Forensic Science International 357 (2024) 111997

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Review article

Striated toolmarks comparison and reporting methods: Review and perspectives

Jean-Alexandre Patteet *, Christophe Champod

School of Criminal Justice, Faculty of Law, Criminal Justice and Public Administration, University of Lausanne, Lausanne 1015, Switzerland

ARTICLE INFO

Keywords:
Toolmarks
Review

ABSTRACT

Forensic toolmark examiners have been comparing features observed in toolmarks to help determine their source for over a century. However, in the past decade, the holistic process of comparing toolmarks and presenting findings in court have faced intense scrutiny. This paper provides a summary of the voiced criticisms, primarily concerning the scientific reliability and validity of the comparison methods employed by examiners and the conclusions they testify to. The focus of this review is specifically on the examination of striated toolmarks. We assess the comparison methods and reporting practices currently in use, while also delving deeper into research aligned with current recommendations, such as PCAST (The President's Council of Advisors on Science and Technology). Throughout the review, we examine both the strengths and weaknesses of existing practices, aiming to assist practitioners in identifying key research needs and addressing the concerns raised by critics. By doing so, we seek to enhance the credibility and effectiveness of toolmark analysis in the field of forensic science.

1. Introduction – Toolmarks, practice and questioning

1.1. What is a toolmark and how it is used in forensic investigations

The definition of the toolmark has always been understood and depicted the same way over time [1–5].

A toolmark derives from the interaction between a harder tool and a softer surface, resulting in distinctive patterns. There are two main types of toolmarks based on this interaction. One type involves the tool pressing against the surface, creating impressed marks like those left by a hammer head. The other type occurs when the tool scratches the surface, producing striated marks as seen with cutting pliers. Often, tools can leave a combination of both impressed and striated marks, such as a screwdriver used to open a metallic box.

This paper focuses specifically on striated toolmarks, which are commonly encountered in forensic investigations. Striated toolmarks consist of continuous peaks and valleys, formed by the tool's sliding motion on the substrate surface. The characteristics of these striations, such as depth and width, depend on the surface roughness of the tool or its blade. The same tool can create marks that vary depending on how it is used, including the angle at which it is held.

In forensic analysis, questioned striated marks are not directly compared to the submitted tools of interest. Instead, examiners generate reference marks by replicating the alleged activities and adjusting how the tool interacts with a reference surface resembling the substrate of the questioned marks.

Toolmarks play a crucial role in various investigative contexts where tools are used to commit specific activities. This can range from crimes against property, attacks on safes and ATMs, to bomb manufacturing. Tools like crowbars, screwdrivers, and tongue and groove pliers may leave marks on doors, window frames, cylinders, safes, and other surfaces that forensic examiners secure and analyze.

For cases involving striated toolmarks, the primary objective of the forensic toolmark examiner (FTE) is to compare questioned marks to reference marks made with a tool of interest (TOI). This comparison helps determine if the TOI was used to produce the questioned marks or not. The task is essentially an evaluative one, as explained by Jackson [6]. Striated toolmarks are less commonly used for intelligence purposes, particularly in the absence of a specific tool of interest, due to the high level of variability between marks made by the same tool. Impressed marks, on the other hand, are more likely to be employed for linking cases due to their greater reproducibility.

Despite being widely used, the scientific foundation of forensic science disciplines, and especially toolmark analysis, has faced criticism from various quarters. Reports, papers and commentators have questioned the methods and conclusions of the discipline. In the following

* Corresponding author.
E-mail addresses: jean-alexandre.patteet@unil.ch (J.-A. Patteet), christophe.champod@unil.ch (C. Champod).

https://doi.org/10.1016/j.forsciint.2024.111997
Received 12 September 2023; Received in revised form 2 February 2024; Accepted 17 March 2024
Available online 18 March 2024

section, we will summarize some of these critiques and explore their potential impact on the practice of toolmark analysis.

1.2. Criticisms and challenges to the discipline

The 2016 President's Council of Advisors on Science and Technology (PCAST) report and its addendum [7,8] provide a comprehensive overview of the daily challenges faced by toolmark analysis examiners, including concerns regarding comparisons and court testimony. The PCAST report was commissioned after the, National Research Council (NRC), initial critical wake-up call directed at all traditional forensic disciplines, including toolmark examination, in 2009 [9].

The toolmark examination method is primarily subjective, relying on the examiner's training and experience to observe and form an opinion based on their interpretation of the observed features. Years of experience remain a commonly used yardstick of examiners' competence [10]. However, several studies have highlighted that forensic toolmark examiners (FTEs) may be prone to overconfidence, leading to exaggerated claims and errors [11–14]. While the overall skills of FTEs are not in question, it is crucial for forensic evidence presented in court to be supported by sound scientific methods, principles, and demonstrable performance under controlled conditions [7–9,15]. The ongoing Genrich case [16], where cutting pliers were used to manufacture explosive devices, exemplifies how reports such as the NRC and PCAST have influenced the debate over the validity of toolmark evidence, leading to increased scrutiny of the discipline's methods and conclusions.

These issues raised by PCAST were not new. US jurisprudence, both at the federal and state levels, under the Frye or Daubert standard, has addressed challenged toolmark cases [17–19]. A few cases have established jurisprudence supporting the lack of reliable and admissible testimony [20]. The lack of demonstrated reproducibility and reliability in published studies has also raised questions about the validity of toolmark examination methods [9].

Some have argued that the subjective nature of toolmark examination, even when examiners count consecutive striations, might not meet the basic admissibility criteria established by the Daubert decision [21]. Consequently, there is a growing demand for more objective methods that rely less on the individual examiner's expertise and more on measurable features and statistical analysis. Many studies often focus on comparing toolmarks produced by different tools, yet they tend to overlook their counterpart: marks produced by the same tool. Consequently, research tends to emphasize between-source variability while not adequately considering within-source variability [20]. For example, Miller's study [22] examines the reproducibility of bolt cutter marks over ten years. However, the study heavily relied on the subjective comparison

method used by FTEs, raising questions about its objectivity and reliability. Criticisms have also been made regarding how examiners render their reports and testify in court.

To provide some context, the Association of Firearm and Tool Mark Examiners (AFTE) **Theory of Identification** [2,23] serves as a widely used reference in the field of toolmark examination. It relies on the examiner's ability to identify "sufficient agreement" between marks to conclude that they were made by the same tool. However, the PCAST report criticized this method for circular reasoning and the absence of objective criteria for determining "sufficient agreement".

Spiegelman and Tobin have argued that the AFTE theory does not meet the criteria of a scientific protocol for various reasons. Notably, one of the main criticisms is that the theory has never been subjected to testing for its error rate, a fundamental requirement for any scientific protocol [24]. The categorical conclusions supported by this reasoning have faced significant criticism from courts, which have strongly advised against phrasing such as "identification to the exclusion of all other tools".

Another concern is that the theory lacks specific protocols on how to conduct comparisons and evaluate findings. Without such protocols, examiners cannot reliably assess whether a tool may or may not be the origin of a toolmark [25].

The field of toolmark examination significantly lacks research based on systematic acquisition of features and comparison algorithms (error rate studies). Moreover, there is a lack of studies that combine method results with guidance on how to interpret and present these results in a report. Such research gaps hinder the advancement and validation of toolmark examination practices.

The next part will review the different methods used today by FTEs.

Toolmarks examination encompasses firearms examination and while reviews have already been conducted on this matter [26,27], we will still discuss the most relevant publications which, in our opinion, inform the issues associated to toolmarks as a whole and not only for firearms. Moreover, we would like to draw attention to the correspondence by Stamouli et al. in 2021, where they published a list of innovations that they believe should be implemented in the field of firearms examination. These requirements, including developing methods for the evaluation of findings, incorporating 3D surface methods, and objectifying the comparison process, are equally relevant and applicable to other striated toolmarks beyond firearms examination [28].

2. FTEs comparison methods

The following methods are presented in order of appearance over the years.

2.1. The visual pattern matching approach

FTEs have always utilized microscopes and comparison microscopes to observe toolmarks. This approach, called pattern matching, is based on describing and aligning striated patterns under controlled magnification. The method involves a side-by-side comparison between a questioned and a reference mark, directly used to assess the origin of the mark [29–31]. The reliability of conclusions derived from this comparison method is often justified by the FTEs' extensive experience, having seen numerous similar images, and their thorough training, which covers tool production and surface characteristics. Additionally, months of work under the supervision of seasoned experts contribute to their expertise. However, it is essential to acknowledge the subjective nature of this task.

To reduce subjectivity, researchers have sought quantitative methods to measure observations. As a result, they have developed and tested some numerical criteria.

2.2. The numerical criteria: manually counting corresponding striae

Three quantitative methods were initially proposed, but only one was successful and accepted by the FTE community. Researchers first tested the accuracy of counting the quantity of striae that were in agreement between the questioned and reference marks. However, they quickly realized that it was not an appropriate method, as the mark's size directly impacted the absolute count. Similarly, calculating the proportion of corresponding striae to the total number of striae led to high error rates [32–34], so this approach was also abandoned.

Biasotti came up with the idea, which was later accepted, that counting "consecutive matching striae" (CMS) was the best way to establish a numerical criterion [35–39]. If a FTE observes 6 CMS or more, the conclusion of an identification can be claimed, as no comparison between different sources has shown that amount of CMS. However, CMS faced skepticism from the scientific community [21,40].

Indeed, CMS, like the pattern matching approach [41], still exhibits subjectivity. Although the CMS numerical criterion provides a threshold for reporting conclusions, the act of recognizing and counting consecutively corresponding striae depends on the examiner. The method was also designed to avoid finding false positives, making it more inclined

towards “identification” according to the AFTE guidelines, rather than “exclusion” and “inconclusive” findings. This 6 CMS threshold for “identification” introduces bias, as examiners may be more likely to look for a 6th striation if they only have 5 [42], similar to fingerprint examiners relying on a 12 or 16 point criteria for identification [43,44].

Regardless of the numerical criteria used, these approaches lack objectivity. To address this issue and meet scientific requirements, which will later be discussed by PCAST, many authors recommended the implementation of automated techniques [45–49].

2.3. The use of automatic systems

The following paragraphs will present current toolmarks research that is dedicated to automatic methods used to analyze, compare, and evaluate toolmarks.

The use of automated systems has been a part of research for a long time. In 1981 already, Deinet developed and tested different models to compare striated marks, laying the groundwork for these types of studies [50,51]. Unfortunately, there were not many follow-ups in the 80 s. The underlying idea behind using automatic methods is to capture images or topographies in a reproducible manner, reducing dependency on the operator, and leveraging computer algorithms to systematically compare features (or a simplified version of them).

In the 1990s, new methods for collecting images from toolmarks emerged, incorporating techniques used in other scientific domains such as medicine. For example, the video microscope was implemented in toolmark examination in 1990 [52,53]. Database systems like TRAX were also developed and used to standardize the illumination process [54–56]. Nowadays, the illumination conditions remain one of the drawbacks of manual comparison microscopes, as they highly depend on the examiner and influence the appearance of the visualized toolmark [57,58].

An illustration of this phenomenon is shown in Fig. 1.

The TRAX system also calculated a score based on gray level values and provided a list of potential associated marks in a database of previously acquired images [54,56]. The search against a database involved one-to-many comparisons, but it can also apply to one-to-one comparisons.

A significant step forward was made when publications started to provide performance metrics as the lack thereof was a recurrent concern raised by the NRC report in 2009 [9]. The reader can refer to Mattijssen’s paper for an explanation of the different ways of presenting error rates [59]. In combination with these metrics, examiners started to use instruments which capture reproducible images such as 3D microscopes.

De Kinder and Bonfanti were one of the first to use a laser profilometer and extract profiles from striated bullet marks [57]. Leading the way for numerous articles that will form the basis of their research on the use of 3D images of marks reduced as profiles and compared with simple correlation coefficients [60–64].

Heizmann presented the GE/2 system, which uses various illuminations combined to produce a final result [58,65]. The system automatically straightens 2D images, enabling the study of curved or damaged samples. Profiles are extracted for comparison, and the cross-correlation function (CCF) is applied for the comparison process. The system’s performance is evaluated by calculating the percentage of hits in a given sized database. Additionally, emphasis was given to the necessary data pre-processing of 2D and 3D images [66]. It also shows that groove-shaped toolmarks can have irrelevant parts in-between striated parts of the mark, which brings artifacts into the profiles extracted and increases variability. Heizmann goes in depth into methods to detect these irrelevant parts and segment striated marks to reduce the effect of those unwanted parts.

Baldwin et al. [67] compared types of tools distinguished by their different manufacturing processes. They acquired 2D images of marks and developed an alignment algorithm that simulated the work of FTEs under a comparison microscope. A simple correlation function was used to obtain comparison scores. They reported error rates ranging from 3% to 50%, depending on the type of tools considered. This illustrates that some manufacturing processes lead to reproducible characteristics that are more difficult to distinguish (molding), while others create more variability between marks (milling). A process like molding will create repetitive characteristics on marks made by tools of the same model which is detrimental when examiners are trying to assess the source (one single tool) of a toolmark. Moreover, the authors explained that their algorithm was specifically based on a given set of tools and had less discrimination power on other types of tools not used in the algorithm’s development. This highlights the importance of adopting a method as specific as possible to the tool under examination because the methods developed may lack generalization capabilities across tools.

A few years later, Bachrach et al. published a study on screwdriver and tongue and groove plier marks [68]. They used a confocal microscope to obtain 3D topographies of the marks. The 3D information was then reduced to a 1D mean profile. The similarity measure used here is called ‘relative distance’ and is based on locally normalized squared distances. Toolmarks from screwdrivers were compared, and parameters such as the angle of the tool and the material of the substrate were

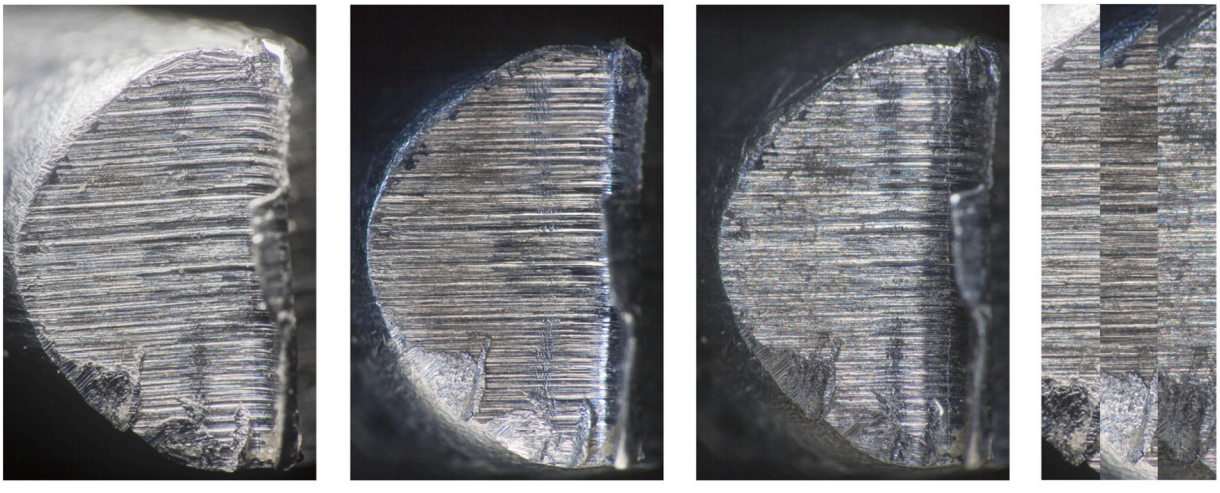

Fig. 1. The figure illustrates the variation in observation of the same cutting plier toolmark under different illumination conditions. The toolmark images are presented side by side in (d). As a result, the appearance of the striated patterns on the toolmark exhibits noticeable differences.

varied. They showed that the error rates increased when toolmarks from different angles were compared or if different substrates were used. Error rates below 0.5% were reported when comparisons were carried out under similar angles and materials, whereas error rates could rise up to 50% for comparisons with differences in angles or substrates. This demonstrates that parameters such as angle and substrate influences toolmarks variability, and FTEs should be aware of how they create references. For example, when creating reference toolmarks from the same source, examiners should control and reuse the same parameters as precisely as possible. Similar recommendations were later supported by other research [69,70]. Different substrates interact differently with the tool; therefore, it is best practice to use a reference material as close as possible to the one bearing the marks under investigation.

Chumbley et al. also acquired 3D data of screwdriver toolmarks using a confocal microscope [71,72]. As in Bachrach et al., topographies were reduced to 1D profiles. Using profiles is a way to represent the average topography as opposed to retaining noisy 3D or 2D information. Striated toolmarks are created by the sliding motion, which is not always linear and will create variation in 2D. The length of the mark will also depend on how the tool is used; a sliding screwdriver toolmark could be a few millimeters or a few centimeters long. The case is, of course, different for impressed toolmarks, which will reproduce size more easily. More thought was put into the development of the metric, which is based on correlations of various portions of profiles. A final 'Mann-Whitney U-Statistic' is performed, and a metric called 'T1' is used. The authors state that a high 'T1' value supports the hypothesis of a common origin. Depending on a set threshold value, the measures of false positive and false negative probabilities will vary, but the reported overall error rate is around 11%. Hadler and Morris based their research on Chumbley's method and improved the algorithm, resulting in a total error rate of 3% on their dataset [73].

As in Bachrach's article, the angle of the screwdriver is modified and marks made with different angles are compared. It is stated that with a 10° difference, marks from the same tool start to be distinguishable from each other and that around 25°, it is like comparing marks made by different tools. These result mean that examiners should keep the angle as steady as possible when building a dataset representing within-source variability. It also means that an extensive reference production (allowing for many angles) should be performed to make sure that the whole range of marks from a single tool is covered Additional publications of interest followed Chumbley et al., aiming at improving the algorithm, but the authors did not report error rates [74,75].

In 2014, Baiker et al. used 2D and 3D screwdriver toolmarks data for their research [76]. They were one of the first to implement the calculation of a likelihood ratio (LR), a metric used in forensic science to describe the strength of the observations in support of a common source as opposed to different sources. The LR was computed using the frequency distributions of cross-correlation scores. Error rates were below 1% when toolmarks were made with the same angle but started to increase (4%-30%) when different angles were compared to each other. This confirmed what was previously discussed in screwdrivers articles and is expected. The tilting angle changes the part of the tool in contact with the substrate, creating variation in the toolmarks. They also managed to demonstrate that, in such cases, the algorithm performed better than examiners, and overall profiles from 3D topographies gave better results than the ones from 2D images.

In 2019, Chen et al.introduce the Congruent Matching Profile Segments (CMPS) method [77]. It divides profiles into segments and correlates these segments to a base profile. The number of segments with high correlation and the position at which the correlation function has the best result is what forms the CMPS metric. To implement this method, a 3D confocal microscope was used to capture land engraved areas (LEA) on bullets. Images are processed and filtered to remove the general topography of the bullet curvature and the noise to segment the striated pattern. A bullet is then reduces as a set of profiles, one for each LEA. For comparison, each LEA profile of a bullet is compared to each LEA profile of another. The number of segments is one of the CMPS method parameters and its value will influence the result. This parameter has to be adjusted for each dataset. For the datasets composed of 45 same source and 549 different-sources comparisons, the CMPS method showed a better separation than the cross correlation score only. This method is better suited when profiles have varying lateral and vertical scales or when some parts of the pattern is not well defined.

Roberge et al. used the similar approach with a new method for bullet comparison [27]. It is made of two scores, the Line Counting Score LCS) and the Pattern Matching Score (PMS). The profile extraction process from 3D images is similar with other studies where a flat view of the 360° topography is used and the algorithm automatically detects lands and grooves. PMS is a linear combination of the CCF and the absolute normalized difference (AND). AND is used for its versatility with vertical scale variance in profiles. LCS takes advantage of the normalized number of corresponding striae positions for peaks and valleys as defined in the article and in [78]. The authors also explain the importance of resolution when capturing images, a lower resolution resulted in higher error rates with the same method.

As examplified with the studies presented above, the objectivity of the methods is increased by the adoption of a statistical approach based on acquisitions and comparisons that are independent from the observer. These comparisons refer respectively to same source and from different sources transactions.

It is important to understand that marks from the same tool cannot be identical to themselves for various reasons. Parameters such as the tilting angle or axial rotation of a tool or the specific part of the tool used will result in marks that cannot be associated using their striated pattern. Contrasting with firearms where the striated marks on bullets are not drastically affected by consecutive shoots (but we acknowledge that they are to some degree), marks left by tools can be highly impacted by the way the tool is used. For some tools, the location on the tool (such as the position of the cut on a blade) can vary from one mark to another. In such cases, the source is not simply the tool that was used but more specifically the part/area/location of the tool that was used.

The next set of articles takes further advantage of machine learning (ML) techniques as a way to retrieve and associate toolmarks.

Gambino et al. in 2011 compared primer shear marks caused by sliding breach faces [79]. Even though the research dealt with marks left by firearms, it marked the early use of ML techniques in this field, taking advantage of more elaborated comparison metrics. As adopted by previous authors [68,71,74,75,80], a confocal microscope was used to capture the topography with enough resolution and the capacity to deal with the angle variations of the surface. After extracting 1D profiles from topographies and aligning them using the CCF, principal component analysis is applied ('PCA') to reduce dimensionality, followed by support vector machines ('SVM') as a classification tool. Conformal prediction theory ('CPT') allowed the authors to reach an error rate of 3.5% on their limited dataset.

Petraco et al. in 2011 and 2013 used a very similar approach to Gambino but on screwdriver toolmarks [81,82] as well as bullets [83]. By using slightly different validation methods such as hold-one-out cross- validation ('HOO-CV') or 'bootstrap', the authors reported error rates between 1% and 10%.

Hare et al.'s publication in 2017 conduct LEA to LEA comparisons using random forest machine learning [84]. It focuses at first on the detection and location of the LEAs and the removal of bullet curvature. Once aligned with CCF, multiples comparison metrics are extracted (notably CMS, CCF, average difference, sum of peaks). All these metrics are predictors of the random forest classifier. The results for the considered dataset is promising with an error rate under 0.1%. Random forest also allowed to contrast the relative importance of the predictors and CCF outperformed all others metrics.

In 2020, Riva et al. implemented machine learning methods using PCA [85]. Even though this study was done on cartridge cases, the methodology could be applied to striated marks left by tools. Three

different metrics were used for all comparisons and reduced using PCA. The reduced principal components (PC1 and PC2) are thus used as new similarity scores. The same-source and different-sources distributions (probability densities) are then represented in a bi-dimensional space. The same-source variability is specific to a given firearm and ammunition. The LR is obtained by dividing the probability densities of the obtained scores under both distributions (same-source and different-sources respectively). The errors are expressed as rates of misleading evidence (RMEP and RMED) which means that if the value of the LR supports the hypothesis that is not the ground truth (a LR of 100 for a different source comparison for example), it is considered as an error. The authors show how the ammunition influences the error rates. Typically Winchester ammunition lead to 30–34% RMED and 7–10% RMEP whereas Geco Sintox ammunication had 0–0.3% RMED and 0.1% RMEP.

We believe that the machine learning methods are good alternatives to the variety of single score methods that have been published. ML techniques enable to use all metrics at the same time and take advantage of their synergies and complementarity.

Other papers have been published without fully reporting the error rates but are showing the efforts made to systematically compare toolmarks. Ahvenainen showed what interferometry could display and compared profiles of marks made by the same cutting plier [86]. Heikkinen used the same acquisition technology and compared profiles visually using the CMS criteria. The performance is given by the amount of marks that were correctly linked to the originating tool (74/80) [87].

Keglevic implemented the use of 'neural networks' to compare screwdriver marks from the Netherlands Forensic Institute (NFI) database [88,89]. The performance is described by different metrics, such as the mean average precision ('MAP').

Most articles cited above do not delve into the reporting of findings according to the method that was developed. Some of them support the AFTE theory of identification, which means the reporting for court purposes is made in terms of identification, exclusion, or inconclusive results. However, this model has been criticized over the years, and researchers need to adapt and discuss their results in another way. Baiker et al. addressed this issue by implementing a proper and recognized statistical framework based on the likelihood ratio [76]. Others also indicated in their perspectives that it would be beneficial to present the results according to a probabilistic framework, such as using likelihood ratios (LRs) [73,77,90,91].

Automatic comparison systems used to assign a likelihood ratio are underpinned by databases obtained from defined datasets. To implement such comparison and assessment method, researchers need to build two distinct datasets. The first is made of comparisons between marks from the same source (aka within-source variability) and the second one comes from comparisons between marks from different sources (aka between- sources variability). By creating such datasets and using metrics (measures of the closeness or similarity between marks) for each comparison, examiners are able to build metrics distributions. The within-source and between-sources distributions form jointly a model that will underpin the LR calculation taken at the point obtained for the comparison between a questioned mark and a TOI reference mark. It is important to articulate how to build these distributions. Even though most articles described empirical distribution obtained from hundreds or even thousands of comparisons, they do not detail how to construct those datasets/distributions in operational conditions. Indeed, it has been shown that the results depend on the tool, its angle or the substrate. It is thus necessary in a real case to construct distributions that are specific to the case. Such an approach has been presented for firearms by Riva et al. [85]. In their publication, the authors started with a datasets of 60 cartridge cases for the within-source variability, which means recovering 60 ammunition components shot with the firearm of interest. They then lowered this number down to 7 cartridge cases showing that the error rates and associated LR values were not significantly affected. It does not mean that the number seven can be used for all firearms and toolmarks cases. It however opens an operational mechanism to control the number of marks required to approach the within-source distribution. Unfortunately, studies exploring how to construct within and between distributions for tools do not exist currently. However, we believe that the method used to approach the question for firearms can apply to striated toolmark examination.

Some additional toolmark specificities have to be considered though. As indicated by Riva et al. distributions cannot be constructed with marks from any firearm and ammunition. The same must apply to tools with the addition of the tool's usage parameter. For example, the within-source distribution of a cutting plier will have to be built with reference marks that have been produced in the same material and shape as the questioned mark, but also, and more importantly, by using only one location along the blade of the cutting plier. It means that for one cutting plier, multiple within-source distributions can be created for each location along the blade. Similarly, it means that multiple within-source distributions can be created for a single screwdriver for each rotating and tilting angle. To mitigate these possibilities in a given case, we advise examiners to determine which area (for cutting plier in our above example) or which angle (for screwdrivers in our above example) is the most suitable.

From the above review, we observe that technological advances in acquiring and comparing toolmarks open new avenues for the discipline, offering independent and measurable techniques that can compensate for subjectivity. In our opinion, these advancements should be promoted and incorporated into examination practices. These systems will be characterized by measured error rates and metrics describing the weight to be assigned to the correspondence (or lack thereof) between a mark and reference marks. Among these metrics, LR is increasingly being used. They will also present results on within and between variabilities and discuss how are these variabilities influenced by the toolmarks deposition process.

The LR is a probabilistic measure that contrasts with the current reporting practice, which is centered on only three options (identification, exclusion, and inconclusive), with two of them being categorical. In the next section, we review the current reporting practices and link them with the essentially probabilistic outputs of future comparison systems.

3. FTEs reporting practices

3.1. The AFTE theory of Identification

In the 1990s, the AFTE promulgated guidelines to FTEs defining the range of conclusions that can be reached following the comparison of toolmarks [2,23]. Three distinct conclusions were defined:

- *Identification* means there is 'sufficient agreement' between characteristics observed on the mark and the references such that it can be concluded, in the opinion of the examiner, that the TOI produced the questioned mark.
- *Exclusion* indicates significant disagreement of characteristics leading to the conclusion that the TOI did not produce the questioned mark.
- If there is insufficient similarities or differences to reach of the two-above definitive conclusions, the reached conclusion is said to be *inconclusive* which does not indicate if the questioned mark was made by the TOI or not.

The AFTE further defined the concept of "sufficient agreement" as follows (p.86): *This "sufficient agree- ment" is related to the significant duplication of random toolmarks as evidenced by the correspondence of a pattern or combination of patterns of surface contours. Significance is determined by the comparative examination of two or more sets of surface contour patterns comprised of individual peaks, ridges and furrows. Specifically, the relative height or depth, width, curvature and spatial relationship of the individual peaks, ridges and furrows within one set of surface contours*

are defined and compared to the corresponding features in the second set of surface contours. Agreement is significant when it exceeds the best agreement demonstrated between toolmarks known to have been produced by different tools and is consistent with agreement demonstrated by toolmarks known to have been produced by the same tool. The statement that "sufficient agreement" exists between two toolmarks means that the likelihood another tool could have made the mark is so remote as lo be considered a practical impossibility.

As previously mentioned by PCAST and other commentators, the AFTE definition of identification lacks clarity regarding what constitutes the "best agreement" and the meaning of "practical impossibility" from a probabilistic perspective. The current definition encourages FTEs to make identifications based on their personal satisfaction with the level of agreement. The subjective nature of this conclusion is acknowledged by the AFTE. In fact, their statement reads: *Currently, the interpretation of individualization/identification is subjective in nature, founded on scientific principles and based on the examiner's training and experience.*

However, we believe that the reference to "scientific principles" in this context is misplaced.

If the FTE does not feel that the observations enabled him or her to choose between exclusion and identification, the conclusion reached will be inconclusive. Hence this conclusion covered the wide spectrum between the two categorical conclusions without offering any nuances or appreciation of the weight provided by the observation for one or the other side.

The United States Department of Justice (DOJ), while attempting to harmonize conclusions, kept the same terms as the AFTE but advised FTEs to avoid complementary formulations that may suggest their opinions are facts. According to the DOJ, FTEs shall not use the term "individualize" or assert that two toolmarks originated from the same source "to the exclusion of all other sources" [92]. Although well-intentioned, this just adds a confusing layer to the understanding of the terms identification or exclusion. In our opinion, it merely postpones the problem.

Adopting a probabilistic framework overcomes most of the above difficulties and is presented below.

3.2. A probabilistic framework

Let's explore the concept with an example. A safe was forced during a burglary, and a striated toolmark is found on the edge of the safe door. Security CCTV footage shows that the perpetrator used what appears to be a flat head screwdriver. Thanks to information put forward by a witness, a person of interest is apprehended the next day, and a toolbox is seized from his apartment. Multiple tools are found, including a flat head screwdriver (the TOI).

The FTE prepared reference marks with this tool and compare them to the questioned toolmark. The forensic examination aims at discriminating the following set of propositions (*Hp* and *Hd*):

Hp. **:** *The screwdriver of interest is at the origin of the questioned mark*

Hd. **:** *Another unknown flat head screwdriver is at the origin of the questioned mark*

The propositions are derived from the framework of circumstances and in this case the alternative (*Hd*) invokes another unknown flat head screwdriver because of the technical observations obtained from the CCTV footage.

The probabilistic framework to help FTEs assess the strength of their findings in light of both propositions is based on the assignment of a LR, a.k.a. a Bayes factor [93]. The likelihood ratio derives from Bayes' Theorem (Eq. 1) and expresses how forensic findings (*E*) will probabilistically impact the probabilities associated with the propositions. The letter *I* refers to the technical background information (in our example, the information obtained regarding the tool used from the CCTV footage).

$$\frac{P(H_p|E,I)}{P(H_d|E,I)} = \frac{P(E|H_p,I)}{P(E|H_d,I)} \cdot \frac{P(H_p,I)}{P(H_d,I)} \quad (1)$$

When adopting a probabilistic framework and taking advantage of automatic systems, the FTE will compute a LR using the results of an algorithmic comparison system that has been qualified with known forensic error rates. The numerator of the LR will be informed by the 'reproducibility' of the characteristics of the marks under the proposition *Hp* (the hypothesis that the marks were made by the same tool) and the de- nominator of the LR will be informed by the 'selectivity' of the features under *Hd* (the hypothesis that the marks were made by different tools).

The 'reproducibility' can be determined by studying the variability of a metric between marks that have the same origin. The same metric will be used to assess the variability between marks of different origins, representing the 'selectivity' of the characteristics being compared.

The previous automatic systems we reviewed were based on a comparison metric (such as the correlation between images) to express the similarity between the questioned mark and the reference mark. Such metrics, commonly referred to as scores, can also be the basis of an LR-based system, as shown in the work of Jacquet and Champod [94]. In this case, we talk about score-based likelihood ratios, and they are not new in the area of striated marks. Recent developments dealing with marks left on ammunition elements are also computing score-based likelihood ratios [14,27,85,95–97].

There are fundamental differences with the AFTE approach when adopting the probabilistic approach. Firstly, with the probabilistic approach, the FTE does not provide a definite conclusion but expresses a LR that is essentially probabilistic in nature. It is crucial to avoid misunderstanding large LRs as certain evidence that the TOI is the source of the mark. It provides a probability of evidence given *Hp* divided by a probability of evidence given *Hd* (Eq. 1). Therefore, the factfinder should be fully informed of the probabilistic nature of the forensic findings, without creating an illusion of absolute certainty. Concluding to an "identification" based on an LR requires a decision mechanism that is beyond the scope of the FTE's duty [98]. When adopting an LR-based reporting scheme, the FTE will only express the strength of their findings in relation to the competing propositions.

In addition, the LR-based approach allows for the proper weighting of findings that are currently described as "inconclusive" under the AFTE framework, which often does not provide helpful information to the court [99]. Instead, the term "inconclusive" will now be reserved for likelihood ratios close to 1. An LR above 1 will provide support for the hypothesis *Hp*, while an LR below 1 will provide support for the hypothesis *Hd*. The magnitude of the LR simply indicates the strength of the support for the respective propositions.

The fundamental difference when using an LR-based reporting system is that the FTE takes a position regarding the observations or findings, rather than directly making conclusions about the propositions as done in the AFTE approach. This approach avoids a common syllogistic error known as the "prosecutor fallacy" [100,101] or transposing the conditional [102]. By presenting the LR, the FTE provides a quantitative measure of the strength of the evidence, allowing the factfinder to make more informed decisions.

We encourage the adoption of a LR-based probabilistic framework for reporting forensic findings as pro- posed by the European Network of Forensic Sciences (ENFSI) in their guideline for evaluative reporting [103].

We observe that, lately, the Organization of Scientific Area Committees for Forensic Science (OSAC) proposed a standard scale of conclusions that lies somewhere between the AFTE approach and a fully probabilistic framework. This standard was subsequently reviewed and revised by the Academy Standards Board (ASB), resulting in the first edition of the "Standard Scale of Source Conclusions Criteria for Toolmark Examinations" [104]. In this document, ASB presents three

conclusions:

- *Opinion of Same Source (Identification)* means that the observed characteristics of the questioned toolmark and the TOI references provide very strong support that they were marked by the same tool and very weak or no support that they were marked by different tools.
- *Inconclusive* indicates that the observed characteristics of the items in question are insufficient to support, either, that the items were marked by the same tool, or, that the items were marked by different tools.
- *Opinion of Different Source (Exclusion)* means that the observed characteristics of the questioned toolmark and the TOI references provide very strong support that they were marked by different tools and very weak or no support that they were marked by the same tool.

By choosing the same terms as the AFTE, but introducing the concept of support, the ASB merge explanations of support towards hypothesis with the decisions of identification and exclusion. One critical limitation of the ASB document is the lack of clarity on what constitutes "very strong support". The document does not specify whether the support is represented by a probabilistic value such as a LR, a comparison score, a count of CMS, or subjective observations. The ASB's proposed scale of conclusions provides some level of standardization in reporting toolmark examination results, but it still falls short of adopting a fully probabilistic framework.

Using this standard, there is a risk that the a "support" will be misconstrued as the probability of the proposition. A typical error knows as transposing the conditional [102]. As already suggested by [105] and [106], FTE should report on the value of the observations and not on the propositions themselves.

The association between propositions and decisions has been discussed by Biedermann et al. [107,108] in what is known as the "decision theory" and involves concepts such as risk, benefit, or consequences of the decisions, which are not discussed neither by the OSAC nor by the ASB.

In our opinion, efforts should be made to fully align the reporting practice with the probabilistic framework described above. It inevitably means that the terminology used by AFTE and other bodies will be revoked in favor of terms that truly express the degree of support the findings provide to the propositions under examination. The ENFSI guideline [103] paved the way.

4. Conclusion

PCAST and other organizations have addressed the major challenges regarding forensic toolmark examinations, highlighting concerns about the scientific validity and reliability of current methods. However, recent research has shown a positive response to these concerns.

The implementation of automatic methods coupled with statistical analyses is promising and addresses the issues of objectivity, reliability, and validity. To establish the reliability of a new method, it should be mandatory to study both same-source and different-source datasets.

Constructing datasets that can be used operationally is currently lacking and such an effort would reinforce the methods presented in most publication and increase the potential for them to be used in operational casework.

Moreover, the validity of the methods should be supported by measured error rates. Efforts should be made to clearly link the results to their interpretation and reporting, providing transparency in the decision- making process.

The current views in line with the 1998 AFTE theory of identification bring forth too many problems, mainly linked to the lack of transparency in justifying the conclusions or decisions. The mixing of hypotheses and decisions in the current standards or drafts can lead to confusion and misinterpretation.

To overcome these issues, a probabilistic framework should be adopted. Automatic methods are compatible with such frameworks and have already been successfully developed in firearms examination, for example. By adopting a probabilistic approach, the field of toolmark examination can enhance its scientific basis and provide more informative and transparent conclusions to the courts.

CRediT authorship contribution statement

Christophe Champod: Supervision, Writing – original draft, Writing – review & editing. **Jean-Alexandre Patteet:** Conceptualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

None

References

[1] D.Q. Burd, R.S. Greene, Tool mark comparisons in criminal investigations, J. Crim. Law Criminol. 39 (3) (1948) 379–391.
[2] AFTE, Theory of identification, range of striae comparison reports, and modified glossary definitions-an afte criteria for identification committee report, AFTE J. 24 (3) (1992) 336–340.
[3] C.R. Meyers, Firearms and toolmark identification - an introduction, AFTE J. 25 (4) (1993) 281–285.
[4] H. Katterwe, Toolmarks, Vol. 5 of Major Reference Works, Wiley, Chichester, UK, 2009, book section Toolmarks, pp. 1–10. doi:10.1002/9780470061589.fsa36510.1002/9780470061589.fsa365.URL ⟨https://doi.org/10.1002/9780470061589.fsa365⟩.
[5] N. Petraco, Color Atlas of Forensic Toolmark Identification, CRC Press, Boca Raton, 2011.
[6] G. Jackson, S. Jones, G. Booth, C. Champod, I.W. Evett, The nature of forensic science opinion – a possible framework to guide thinking and practice in investigations and in court proceedings, Sci. Justice 46 (1) (2006) 33–44, https://doi.org/10.1016/S1355-0306(06)71565-9.
[7] PCAST, Report to the president, forensic science in criminal courts: Ensuring scientific validity of feature comparison methods, Report, Executive Office of the President President's Council of Advisors on Science and Technology (2016). URL ⟨https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf⟩.
[8] PCAST, An addendum to the pcast report on forensic science in criminal courts, Report, Executive Office of the Presi- dent (January 6 2017).URL ⟨https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_addendum_finalv2.pdf⟩.
[9] National Research Council, Strengthening Forensic Science in the United States: A Path Forward, The National Academies Press, Washington, D.C., 2009.
[10] U.S. v Marlon, case No.13 - CF - 1312 (Jan 21 2016 2016).
[11] E.J. Mattijssen, C.L.M. Witteman, C.E.H. Berger, N.W. Brand, R.D. Stoel, Validity and reliability of forensic firearm examiners, Forensic Sci. Int. 307 (2020) 110112, https://doi.org/10.1016/j.forsciint.2019.110112. ⟨http://www.sciencedirect.com/science/article/pii/S0379073819305249⟩ (URL).
[12] A. Nordgaard, R. Ansell, W. Drotz, L. Jaeger, Scale of conclusions for the value of evidence, Law, Probab. Risk 11 (1) (2012) 1–24, https://doi.org/10.1093/lpr/mgr020. ⟨http://lpr.oxfordjournals.org/content/11/1/1.abstract⟩ (URL).
[13] A. Biedermann, P. Garbolino, F. Taroni, The subjectivist interpretation of probability and the problem of individualisa- tion in forensic science, Sci. Justice 53 (2) (2013) 192–200, https://doi.org/10.1016/j.scijus.2013.01.003. ⟨http://ac.els-cdn.com/S135503061300004X/1-s2.0-S135503061300004X-main.pdf?_tid=570a38bc-79f5-11e3-8a6c-00000aacb35e&acdnat=1389358218_1cf3347542f5e5240588abb3c25267dc⟩.
[14] E.J. Mattijssen, C.L.M. Witteman, C.E.H. Berger, X.A. Zheng, J.A. Soons, R. D. Stoel, Firearm exami- nation: Examiner judgments and computer-based comparisons, J. Forensic Sci. 66 (1) (2021) 96–111, https://doi.org/10.1111/1556-4029.14557, doi:10.1111/1556-4029.14557. URL https://doi.org/10.1111/1556-4029.14557 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7821150/pdf/JFO-66-96.pdf.
[15] M.J. Saks, D.M. Risinger, R. Rosenthal, W.C. Thompson, Context effects in forensic science: A review and application of the science of science to crime laboratory practice in the united states, Sci. Justice 43 (2) (2003) 77–90.
[16] People v. Genrich, case 1019COA132 (2019).
[17] R. Grzybowski, J. Miller, B. Moran, J. Murdock, R. Nichols, R. Thompson, Firearm/toolmark identification: passing the reliability test under federal and state evidentiary standards, AFTE J. 35 (2) (2003) 209–241.
[18] D.L. Faigman, D.H. Kaye, M.J. Saks, J. Sanders, E.K. Cheng, Firearms and Toolmark Identification, 2006th Edition, Vol. 4, Thomson/West, 2007, book section 36, pp. 525–584.
[19] State of Florida v. Ramirez, case SC92975 (December 20, 2001 2001). [link].URL ⟨https://casetext.com/case/ramirez-v-state-237⟩.

[20] A. Schwartz, A systemic challenge to the reliability and admissibility of firearms and toolmark identification, Columbia Sci. Technol. Law Rev. 6 (2005) 1–42.
[21] A. Schwartz, A challenge to the admissibility of firearms and toolmark identifications: Amicus brief prepared on behalf of the defendant in united states v. kain, crim. 03-573-1 (e.d. pa. 2004), The Journal of Philosophy, Science and Law 4 (December 7) (2004) ⟨http://www.psljournal.com/archives/all/kain.cfm⟩.
[22] J. Miller, An evaluation of the persistence of striated and impressed toolmarks encompassing a ten-year period of tool application, and a summary of forensic research on bolt cutters, AFTE J. 38 (4) (2006) 310–326.
[23] AFTE, Theory of identification as it relates to toolmarks, AFTE J. 30 (1) (1998) 86–88.
[24] C. Spiegelman, W.A. Tobin, Analysis of experiments in forensic firearms/ toolmarks practice offered as support for low rates of practice error and claims of inferential certainty, Law Probab. Risk (2012). ⟨http://lpr.oxfordjournals.org/content/early/2012/10/01/lpr.mgs028.abstract⟩ (URL).
[25] W.A. Tobin, P.J. Blau, Hypothesis testing of the critical underlying premise of discernible uniqueness in firearms- toolmarks forensic practice (Winter), Jurimetr. J. 53 (2013) 121–142.
[26] E.J. Mattijssen, W. Kerkhoff, R. Hermsen, R.A.G. Hes, Interpol review of forensic firearm examination 2019–2022, Forensic Sci. Int.: Synerg. 6 (2023) 100305, https://doi.org/10.1016/j.fsisyn.2022.100305.URL. ⟨https://www.sciencedirect.com/science/article/pii/S2589871×22000900⟩.
[27] D. Roberge, A. Beauchamp, S. Levesque, Objective identification of bullets based on 3d pattern matching and line counting scores, 1940021–1 – 1940021–34, Int. J. Pattern Recognit. Artif. Intell. 33 (11) (2019), https://doi.org/10.1142/S0218001419400214, ⟨https://doi.org/10.1142/S0218001419400214⟩, 1940021–1 – 1940021–34.
[28] A. Stamouli, A. Walters, Correspondence: firearms and gunshot residue-description of the fields and future perspectives, AFTE J. 53 (1) (2021) 3–8.
[29] P. Lane, Toolmarks on battery terminals, AFTE J. 20 (2) (1988) 151–153.
[30] C.R. Meyers, Toolmarks on a plastic bag, AFTE J. 20 (1) (1988) 55–56.
[31] F.H. Cassidy, An unusual toolmark from a bolt cutter, AFTE J. 26 (1) (1994) 21–22.
[32] J. Miller, Criteria for identification of toolmarks, AFTE J. 30 (1) (1998) 15–61.
[33] S.J. Butcher, P.D. Pugh, A study of marks made by bolt cutters, J. Forensic Sci. Soc. 15 (2) (1975) 115–126.
[34] J. Miller, M. Neel, Criteria for identification of toolmarks part iii* supporting the conclusion, AFTE J. 36 (1) (2004) 7–38.
[35] A.A. Biasotti, J. Murdock, "criteria for identification" or "state of the art" of firearm and toolmark identification, AFTE J. 16 (4) (1984) 16–34.
[36] A.A. Biasotti, F.A. Tulleners, California doj training syllabus – forensic firearms and toolmark identification modules 1 and 2, AFTE J. 16 (1) (1984) 30–118.
[37] A.A. Biasotti, F.A. Tulleners, Training syllabus for forensic firearms and toolmark identification modules 3 and 4, AFTE J. 16 (2) (1984) 29–102.
[38] T. Uchiyama, The probability of corresponding striae in toolmarks, AFTE J. 24 (3) (1992) 273–290.
[39] R.G. Nichols, Consecutive matching striations (cms): Its definition, study and application in the discipline of firearms and tool mark identification, AFTE J. 35 (3) (2003) 298–306.
[40] M. Arosio, Adina schwartz est-elle crédible ?, Bachelor thesis, Sch. Crim. Justice (2014). ⟨https://esc-app.unil.ch/resourcespace/pages/view.php?ref=696&search=adina&order_by=relevance&sort=DESC&offset=0&archive=0&k=&curpos=0&restypes=2⟩.
[41] G. Wevers, M. Neel, J. Buckleton, A comprehensive statistical analysis of striated tool mark examinations, part 2: Comparing known matches and known non-matches using likelihood ratios, AFTE J. 43 (2) (2011) 137–145.
[42] S.A. Cole, Implementing counter-measures against confirmation bias in forensic science, J. Appl. Res. Mem. Cogn. 2 (1) (2013) 61–62, https://doi.org/10.1016/j.jarmac.2013.01.011.URL. ⟨http://www.sciencedirect.com/science/article/pii/S2211368113000120⟩.
[43] I.W. Evett, R. Williams, A review of the sixteen points fingerprint standard in england and wales, J. Forensic Identif. 46 (1) (1996) 49–73.
[44] B.T. Ulery, R.A. Hicklin, G.I. Kiebuzinski, M.A. Roberts, J. Buscaglia, Understanding the sufficiency of information for latent fingerprint value determinations, Forensic Sci. Int. 230 (1-3) (2013) 99–106.
[45] S.G. Bunch, Consecutive matching striation criteria: a general critique, J. Forensic Sci. 45 (5) (2000) 955–962.
[46] D. Meuwly, P. Margot, Fingermarks, shoesole and footprint impressions, tire impressions, ear impressions, toolmarks, lipmarks, bitemarks - a review (sept 1998 - aug 2001), 13th Interpol Forensic Science Symposium D1 (2001) 1–52.
[47] B. Moran, Toolmark criteria for identification: Pattern match, cms, or bayesian, INTERfaces 28 (2001) 9–10.
[48] B. Moran, A report on the afte theory of identification and range of conclusions for tool mark identification and resulting approaches to casework, AFTE J. 34 (2) (2002) 227–235.
[49] C. Champod, D. Baldwin, F. Taroni, J.S. Buckleton, Firearm and tool marks identification: The bayesian approach, AFTE J. 35 (3) (2003) 307–316.
[50] W. Deinet, Studies of models of striated marks generated by random processes, J. Forensic Sci. 26 (1) (1981) 35–50.
[51] H. Katterwe, W. Deinet, Anwendung eines wahrscheinlichkeitstheoretischen modells zur bewertung des Übereinstim- mungsgrades von spurenmustern, Arch. F. üR. Kriminol. 171 (1983) 78–88.
[52] E.E. Hueske, A preliminary report on the application of fiber optic videomicroscopy to firearm and tool mark examination, AFTE J. 22 (3) (1990) 280–287.
[53] E.E. Hueske, The application of fiber optic videomicroscopy to firearm and tool mark examination – a further look, AFTE J. 25 (2) (1993) 132–139.
[54] Z. Geradts, J. Keijzer, I. Keereweer, A new approach to automatic comparison of striation marks, J. Forensic Sci. 39 (4) (1994) 974–980.
[55] Z. Geradts, J. Keijzer, I. Keereweer, Automatic comparison of striation marks and automatic classification of shoe prints, SPIE's 1995 International Symposium on Optical Science, Engineering, and Instrumentation 2567 (1995) 151–164. URL ⟨https://doi.org/10.1117/12.218471⟩.
[56] Z. Geradts, J. Keijzer, Trax for toolmarks, AFTE J. 28 (3) (1996) 183–190.
[57] J. De Kinder, M. Bonfanti, Automated comparisons of bullet striations based on 3d topography, Forensic Sci. Int. 101 (2) (1999) 85–93.
[58] M. Heizmann, Automated comparison of striation marks with the system ge/2, in: Z. J. Geradts, L. I. Rudin (Eds.), SPIE International Symposium on Law Enforcement Technologies – Investigative Image Processing II, Vol. 4709, SPIE, 2002, pp. 80–91.
[59] E.J. Mattijssen, C.E. Berger, P. Vergeer, W. Kerkhoff, R.D. Stoel, Firearm evaluation at source level: How to define the relevant population and how to apply an unrestrictive alternative proposition, PDF hosted Radboud Repos. Radboud Univ. Nijmegen (2021).
[60] F. Puente León, Automated comparison of firearm bullets, Forensic Sci. Int. 156 (1) (2006) 40–50, https://doi.org/10.1016/j.forsciint.2004.12.016. ⟨https://www.sciencedirect.com/science/article/pii/S0379073805000022⟩.
[61] A. Banno, Estimation of bullet striation similarity using neural networks, J. Forensic Sci. 49 (3) (2004) 500–504.
[62] F. Xie, S. Xiao, L. Blunt, W. Zeng, X. Jiang, Automated bullet-identification system based on surface topography techniques, Wear 266 (5-6) (2009) 518–522.
[63] B. Bachrach, A statistical validation of the individuality of guns using 3d images of bullets, Report 213674, Natl. Inst. Justice (2006).
[64] W. Chu, J. Song, T. Vorburger, J. Yen, S. Ballou, B. Bachrach, Pilot study of automated bullet signature identification based on topography measurements and correlations*†, J. Forensic Sci. 55 (2) (2010) 341–347, https://doi.org/10.1111/j.1556-4029.2009.01276.x. ⟨https://doi.org/10.1111/j.1556-4029.2009.01276.x⟩.
[65] M. Heizmann, F.P. Leon, Imaging and analysis of forensic striation marks, Opt. Eng. 42 (12) (2003) 3423–3432, https://doi.org/10.1117/1.1622389.
[66] M. Heizmann, Techniques for the segmentation of striation patterns, IEEE Trans. Image Process. 15 (3) (2006) 624–631, https://doi.org/10.1109/TIP.2005.863038.
[67] D. Baldwin, M. Morris, S. Bajic, Z. Zhou, J. Kreiser, Statistical tools for forensic analysis of toolmarks, Rep., Ames Lab., IA (US) (2004).
[68] B. Bachrach, A. Jain, S. Jung, R.D. Koons, A statistical validation of the individuality and repeatability of striated tool marks: Screwdrivers and tongue and groove pliers, J. Forensic Sci. 55 (2) (2010) 348–357.
[69] M. Baiker, R. Pieterman, P. Zoon, Toolmark variability and quality depending on the fundamental parame- ters: Angle of attack, toolmark depth and substrate material, Forensic Sci. Int. 251 (2015) 40–49, https://doi.org/10.1016/j.forsciint.2015.03.003. ⟨http://www.sciencedirect.com/science/article/pii/S0379073815001050Xhttps://ac.els-cdn.com/S037907381500105X/1-s2.0-S037907381500105X-main.pdf?tid=5e8bdfcc-294b-45b4-9b3c-5a27854405b6&acdnat=1539596432_7cd3429413c0d69e6d91a499ed328fcc⟩.
[70] D.L. Garcia, R. Pieterman, M. Baiker, Influence of the axial rotation angle on tool mark striations, Forensic Sci. Int. 279 (2017) 203–218, https://doi.org/10.1016/j.forsciint.2017.08.021. ⟨http://www.sciencedirect.com/science/article/pii/S0379073817303201https://www.sciencedirect.com/science/article/pii/S0379073817303201?via%3Dihub⟩.
[71] L.S. Chumbley, M.D. Morris, M.J. Kreiser, C. Fisher, J. Craft, L.J. Genalo, S. Davis, D. Faden, J. Kidd, Validation of tool mark comparisons obtained using a quantitative, comparative, statistical algorithm, J. Forensic Sci. 55 (4) (2010) 953–961, https://doi.org/10.1111/j.1556-4029.2010.01424.x.
[72] L.S. Chumbley, M. Morris, Significance of association in tool mark characterization, Report 243319, Natl. Inst. Justice (2013). ⟨https://www.ncjrs.gov/pdffiles1/nij/grants/243319.pdf⟩.
[73] J.R. Hadler, M.D. Morris, An improved version of a tool mark comparison algorithm, J. Forensic Sci. 63 (3) (2018) 849–855, https://doi.org/10.1111/1556-4029.13640. URL https://doi.org/10.1111/1556-4029.13640.
[74] T. Grieve, L.S. Chumbley, J. Kreiser, M. Morris, L. Ekstrand, S. Zhang, Objective comparison of toolmarks from the cutting surfaces of slip-joint pliers, AFTE J. 46 (2) (2014) 176.
[75] R. Spotts, L.S. Chumbley, L. Ekstrand, S. Zhang, J. Kreiser, Optimization of a statistical algorithm for objective com- parison of toolmarks, J. Forensic Sci. 60 (2) (2015) 303–314, https://doi.org/10.1111/1556-4029.12642.URL. ⟨https://doi.org/10.1111/1556-4029.12642http://onlinelibrary.wiley.com/store/10.1111/1556-4029.12642/asset/jfo12642.pdf?v=1&t=i7nbpba8&s=8be5b73419dd0c1559d36433e83a20ca35e4bcda⟩.
[76] M. Baiker, I. Keereweer, R. Pieterman, E. Vermeij, J. van der Weerd, P. Zoon, Quantitative comparison of striated tool- marks, Forensic Sci. Int. 242 (2014) 186–199, https://doi.org/10.1016/j.forsciint.2014.06.038. ⟨http://www.sciencedirect.com/science/article/pii/S0379073814002746https://ac.els-cdn.com/S0379073814002746/1-s2.0-S0379073814002746-main.pdf?_tid=401e0f6d-aba2-46cf-85c5-59a90fb6a27e&acdnat=1539596427_219a15d69e76f9cda439972514bacd60⟩ (URL).
[77] Z. Chen, W. Chu, J.A. Soons, R.M. Thompson, J. Song, X. Zhao, Fired bullet signature correlation using the congruent matching profile segments (cmps) method, Forensic Sci. Int. (2019) 109964, https://doi.org/10.1016/j.forsciint.2019.109964.URL. ⟨http://www.sciencedirect.com/science/article/pii/S0379073819303767⟩.

[78] W. Chu, R.M. Thompson, J. Song, T.V. Vorburger, Automatic identification of bullet signatures based on consecutive matching striae (cms) criteria, Forensic Sci. Int. 231 (1–3) (2013) 137–141, https://doi.org/10.1016/j.forsciint.2013.04.025.URL. ⟨http://www.sciencedirect.com/science/article/pii/S037907381300248Xhttp://ac.els-cdn.com/S037907381300248X/1-s2.0-S037907381300248X-main.pdf?_tid=de74fe54-406b-11e3-8fd2-00000aacb35f&acdnat=1383031959_b139b24785d378f301459907d2086488⟩.
[79] C. Gambino, P. McLaughlin, L. Kuo, F. Kammerman, P. Shenkin, P. Diaczuk, N. Petraco, J.E. Hamby, N.D.K. Petraco, Forensic surface metrology: Tool mark evidence, Scanning 33 (2011) 272–278.
[80] R. Bolton-King, J.P.O. Evans, C.L. Smith, J.D. Painter, D.F. Allsop, W.M. Cranton, What are the prospects of 3d profiling systems applied to firearms and toolmark identification? AFTE J. 42 (1) (2010) 23–33.
[81] N.D.K. Petraco, C. Gambino, F.L. Kammerman, Application of machine learning to toolmarks: Statistically based methods for impression pattern comparisons, Report, U.S Department of Justice (December 2011).URL ⟨https://www.ncjrs.gov/pdffiles1/nij/grants/239048.pdf⟩.
[82] N.D.K. Petraco, L. Kuo, H. Chan, E. Phelps, C. Gambino, P. McLaughlin, F. Kammerman, P. Diaczuk, S. Peter, N. Petraco, J.E. Hamby, Estimates of striation pattern identification error rates by algorithmic methods, AFTE J. 45 (3) (2013) 235–244.
[83] J. Monkres, C. Luckie, N.D.K. Petraco, A. Milam, Comparison and statistical analysis of land impressions from consecutively rifled barrels, AFTE J. Vol. 45 (2013) 3–20.
[84] E. Hare, H. Hofmann, A. Carriquiry, Automatic matching of bullet land impressions, Ann. Appl. Stat. (2017) 2332–2356.
[85] F. Riva, E.J.A.T. Mattijssen, R. Hermsen, P. Pieper, W. Kerkhoff, C. Champod, Comparison and interpretation of impressed marks left by a firearm on cartridge cases – towards an operational implementation of a likelihood ratio based technique, Forensic Sci. Int. 313 (2020) 110363, https://doi.org/10.1016/j.forsciint.2020.110363.URL. ⟨http://www.sciencedirect.com/science/article/pii/S0379073820302255⟩.
[86] P. Ahvenainen, I. Kassamakov, K. Hanhijärvi, J. Aaltonen, S. Lehto, T. Reinikainen, E. Hæggström, Csi helsinki: Swli in forensic science: Comparing toolmarks of diagonal cutting pliers, in: C.S.I. Helsinki: SWLI in Forensic Science: Comparing Toolmarks of Diagonal Cutting Pliers, Vol. 1211, AIP Conference Proceedings, 2010, pp. 2084–2091.
[87] V.V. Heikkinen, I. Kassamakov, C. Barbeau, S. Lehto, T. Reinikainen, E. Hæggström, Identifying diagonal cutter marks on thin wires using 3d imaging, J. Forensic Sci. 59 (1) (2014) 112–116, https://doi.org/10.1111/1556-4029.12291. ⟨https://doi.org/10.1111/1556-4029.12291http://onlinelibrary.wiley.com/store/10.1111/1556-4029.12291/asset/jfo12291.pdf?v=1&t=hqj9iqjz&s=43fe53f08c384c1c19f1a40546698ae79f64afa7⟩.
[88] M. Keglevic, R. Sablatnig, Learning a similarity measure for striated toolmarks using convolutional neural networks, in: 7th International Conference on Imaging for Crime Detection and Prevention (ICDP 2016), 2016, pp. 1–6. ⟨doi:10.1049/ic.2016.0069⟩.
[89] M. Keglevic, R. Sablatnig, Retrieval of striated toolmarks using convolutional neural networks, IET Comput. Vis. 11 (7) (2017) 613–619, https://doi.org/10.1049/iet-cvi.2017.0161.
[90] E.F. Law, K.B. Morris, C.M. Jelsema, Determining the number of test fires needed to represent the variability present within firearms of various calibers, Forensic Sci. Int. 290 (2018) 56–61, https://doi.org/10.1016/j.forsciint.2018.06.010. ⟨http://www.sciencedirect.com/science/article/pii/S037907381830330X⟩.
[91] A.H. Dorfman, R. Valliant, Inconclusives, errors, and error rates in forensic firearms analysis:three statistical perspectives, Forensic Sci. Int.: Synerg. (2022) 100273, https://doi.org/10.1016/j.fsisyn.2022.100273. ⟨https://www.sciencedirect.com/science/article/pii/S2589871×22000584⟩.
[92] US Department of Justice, Approved uniform language for testimony and reports for the forensic firearms/toolmarks discipline pattern match examination, Report, U.S. Department of Justice (2018).URL ⟨https://www.justice.gov/olp/page/file/1083671/download⟩.
[93] C.G.G. Aitken, F. Taroni, S. Bozza, Statistics and the Evaluation of Evidence for Forensic Scientists. Statistics in Practice, 3rd Edition, John Wiley and Sons, Chichester, 2020.
[94] M. Jacquet, C. Champod, Automated face recognition in forensic science: Review and perspectives, Forensic Sci. Int. 307 (2020) 110124, https://doi.org/10.1016/j.forsciint.2019.110124.URL. ⟨https://www.sciencedirect.com/science/article/pii/S0379073819305365⟩.
[95] F. Riva, C. Champod, Automatic comparison and evaluation of impressions left by a firearm on fired cartridge cases, J. Forensic Sci. 59 (3) (2014) 637–647, https://doi.org/10.1111/1556-4029.12382. ⟨https://doi.org/10.1111/1556-4029.12382http://onlinelibrary.wiley.com/store/10.1111/1556-4029.12382/asset/jfo12382.pdf?v=1&t=hsx7lepj&s=274f4925a6c9b73558f8f6ce3bd68047245374fe⟩.
[96] F. Riva, R. Hermsen, E. Mattijssen, P. Pieper, C. Champod, Objective evaluation of subclass characteristics on breech face marks, J. Forensic Sci. 62 (2) (2017) 417–422, https://doi.org/10.1111/1556-4029.13274. ⟨https://doi.org/10.1111/1556-4029.13274http://onlinelibrary.wiley.com/store/10.1111/1556-4029.13274/asset/jfo13274.pdf?v=1&t=j00mday6&s=64823767c4425e1ba1314dd17692d4d832534c05⟩.
[97] J. Song, H. Song, Reporting likelihood ratio for casework in firearm evidence identification, J. Forensic Sci. (2022), https://doi.org/10.1111/1556-4029.15186. ⟨https://doi.org/10.1111/1556-4029.15186https://onlinelibrary.wiley.com/doi/pdfdirect/10.1111/1556-4029.15186?download=true⟩. doi: https://doi.org/10.1111/1556-4029.15186. URL.
[98] C. Champod, A. Biedermann, Overview and Meaning of Identification/Individualization, 3rd Edition, Vol. 4, Elsevier, Oxford, 2023, pp. 53–62 (book section Overview and Meaning of Identification/Individualization).
[99] G. Dutton, Considerations for adoption of an evaluative reporting framework for the interpretation of firearms and toolmarks evidence, AFTE J. 49 (4) (2017) 239–251.
[100] W.C. Thompson, E.L. Schumann, Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defence attorney's fallacy, Law Hum. Behav. 11 (3) (1987) 167–187, https://doi.org/10.1007/BF01044641.
[101] W.-C. Leung, The prosecutor's fallacy – a pitfall in interpreting probabilities in forensic evidence, Med., Sci. Law 42 (1) (2002) 44–50.
[102] I.W. Evett, Avoiding the transposed conditional, Sci. Justice 35 (2) (1995) 127–131.
[103] S. Willis, L. McKenna, S. McDermott, G. O'Donnell, A. Barrett, B. Rasmusson, T. Hoglund, A. Nordgaard, C. Berger, M. Sjerps, J. Molina, G. Zadora, C. Aitken, T. Lovelock, L. Lunt, C. Champod, A. Biedermann, T. Hicks, F. Taroni, Enfsi guideline for evaluative reporting in forensic science, Rep., Eur. Netw. Forensic Sci. Inst. (2015). ⟨http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf⟩.
[104] Academy Standards Board (ASB), Firearms and toolmarks 3d measurement systems and measurement quality control, Report, Academy Standards Board (2021).
[105] S. Bunch, G. Wevers, Application of likelihood ratios for firearm and toolmark analysis, Sci. Justice 53 (2) (2013) 223–229, https://doi.org/10.1016/j.scijus.2012.12.005. ⟨http://www.sciencedirect.com/science/article/pii/S1355030612001529http://ac.els-cdn.com/S1355030612001529/1-s2.0-S1355030612001529-main.pdf?_tid=ca1d54c6-e0f5-11e3-8911-00000aab0f6c&acdnat=1400683381_cf8331944cba339d9d93a6f71a2c1754⟩ (URL).
[106] W. Kerkhoff, R. Stoel, E.J. Mattijsen, H. R, P. Hertzman, D. Hazard, M. Gallidabino, T. Hicks, C. Champod, Cartridge case and bullet comparison: Examples of evaluative reporting, AFTE J. 49 (2) (2017) 111–121.
[107] A. Biedermann, S. Bozza, F. Taroni, Decision theoretic properties of forensic identification: Underlying logic and argu- mentative implications, Forensic Sci. Int. 177 (2-3) (2008) 120–132, https://doi.org/10.1016/j.forsciint.2007.11.008. ⟨http://www.sciencedirect.com/science/article/B6T6W-4RJJRH8-1/2/f04616f4b895efbf24d74d57d2073cc3⟩.
[108] A. Biedermann, J. Vuille, Understanding the logic of forensic identification decisions (without numbers), Sui Generis (2018) S397–S413, https://doi.org/10.21257/sg.83. ⟨https://sui-generis.ch/83⟩.