizons of soils, due to the decreasing porosity of the soil from the surface to deep layers. The fauna of deeper layers of soil is typically scarce and opportunistic (60). This case highlights the value of *P. loricatus* as markers of deep burials. The corpse in this case was undergoing advanced decay with some skeletal remains, similar stage to case study, grave 2. With further studies on the species, it might be possible to define its role in advanced and/or late decomposition within the deep grave environment.

The three case studies confirm the association of mites with decomposing human remains in graves, shallow, and deep and at different stages of decomposition. Exposure to a variety of environments, such as garden soil or dung, allows more information on specificity to habitats, which helps identify specific markers of decomposition, locations, or a stage of decay. This is particularly important when investigating homicide cases and there is little or no insect activity.

Acknowledgments

We want to thank Dr Marta Saloña-Bordas for identifying the family and genus of the fly (Diptera, Fig. 8).

References

1. Pringle JK, Jervis J, Cassella JP, Cassidy NJ. Time-lapse geophysical investigations over a simulated urban clandestine grave. J Forensic Sci 2008;53(6):1405–16. https://doi.org/10.1111/j.1556-4029.2008.00884.x.
2. Pastula EC, Merritt RW. Insect arrival pattern and succession on buried carrion in Michigan. J Med Entomol 2013;50(2):432–9. https://doi.org/10.1603/ME12138.
3. VanLaerhoven SL, Anderson GS. Insect succession on buried carrion in two biogeoclimatic zones of British Columbia. J Forensic Sci 1999;44(1):32–43. https://doi.org/10.1520/JFS14409J.
4. Gaudry E. The insects colonisation of buried remains. In: Amendt J, Campobasso CP, Goff ML, Grassberger M, editors. Current concepts in forensic entomology. Dordrecht, Germany: Springer, 2009;273–311. https://doi.org/10.1007/978-1-4020-9684-6_13.
5. Catts EP, Goff ML. Forensic entomology in criminal investigations. Annu Rev Entomol 1992;37:253–72. https://doi.org/10.1146/annurev.en.37.010192.001345.
6. Grassberger M, Frank C. Initial study of arthropod succession on pig carrion in a central European urban habitat. J Med Entomol 2004;41(3):511–23. https://doi.org/10.1603/0022-2585-41.3.511.
7. Goff ML. Estimation of postmortem interval using arthropod development and successional patterns. Forensic Sci Rev 1993;5:81–94. https://doi.org/10.1097/00000433-198809000-00009.
8. Braig HR, Perotti MA. Carcases and mites. Exp Appl Acarol 2009;49(1–2):45–84. https://doi.org/10.1007/s10493-009-9287-6.
9. Rodriguez WC, Bass WM. Decomposition of buried bodies and methods that may aid in their location. J Forensic Sci 1985;30(3):836–52. https://doi.org/10.1520/JFS11017J.
10. Lundt H. Ecological observations about the invasion of insects into carcasses buried in soil. Pedobiologia 1964;4:158–80.
11. Turner B, Wiltshire P. Experimental validation of forensic evidence: a study of the decomposition of buried pigs in a heavy clay soil. Forensic Sci Int 1999;101(2):113–22. https://doi.org/10.1016/S0379-0738(99)00018-3.
12. Bornemissza GF. An analysis of arthropod succession in carrion and the effect of its decomposition on the soil fauna. Aust J Zool 1957;5(1):1–12. https://doi.org/10.1071/ZO9570001.
13. Goff ML. Use of Acari in establishing a postmortem interval in a homicide case on the island of Oahu, Hawaii. In: Dusbábek F, Bukva V, editors. Modern acarology. vol. 1. The Hague, The Netherlands: SPB Academic Publishng, 1991;A439–A42.
14. Russell DJ, Schulz MM, OConnor BM. Mass occurrence of astigmatid mites on human remains. Abh Ber Naturkundemus Gorlitz 2004;76:51–6.
15. Perotti MA. Megnin re-analysed: the case of the newborn baby girl, Paris, 1878. Exp Appl Acarol 2009;49(1–2):37–44. https://doi.org/10.1007/s10493-009-9279-6.
16. Salona-Bordas MI, Perotti MA. First contribution of mites (Acari) to the forensic analysis of hanged corpses: a case study from Spain. Forensic Sci Int 2014;244:e6–11. https://doi.org/10.1016/j.forsciint.2014.08.005.
17. Salona MI, Moraza ML, Carles-Tolra M, Iraola V, Bahillo P, Yelamos T, et al. Searching the soil: forensic importance of edaphic fauna after the removal of a corpse. J Forensic Sci 2010;55(6):1652–5. https://doi.org/10.1111/j.1556-4029.2010.01506.x.
18. Medina AG, Herrera LG, Perotti MA, Rios GJ. Occurrence of *Poecilochirus austroasiaticus* (Acari: Parasitidae) in forensic autopsies and its application on postmortem interval estimation. Exp Appl Acarol 2013;59(3):297–305. https://doi.org/10.1007/s10493-012-9606-1.
19. Szelecz I, Losch S, Seppey CVW, Lara E, Singer D, Sorge F, et al. Comparative analysis of bones, mites, soil chemistry, nematodes and soil micro-eukaryotes from a suspected homicide to estimate the post-mortem interval. Sci Rep 2018;8(1):25. https://doi.org/10.1038/s41598-017-18179-z.
20. Kamaruzaman NAC, Masan P, Velasquez Y, Gonzalez-Medina A, Lindstrom A, Braig HR, et al. *Macrocheles* species (Acari: Macrochelidae) associated with human corpses in Europe. Exp Appl Acarol 2018;76(4):453–71. https://doi.org/10.1007/s10493-018-0321-4.
21. Hani M, Thieven U, Perotti MA. Soil bulb mites as trace evidence for the location of buried money. Forensic Sci Int 2018;292:E25–E30. https://doi.org/10.1016/j.forsciint.2018.09.016.
22. Ducarme X, Wauthy G, Andre HM, Lebrun P. Survey of mites in caves and deep soil and evolution of mites in these habitats. Can J Zool 2004;82(6):841–50. https://doi.org/10.1139/z04-053.
23. Perotti MA, Braig HR. Phoretic mites associated with animal and human decomposition. Exp Appl Acarol 2009;49(1–2):85–124. https://doi.org/10.1007/s10493-009-9280-0.
24. Lesne P. Moeurs du *Limosina sacra* Meig. (famille Muscidae, tribu Borborinae). Phénomènes de transport mutuel chez les animaux articulés. Origines du parasitisme chez les insectes diptères [Manners of Limosina sacra Meig. (family Muscidae, Borborinae tribe). Mutual transport phenomena in articulated animals. Origins of parasitism in Diptera insects Mores of Limosina sacra Meig. (family Muscidae, Borborinae tribe). Bull Soc Entomol Fr 1896;1(6):162–5.
25. Camerik AM. Phoresy revisited. In: Sabelis MW, Bruin J, editors. Trends in acarology. Dordrecht, the Netherlands: Springer, 2010;333–6. http://doi.org/10.1007/978-90-481-9837-5_53.
26. Houck MA, Oconnor BM. Ecological and evolutionary significance of phoresy in the Astigmata. Ann Rev Entomol 1991;36:611–36. https://doi.org/10.1146/annurev.ento.36.1.611.
27. Prichard JG, Kossoris PD, Leibovitch RA, Robertson LD, Lovell FW. Implications of trombiculid mite bites: report of a case and submission of evidence in a murder trial. J Forensic Sci 1986;31(1):301–6. https://doi.org/10.1520/JFS11887J.
28. Mégnin PJ. La faune des cadavres [The fauna of carcasses]. Ann Hyg Publ Med Leg Serie 1895;3(33):64–7.
29. Motter MG. A contribution to the study of the fauna of the grave. A study of on hundred and fifty disinterments, with some additional experimental observations. J N Y Entomol Soc 1898;6(4):201–31.
30. Bourel B, Tournel G, Hedouin V, Gosset D. Entomofauna of buried bodies in Northern France. Int J Legal Med 2004;118(4):215–20. https://doi.org/10.1007/s00414-004-0449-0.
31. Krantz GW. Collection rearing, and preparation for study. In: A manual of acarology, 2nd edn. Corvallis, OR: Oregon State University Bookstores Inc, 1978;77–98.
32. Hughes AM. The mites of stored food and houses, 2nd edn. London, U.K.: Ministry of Agriculture and Fisheries/Her Majesty's Stationery Office, 1976; Technical Bulletin 9, iv.
33. Manson DCM. A contribution to the study of the genus Rhizoglyphus Claparede, 1869 (Acarina: Acaridae). Acarologia 1972;13(4):621–50.
34. Gerson U, Yathom S, Capua S, Thorens D. *Rhizoglyphus robini* Claparede (Acari, Astigmata, Acaridae) as a soil mite. Acarologia 1985;26(4):371–80.
35. Diaz A, Okabe K, Eckenrode CJ, Villani MG, OConnor BM. Biology, ecology, and management of the bulb mites of the genus Rhizoglyphus (Acari: Acaridae). Exp Appl Acarol 2000;24(2):85–113. https://doi.org/10.1023/A:1006304300657.
36. Fan QH, Zhang ZQ. Revision of *Rhizoglyphus* Claparède (Acari: Acaridae) of Australasia and Oceania. London, U.K.: Systematic and Applied Acarology Society, 2004;216–70.

37. Lindquist EE, Krantz GW, Walter DE. Order mesostigmata. In: Krantz GW, Walter DE, editors. A manual of acarology, 3rd edn. Lubbock, TX: Texas Tech University Press, 2009;124–232.
38. Hyatt KH. Mites of the subfamily Parasitinae (Mesostigmata: Parasitidae) in the British Isles. Bull. Br. Museum (Natural History) 1980;38 (5):237–378.
39. Eyndhoven GV. Artunterschiede beim Genus *Rhizoglyphus* (Acar.) [Differences in the Genus *Rhizoglyphus* (Acar.)]. *Proc XI Intl Cong Entomol (Vienna)*1960;274–6.
40. Eyndhoven GV. The *Rhizoglyphus echinopus* of Fumouze and Robin. Mitt Schweiz Ent Ges 1963;36:48–9.
41. Eyndhoven GV. *Rhizoglyphus engeli* nov. spec., with notes on the genus Rhizoglyphus (Acari, Acaridae). Beaufortia 1968;15(193):95–103.
42. Radwan J, Unrug J, Snigorska K, Gawronska K. Effectiveness of sexual selection in preventing fitness deterioration in bulb mite populations under relaxed natural selection. J Evol Biol 2004;17(1):94–9. https://doi.org/10.1046/j.1420-9101.2003.00646.x.
43. Fan QH, Zhang ZQ. *Rhizoglyphus echinopus* and *Rhizoglyphus robini* (Acari: Acaridae) from Australia and New Zealand: identification, host plants and geographical distribution. Syst Appl Acarol (Special Publ) 2003;16(1):1–16. https://doi.org/10.11158/saasp.16.1.1.
44. Deere JA, Coulson T, Smallegange IM. Life history consequences of the facultative expression of a dispersal life stage in the phoretic bulb mite (*Rhizoglyphus robini*). PLoS One 2015;10(9):e0136872. https://doi.org/10.1371/journal.pone.0136872.
45. Zanetti NI, Visciarelli EC, Centeno ND. Associational patterns of scavenger beetles to decomposition stages. J Forensic Sci 2015;60(4):919–27. https://doi.org/10.1371/journal.pone.0136872.
46. Silahuddin SA, Latif B, Kurahashi H, Heo CC. The Importance of habitat in the ecology of decomposition on rabbit carcasses in Malaysia: implications in forensic entomology. J Med Entomol 2015;52(1):9–23. https://doi.org/10.1093/jme/tju00.
47. Anderson GS, VanLaerhoven SL. Initial studies on insect succession on carrion in southwestern British Columbia. J Forensic Sci 1996;41 (4):617–25. https://doi.org/10.1520/JFS13964J.
48. Panagiotakopulu E, Buckland PC. Forensic archaeoentomology – an insect fauna from a burial in York Minster. Forensic Sci Int 2012;221 (1–3):125–30. https://doi.org/10.1016/j.forsciint.2012.04.020.
49. Wankel H. Beiträge zur österreichischen grotten-fauna. [Contributions to the Austrian grotto fauna]. Sitzber K Akad Wiss Wien Math-naturw Kl 1861;43:251–64.
50. Várfalvyová D, Stanko M, Miklisová D. Composition and seasonal changes of mesostigmatic mites (Acari) and fleas fauna (Siphonaptera) in the nests of *Mus spicilegus* (Mammalia: Rodentia). Biologia 2011;66 (3):528–34. https://doi.org/10.2478/s11756-011-0050-1.
51. Costa M. The mesostigmatic mites associated with *Copris hispanus* (L.) (Coleoptera, Scarabaeidae) in Israel. J Linn Soc Lond Zool 1963;45 (303):25–45. https://doi.org/10.1111/j.1096-3642.1963.tb00485.x.
52. Samsinak K. Acaros en moscas de la familia Spaeroceridae. II [Mites on flies of the family Sphaeroceridae. II]. Acarologia 1989;30(2):85–105.
53. Kirk AA. The effect of the dung pad fauna on the emergence of *Musca tempestiva* [Dipt.: Muscidae] from dung pads in southern France. Entomophaga 1992;37(4):507–14. https://doi.org/10.1007/BF02372320.
54. Perez-Martinez S, Moraza ML, Salona-Bordas MI. Gamasina mites (Acari: Mesostigmata) associated with animal remains in the mediterranean region of Navarra (Northern Spain). Insects 2019;10(1):5. https://doi.org/10.3390/insects10010005.
55. Bajerlein D, Błoszyk J, Gwiazdowicz D, Ptaszyk J, Halliday B. Community structure and dispersal of mites (Acari, Mesostigmata) in nests of the white stork (*Ciconia ciconia*). Biologia 2006;61(5):525–30. https://doi.org/10.2478/s11756-006-0086-9.
56. Lundqvist L. Phoretic Gamasina (Acari) from Southern Sweden: taxonomy, host preferences and seasonality. Acarologia 1998;39(2):111–4.
57. Buck M. Sphaeroceridae (Diptera) reared from various types of carrion and other decaying substrates in Southern Germany, including new faunistic data on some rarely collected species. Eur J Entomol 1997;94(1):137–51.
58. Payne JA, King EW, Beinhart G. Arthropod succession and decomposition of buried pigs. Nature 1968;219(5159):1180–1. https://doi.org/10.1038/2191180a0.
59. Anton E, Niederegger S, Beutel RG. Beetles and flies collected on pig carrion in an experimental setting in Thuringia and their forensic implications. Med Vet Entomol 2011;25(4):353–64. https://doi.org/10.1111/j.1365-2915.2011.00975.x.
60. Petersen H, Luxton M. A comparative analysis of soil fauna populations and their role in decomposition processes. Oikos 1982;39(3):287–388. https://doi.org/10.2307/3544689.
61. Tempfli B, Szabó Á, Ripka G. New records of tydeid, phytoseiid and tenuipalpid (Acari: Tydeidae, Phytoseiidae, Tenuipalpidae) mites from Hungary. Acta Phytopathol Entomol Hung 2014;49(2):275–9. https://doi.org/10.1556/APhyt.49.2014.2.14.
62. Solarz K, Szilman P, Szilman E, Krzak M, Jagła A. Some allergenic species of astigmatid mites (Acari, Acaridida) from different synanthropic environments in southern Poland. Acta Zool Cracov 2004;47(3–4):125–45. https://doi.org/10.3409/173491504783995843.
63. Mehl R. Occurrence of mites in Norway and the rest of Scandinavia. Allergy 1998;53(48):28–35. https://doi.org/10.1111/j.1398-9995.1998.tb04993.x.
64. Lesna I, Sabelis M, Bolland H, Conijn C. Candidate natural enemies for control of *Rhizoglyphus robini* Claparede (Acari: Astigmata) in lily bulbs: exploration in the field and pre-selection in the laboratory. Exp Appl Acarol 1995;19(11):655–69. https://doi.org/10.1007/BF00145254.
65. Solarz K. The allergenic fauna of house-dust mites in some Silesian towns. Wiad Parazytol 1986;32:431–3. https://doi.org/10.1007/BF00145254.
66. Luxton M. Patterns of food intake by some astigmatic mites of beech woodland soil (Acari: Astigmata). Pedobiologia 1995;39(3): 238–42.
67. Hallas TE, Korsgaard J. Annual fluctuations of mites and fungi in Danish house dust – an example. Allergol Immunopathol (Madr) 1983;11 (3):195–200.
68. Wilkin DR, Murdoch G, Woodville HC. Chemical control of mites infesting freesia corms and Narcissus bulbs. Ann Appl Biol 1976;82 (1):186–9.
69. Wasylik A. The mites (Acarina) of potato and rye fields in the environs of Choryn. Pol Ecol Stud 1975;1:83–91.
70. Hughes AM. The mites associated with stored food products. London, U.K.: Ministry of Agriculture and Fisheries/Her Majesty's Stationery Office, 1948; Technical Bulletin 9.
71. Michael AD. British tyroglyphidae, vol. II. London, U.K.: Ray Society, 1903;60–151.
72. Barczyk G, Madej G. Comparison of the species composition of Gamasina mite communities (Acari: Mesostigmata) in selected caves of the Kraków-Częstochowa Upland (southern Poland) and their immediate surroundings. J Nat Hist 2015;49(27–28):1673–88. https://doi.org/10.1080/00222933.2014.976667.
73. Skubała P, Dethier M, Madej G, Solarz K, Mąkol J, Kaźmierski A. How many mite species dwell in subterranean habitats? A survey of Acari in Belgium. Zool Anz 2013;252(3):307–18. https://doi.org/10.1016/j.jcz.2012.09.001.
74. Fabri R. Invertebrati della Grotta del Re Tiberio, di altre cavità naturali attigue e della cava di Monte Tondo [Invertebrates of the Grotta del Re Tiberio, of other adjacent natural cavities and of the quarry of Monte Tondo]. In: Ercolani M, Lucci P, Piastra S, Sansavini B, editors. I gessi e la cava di Monte Tondo. Studio multidisciplinare di un'area carsica nella vena del gesso romagnola [The chalks and the quarry of Monte Tondo. Multidisciplinary study of a karst area in the vein of Romagna plaster]. Vol. II. Bologna, Italy: Memorie dell'Istituto Italiano di Speleologia, 2013;303–34.
75. Minodora M, Stelian I. Characterisic soil mite's communities (Acari: Gamasina) for some natural forests from Bucegi Natural Park-Romania. Period Biol 2014;116(3):303–12.
76. Minodora M. Predatory mites (Acari: Mesostigmata-Gamasina) from soil of some spoilt areas from retezat and tarcu-petreanu mountains. Studia Univ VG SSV 2010;20:89–94.
77. Fend'a, P. Mites (Mesostigmata) inhabiting bird nests in Slovakia (Western Carpathians). In: Sabelis MW, Bruin J, editors. Trends in Acarology. Proceedings of the 12th International Congress. Dordrecht, the Netherlands: Springer, 2010;199–205.
78. Fend'a P, Ciceková J. Soil mites (Acari, Mesostigmata) of oak-hornbeam forest in NR Katarínka, Southwest Slovakia. In: Schlaghamersky K, Pizl V, editors. Contributions to soil zoology in Central Europe III. České Budějovice, Czech Republic: Inst Soil Biol, Acad Sci, 2009;29–32.
79. Haitlinger R. Arthropods (Siphonaptera, Anoplura, Acari) of small mammals of Karkonosze Mts.(Sudetes). Zesz Nauk UP Wroc, Biol Hod Zwierz 2007;55(559):23–43.
80. Fend'a P, Schniererová E. Mites (Acarina, Gamasida) in littoral zone of Jakubov fishponds (Slovakia). In: Tajovsky K, Schlaghamersky J, Pizl V, editors. Contributions to soil zoology in Central Europe I. České Budějovice, Czech Republic: Inst Soil Biol, Acad Sci, 2005;9–14.
81. Fend'a P, Košel V. Mites (Acarina: Mesostigmata) inhabiting caves of the Belianske Tatry Mts (Northern Slovakia). Biologia (Bratislava) 2004;59(15):35–40.

82. Kováč Ľ, Mock A, Ľuptáčik P, Košel V, Fend'a P, Svatoň J, et al. Terrestrial arthropods of the Domica cave system and the Ardovská cave (Slovak Karst) – principal microhabitats and diversity. In: Tajovsky K, Schlaghamersky J, Pizl V, editors. Contributions to soil zoology in Central Europe I. České Budějovice, Czech Republic: Inst Soil Biol, Acad Sci, 2005;61–70.

83. Mock A, Ľuptáčik P, Fend'a P, Svatoň J, Országh I, Krumpál M. Terrestrial arthropods inhabiting caves near Velky Folmar (Cierna hora Mts., Slovakia). In: Tajovsky K, Schlaghamersky J, Pizl V, editors. Contributions to soil zoology in Central Europe I. České Budějovice, Czech Republic: Inst Soil Biol, Acad Sci, 2005;95–101.

84. Lundqvist L, Hippa H, Koponen S. Invertebrates of Scandinavian caves IX. Acari: Mesostigmata (Gamasina), with a complete list of mites. Acarologia 2000;40(4):357–65.

85. Skorupski M, Luxton M. Mesostigmatid mites (Acari: Parasitiformes) associated with yew (*Taxus baccata*) in England and Wales. J Nat Hist 1998;32(3):419–39. https://doi.org/10.1080/00222939800770221.

86. Skalski A. Charakterystyka współczesnej fauny Szczeliny Chochołowskiej w Tatrach [Characteristic of present fauna in Szczelina Chochołowska in Tatry Mountains]. Pr Muz Ziemi 1967;36(11):281–7.

87. Bloszyk J, Gwiazdowicz DJ, Halliday B, Dolata PT, Goldyn B. Nests of the black stork *Ciconia nigra* as a habitat for mesostigmatid mites (Acari: Mesostigmata). Biologia 2009;64(5):962–8. https://doi.org/10.2478/s11756-009-0146-z.

88. Kristofik J, Masan P, Sustek Z. Arthropods (Pseudoscorpionidea, Acarina, Coleoptera, Siphonaptera) in nests of the bearded tit (*Panurus biarmicus*). Biologia 2007;62(6):749–55. https://doi.org/10.2478/s11756-007-0142-0.

89. Cuthbertson AGS, Murchie AK. The presence of *Anystis baccarum* (L.) in Northern Ireland bramley apple orchards. Ir Nat J 2004;27(12): 465–7.

90. Christian A. Colonization of primary sterile soils by epedaphic gamasina mites. In: Bernini F, Nannelli R, Nuzzaci G, De Lillo E, editors. Acarid phylogeny and evolution: adaptation in mites and ticks. Dordrecht, the Netherlands: Springer, 2002;169–73. https://doi.org/10.1007/978-94-017-0611-7_17.

91. Petrova V, Salmane I, Çudare Z. The predatory mite (Acari, Parasitiformes: Mesostigmata (Gamasina); Acariformes: Prostigmata) community in strawberry agrocenosis. Acta Uni Latv Biol 2004;676: 87–95.

92. Tryjanowski P, Baraniak E, Bajaczyk R, Gwiazdowicz DJ, Konwerski S, Olszanowski Z, et al. Arthropods in nests of the red-backed shrike (*Lanius collurio*) in Poland. Belg J Zool 2001;131(1):69–74.

93. Curry JP. The Arthropod fauna associated with cattle manure applied as slurry to grassland. P Roy Irish Acad B 1979;79(2):15–27.

94. Brady J. The mites of poultry litter: observations on the bionomics of common species, with a species list for England and Wales. J Appl Ecol 1970;7:331–48. https://doi.org/10.2307/2401384.

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14519
Available online at: onlinelibrary.wiley.com

CASE REPORT

PATHOLOGY/BIOLOGY

Guendalina Gentile,[1] *B.Sc.; Marta Bianchi*,[1] *M.D.; Michele Boracchi*,[1] *M.D.; Carlo Goj*,[1] *M.D.; Stefano Tambuzzi*,[1] *M.D.; and Riccardo Zoja*,[1] *M.D., Ph.D.*

Forensic Pathological Considerations of a Unique Case of "Complicated Suicide"*,†

ABSTRACT: In the forensic literature, peculiar and uncommon cases of suicides defined as "complicated" are reported. In these circumstances, the suicide method chosen by the victim fails, and death occurs due to a subsequent unforeseen accidental event defined as secondary trauma. Through retrospective examination of 25,512 autopsies in 27 years (1993–2019) at the Bureau of Legal Medicine of Milan, a unique case of complicated suicide was identified from a total of 4497 suicides. It concerns an elderly man who, after killing his wife by inflicting incised wounds to her neck, tried to hang himself by tying a rope to a heater and jumping from the window located over the heater itself. However, the rope suddenly snapped and the man fells to the ground causing fatal traumatic injuries. Death occurred because of an accidental event caused by the failure of the hanging mechanism. Therefore, a peculiar yet characteristic case of complicated suicide is described.

KEYWORDS: forensic pathology, complicated suicide, complex suicide, hanging, fall from height, autopsy, homicide–suicide

A "planned complex suicide" is a very uncommon suicide event (1) in which multiple detrimental methods are used in order to avoid the eventual ineffectiveness of one of them (2), thereby achieving a guaranteed fatal outcome (3,4). In 1974, planned complex suicides were classified as "primary" and "secondary" depending on the application sequence of the chosen method (5), to specify their simultaneous use or their chronological succession, respectively. These suicides were then also defined as "planned" if at least two methods involved were previously established. In contrast, "unplanned" suicides are events in which the victim—still conscious and able to act after the failure of the first chosen suicidal modality—resorts to an alternative and improvised method of death (6,7). This eventuality may also occur if the chosen suicidal methods turn out to be too painful or slow in causing death (8). The term "complicated suicide" refers to rare suicides in which the failure of the initial method chosen by the victim is followed by an accidental trauma, which is different from that planned and is therefore unintentional. This type of trauma is considered a secondary fatal complication of the suicide gesture (9). Although these peculiar forms of suicides are occasionally reported in the literature (9,10), they do not present typical well-defined characteristics; it is for this reason that they are often challenging to deal with. A crucial issue remains to establish the manner of death, that is, whether it was really due to suicide, or was an accidental event, or even homicide (9).

The authors present the unusual medico-legal characteristics of a unique case of complicated suicide, which came to the attention of the Bureau of Legal Medicine of Milan throughout 27 years of autopsies from 1993 to 2019.

Case Report

A 70-year-old man was found dead on the landing of the external steps behind a building, lying on his right side with a snapped rope around his neck. During the police on-site inspection at the victim's apartment, there was no sign of forced entry and the key was in its normally inserted position in the keyhole from the inside. Inside the apartment, the police found the body of a woman in a supine position lying in a remarkably large pool of blood with a blood-stained kitchen knife beside her. The woman was identified as the wife of the elderly man and showed multiple cutting injuries to the arms and in the anterior cervical region. An on-site inspection revealed the presence of one end of a rope fastened to a heater below a window, with the other end hanging loose outside the window. The rope was 155 cm long and the loose end appeared to be frayed. Viewed from the window, the body of the man could be seen on the landing of the external steps.

The neighbors questioned by the police declared that the woman used to spend many hours every day playing video-poker at the nearby bar. That same day, they had heard the couple arguing due to the woman's long-term gambling addiction.

[1]Dipartimento di Scienze Biomediche per la Salute, Sezione di Medicina Legale e delle Assicurazioni, Università degli Studi di Milano, via Luigi Mangiagalli, 37, Milano, 20133, Italy.

Corresponding author: Riccardo Zoja, M.D., Ph.D. E-mail: riccardo.zoia@unimi.it

*This article does not contain any studies with living human participants performed by any of the authors. All the studies conducted followed the guidelines provided by Legislation and the National Bioethical Committee.

†The author(s) declared that the Investigating Magistrate authorized all the forensic investigations performed in this study, in accordance with the Italian Law.

[Correction added 12 August 2020. The Highlights section was removed.]

Received 16 Mar. 2020; and in revised form 21 June 2020; accepted 6 July 2020.

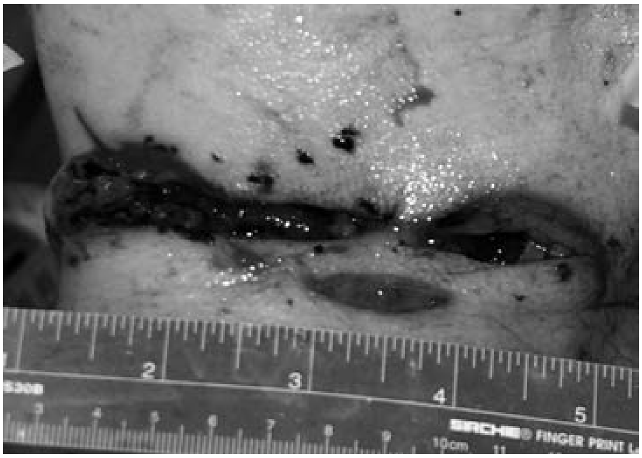


FIG. 1—*Macroscopic overview of the woman's incised wounds in the cervical region. [Color figure can be viewed at wileyonlinelibrary.com]*

They also claimed that over the past few months quarrels used to happen almost every day.

It was hypothesized that at the height of yet another dispute, the man had killed his wife by inflicting multiple incised wounds with the kitchen knife and shortly after had tried to commit suicide by hanging. It appeared that the man tied one end of the rope to the heater and the other end around his neck, and then jumped off the window ledge above the heater. However, the rope must have snapped unexpectedly causing the man to free-fall from a height of 15 m to the landing of the external steps, where he was later found. In order to clarify the circumstances surrounding the deaths, the investigating magistrate immediately ordered the judicial autopsy at the Bureau of Legal Medicine of Milan.

Autopsy of the Woman

The body of a Caucasian woman aged 66 years was well nourished and in a good state of preservation (weight: 66 kg; height: 159 cm), with light hypostasis. At external examination, both hands showed several cutting injuries with regular margins that were characterized by hemorrhagic infiltration separation. From the medial right anterior cervical region to the left latero-cervical region, a 12-cm area of three wide and deep cutting injuries of the soft tissues was evident, exposing the underlying muscular–fascial structures. These cutting injures showed fine, irregular margins, infiltrated with blood and with a maximum separation of 3 cm (Fig. 1).

At dissection, a bilateral hemorrhagic infiltration of the sternocleidomastoid, sternothyroid, and thyrohyoid muscles was evident. Also, there was a linear fracture of the upper horns of the thyroid cartilage of the larynx with hemorrhagic infiltration of the tissue. This lesion was considered as an incised wound due to the incisions of the neck. Full-thickness incision of the muscular–fascial structures and the thyrohyoid membrane to the vertebral level was detected and showed hemorrhagic infiltration of the soft tissues, irregular margins, and a maximum separation of 0.5 cm. The superior edge of the laryngeal shield, on the left, showed a full-thickness cutting injury that included the internal jugular vein, which was completely transected; the margins of the vein were irregular and infiltrated with blood. No other significant findings were noted at autopsy; therefore, death was attributed to incised wounds of the neck.

Autopsy of the Man

The body of the man appeared well nourished and in a good state of preservation (weight: 78 kg; length: 175 cm) with normal intensity hypostasis. At external examination, several bruises and abrasions were noted on the face, trunk, and upper and lower limbs bilaterally, as well as some wounds with hemorrhagic margins at the trunk and the lower left limb. In the right latero-cervical area, a whitish linear streak (5 cm long) resembling a slight cutaneous furrow was detected, and in the left area, a linear abrasion (4.2 cm) surrounded by a cutaneous reddish bruise was observed (Fig. 2). At autopsy, various injuries were revealed: (i) deep hemorrhagic infiltration of the right parietal and temporo-occipital scalp areas with the involvement of the homolateral temporal muscle; (ii) a bilateral parieto-occipital and cerebellar subarachnoid hemorrhage; (iii) blood in the cerebral ventricles; (iv) hemorrhagic infiltration of the sternocleidomastoid, sternothyroid, and thyrohyoid muscles; (v) bilateral fractures of the upper horns of the laryngeal thyroid cartilage; (vi) multiple bilateral displaced rib and clavicle fractures with perilesional hemorrhagic infiltration; (vii) blood in the peritoneal cavity and hemorrhagic infiltration of the omentum and mesenteric tissues; (viii) right renal intra-capsular hematoma with full-thickness renal vein laceration; (ix) fracture of bilateral sacroiliac joints; and (xi) fracture of the right ilio-pubic bone with perilesional hemorrhagic infiltration. The spinal column was free of injuries. Death was attributed to skeletal and visceral injuries due to falling from height.

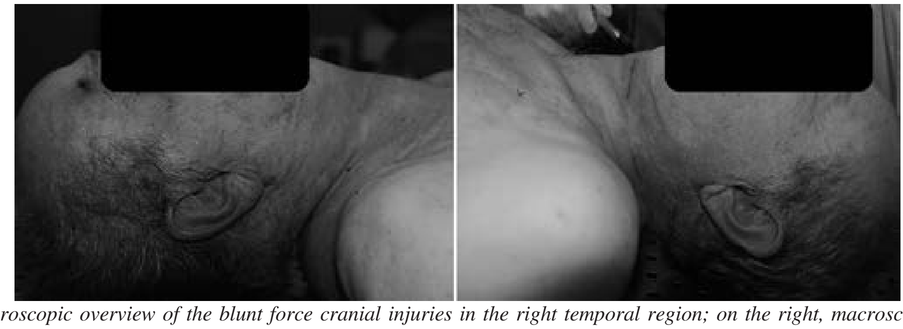


FIG. 2—*On the left, macroscopic overview of the blunt force cranial injuries in the right temporal region; on the right, macroscopic overview of the ligature abrasion. [Color figure can be viewed at wileyonlinelibrary.com]*

During both autopsies, biological fluids (blood, urine, bile, and gastric content) and different samples of viscera (brain, lungs, heart, liver, spleen, kidneys, and skin) were collected for subsequent toxicological and histological examination, as ordered by the inquiring magistrate.

Toxicological Analyses

The inquiring magistrate authorized the analysis of all the biological fluids sampled during the autopsy examinations, but only of the samples of brain, lungs, liver, and kidneys as far as the viscera. All the analyses were carried out at the Laboratory of Forensic Toxicology of the University of Milan. The samples were tested in order to identify the presence of illicit drugs, alcohol, or other substance with pharmacological activity. The man had a blood alcohol content (BAC) of 0.8 g/dL; the woman had a BAC of 0.22 g/dL and had traces of caffeine and cotinine.

Histological Examinations

The inquiring magistrate authorized the analysis of all the tissue samples sampled from both bodies, which were fixed in 10% buffered formalin. The samples were later subjected to postfixative techniques, and the histological slides obtained were stained with hematoxylin and eosin, Masson's trichrome, and Resorcin–Fuchsin staining. Finally, they were observed under a Leica DME optical microscope, and the most significant images were acquired with a Leica DC300F digital camera. The microscopic observation demonstrated the vitality of the injuries localized at the woman's neck and of the man's blunt and asphyxial lesions.

Discussion

In most suicides, only one method is used—a simple suicide; however, in a small percentage of cases (1.5%–5%) (9), suicides occur by multiple methods—complex suicides. Forensic literature has reported suicidal cases in which up to six different methods were applied (11). The strict division between simple and complex suicides is not always clear. Indeed, peculiar forms of suicide are known in which the first method chosen by the victim fails and death occurs accidentally due to a secondary, external, and unpredictable complication. These very uncommon events—complicated suicides—differ from planned and unplanned complex suicides and are generally challenging to deal with in the field of forensics (2) since the manner of death can remain unclear and ambiguous. Indeed, even after a careful on-site inspection at the death scene and a thorough autopsy examination of the body, interpretative difficulties may persist (11). In these cases, the main issue that a forensic pathologist must tackle concerns the establishment of the manner of death: suicide, homicide, or accidental death (9).

In the case presented, the crime scene was particularly complex. The murdered woman was lying on the floor inside her house without any sign of forced entry and with incised wounds on her hands. With the incisions being different in direction, shape, and depth, they were attributed to the woman's attempt to defend herself during the assault. Her husband was found dead on the landing of the external steps, 15 meters under the window of the same apartment, which was wide open. A snapped rope was found around the man's neck, and no sharp-force injuries were detected on his hands and upper limbs.

The forensic on-site inspection at the apartment was crucial, identifying features strongly suggestive of an attempted suicidal hanging by the man. In particular, a piece of snapped rope still tied to a heater and dangling outside through a wide-open window over the heater itself was seized. The rope turned out to be similar in composition, shape, and size to the broken piece of rope found around the man's neck. Both the snapped ends of the rope were characterized by a comparable fraying of the fibers, which was consistent with a spontaneous break; there were no clean cuts of the rope. The rest of the rope appeared intact in its structure and was devoid of any evidence of tampering, making it unlikely that this was a murder concealed as a suicide. Therefore, the conclusion was that the man's death was caused by an accidental event due to the sudden breaking of the rope with which the man had planned to hang himself. The autopsy examination confirmed multiple blunt, traumatic injuries, especially located at the head, as well as neck injuries related to hanging. All the above-mentioned injuries to the man, as demonstrated by histological examination, were characterized by hemorrhagic infiltration, and thus, all occurred when the man was still alive. In particular, the possibility that the fall from height had occurred postmortem following a rope break after the man had died from asphyxiation, was deemed unlikely. With this evidence, we concluded that the almost simultaneous application of two different detrimental means had contributed to the man's death. At autopsy, the traumatic blunt cranial injury was so prominent compared to the other injuries that it was a reasonable hypothesis that the cause of death was attributable to skeletal and visceral lesions due to a fall from height.

This case is an example of suspicious death, in which there is uncertainty regarding the circumstances of the death itself and the relatively equal probability that two or more manner of death exist (12). In order to reconstruct what happened, the synergy between the autopsy data obtained by the forensic pathologist and the circumstantial and investigative information collected by the police officers becomes essential. The latter, in fact, played a crucial role by providing valuable information that allowed us to identify the case as murder–suicide. This conclusion was corroborated by the findings of the autopsy examination. We emphasize that in all forensic cases, especially in complex ones such as the one presented, it is important to integrate the complete investigative and autopsy information to make a determination of cause and manner of death.

References

1. Altun G. Planned complex suicide: report of three cases. Forensic Sci Int 2006;157(2–3):83–6. https://doi.org/10.1016/j.forsciint.2005.04.039.
2. Blanco-Pampín JM, Suárez-Peñaranda JM, Rico-Boquete R, Concheiro-Carro L. Planned complex suicide. An unusual suicide by hanging and gunshot. Am J Forensic Med Pathol 1997;18(1):104–6. https://doi.org/10.1097/00000433-199703000-00020.
3. Bohnert M, Rothschild MA. Complex suicides by self-incineration. Forensic Sci Int 2003;131(2–3):197–201. https://doi.org/10.1016/s0379-0738(02)00449-8.
4. Türk EE, Anders S, Tsokos M. Planned complex suicide. Report of two autopsy cases of suicidal shot injury and subsequent self-immolation. Forensic Sci Int 2004;139(6):35–8. https://doi.org/10.1016/j.forsciint.2003.09.013.
5. Marcinkowski T, Pukacka-Sokolowska L, Wojciechowski T. Planned complex suicide. Forensic Sci 1974;3(1):95–100. https://doi.org/10.1016/0300-9432(74)90013-2.
6. Racette S, Sauvageau A. Planned and unplanned complex suicides: a 5-year retrospective study. J Forensic Sci 2007;52(2):449–52. https://doi.org/10.1111/j.1556-4029.2007.00387.x.
7. Cingolani M, Tsakri D. Planned complex suicide: report of three cases. Am J Forensic Med Pathol 2000;21(3):255–60. https://doi.org/10.1097/00000433-200009000-00015.
8. Bohnert M, Pollak S. Complex suicides – a review of the literature. Arch Kriminol 2004;213(5–6):138–53.

9. Töro K, Pollak S. Complex suicide versus complicated suicide. Forensic Sci Int 2009;184(1–3):6–9. https://doi.org/10.1016/j.forsciint.2008.10.020.

10. Barranco R, Diana C, Ventura F. Forensic pathological study of complex and complicated suicides: a twelve-year case series in Genoa (Italy). J Forensic Leg Med 2019;65:5–8. https://doi.org/10.1016/j.jflm.2019.04.007.

11. Petković S, Maletin M, Durendić-Brenesel M. Complex suicide: an unusual case with six methods applied. J Forensic Sci 2011;56(5):1368–72. https://doi.org/10.1111/j.1556-4029.2011.01821.x.

12. Lacks RD, Westveer AE, Dibble A, Clemente J. Equivocal death investigation: case study analyses. Vict Offender 2008;3(2–3):150–64. https://doi.org/10.1080/15564880801938292.

doi: 10.1111/1556-4029.14526
Available online at: onlinelibrary.wiley.com

CASE REPORT

PATHOLOGY/BIOLOGY

Rayana A. Costa,[1] *M.Sc.; Nayara A. dosSantos,*[1,2] *M.Sc.; Thayná S. M. Corrêa,*[3] *B.Sc.; Nathália L. P. Wyatt,*[1] *B.Sc.; Carlos A. Chamoun,*[4,5] *Ph.D.; Maria T. W. D. Carneiro,*[1] *Ph.D.; and Wanderson Romão,*[1,2,3] *Ph.D.*

Detection of Pb, Ba, and Sb in Cadaveric Maggots and Pupae by ICP-MS*

ABSTRACT: The concentrations of lead (Pb), barium (Ba), and antimony (Sb), characteristic of GSR, were determined in soil sediments and immature (larvae) of cadaveric flies of the family Calliphoridae, by inductively coupled plasma mass spectrometry (ICP-MS). This research refers to a case study from two real crime scenes in which the corpses were in an advanced state of decomposition. In case 1, the victim had holes similar to gunshot wounds, and in case 2, there was no evidence of perforations in the corpse. Soil sediment collection was performed at three different points of the terrain, at a minimum distance of 10 m from the corpse, for cases 1 and 2. In relation to the collection of immatures, larvae were collected in regions of the mouth, nose, and orifices similar to the entry of firearms projectile into the body, for case 1, and collection of larvae and pupae, located on the body and underneath it, for case 2. It was possible to detect and quantify the three elements of interest (Pb, Ba, and Sb) by ICP-MS in both sediment and cadaveric larvae. Concentrations of 4.44, 8.74, and 0.08 µg/g were obtained for Pb, Ba, and Sb, respectively, in the soil for case 1. For the case 2, the concentrations in Pb, Ba, and Sb were from 16.34 to 26.02 µg/g; from 32.64 to 57.97 µg/g and from 0.042 to 0.30 µg/g, respectively. In the larvae, Pb, Ba, and Sb were quantified in cases 1 and 2 with a concentration of 6.28 and 1.78 µg/g for Pb, 1.49 and 2.94 µg/g for Ba, 0.50 µg/g and <LD for Sb, respectively. These new results present the detection of characteristic elements of GSR in cadaveric larvae in humans in a real crime scene, besides highlighting the importance of the study of immature flies, using the ICP-MS technique in forensic analysis.

KEYWORDS: forensic entomology, gunshot residue, inductively coupled plasma mass spectrometry, soil sediment, larvae

Based on the Brazilian scenario composed of high rates of crime and violence, homicide cases are on an upward scale. According to the Atlas of Violence, developed by the Instituto de Pesquisa Econômica Aplicada (IPEA), in 2017, 72% of homicides recorded in Brazil were caused by the use of firearms (1).

The alarming data show the importance of forensic sciences, especially subareas such as ballistics and forensic entomology, in clarifying and technical proof regarding the occurrence of crimes. Ballistics focuses on the study of firearms, ammunition, and the effects caused by shooting, in order to elucidate crimes, especially homicides, since the injuries caused by the use of firearms vary according to the type of weapon, ammunition, and the distance of the shot (2). Forensic entomology, in turn, consists of the study of insects and other arthropods associated with criminal issues to determine the cause of death (3,4), either homicide, suicide, accidents, or natural causes, besides enabling the estimation of PMI – postmortem interval (5,6).

Entomology is not only based on the morphological characteristics of insects, but also in the identification of specific insect species found in the corpse and in studies their development stages such as larvae time of occurrence and pupae of flies, thus allowing to estimate efficiently the moment when death occurred (7).

However, in real cases, in which the corpse is found in an advanced state of decomposition, some factors may make it even more difficult to obtain clarifications about the crime, such as the place of discovery of the body and the conditions of concealment of the corpse (8). Thus, evaluations of the cause of death by analyses with classical methodologies can be complex, since they require adequate conditions as well as the presence of solid tissues, because visual analysis of the bullet wound can be compromised (9–11).

These difficulties are even more evident in cases of firearm homicides, for example, when changes occur at the scene of the crime, concealment of the corpse, activity of insects in or around the wound tract, decomposition of the body, since they interfere in the quantification of GSR, which consists of a small amount of material expelled from a projectile during a shot and which are crucial to determine the conditions of the crime. The action of burying and the decomposition of the body can make less clear the tattoo or residue of obvious shots. On the other hand,

[1]Department of Chemistry, Federal University of Espírito Santo, Vitória, ES, 29075-910, Brazil.
[2]National Institute of Forensic Science and Technology (INCT Forensics), Vitória, ES, 29075-910, Brazil.
[3]Federal Institute of Espírito Santo, Vila Velha, ES, 29106-010, Brazil.
[4]Federal Institute of Espírito Santo, Viana, ES, 29135-000, Brazil.
[5]Department of Criminology, Superintendence of Technical and Scientific Police of Espírito Santo, Vitória, ES, 29045-402, Brazil.
Corresponding author: Wanderson Romão, Ph.D. E-mail: wandersonromao@gmail.com

*Funding provided by the Conselho Nacional de Desenvolvimento Científico e Tecnológico-CNPq (422555/2018-5, and 305359/2017-7).
[Correction added 11 August 2020. The Highlights section was removed.]
Received 27 April 2020; and in revised form 8 July 2020; accepted 10 July 2020.

the activity of insects can supplant existing tracts or even create new ones by altering the morphology of the wound, consequently identify and quantify the residues in a gunshot wound by firearm in a corpse in a state of decomposition is extremely important (8).

These residues come from the primer compound present in the ammunition cartridge and have characteristic inorganic components, lead (Pb), barium (Ba), and antimony (Sb), which are detectable (8). With the action of interfering factors, it is necessary to use analytical methodologies such as inductively coupled plasma optical emission spectroscopy (ICP OES) (12–14) in place of traditional qualitative techniques, such as colorimetric assays (Feigl–Suter reaction) in the recognition of GSR, since this test has the limitation of detecting only one of the characteristic components, in this case Pb (15). In addition, contamination with other materials containing Pb may induce the production of false-positive results (16,17).

The ICP OES and ICP-MS technique are promising in the detection of inorganic GSR of inorganic nature (8,12–14,18,19) due to its high sensitivity in the detection of trace elements because they are multielementary, simple, fast, and cost-effective compared to the most common techniques for this type of analysis, such as dispersive energy spectroscopy coupled with scanning electron microscopy (SEM/EDS) (17,20) and X-ray fluorescence (XRF) (21).

The SEM/EDS technique is used for GSR analysis and defended by many criminal experts because it relates particle morphology to chemical composition (22). LaGoo *et al.* (2010) evaluated wound samples caused by firing a firearm in the period of late summer (days 1, 2, 5, and 8) and during winter (days 1, 2, 5, and 44). The authors reported that unreliable results were obtained by SEM/EDS after day 1 for both periods. Thus, the SEM/EDS technique presents a certain limitation for the samples of decomposing tissue, may be affected by environmental conditions (rains) in addition to the nature of the wounds (oily) that prevents proper collection using the adhesive stub method. LaGoo *et al.* (8) concluded that ICP-MS has potential in the chemical identification of GSR in larvae of flies and tissues in addition to detecting of GSR in several advanced stages of decomposition and environmental conditions not affecting the results.

In 2003, Roeterdink *et al.* (18) investigated the detection Pb, Ba, and Sb by ICP-MS in fly larvae present in contaminated beef in a closed environment and under controlled conditions. LaGoo *et al.* (8), in 2010, conducted a study for the detection of GSR in larvae of flies and swine tissues outdoors according to climatic variations during the pig decomposition process, using the ICP-MS technique. The authors concluded that digestion of larvae and tissues may be of great importance for forensic pathology in identifying suspected firearm injuries in an advanced stage of decomposition. In 2015, Motta *et al.* (14), quantified Pb, Ba, and Sb in GSR through larvae of cadaveric flies, of the family Calliphoridae, collected in the corpse of a pig hit by firearms, simulating a real case, through the analytical technique of ICP OES.

From the remarkable evolution of the application of the ICP technique in entomological studies, the present case study brings a real approach of this remarkable analytical technique for detection of Pb, Ba, and Sb, characteristic elements of GSR, in two homicide crime scenes, one located in a city 80 km away and another 28 km from the city of Vitória, capital of the state of Espírito Santo, Brazil, through the analysis of larvae and pupae of cadaveric flies of the family Calliphoridae, collected from human corpses, in an advanced stage of decomposition, during forensic examinations, with one of the victims buried in a shallow excavation, <1.0 m deep, and another victim not buried, already in a final state of decomposition, that is, in skeletonization.

Experimental Procedure

Materials and Reagents

The analytes were extracted from the soil and immature larvae using nitric acid, HNO_3 P.A. (Synth, São Paulo, Brazil), previously purified in sub-boiling DestillAcid, model BSB-939-IR (Berghof, Germany), ultra pure water (18.2 MΩ resistivity) prepared by a reverse osmosis system (Purelab Ultra Mk2, U.K.), hydrogen peroxide 30% w/w P.A (H_2O_2; Sigma Aldrich, St. Louis, MO), and microwave, Multiwave GO model (Anton Paar, Ankerstraße 6, Austria). For the construction of the analytical curve, a stock solution standard (10,000 μg/L) of Pb (Inorganic Ventures, Christiansburg, VA) Ba, and Sb (Absolute Standards INC, Hamden, CT) was used. The detection of the trace elements was performed using an ICP-MS NexIon 300D equipment (PerkinElmer do Brasil Ltda, São Paulo, SP, Brazil).

Sample Preparation

Experimental Control

Case 1—Corpse of a male individual, buried in shallow excavation (<1.0 m deep) in a forest region of the Atlantic forest type and fairly irregular terrain. For the experimental control, soil sediments were collected at three distinct points in the terrain, at a distance of 10.0 m from the corpse.

Case 2—Corpse of male individual, not buried, found in the final stage of decomposition, already in skeletonization, in a very wide and abandoned terrain, with several points of rubble of works and garbage. The experimental control of the soil followed the same criteria of case 1, mentioned above.

The sediment samples, in both cases, were dried for 72 h at 60°C and then submitted to acid decomposition assisted by microwave radiation, were added 10.0 mL of HNO_3 P.A. to 0.200 g of sample. The mixture was subjected to a heating program according to the EPA 3051a method. After cooling, the samples were increased to 15.0 mL with ultrapure water.

Entomological Samples

Case 1—Larvae samples were collected from the corpse's natural orifices (mouth, nose, and ears), from openings in epithelial tissue, similar to gunshot firearm projectile, on the body, and around the point where it was buried.

Case 2—The samples of larvae and pupae were collected on the corpse (between the clothes) and underneath it, which presented large fractures in the cephalic region similar to those caused by a blunt instrument.

All larvae, case 1 and 2, were stored in a polypropylene tube containing 70% alcohol and were then washed externally, with distilled water and 70% alcohol, twice, and thus sent to the laboratory for analysis, where they were dried for 24 h at 60°C, macerated and submitted to microwave-assisted acid decomposition.

TABLE 1—*Operating conditions for GSR analysis by ICP-MS.*

Parameter	Operating Conditions
Plasma gas flow rate (L/min)	16
Nebulizer gas flow rate (L/min)	0.96
Auxiliary gas flow rate (L/min)	1
RF power (W)	1350
Spray chamber	Baffled Cyclonic, Quartz
Nebullizer type	Concentric (Meinhard Type C), Quartz
Torch	EasyGlide™, Quartz
Number of replicates	3
Isotopes	^{121}Sb, ^{138}Ba, ^{208}Pb
Internal standards	^{103}Rh

Were added in a decomposition vessel, 0.500 g of sample was weighed in which 6.0 mL of concentrated HNO_3 previously purified and 4.0 mL of H_2O_2 30% w/w P.A. The heating program consisted of a ramp of 10 min, and a stay time of 5 min, at 100°C. Subsequently, the samples were increased to 15.0 mL with ultrapure water.

ICP-MS

The analytical curve was constructed with six points, and the concentrations were from 0 to 10 µg/L. For evaluate the quantification of trace elements by method, we used the limit of detection (LOD) and limit of quantification (LOQ) according to the International Union of Pure and Applied Chemistry (IUPAC), Eqs 1 and 2. The LOD and LOQ were calculated from 15 readings of the preparation blank. The experimental conditions for analysis by ICP-MS can be seen in Table 1.

$$\mathrm{LOD} = 3 * s/a \tag{1}$$

$$\mathrm{LOQ} = 10 * s/a \tag{2}$$

Where, s = standard deviation and a = slope.

Results and Discussion

The criminal forensics team cannot accurately infer from a simple visual analysis the amount and types of injuries in the corpse, due to the corpse conditions and the location that was found, which made it difficult to verify more specifically possible injuries at the scene for the expert record. Thus, it was necessary, a more in-depth study to conclude the facts.

Tests inserted in the routine of the technical scientific police for the detection of GSR, such as colorimetric test using sodium rhodizonate reagent (Feigl–Suter reaction) can introduce false-negative and false-positive results, because it detects the presence of Pb, even if it originates from other materials and or activities that have contact with this element (17,23). Hence, this test enables the achievement of unreliable results, which require more sensitive complementary techniques, for possible scientific proof of the case, such as ICP-MS, since it is used for GSR analysis since 1998, where it was first used by the laboratory of the Federal Bureau of Investigation (FBI). Since then, it has been widely used for quantitative analysis of GSR (24–27).

The literature presents several studies that detect or quantify GSR, either in body parts or clothing, which can be used to aid criminal expertize to elucidate crimes (27–29). Research also points to the use of larvae in the determination of Pb, Ba, and Sb, from GSR, in tissue samples and corpse surfaces associated with sensitive analytical techniques, showing the potential of the use of larvae for the observation/detection of Pb, Ba, and Sb (8,18,19,30).

Figure 1 shows a photo of the larvae of cadaveric flies of the Calliphoridae family used for the present study. Motta *et al.* (14), indicated that this species of fly upon contact with regions of firearm injuries feed and ingest residues primers, or GSR, characteristic of firearm shooting.

The concentrations of Pb, Ba, and Sb in sediment were calculated in order to evaluate whether the residues and primers characteristic of gunpowder would also be present in the soil, since these three associated elements are not naturally found in the earth (31).

For method verification, 15 successive readings of the preparation blank were obtained. The LOD and LOQ, Eqs 1 and 2, were calculated for sediment and immatures of each element (Pb, Ba, and Sb) listed in Table 2. It is observed that considerably low values were obtained by the ICP-MS technique for both sediments, used in the control (LOD from 0.0612 to 0.630 ng/L), as well as for larvae (LOD ranging from 0.087 to 5.7 ng/L). LaGoo *et al.* (8) evaluated the detection of GSR (Glock 9 mm) in larvae (*Phaenicia sericata*) collected during the summer and winter using the ICP-MS technique obtaining LOD values ranging from 0.017 to 0.106 ng/mL and LOQ from 0.10 to 1.0 ng/mL for the Pb, Ba, and Sb. Motta *et al.* obtained LODs in a range of 0.15 to 4.79 µg/L and LOQs ranging from 0.50 to 15.97 µg/L when analyzing *Crysomya albiceps* cadaveric larvae, by ICP OES (14). Duarte evaluated the estimation of the firing distance by means of the quantification of GSR, using fly larvae of the family Calliphoridae from contamination after gunshots (Glock 9 mm). In this study, LOD values ranged from 0.01 to 0.05 µg/L obtained through an ICP-MS and LOQ values were not reported (32). The LOD and LOQ values obtained in the present study are much lower than the previous reports, indicating the greater sensitivity of this method. It is noteworthy that all comparative studies that used larvae were collected from pig tissue, since this animal has similar decomposition to that observed in the human species.

The results obtained by sediment samples (Table 3) demonstrate considerable concentrations of Pb, Ba, and Sb, especially in case 2. These values may be associated with the characteristic of the site, since the land had a lot of garbage and several discharge points of construction debris, which can cause soil contamination.

In the detection of the elements in the larvae collected in the corpse, Pb, Ba, and Sb were observed and their concentrations are available in Table 4.

Generally, in shots fired at close range, GSR sits impregnated in the human skin targeted (12,13,17,22). Furthermore, studies show that the use of immature larvae for GSR detection may be influenced by larval stage period and climatic conditions. In the early stages of the larval period, immature larvae is more likely to feed on tissue on the body surface; consequently, they have higher intake of inorganic GSR, and therefore, high concentration of Pb, Ba, and Sb can be quantified.

It is observed that, in case 1, the results obtained by ICP-MS corroborate with the results issued in the expert report, since there is a higher concentration of Pb (6.28 ± 1.41 µg/g) in the larvae than in the sediment (4.44 ± 0.25 µg/g). This is also observed for Sb element, with a concentration of 0.08 ± 0.3 µg/g in the soil and 0.50 ± 0.032 µg/g in the larva, indicating that the larva ingested Sb in the process of decomposition of the

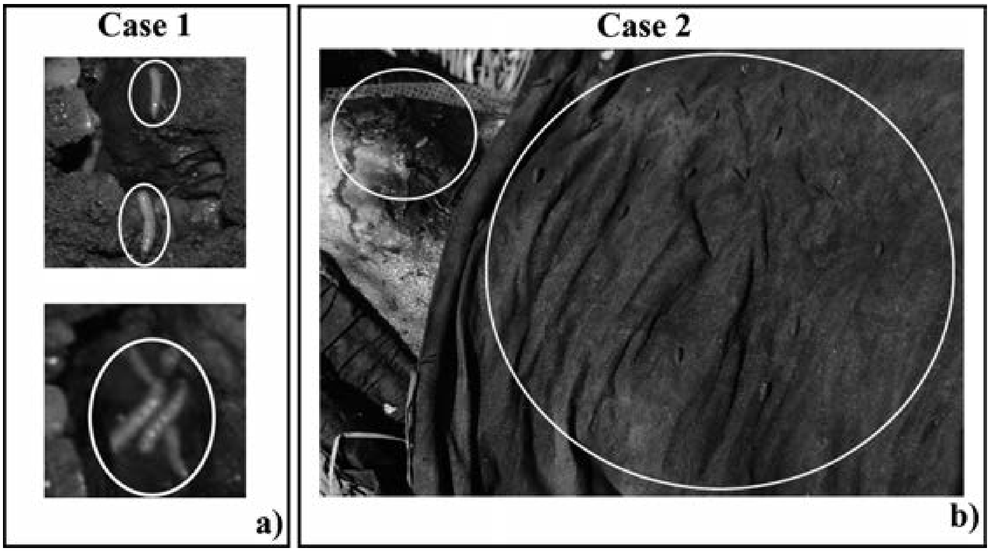


FIG. 1—*Immature flies of the family Calliphoridae: (a) Case1 – larvae collected in the region of the mouth, nose, ears, and body orifices similar to the entrance of firearm projectile (FAP); (b) Case 2 – larvae and pupae collected on the body and between garments.* [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2—*Limits of detection and quantification for sediment and larvae related to the cases under study.*

	Sediment		Larvae	
Elements	LOD (ng/L)	LOQ (ng/L)	LOD (ng/L)	LOQ (ng/L)
Case 1				
Pb	0.130	0.430	0.087	0.290
Ba	0.0929	0.310	0.230	0.760
Sb	0.0612	0.200	0.100	0.370
Case 2				
Pb	0.290	0.970	0.170	0.580
Ba	0.430	1.400	0.450	1.500
Sb	0.630	2.100	5.700	19.00

TABLE 3—*Concentration (μg/g) of Pb, Ba, and Sb in sediment samples obtained by ICP-MS.*

Concentration (μg/g)			
	Pb	Ba	Sb
Case 1			
Soil	4.44 ± 0.25	8.74 ± 0.53	0.08 ± 0.03
Case 2			
Soil 1	16.34 ± 0.12	32.64 ± 4.79	0.044 ± 0.01
Soil 2	7.98 ± 0.90	48.64 ± 8.20	0.042 ± 0.01
Soil 3	26.02 ± 9.73	57.97 ± 3.71	0.30 ± 0.08

TABLE 4—*Concentration (mg/kg) of Pb, Ba, and Sb in larvae samples obtained by ICP-MS.*

Concentration (μg/g)				
		Pb	Ba	Sb
Case 1	Larvae	6.28 ± 1.41	1.49 ± 0.34	0.50 ± 0.032
Case 2	Larvae	1.78 ± 0.40	2.94 ± 0.53	<LD

corpse, which had been shot by gunshots at close range. For the Ba element, a decrease is observed when we evaluated the concentration in the larva and sediment, this can be explained by the reason that Ba is naturally found in sedimentary rock, matching with higher concentrations in the soil (33). To evaluate the accuracy of quantification/detection, tests were performed with analyte additions and it was found that the recoveries were from 88.0 to 132.1%, showing that the method is suitable for quantification/detection. Despite the conditions of the corpse (its state of decomposition and without the on-site determination that it was shot), there is an indication that it was hit by a FAP at a short distance (less than 50.0 cm between the barrel of the weapon and the target), a common fact in homicide cases with summary execution. It is possible to state unequivocally that the crime in question (case 1) presents minimum requirements of those typically observed in intentional homicides, with nonimmediate concealment of the corpse, since the results found by ICP-MS confirm the observations made at the place where the corpse was found.

Regarding case 2, there were higher concentrations of the three elements (Pb, Ba, and Sb) in the soil against the larvae (Tables 3 and 4). There is a concentration ranging from 7.98 to 26.02 μg/g for Pb, from 32.64 to 57.97 μg/g for Ba, and from 0.042 to 0.3 μg/g for Sb in the sediments of the three soil points collected, while in the larvae as concentrations for Pb, Ba, and Sb were 1.78 ± 0.40 μg/g, 2.94 ± 0.53 μg/g and < LD, respectively. Recovery tests demonstrate that the method is suitable for analysis since good values were obtained. Thus, the results obtained by ICP-MS confirm what is specified in the expert report, since the victim of this homicide did not present lesions typical of those caused by FAP, but presented characteristics of violent death with fractures in the cephalic region, similar to those caused by a blunt instrument.

Boracchi *et al.* evaluated whether GSR could be confused with toxic elements of air or soil. The analyses were performed by ICP-MS, and the results were obtained from cadaveric skin. The corpses used were divided into two groups, the first (A) consisting of 25 victims who had no gunshot wounds and had been found in an open environment between 10 days and 3 years. The second group (B) consisted of 16 corpses from a cemetery in Milan and none presented gunshot wounds. The second group (B) consisted of 16 bodies from a cemetery in Milan and none had gunshot wounds. Values of LOD and LOQ were not expressed and the authors report that the LOQ confirmed the absence of deposition/percolation of toxic elements. Groups A and B and negative control presented values close to zero or very low being lower than the body values of the equipment. The positive control group presented high levels of Pb (2.618 and 82.97 μg/L), Ba (0.049 and 1.253 μg/L), and Sb 0.018 and 2.169 μg/L (34).

The study by Boracchi *et al*. (31) presented a more quantitative approach when compared with others mentioned, but there is no way to perform a comparative analysis between the results, since sampling for detection was in corpse skin and in the present study cadaveric larvae of corpses in advanced state of decomposition were employed. Due to the lack of research with the proposed methodology, the comparison of the results is limited and reinforces the importance of the unpublished results reproduced from the study of cases 1 and 2.

Conclusion

The results obtained by perinecroscopic observation of the victims (case 1 and 2), due to the poor conditions of the site and poor lighting, in addition to the cadaveric decomposition state and the great dirt observed in the corpse, it was necessary to investigate by analytical technical means ICP-MS of the soil and immature found in the respective corpse. The concentrations of the characteristic elements of GSR in sediment near the cadaver were 4.44 μg/g for Pb, 8.74 μg/g for Ba, and 0.08 μg/g for Sb in case study 1; In case 2, 7.98–26.02 μg/g, 32.64–57.97 μg/g, and 0.042–0.30 μg/g for Pb, Ba, and Sb, respectively, were observed. Thus, the results indicated that, for case 1, there was firing of a firearm at close range against the corpse found, since Pb (6.28 ± 1.41 μg/g), Ba (1.49 ± 0.34 μg/g), and Sb (0.50 ± 0.32 μg/g) concentrations in larvae were higher than in the control, characterizing that these larvae consumed significant amounts of the elements that are present in particles characteristic of GSR corroborating the data of the criminal expert report. For case 2, it was observed that the results were not conclusive regarding the presence of characteristic elements of primers, or GSR, inside the larvae, making it impossible to indicate any indication of summary execution with a firearm, with shots short distance (<50.0 cm between the barrel of the gun and the target), a common fact in cases of summary execution, since the concentrations of Pb (1.78 ± 0.40 μg/g), Ba (2.94 ± 0.53 μg/g), and Sb (<LOD) were lower in the larvae in front of the sediments. Therefore, the present case study demonstrates new results of detection characteristic elements of GSR in real crime scenes, besides pointing out the potentiality of the ICP-MS technique associated with forensic entomology as an application tool, even presenting limitations of mass analysis. However, it is important to emphasize that conventional sample collection for GSR analysis was not the most appropriate given the circumstances of the crime scene. Thus, the article presents results that help to elucidate crimes, even though the analytes/traces are arranged in low concentrations, proven by the excellent values of LOD and LOQ obtained by this multi-element, fast, and robust technique.

Acknowledgments

The authors would like to thank the Núcleo de Competências em Química do Petróleo e ao LabPetro for the use of their facilities and the Polície do Espírito Santo for providing the samples.

References

1. Instituto de Pesquisa Econômica Aplicada (Ipea) atlas da violência [Institute for Applied Economic Research (Ipea) atlas of violence]. 2017. https://www.ipea.gov.br/atlasviolencia/dados-series/31 (accessed September 4, 2019).
2. Wallace JS. Chemical analysis of firearms, ammunition, and gunshot residue, 2nd edn. London, U.K.: CRC Press, 2008;3–10.
3. Wolff M, Builes A, Zapata G, Morales G, Benecke M. Detection of parathion (0,0- diethyl 0-(4nitrophenyl) phosphorothioate) by HPLC in insects of forensic importance in Medellín, Colombia. J Forensic Med Toxicol 2004;5(1):6–11.
4. Introna F, Campobasso CP, Goff ML. Entomotoxicology. Forensic Sci Int 2001;120(1–2):42–7. https://doi.org/10.1016/s0379-0738(01)00418-2.
5. Arnaldos MI, García MD, Romera E, Presa JJ, Luna A. Estimation of postmortem interval in real cases based on experimentally obtained entomological evidence. Forensic Sci Int 2005;149(1):57–65. https://doi.org/10.1016/j.forsciint.2004.04.087
6. Arnaldos MI, Sanchéz F, Álvarez P, García MD. A forensic entomology case from the Southeastern Iberian Peninsula. J Forensic Med Toxicol 2004;5(1):22–5.
7. Pai C, Jien M, Cheng Y, Yang C. Application of forensic entomology to postmortem interval determination of a burned human corpse: a homicide case report from southern Taiwan. J Formos Med Assoc 2007;106 (9):792–8.
8. LaGoo L, Schaeffer LS, Szymanski DW, Smith RW. Detection of gunshot residue in blowfly larvae and decomposing porcine tissue using inductively coupled plasma mass spectrometry (ICP-MS). J Forensic Sci 2010;55(3):624–32. https://doi.org/10.1111/j.1556-4029.2010.01327.x.
9. Beyer J, Enos W, Stajic M. Drug identification through analysis of maggots. J Forensic Sci 1980;25(2):411–2. https://doi.org/10.1520/JFS12147J
10. Goff M, Miller M, Paulson J, Lord W, Richards E, Omori A. Effects of 3,4-methylenedioxymethamphetamine in decomposing tissues on the development of Parasarcophaga ruficornis (Diptera: Sarcophagidae) and detection of the drug in postmortem blood, liver tissue, larvae and puparia. J Forensic Sci 1997;42(2):276–80. https://doi.org/10.1520/JFS14110J
11. Bourel B, Fleurisse L, Hédouin V, Cailliez J-C, Creusy C, Gosset D, et al. Immunohistochemical contribution to the study of morphine metabolism in calliphoridae larvae and implications in forensic entomotoxicology. J Forensic Sci 2001;46(3):596–9. https://doi.org/10.1520/JFS15009J
12. Vanini G, Souza RM, Destefani CA, Merlo BB, Piorotti TM, de Castro EVR, et al. Analysis of gunshot residues produced by.38 caliber handguns using inductively coupled plasma-optical emission spectroscopy (ICP OES). Microchem J 2013;115(1):106–12. https://doi.org/10.1016/j.microc.2014.03.003
13. Vanini G, Destefani CA, Merlo BB, Carneiro MTWD, Filgueiras PR, Poppi RJ, et al. Forensic ballistics by inductively coupled plasma-optical emission spectroscopy: quantification of gunshot residues and prediction of the number of shots using different firearms. Microchem J 2015;118 (1):19–25. https://doi.org/10.1016/j.microc.2014.07.016
14. Motta LC, Vanini G, Chamoun CA, Costa RA, Vaz BG, Costa HB, et al. Detection of Pb, Ba, and Sb in blowfly larvae of porcine tissue contaminated with gunshot residue by ICP OES. J Chem 2015;2015 (1):737913. https://doi.org/10.1155/2015/737913
15. Bartsch MR, Kobus HJ, Wainwright KP. An update on the use of the Sodium Rhodizonate test for the detection of lead originating from firearm discharges. J Forensic Sci 1996;41(6):1046–51. https://doi.org/10.1520/JFS14047J
16. Garofano L, Capra M, Ferrari F, Bizzaro GP, Di Tullio D, Dell'Olio M, et al. Gunshot residue – further studies on particles of environmental and occupational origin. Forensic Sci Int 1999;103(1):1–21. https://doi.org/10.1016/S0379-0738(99)00035-3
17. Laflèche DJN, Brière SJJ, Faragher NF, Hearns NGR. Gunshot residue and airbags: part I. Assessing the risk of deployed automotive airbags to produce particles similar to gunshot residue. Can Soc Forensic Sci J 2018;52(1):26–32. https://doi.org/10.1080/00085030.2018.1463202
18. Koons RD, Havekost DG, Peters CA. Analysis of gunshot primer residue collection swabs using flameless atomic absorption spectrophotometry and inductively coupled plasma atomic emission spectrometry: effects of a modified extraction procedure and storage of standards. J Forensic Sci 1989;34(1):218–21. https://doi.org/10.1520/JFS12624J
19. Roeterdink EM, Dadour IR, Watling RJ. Extraction of gunshot residues from the larvae of the forensically important blowfly *Calliphora dubia* (Macquart) (Diptera: Calliphoridae). Int J Legal Med 2004;118 (2):63–70. https://doi.org/10.1007/s00414-003-0408-1

20. Taborelli A, Gibelli D, Rizzi A, Andreola S, Brandone A, Cattaneo C. Gunshot residues on dry bone after decomposition – a pilot study. J Forensic Sci 2012;57(5):1281–4. https://doi.org/10.1111/j.1556-4029.2012.02119.x
21. Berendes A, Neimke D, Schumacher R, Barth M. A versatile technique for the investigation of gunshot residue patterns on fabrics and other surfaces: m-XRF. J Forensic Sci 2006;51(5):1085–90. https://doi.org/10.1111/j.1556-4029.2006.00225.x
22. Romolo FS, Margot P. Identification of gunshot residue: a critical review. Forensic Sci Int 2001;119(2):195–211. https://doi.org/10.1016/S0379-0738(00)00428-X
23. Dalby O, Butler D, Birkett JW. Analysis of gunshot residue and associated materials – a review. J Forensic Sci 2010;55(4):924–43. https://doi.org/10.1111/j.1556-4029.2010.01370.x
24. Koons RD. Analysis of gunshot primer residue collection swabs by inductively coupled plasma-mass spectrometry. J Forensic Sci 1998;43(4):748–54. https://doi.org/10.1520/JFS14301J
25. Santos A, Ramos P, Fernandes L, Magalhães T, Almeida A, Sousa A. Firing distance estimation based on the analysis of GSR distribution on the target surface using ICP-MS—an experimental study with a 7.65mm×17mm Browning pistol (.32 ACP). Forensic Sci Int 2015;247(1):62–8. https://doi.org/10.1016/j.forsciint.2014.12.006
26. Heringer RD, Ranville JF. Gunshot residue (GSR) analysis by single particle inductively coupled plasma mass spectrometry (spICP-MS). Forensic Sci Int 2018;288(1):e20–5. https://doi.org/10.1016/j.forsciint.2018.05.010
27. Costa RA, Motta LC, Destefani CA, Rodrigues RRT, do Espírito Santo KS, Aquije GMFV, et al. Gunshot residues (GSR) analysis of clean range ammunition using SEM/EDX, colorimetric test and ICP-MS: a comparative approach between the analytical techniques. Microchem J 2016;129(1):339–47. https://doi.org/10.1016/j.microc.2016.07.017
28. Weber IT, de Melo AJG, Lucena MAM, Rodrigues MO, Junior SA. High photoluminescent metal-organic frameworks as optical markers for the Identification of gunshot residues. Anal Chem 2011;83(12):4720–3. https://doi.org/10.1021/ac200680a
29. Tucker W, Lucas N, Seyfang KE, Kirkbride KP, Popelka-Filcoff RS. Gunshot residue and brakepads: compositional and morphological considerations for forensic casework. Forensic Sci Int 2017;270(1):76–82. https://doi.org/10.1016/j.forsciint.2016.11.024
30. Oliveira-Costa J. Entomologia forense. Quando os insetos são vestígios [Forensic entomology. When insects are traces], 3th edn. Rio de Janeiro, Brazil: Millenium, 2011;342–9.
31. Wastowski AD, da Rosa GM, Cherubin MR, Rigon JPG. Characterization of chemical elements in soil submitted to different systems use and management by energy dispersive x-ray fluorescence spectrometry (EDXRF). Quim Nova 2010;33(7):1449–52. https://doi.org/10.1590/S0100-40422010000700005
32. de Prata Neves Duarte M. Firing distance estimation through the quantification of gunshot residues in blowfly larvae (Calliphoridae family) using inductively coupled plasma-mass spectrometry [dissertation]. Porto, Portugal: Universidade de Porto, 2015.
33. Lima ESA, do Amaral Sobrinho NMB, Magalhães MOL, do Nascimento Guedes J, Zonta E. Barium absorption by rice plants (*Oryza sativa* L.) and mobility in soil treated with barite under different redox potential conditions. Quim Nova 2012;35(9):1746–51. https://doi.org/10.1590/S0100-40422012000900008
34. Boracchi M, Andreola S, Collini F, Gentile G, Lucchini G, Maciocco F, et al. Can cadaverous pollution from environmental lead misguide to false positive results in the histochemical determination of gunshot residues? In-depth study using ultra-sensitive ICP-MS analysis on cadaveric skin samples. Forensic Sci Int 2018;292(1):23–6. https://doi.org/10.1016/j.forsciint.2018.08.041

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14532
Available online at: onlinelibrary.wiley.com

CASE REPORT

PATHOLOGY/BIOLOGY

Kana Unuma,[1] *M.D.; Ryo Watanabe*,[1] *M.D.; Naho Hirayama*,[1] *M.D.; and Koichi Uemura*,[1] *M.D.*

Autopsy Identification of Viable *Mycobacterium Tuberculosis* in the Lungs of a Markedly Decomposed Body

ABSTRACT: Various infectious diseases, including COVID-19, MERS, and tuberculosis, are global public health issues. Tuberculosis, which is caused by *Mycobacterium tuberculosis* (MTB), is highly contagious and can be transmitted through inhalation of the bacteria. However, it has been assumed that the infectiousness of bacteria and viruses in dead bodies weakens as the time from death increases. In particular, there is little awareness of infection control measures concerning decomposed bodies or even the need for such measures. The deceased, in whom we discovered MTB 3 months following her death, was a woman in her 80s who died at home. We performed judicial autopsy, because police suspected homicide when her husband hanged himself. Obtained organs were used for microscopic examination by hematoxylin–eosin staining and Ziehl–Neelsen staining. In addition, real-time PCR and mycobacterial culture testing using Ogawa's medium were performed for the detection of MTB. We found that the MTB in the decomposed body remained viable and potentially infectious. To identify the bacterial strain further, we performed DNA-DNA hybridization and identified the strain as MTB complex. Potentially infectious live MTB survived in the dead body far longer than had been previously reported. Pathologists should consider microbial culture tests for all autopsied cases in which the decedent's medical history or macro-examination suggests possible infection, even when a long duration of time has passed since death. Pathologists and specialists who perform autopsies should recognize that all dead bodies are potentially infectious, including those in which long periods have elapsed since death.

KEYWORDS: *Mycobacterium tuberculosis*, tuberculosis, biosafety, forensic pathology, autopsy, decomposition

Tuberculosis is not a disease of the past. According to the World Health Organization, approximately one-third of the world's population has tuberculosis, and 1.7 million deaths per year are attributed to the disease, making it one of the top 10 causes of death in the world (1). While new infections of tuberculosis are consistently low in certain countries, large population movements, the increasing incidence of drug-resistant strains, and the association with human immunodeficiency virus (HIV) infection make it a disease that is commonly dealt with in forensic medicine (2).

When performing forensic autopsies, the deceased patient's medical history is often unknown at the time of autopsy. *Mycobacterium tuberculosis* (MTB), an airborne pathogen that causes tuberculosis, is one of the most contagious pathogens and therefore requires the highest level of precautions. However, there is an assumption that the infectiousness of bacteria and viruses in dead bodies weakens as the duration of time since death increases and that potential pathogens have nearly no likelihood of surviving in a decomposed body. In particular, there is little awareness of infection control measures regarding decomposed bodies or even the need for such measures. Due to such assumptions and lack of awareness, forensic pathologists usually do not perform microbial cultures and do not always strictly adhere to biosafety preventive measures when dealing with dead bodies in which long periods have elapsed since the patient's death.

We identified viable MTB 3 months after death in a setting of marked decomposition. We describe the detection of active MTB that survived in a dead body for far longer than that previously reported. Pathologists and specialists who perform autopsies should recognize that all dead bodies are potentially infectious, including those in which long periods have elapsed since death. We recommend that pathogen detection, using several methodologies, should be attempted even in the setting of decomposition.

Case Report

The deceased was a woman in her 80s who resided with her husband. She had difficulty walking after sustaining a hip fracture due to a fall and reportedly lost weight because of anorexia. After a few months, the neighbors of the woman noticed an infestation of maggots in her house and informed the police. The police discovered the dead body of the woman lying on a bed mat with a white sheet over her face. The husband was also discovered to have died after hanging himself with an electrical cord tied to a pipe running along the ceiling of their bedroom. A suicide note left by the husband indicated the date and time that his wife had died (approximately 3 months prior to discovery) and his plan to commit suicide after his wife had died.

[1] Department of Forensic Medicine, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, 1-5-45, Yushima, Bunkyo-ku, Tokyo, 113-8519, Japan.

Corresponding author: Kana Unuma, M.D. E-mail: unuma.legm@tmd.ac.jp

Received 6 June 2020; and in revised form 8 July 2020, 12 July 2020; accepted 14 July 2020.

According to the medical history, she had reflux esophagitis and osteoporosis and had been seeing her family physician regularly until 1 month prior to the presumed date of death. She had a prescription for proton pump inhibitor. She had no history of MTB. Further investigation revealed that the average monthly temperature in the region during the month from her death until the discovery of her body ranged from 3.8 to 8.5°C.

The postmortem examination indicated a height of 148 cm, weight of 23.9 kg, and an emaciated body in an advanced state of decomposition. The face and all four extremities were desiccated, and both maggots and pupae were locally present on the trunk (Fig. 1A). The weight of the left lung was 410 g and that of the right lung was 300 g. Examination of the lungs revealed one purulent nodule measuring approximately 8 cm in the left lower lobe. The nodule was yellowish-white and had relatively distinct contours. A second purulent nodule measuring approximately 3 cm with relatively distinct contours and a greenish-white color was identified in the right upper lobe, and another with the same characteristics was identified in the right lower lobe (Figs 1B and 2). Histological examination using hematoxylin and eosin staining showed decomposition-related changes throughout the lungs and a breakdown of normal lung tissue. However, lymphocytes and other cells that had lost their shape due to decomposition surrounded the granulomatous part of the lung, forming a nodular-like lesion (Fig. 3A). It was unclear whether the center of this lesion had undergone caseous necrosis or not because of the decomposition process. Using Ziehl–Neelsen staining, which detects MTB, we observed rod-shaped bacilli in the left upper lobe and left lower lobe that were stained red (Fig. 3B).

We performed real-time polymerase chain reaction on the nodules in the left lower lobe to identify MTB complex nucleic acids using TRCReady MTB (Tosoh Bioscience, Tokyo, Japan), an automated real-time nucleic acid amplification test that uses transcription reverse-transcription concerted reaction technology, and the results were positive. We also performed mycobacterial

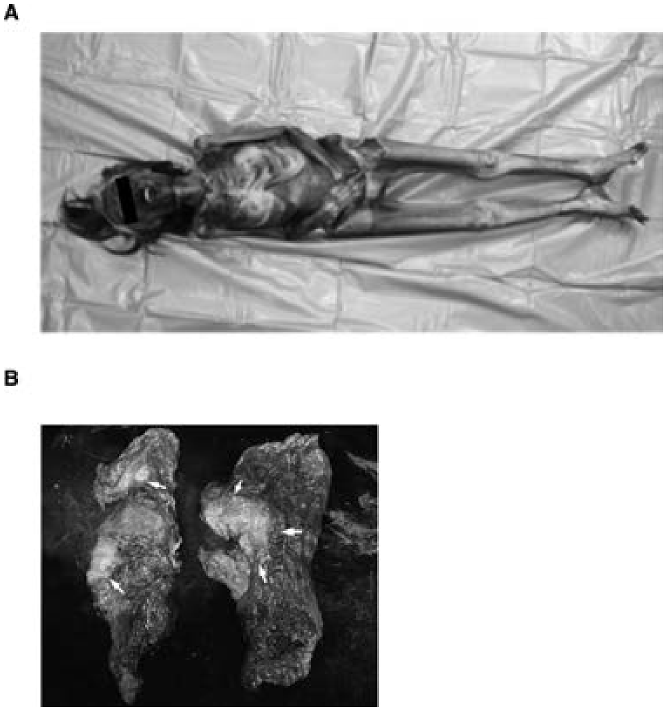


FIG. 1—*(A) External appearance of the body showing emaciation and advanced decomposition with dessication. (B) Images of the tubercular nodule-like lesion on the divided surface of the right and left lungs (arrows). [Color figure can be viewed at wileyonlinelibrary.com]*

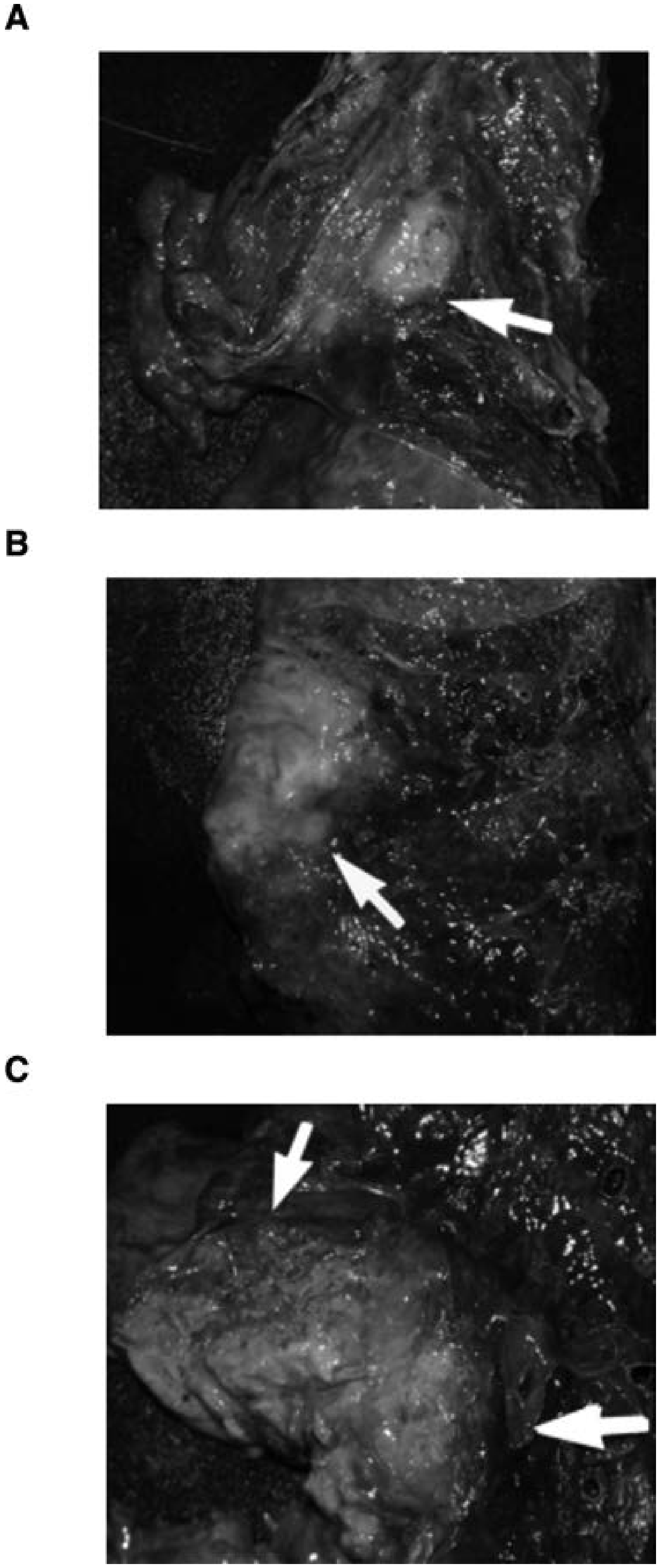


FIG. 2—*An enlargement of the tubercular lesion of the (A) right upper lung (arrow), (B) right lower lung (arrow), and (C) left upper lung (arrow). [Color figure can be viewed at wileyonlinelibrary.com]*

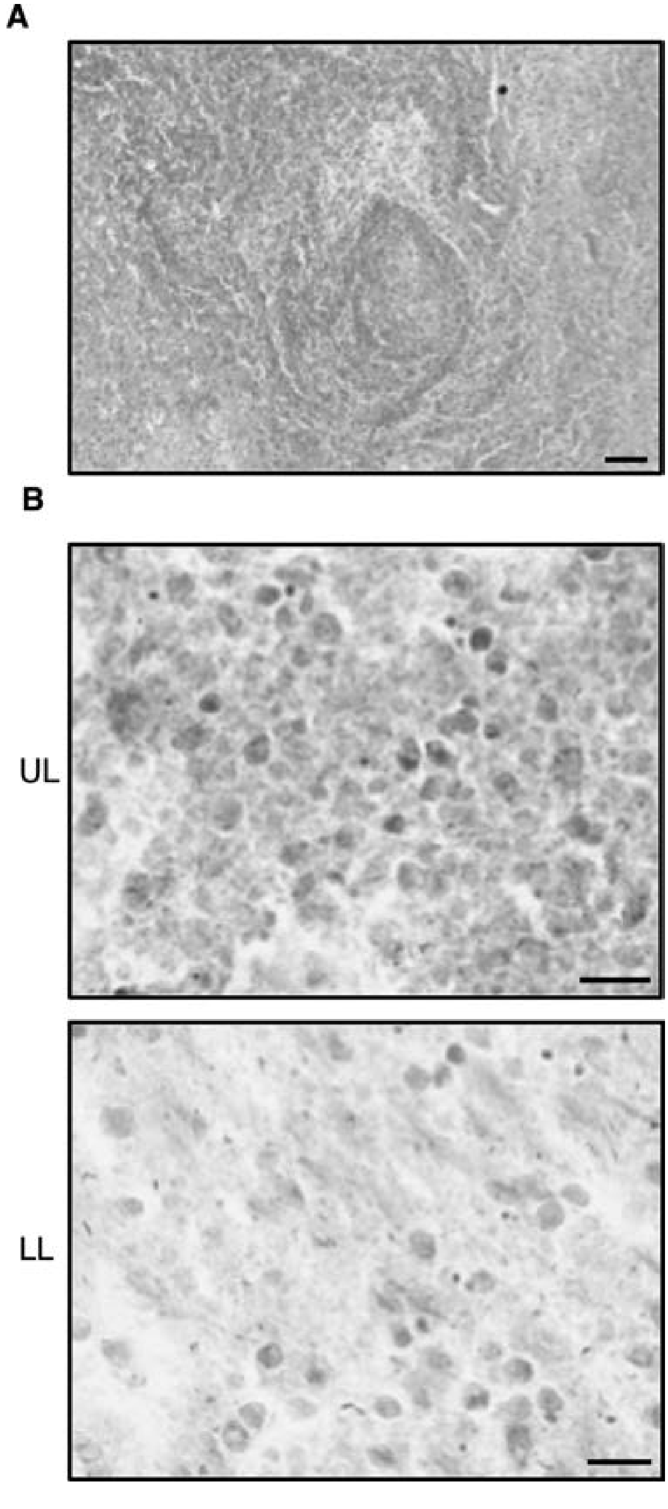


FIG. 3—*(A) Hematoxylin and eosin staining of the tubercular lesion. Bar = 100 μm. (B) Ziehl–Neelsen staining of the tubercular lesion of the left lung. Rod-shaped bacilli (*Mycobacterium tuberculosis*) that are stained red are visible. UL: upper lobe, LL: lower lobe. Bar = 20 μm. [Color figure can be viewed at wileyonlinelibrary.com]*

culture testing using Ogawa's medium for a period of 4 weeks, the results of which were also positive. To further identify the bacterial strain, we performed a DNA-DNA hybridization protocol and identified the strain as MTB complex. The MTB strain was not resistant to common anti-tuberculosis drugs at minimum inhibitory concentrations.

The heart weighed 240 g, and mild stenosis of the left anterior descending artery was observed. There was no intracranial bleeding. We detected a low concentration of acetaminophen in the liver on drug analysis using liquid chromatography–mass spectrometry; however, this was not found to be the cause of death. There were no other significant findings.

The cause of death was determined to be pulmonary tuberculosis.

Discussion

This is a case report of the successful isolation of MTB from a decomposed body with a postmortem interval of 3 months.

There is ongoing debate among forensic pathologists over the significance of postmortem microbiological culture testing due to issues such as postmortem contamination and proliferation of putrefactive bacteria (3). As the time elapsed since death increases, bacterial cultures of postmortem samples are more likely to yield polymicrobial growth than pure growth of a single species (4). The usefulness of postmortem bacterial cultures has identified a window of between 15–48 h after death as the time at which cultures are useful (5–7).

Thus, many forensic pathologists tend not to perform postmortem cultures on cases in which ≥2 days have elapsed since death. Our case showed that specific microorganisms such as MTB were able to survive, remain viable, and be potentially infectious for several months after death. For highly virulent pathogens, such as MTB, detection combined with clinical and pathological findings might be able to determine whether the pathogen contributed to the individual's death even if the test sample has decomposed.

There has been a wide variety of data regarding the survival of MTB. One study found that MTB contained within sputum lost all ability to survive after 4–8 days at a temperature of 37°C but was able to survive for at least 14 days at a temperature of 2–4°C (8). A study conducted in the state of New Mexico in the United States—which is characterized by a dry climate and had temperatures ranging from 0 to 17°C at the time of the study—showed that a culture performed on lung tissue obtained from a dead body that had been exhumed 8 days after burial detected viable MTB (9). Another study reported that an MTB test performed on lung tissue fixed with 10% buffered formalin for 45 days detected MTB (10). These studies suggest that low temperatures and desiccation can slow down decomposition, which may contribute to the long-term survival of mycobacteria. Similarly, in our case, the survival of MTB could be attributed to the fact that the patient died in winter, when the external temperature and humidity in the region were low.

Mycobacterium tuberculosis is transmitted through the air (airborne transmission); therefore, when processing a dead body in an autopsy room, MTB may be transmitted via inhalation of air containing the bacteria. Thus, during an autopsy, forensic pathologists and all other staff are placed at an increased risk of infection due to the increased exposure to MBT (11). There have been occasional reports of mass outbreaks of tuberculosis originating from postmortem examinations in all regions of the world, including in Japan (12–14). Regarding the risk of infection when handling dead bodies recovered after a disaster, the *Management of Dead Bodies after Disasters 2006* guideline indicates that, "Most infectious organisms do not survive beyond 48 h in a dead body. An exception is HIV, which has been found 6 days postmortem." This guideline has served as the basis for performing autopsies around the world during times of disaster, including the Great East Japan Earthquake and Tsunami (15–17). However, this case revealed that MTB could survive in a dead body as a living organism for a significantly longer period than has been reported. Thus, forensic pathologists and all specialists who perform autopsies should recognize that all dead bodies are potentially infectious, including those in which long periods have elapsed since the patient's death. In addition, in cases where signs of infection are present, forensic pathologists should use a variety of methods to identify microorganisms even if the remains are decomposed. In the case of highly infectious diseases such as TB, these results may be of benefit to living individuals who had contact with the deceased prior to their death.

Acknowledgment

The authors would like to thank A. Komatsu for technical assistance with the autopsy.

References

1. World Health Organization. World Health Organization, global health estimates 2016: deaths by cause, age, sex, by country and by region, 2000-2016. Geneva, Switzerland: World Health Organization, 2018.
2. Stephenson L, Byard RW. Issues in the handling of cases of tuberculosis in the mortuary. J Forensic Leg Med 2019;64:42–4. https://doi.org/10.1016/j.jflm.2019.04.002
3. Wilson SJ, Wilson ML, Reller LB. Diagnostic utility of postmortem blood cultures. Arch Pathol Lab Med 1993;117:986–8.
4. Morris JA, Harrison LM, Partridge SM. Practical and theoretical aspects of postmortem bacteriology. Curr Diagn Pathol 2007;13:65–74.
5. Wood WH, Oldstone M, Schultz RB. A re-evaluation of blood culture as an autopsy procedure. Am J Clin Pathol 1965;43:241–7.
6. Reinhardt G, Zink P, Legler F. Bakteriologische untersuchungsbefunde am herzblut der leiche. [Bacteriological findings in cardiac blood of a cadaver]. Beitr Gerichtl Med 1973;31:311–4.
7. Tang RK, Liu Y, Liu YZ, Zhu SM, Huang W, Zhao P, et al. Evaluation of post-mortem heart blood culture in a Chinese population. Forensic Sci Int 2013;231:229–33. https://doi.org/10.1016/j.forsciint.2013.05.020
8. Traore I, Slosárek M. Survival of mycobacteria in sputum at different temperatures. Czech Med 1981;4:203–8.
9. Nolte KB. Survival of *Mycobacterium tuberculosis* organisms for 8 days in fresh lung tissue from an exhumed body. Hum Pathol 2015;36:915–6. https://doi.org/10.1016/j.humpath.2005.04.010
10. Gerston KF, Blumberg L, Tshabalala VA, Murray J. Viability of mycobacteria in formalin-fixed lungs. Hum Pathol 2014;35:571–5. https://doi.org/10.1016/j.humpath.2004.01.009
11. Nolte KB, Taylor DG, Richmond JY. Biosafety considerations for autopsy. Am J Forensic Med Pathol 2002;23:107–22. https://doi.org/10.1097/00000433-200206000-00001
12. Okochi Y. Hospital outbreak of *Mycobacterium tuberculosis* resulting from autopsy exposure. Kansenshogaku Zassi 2005;79:534–42. https://doi.org/10.11150/kansenshogakuzasshi1970.79.534
13. Burton JL. Health and safety at necropsy. J Clin Pathol 2003;56:254–60. https://doi.org/10.1136/jcp.56.4.254
14. Flavin RJ, Gibbsons N, O'Briain DS. *Mycobacterium tuberculosis* at autopsy–exposure and protection: an old adversary revisited. J Clin Pathol 2007;60:487–91. 15.1136/jcp.2005.032276
15. Morgan O, Tidball-Binz M, van Alphen D. Management of dead bodies after disasters: a field manual for first responders. Washington, DC: Pan American Health Organization, 2009;241–7.
16. Roy N. The Asian tsunami: PAHO disaster guidelines in action in India. Prehosp Disaster Med 2006;21:310–5. https://doi.org/10.1017/s1049023x00003939
17. Ballera JE, de Los Reyes VC, Sucaldito MN, De Guzman A, Sy L Jr, Zapanta MJ, et al. Management of the dead in Tacloban City after Typhoon Haiyan. Western Pac Surveill Response J 2015;6:44–7. https://doi.org/10.5365/WPSAR.2015.6.2.HYN_004

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14510
Available online at: onlinelibrary.wiley.com

CASE REPORT

TOXICOLOGY

Patrick Allan Kosecki,[1] *Ph.D.; Erika Canonico,*[1] *M.S.; and Phillip Brooke,*[1] *M.S.*

Testing Antemortem Blood for Ethanol Concentration from a Blood Kit in a Refrigerator Fire

ABSTRACT: The stability of ethanol in antemortem blood stored under various conditions has been widely studied. Antemortem blood samples stored at refrigerated temperature, at room temperature, and at elevated temperatures tend to decrease in ethanol concentration with storage. It appears that the stability of ethanol in blood exposed to temperatures greater than 38°C has not been evaluated. The case presented here involves comparison of breath test results with subsequent analysis of blood drawn at the time of breath testing. However, the blood tubes were in a refrigerator fire followed by refrigerated storage for 5 months prior to analysis by headspace gas chromatography. The subject's breath was tested twice using an Intoxilyzer 8000. The subject's blood was tested in duplicate using an Agilent headspace gas chromatograph. The measured breath ethanol concentration was 0.103 g/210 L and 0.092 g/210 L. The measured blood ethanol concentration was 0.0932 g/dL for both samples analyzed. Although the mean blood test result was slightly lower than the mean breath test result, the mean breath test result was within the estimated uncertainty of the mean blood test result. Even under the extreme conditions of the blood kit being in a refrigerator fire, the measured blood ethanol content agreed well with the paired breath ethanol test.

KEYWORDS: blood alcohol, headspace, gas chromatography, blood alcohol stability, breath test, toxicology, blood ethanol

In Arizona, a person who operates a motor vehicle while under the influence of intoxicating liquor upon the request of a law enforcement officer shall be asked to take a chemical test of the person's blood, breath, urine, or other bodily substance for the purpose of determining alcohol concentration. The test or tests are chosen by the law enforcement agency. Failure to comply with such a request can result in a one-year suspension of the person's driver's license. The current driving under the influence (DUI) program in the City of Scottsdale, AZ, requires people suspected of DUI to submit to a breath test to determine breath ethanol content. In addition, blood samples are drawn, typically during the fifteen-minute deprivation period of the breath testing process. When the case is not resolved based on the breath test, the forensic laboratory will analyze the blood sample for ethanol concentration. This two-pronged testing approach has greatly reduced the number of blood ethanol tests required of the forensic lab. This approach also provides measurements of ethanol content using two different techniques for legal purposes.

Breath testing provides an immediate result, whereas testing blood usually involves a delay during which time the blood is stored prior to analysis. In DUI cases involving analysis of blood, defense arguments often include various factors that could affect the ethanol content in blood samples stored under different conditions and time periods. Therefore, it is important to understand how the storage of antemortem blood samples can affect their ethanol content. Common forensic storage conditions include refrigeration (4°C), sealed blood tubes, and the use of preservatives and anticoagulants. Studies of antemortem blood stored refrigerated have consistently shown small decreases in ethanol concentration if any change was measured (1–5). Frozen samples have also shown a decrease in ethanol content with storage (6). Under nonstandard forensic storage conditions, room temperature, and elevated temperature, ethanol concentration has been shown to decrease in antemortem blood samples (3,4,7–9).

Upon review of the literature, it appears that no studies exist regarding the stability of ethanol in blood for temperatures higher than 38°C. The blood kit in the case presented in this report was involved in a fire that occurred in a refrigerator in which officers impound blood kits for temporary storage until the blood kits are picked up by Property Technicians to transfer to the Property and Evidence Building (see Fig. 1).

Methods

Blood was collected from a subject suspected of DUI on July 12, 2019, into two 10-mL gray-top Vacutainer® tubes containing 100 mg sodium fluoride and 20 mg potassium oxalate (Becton, Dickinson and Company, Franklin Lakes, NJ). The blood tubes were sealed inside a plastic clamshell box which was sealed inside a cardboard box. The blood kit was placed in a refrigerator for temporary storage. On Saturday, July 13, 2019, there was a fire in the refrigerator. Based on the Fire Department Incident Report, it is estimated that the fire lasted 15–

[1] Scottsdale Police Department Crime Laboratory, 7601 E. McKellips Rd., Building B, Scottsdale, AZ, 85257.

Corresponding author: Phillip Brooke, M.S. E-mail: pbrooke@scottsdaleaz.gov

Received 28 April 2020; and in revised form 18 June 2020; accepted 23 June 2020.

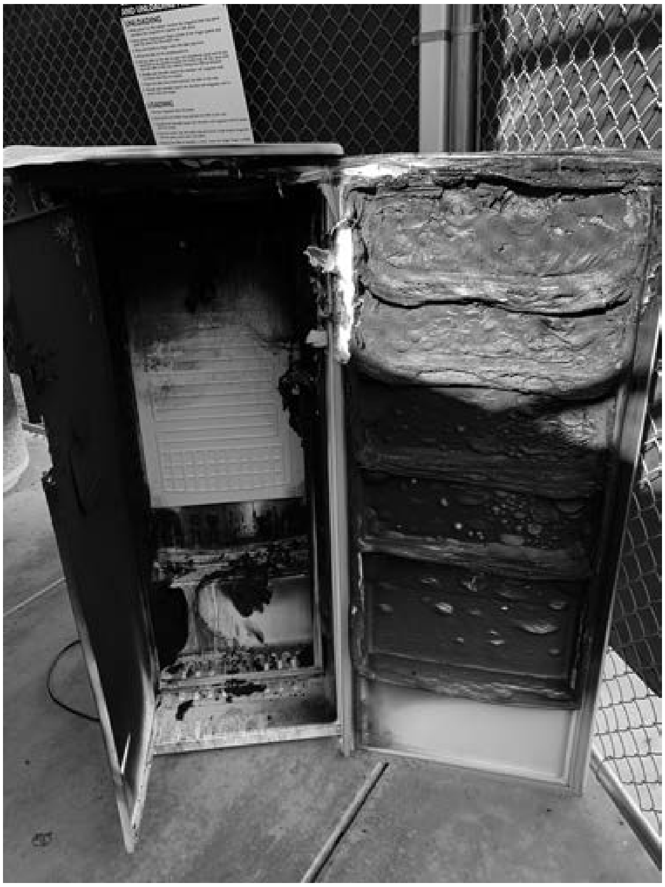


FIG. 1—*Interior of the refrigerator in which the blood kit was stored. The refrigerator was equipped with a secondary metal door inside to secure the evidence deposited in the refrigerator. [Color figure can be viewed at wileyonlinelibrary.com]*

20 min. As received by the laboratory, the outer box was partially burned (see Fig. 2). The inner plastic bag was melted to the cardboard box, and the inner plastic box had some melting on one end (see Fig. 3). The blood tubes were intact with no visible signs of damage. The blood was analyzed December 2, 2019, using an Agilent 7890B gas chromatograph connected to an Agilent 7697A Headspace Sampler (Agilent Technologies, Santa Clara, CA). The gas chromatograph was equipped with a dual-column system and two flame ionization detectors. The 30-m columns were the Agilent DB-ALC1 and DB-ALC2. The data from the DB-ALC1 column were used for the quantification of ethanol. The data from the DB-ALC2 column were used for confirmation of the identification of ethanol. Samples were incubated for 23 min at 60°C. Hydrogen was used as the carrier gas. The gas chromatograph held the oven temperature at 40°C for the analysis.

The gas chromatograph was calibrated using four calibrators with known ethanol concentrations: 0.02, 0.10, 0.20, and 0.40 g/dL. Following calibration and prior to cases samples being analyzed, the instrument was verified to be accurate using positive and negative controls. The case blood sample was tested in duplicate along with eleven other cases. Duplicate testing for each case consisted of one analyst preparing two separate samples from one blood tube and testing the two samples sequentially on the same instrument. Duplicate agreement was required to be within 2% of the mean of the two tests. A positive control was run between every five cases. Additional positive controls and a negative control were analyzed after all cases were analyzed.

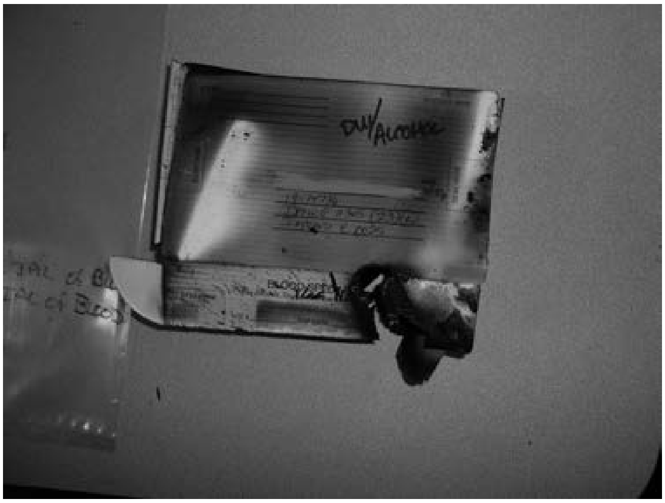


FIG. 2—*Exterior of the blood kit as received by the laboratory. The subject's name has been blurred in the image. [Color figure can be viewed at wileyonlinelibrary.com]*

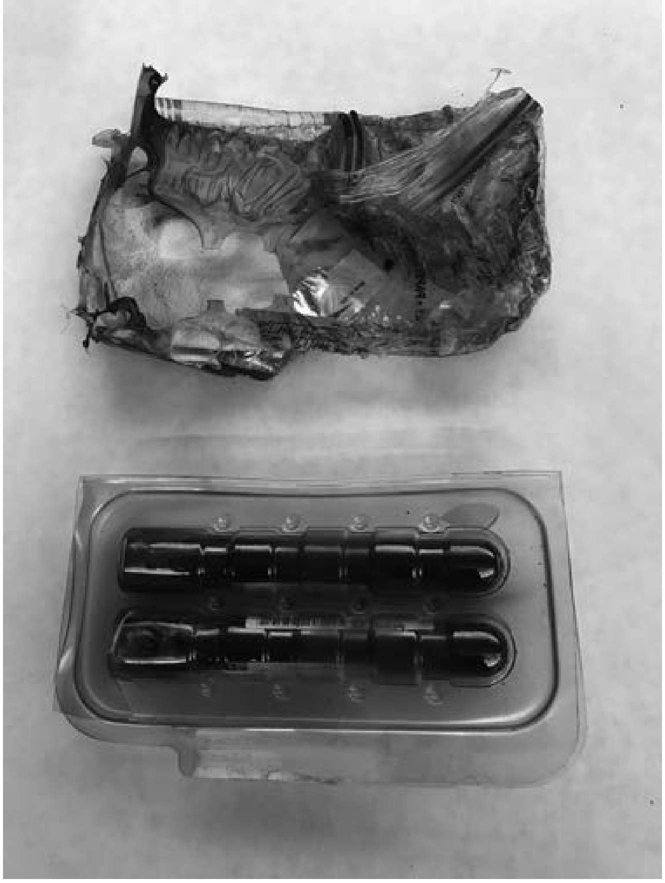


FIG. 3—*Interior of blood kit showing the melted bag and partial melting on the inner plastic box. [Color figure can be viewed at wileyonlinelibrary.com]*

A breath sample was collected and tested on July 12, 2019, in duplicate using an Intoxylizer 8000 (CMI, Inc., Owensboro, KY) following a fifteen-minute deprivation period during which the subject was watched to ensure that he did not belch or place anything into his mouth. The instrument was checked prior to the first test and after the second test using a 0.100 g/210 L dry gas standard. An air blank was tested prior to and after each of the four tests. The two subject tests were taken at an interval not less than 5 min nor more than 10 min apart. The two subject tests were required to agree within 0.020 g/210 L.

Results and Discussion

The blood ethanol concentration was measured on the headspace gas chromatograph to be 0.0932 and 0.0932 g/dL. The uncertainty for blood ethanol measurement in our laboratory was calculated to be five percent at a level of confidence greater than 99.73 percent. The two results from the breath test were 0.103 g/210 L and 0.092 g/210 L.

There is good agreement between the breath test results and blood test results. On average, a blood result is expected to be higher than a breath result for a breath test instrument using a 1:2100 ratio (10–12). Acetaldehyde was also detected in the blood sample following analysis. Acetaldehyde may have been present due to the metabolism of ethanol in the subject. The presence of acetaldehyde has also been reported for blood samples heated in head space vials with a corresponding decrease in ethanol concentration (13). It is possible that the elevated temperatures in the refrigerator fire led to some reduction in ethanol concentration through oxidation of the ethanol to acetaldehyde. Additionally, the five months of refrigerated storage prior to analysis could also account for a slight decrease in ethanol concentration and corresponding increase in acetaldehyde.

Even under the extreme conditions of the blood kit being in a refrigerator fire, the measured blood ethanol content agreed well with the paired breath ethanol test.

References

1. Shan X, Tiscione NB, Alford I, Yateman DT. A study of blood alcohol stability in forensic antemortem blood samples. Forensic Sci Int 2011;211(1–3):46–50. https://doi.org/10.1016/j.forsciint.2011.04.012.
2. Jones AW, Ericsson E. Decreases in blood ethanol concentrations during storage at 4°C for 12 months were the same for specimens kept in glass or plastic tubes. Pract Lab Med 2016;4:76–81. https://doi.org/10.1016/j.plabm.2016.02.002.
3. Winek CL, Paul LJ. Effect of short-term storage conditions on alcohol concentrations in blood from living human subjects. Clin Chem 1983;29(11):1959–60. https://doi.org/10.1093/clinchem/29.11.1959. https://doi.org/10.1093/clinchem/29.11.1959.
4. Vance CS, Carter CR, Carter RJ, Del Valle MM, Peña JR. Comparison of immediate and delayed blood alcohol concentration testing. J Anal Toxicol 2015;39(7):538–44. https://doi.org/10.1093/jat/bkv061.
5. Jones AW. Are changes in blood-ethanol concentration during storage analytically significant? Importance of method imprecision. Clin Chem Lab Med 2007;45(10):1299–304. https://doi.org/10.1515/CCLM.2007.289.
6. Stojiljkovic G, Maletin M, Stojic D, Brkic S, Abenavoli L. Ethanol concentration changes in blood samples during medium-term refrigerated storage. Eur Rev Med Pharmacol Sci 2016;20(23):4831–6.
7. Chang RB, Smith WA, Walkin E, Reynolds PC. The stability of ethyl alcohol in forensic blood specimens. J Anal Toxicol 1984;8(2):66–7. https://doi.org/10.1093/jat/8.2.66.
8. Glendening BL, Waugh TC. The stability of ordinary blood alcohol samples held various periods of time under different conditions. J Forensic Sci 1965;10(2):192–200.
9. Winek T, Winek CL, Wahba WW. The effect of storage at various temperatures on blood alcohol concentration. Forensic Sci Int 1996;78(3):179–85. https://doi.org/10.1016/0379-0738(95)01884-0.
10. Cowan JM, Burris JM, Hughes JR, Cunningham MP. The relationship of normal body temperature, end-expired breath temperature, and BAC/BrAC ratio in 98 physically fit human test subjects. J Anal Toxicol 2010;34(5):238–42. https://doi.org/10.1093/jat/34.5.238.
11. Harding PM, Laessig RH, Field PH. Field performance of the Intoxilyzer 5000: a comparison of blood- and breath-alcohol results in Wisconsin drivers. J Forensic Sci 1990;35(5):1022–8. https://doi.org/10.1520/JFS12925J.
12. Jones AW, Andersson L. Comparison of ethanol concentrations in venous blood and end-expired breath during a controlled drinking study. Forensic Sci Int 2003;132(1):18–25. https://doi.org/10.1016/s0379-0738(02)00417-6
13. Chiarotti M, De Giovanni N. Acetaldehyde accumulation during headspace gas-chromatographic determination of ethanol. Forensic Sci Int 1982;20(1):21–5. https://doi.org/10.1016/0379-0738(82)90101-3

J Forensic Sci, November 2020, Vol. 65, No. 6
doi: 10.1111/1556-4029.14555
Available online at: onlinelibrary.wiley.com

BOOK REVIEW

Michael S. Adamowicz,[1] *Ph.D.*

Review of: *The Scientific Method in Forensic Science: A Canadian Handbook*

REFERENCE: Illes M, Wilson P. The scientific method in forensic science: a Canadian handbook. Toronto, Canada: Canadian Scholars, 2020, 256 pp.

While so many books published in the field of forensic science focus on new methods, new scientific applications to specific casework, and new approaches to technical problems, *The Scientific Method in Forensic Science: A Canadian Handbook* is refreshing in that it is aimed at examining *how* to approach doing science. The book is written from a Canadian point of view and contains many interesting references to salient Canadian casework and court decisions; however, it should not be viewed as relevant to only a Canadian audience. On the contrary, it is very enlightening to read about the Canadian justice system, its court decisions, and their parallel documents to such things as the American NAS and PCAST reports on forensic science. A clear subtext throughout the book is that the various issues and problems that forensic science faces as a discipline transcend borders and are shared by all practitioners. The authors offer that applying the scientific method, using evidence-based practice and analysis, employing critical thinking, and continually evaluating and using scientific literature can all be highly beneficial tools to address many of those problems.

The book is well organized, and the writing is clear and accessible to readers from first-year college students to seasoned practitioners. In addition to the main content, each chapter contains an introduction to the specific topic, a glossary of important terms, a list of further readings, and a bibliography. There are also some thought-provoking discussion questions, instructional “pop-outs” (titles and web addresses for further study), and career profiles of selected Canadian forensic scientists of note. These last areas make this book very appealing as a resource for educators in forensic science programs. The authors are faculty members themselves, and this work nicely fits the need for a tool to explain how science is actually done, within a forensic context. Undergraduate-level students will find this book of particular use. It provides succinct definitions and examples of the scientific method, what critical thinking is and how it is employed, definitions of evidence-based practice and analysis, as well as some of the different types of reasoning (deductive, inductive, abductive, etc.). Explanations of these concepts can often be overlooked in the classroom during the rush to cram students’ heads full of specific domain knowledge, having a resource that captures all of them and provides forensic context as well is very useful.

The book opens with a discussion of the paradigm shift that the authors identify happening around 2008–2012 with the release of the NAS report in the United States and the Goudge and Hart House reports in Canada. The authors relate these critical documents to the deep level of self-examination that forensic scientists have been engaged in since that time in order to better the science that we practice and improve the quality of information that we provide to our justice systems. Referencing documents from both countries broadens the scope and appeal of the book, as well as providing new information to its readers, especially those in the United States who may not be familiar with the Canadian documents. The cited examples all point to the need for more application of rigorous and robust science in the forensic disciplines and list some of the relevant recommendations. The next several chapters are focused on what makes for rigorous and robust science. Topics of discussion cover critical thinking and its application, the parts and significance of a scientific paper, and what a literature review is and how to make use of one. While scientists who read and write scientific literature on a daily basis may not find these sections especially informative, undergraduate-level students and nonscientists will benefit greatly from them. Reading scientific literature, especially doing so critically, is a challenging skill and the chapters in this book should assist students, law enforcement officers, attorneys, and others not familiar with it by demonstrating how to begin to make sense of scientific literature and its peculiarities. There are even some bits of information that veteran scientific readers may find enlightening, such as clearly defining low- and high-level sources, primary and secondary material, and identifying 14 different types of published literature reviews. Again, forensic examples and context are provided to keep all of the topics focused on the authors’ primary thesis.

The book then moves to discussing the proper use of statistics and research project design in forensic examinations. It should be noted that the book does not attempt to explain how to *perform* statistical calculations; however it reviews some basic concepts and how to appropriately employ statistics in casework and then properly communicate their importance. Specific case examples are used to illustrate the authors’ points. The material on research project design will also be of limited use to the experienced laboratory scientist, but for those who have not previously been involved in planning a research study, it is excellent information. Interestingly, the text on project design reads very similarly to some of the types of literature reviews detailed in a previous chapter, covering many different approaches to creating a logical, robust project that will address the central

[1] Forensic Science Program, College of Agricultural Sciences and Natural Resources, University of Nebraska, 103 Agriculture Hall, Lincoln, NE, 68583.

Received 5 Aug. 2020; accepted 5 Aug. 2020.

research question and provide reliable data sufficient to make appropriate scientific conclusions. Different research methods are defined, strengths and weaknesses for each type are described, and publication examples of each type are provided. These sections should be default reading for undergraduate students who want to create research/internship projects of their own, as well as attorneys and judges who seek understanding about how validation studies are created.

The book concludes with examinations of bias, ethics, and communications. The major categories of bias that have been related to forensic science and research are covered, with specific examples from the Canadian justice system used to show how cognitive, conformational, and contextual biases can affect the outcome of a court case. While the text is not an in-depth discussion on the nuances of bias, it is a good introduction to the topic and cites many additional resources for further study. Ethics and ethical standards are also briefly discussed, with the Canadian Society of Forensic Science rules of professional conduct used as an example of published professional ethical standards. Oddly, in a book with so many case examples used to demonstrate and reinforce each of the topics, there are no descriptions of real ethical violations and the enormous problems these situations have caused in court, as well as to the forensic community as a whole. Such an example would have strengthened the treatment of ethics and provided continuity with the rest of the book.

The Scientific Method in Forensic Science: A Canadian Handbook offers readers many recommendations for using sound logic, performing good science, and properly supporting that science in the forensic community. The individual reader may agree or disagree with each of the authors' recommendations, but the book certainly gives an excellent introduction to scientific thought and organization for students and practitioners of all levels and provides many topics for consideration and discussion. Forensic science educators should also find the text and supporting materials in each chapter very useful tools in designing course units dedicated to familiarizing their students with the background, complexities, and methodological approaches of performing science in our unique application regardless of where they live.