Forensic Science International 313 (2020) 110363

Contents lists available at ScienceDirect

Forensic Science International

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

Comparison and interpretation of impressed marks left by a firearm on cartridge cases – Towards an operational implementation of a likelihood ratio based technique

Fabiano Riva[a,b,*], Erwin J.A.T. Mattijssen[c,d], Rob Hermsen[c], Pascal Pieper[c], W. Kerkhoff[c], Christophe Champod[a]

[a] School of Criminal Justice, Faculty of Law, Criminal Justice and Public Administration, University of Lausanne, Switzerland
[b] University Center of Legal Medicine, University of Lausanne, Switzerland
[c] Netherlands Forensic Institute, The Hague, The Netherlands
[d] Radboud University Nijmegen, Behavioural Science Institute, The Netherlands

ARTICLE INFO

Article history:
Received 20 March 2020
Received in revised form 2 June 2020
Accepted 9 June 2020
Available online 10 June 2020

Keywords:
Forensic science
Firearms identification
3D topographies
Datasets
Evaluation

ABSTRACT

Firearm examination is subject to increased scrutiny regarding its foundational validity and inherent subjective nature. The increased use of automatic comparison systems may help to reduce subjectivity. In this paper, we present the performance and limits of an automatic comparison system that assigns a weight to the forensic findings for the comparisons between firing pin marks, breechface marks, or a combination of the two. This weight is expressed by a likelihood ratio (LR) based on 3D topographical measurements coupled with a bi-dimensional statistical model.

As the performance of such systems may depend on the reference databases used to inform the model, we investigated the impact of the brand of ammunition and the number of samples.

We show that reference databases used to calculate LRs should ideally consist of the same type of ammunition as is seen in the case under investigation and that 7 specimens fired by the same firearm are enough to obtain rates of misleading evidence of a similar magnitude compared to those obtained when far more specimens (60) are used.

Additionally, the automatic system was used to assess the outcomes of 7 cases with known same-source or different-source ground truths. These cases were also examined by 8 qualified firearm examiners. In all cases, the experts' appraisals were in line with the ground truth. The automatic system showed some limitations in cases were the data were not sufficient to calculate a robust LR, but also that it can assist and enhance the examiners in their decision process.

1. Research purpose

The traditional forensic feature comparison disciplines, such as firearm comparison, have been practiced for decades, but their foundational validity have been scrutinised in the past years [1–4]. A part of this scrutiny focuses on the scientific basis of these disciplines, but also on the potential risks inherent to the use of a human expert (a firearm examiner) as the instrument of observation and interpretation [3–7]. While some risks of using human experts might be minimised by implementing appropriate procedures to deal with bias [6,8–13], others are harder to overcome. The precise content of their internal database, based on training and experience, and how those databases are employed in practice are hard to study and will probably vary between experts. In that regard, the experts can be regarded as black boxes. Furthermore, it has been shown that training and experience alone are insufficient to result in judgments of the weight of forensic findings that are well-calibrated [14]. To mitigate the risks posed by the subjectivity of human experts, the added value of automatic comparison systems is studied [15–26]. The PCAST report even goes as far as to suggest that (p.47): "Subjective methods can evolve into or be replaced by objective methods".

In a previous paper [23] we presented the basis of such an automatic comparison system that allows assigning a weight to the forensic findings associated with the comparison between firing pin marks (FPM) or breechface marks (BFM). The system allows the assignment of likelihood ratios (LRs) based on 3D topographical measurements coupled with a bi-dimensional statistical model. The main strength of this approach is the very limited intervention

* Corresponding author at: Ecole des Sciences Criminelles, Université de Lausanne, Batochime, CH 1015, Lausanne-Dorigny, Switzerland.
E-mail address: fabiano.riva@unil.ch (F. Riva).

http://dx.doi.org/10.1016/j.forsciint.2020.110363

of the operator that ensures a reproducible procedure to evaluate the comparison results between two impressed marks. The likelihood ratios are obtained by comparing the results (expressed in the form of a series of scores) of a specific comparison with the scores obtained from cartridge cases fired by the same firearm and cartridge cases fired by different firearms respectively. The scores obtained by these comparisons allow the construction of the so-called *within* and *between* distributions (Fig. 1). The *within* distribution is highly dependent on the firearm under consideration and the reproducibility of its marks. The *between* distribution depends on the specificity of the features considered among the firearms constituting the relevant population. These two distributions represent the supporting data of the interpretation model.

This system was designed to support an expert, by means of the calculated LR, during the evaluation step of his/her work. This does not mean that the expert will blindly use the assigned LR and ignore the outcome of the traditional visual comparison procedure.

During the development of the system, operational constraints had not been assessed and the procedures necessary to implement this system in casework had not been set up. In addition, the results reported in [23] showed that the assignment of the likelihood ratio was largely dependent on the firearm and ammunition brand under consideration and of the build-up and size of the relevant population considered in the case. It means that the *within* and the *between* distribution established using one ammunition brand for a specific firearm type (e.g., firearms of similar models having the same class characteristics) cannot be directly applied by generalisation and used for other (firearm / ammunition) combinations.

The purpose of this study is to investigate how the limits of the previous study can be overcome in practice and how such a system may be deployed in operational casework. To approach real cases, the versatility of the database – a term we use to describe the data used to construct the *within* and the *between* distribution on a case by case basis – should be adapted to cover a range of scenarios that is adequate for a fair amount of casework situations. To increase the scope of usage of this interpretative procedure and to determine the impact of the database features on the system's performance, different lines of inquiry have been pursued including the effect of using different types of ammunition brands on the LR calculation, the effect of the number of specimens (test fires) on the LR calculation, and a procedure to incorporate the LR assigned by the system into operational casework. Finally, the implementation of the optimised system has been tested by means of blind test experiments.

2. Methods and material

2.1. D acquisition and scores computation

The same acquisition procedure as in [23] has been used to acquire the 3D measurements on the cartridge cases. The three-dimensional measurements have been performed by using a laser profiler (μScan) equipped with a confocal detector (CF4) using a spatial resolution of 3 μm.

Further processing steps allow segmenting the features to be further compared. They mainly include the automatic primer cup segmentation, the automatic separation between the firing pin marks and the breechface marks on the primer cup and the application of imaging filters to simultaneously reduce the 3D image noise and the prominent shape of the marks. As a result, more weight will be assigned to fine characteristics. Compared to [23], the comparison algorithm used to align two firing pin marks was modified. Previously an ICP (Iterative Closest Point) algorithm was used to align firing pin marks, whereas a optimisation-based algorithm (Simplex method) was used to align breechface marks. This ICP algorithm proved to be less robust for the alignment of firing pin marks showing large reliefs and rough shapes. To cope with this, the ICP algorithm was replaced by the optimisation-based algorithm. More details on the algorithm can be found in [23]. All algorithmic developments and statistical analyses were carried out in Matlab® 2008.

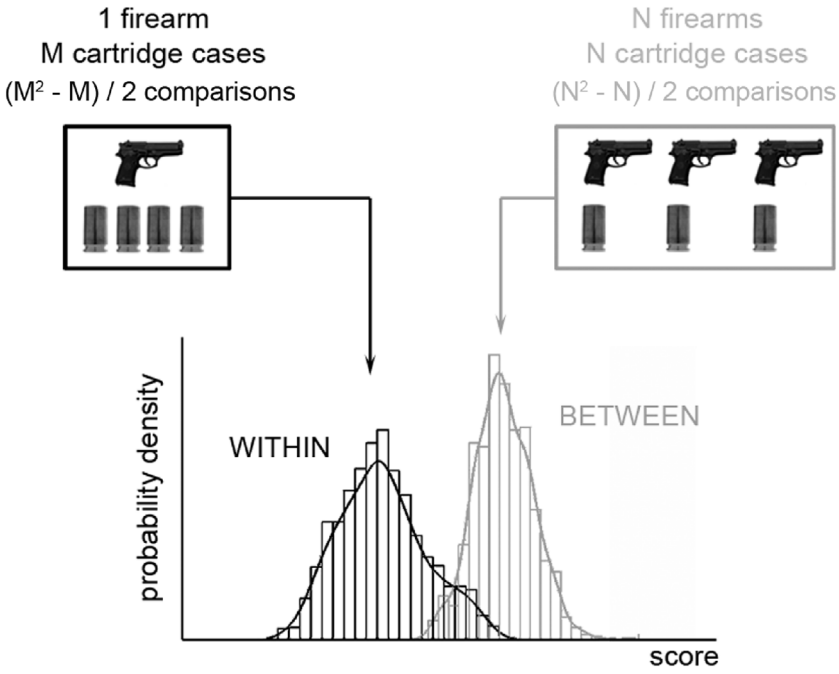

Fig. 1. The procedure used to build the *within* (black) and the *between* (grey) distributions. The *within* distribution is created comparing cartridge cases fired by the same pistol. The *between* distribution is the result of comparisons between samples fired by different firearms. These distributions are composed of scores provided by the automatic comparison system and are illustrated by the mean of histograms. In this example, the x-axis represents a distance-based score. This means that a higher score will represent a lower degree of correspondence between the compared marks.

For each compared impression (the breechface marks and the firing pin marks), three quantitative scores representing the amount of similarities (or differences) between the two marks were computed as in [23]. These similarity scores are based on different morphological features: the first is the correlation index, the second is based on the Euclidean distance and the last is based on the properties of the normal vectors to the surface.

2.2. Bi-dimensional interpretative model and rates of misleading evidence

When two cartridge cases are compared, six scores are obtained: three resulting from the firing pin marks comparison (FPM) and three from the breechface marks comparison (BFM), from which an LR is calculated using a statistical method described in [23]. The method is briefly recalled here with the addition of a few illustrations. To reduce the dimensionality of the problem, Principal Component Analysis (PCA) allows focusing on the two principal components (PC1 and PC2) that offer the highest contribution to the variability. For the PCA to be considered effective, the first two components should represent at least 80% of the total variation of the considered scores. The problem is then reduced to these two values acting as similarity scores (one for each principal component, PC1 and PC2). The PCA procedure can be applied either to the three scores associated with each respective mark (hence a reduction from 3 to 2 variables) or to the six scores considered jointly (hence a reduction from 6 to 2 variables). The result from the *within* and *between* comparison can be represented by two datasets projected on a two-dimensional plane (Fig. 2).

Once the *within* and the *between* data points have been acquired, it is necessary to model the data in this bi-dimensional space to calculate the LR for a new comparison between two cartridge cases. Several methods can be used to model the data, which include supervised and unsupervised machine learning techniques to classify samples between within and between classes. For this project, we chose to estimate the probability density for each class and use non-parametric continuous functions obtained by *Kernel Density Estimation* (KDE) – as shown in Fig. 3.

The LR for a comparison between two cartridge cases is obtained by the ratio of the probability densities of both modelled bi-dimensional distributions (*within* and *between*) at the coordinates X, Y dictated by the transformed scores PC1 and PC2, which now represent the degree of similarity between the compared marks. The LR is the results of two probability densities. The numerator of the LR represents the probability density to observe PC1 and PC2 under the prosecutor's proposition (H_p), that the cartridge cases have been fired by the same firearm (hence the probability density for the findings under the *within* distribution). The denominator of the LR represents the probability density of PC1 and PC2 under the defence's proposition (H_d), that the cartridge cases have been fired by different firearms (hence the probability density for the findings under the *between* distribution). Formally, the likelihood ratio is:

$$LR = \frac{p(PC1, PC2|H_p)}{p(PC1, PC2|H_d)}$$

Where *PC*1 and *PC*2 represent the transformed "similarity scores" obtained by PCA for the comparison between two marks (or the fusion of the marks).

According to this definition, an LR > 1 indicates forensic findings that support the same-source proposition (H_p). A likelihood ratio < 1 conveys forensic findings that support the different-source proposition (H_d). An LR of 1 means that the findings do not give support for either of the considered propositions. Fig. 4 illustrates a case leading to an LR of $1.11e^{11}$, meaning that the forensic findings are $1.11e^{11}$ times more probable under H_p than under H_d.

The forensic performance of the system is assessed using two rates of misleading evidence obtained by the systematic computation of LRs using simulated cases for which a known source is established. The first, the rate of misleading evidence in favour of the defence (RMED) is defined as the percentage of LRs < 1 when the prosecutor's proposition H_p is true, i.e., false negatives. The second, the rate of misleading evidence in favour of the prosecution (RMEP) is defined as the percentage of LRs > 1 when the defence's proposition H_d is true, i.e., false positives.

2.3. Material – an enlarged database of specimens obtained with additional brands of ammunitions

In [23], two distinct *within* distributions were established using two groups of 60 Geco Sintox cartridge cases (124 gr bullet, nickel primer) fired by two SIG Sauer pistols (P226 and P228) calibre 9 mm Luger (in total 1770 possible comparisons per firearm under H_p, each sample being compared with all the others). The *between* distribution was constructed using 79 Geco Sintox cartridge cases fired by 79 different calibre 9 mm Luger SIG Sauer pistols of similar models (3081 possible comparisons under H_d).

For this study, the same firearms have been used, but additional test fires using three different ammunition brands were made: Geco (124 gr. bullet, brass primer), Fiocchi (123 gr. bullet, nickel primer) and Winchester (124 gr. bullet, brass primer). Note that cartridges of the same brand can be from different production batches. This represents in total $2 \times 4 \times 60 = 480$ cartridge cases to investigate the *within* distributions (2 firearms / 4 brands of ammunition / 60 cartridge cases per firearm and ammunition brand) and $79 \times 4 = 316$ cartridge cases for the *between* distributions (79 firearms / 4 brands of ammunition) (Table 1).

2.4. Lines of inquiry towards implementation

2.4.1. Production of simulated cases and underlying within and between distributions

Using the specimens from Table 1, considering the relevant pairwise comparisons, we can construct the following simulated cases under respectively H_p and H_d:

Under H_p: The simulated cases are prepared conditioned on a given firearm and on a given brand of ammunition. Hence for each firearm and brand of ammunition, by drawing from the 60

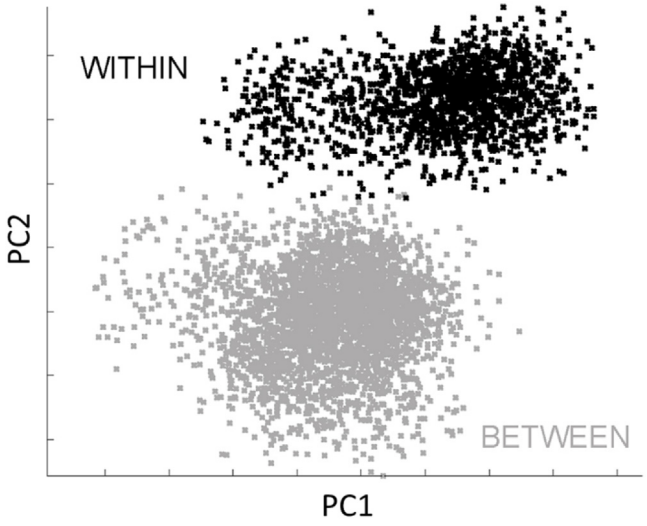

Fig. 2. The *within* (black dots) and the *between* (grey dots) distributions in two dimensions. When the automatic comparison system provides two scores for each comparison, the results can be illustrated using a bi-dimensional plot.

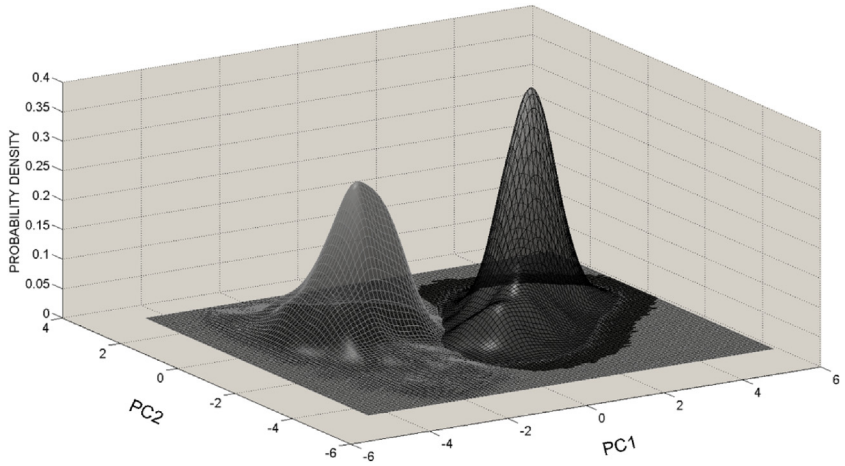

Fig. 3. The bi-dimensional data of Fig. 2 have been modelled using a bi-dimensional non-parametric function: the Kernel Density Estimation.

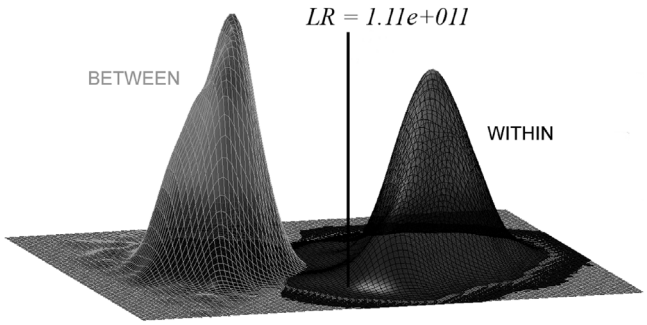

Fig. 4. In this example, one questioned cartridge case has been compared to a test fire. The resulting "similarity scores" for the single comparison are then reduced to PC1 and PC2 by applying the same reduction procedure (PCA spatial transformation) used previously to build the *within* and *between* distributions. The coordinates of the vertical black line thus represent the coordinates given by PC1 and PC2 of a new comparison between one questioned cartridge case and one test fire. The ratio of the probability densities extrapolated for both functions (*within* and *between*) represents the so-called likelihood ratio (LR).

cartridge cases, a total of 1770 comparisons arising from the same source are obtained. The *within* distributions will be based on these 1770 data points. In total, for 2 firearms and 4 brands of ammunition, it amounts to 8 distinct *within* distributions, each with a total of 1770 data points.

Under H_d: The simulated cases are obtained using all the pairwise comparisons among the cartridge cases fired using the 79 different pistols, conditioned on the brand of ammunition. For each brand of ammunition, a total of 3081 comparisons are obtained. All brands of ammunition together represent 4×3081 comparisons. The *between* distributions are either conditioned on the brand of ammunition (4 distinct distributions, with a total of 3081 comparisons per distribution), or are pooled together leading to what we call the *general between* distribution (with a total of 12,324 comparisons, see Fig. 10 for an example).

2.4.2. The influence of the brand of ammunition on the LRs

To study the impact of the brand of ammunition, the LRs obtained for the simulated cases of each brand of ammunition have been compared to the LRs obtained when the denominator of the LR is computed against the *general between* distribution containing all brands of ammunition. The comparison is performed by correlating the LRs obtained from both configurations. The more we deviate from a linear relationship between the configurations, the more the "type of ammunition" is shown to affect the calculated LRs. Strong deviations will limit the possibility of estimating the *between* distribution by a *general between* distribution. On the other side, if the deviations are limited we could entertain the prospect of modelling the specific *between* distribution for a given ammunition using the *general between* distribution. The key question here is whether or not LRs can be robustly assigned without conditioning on the brand of ammunition.

2.4.3. The influence of the number of specimens on the LRs

All the *within* distributions described above are based on a large number of specimens (60). Keeping in mind that such distributions have to be obtained on a case-by-case basis, where it is not realistic to test fire 60 cartridge cases in each case. The results obtained with 60 cartridge cases can serve as a baseline, but there is a need to explore more operational settings. The effect of the number of specimens is studied here by resampling from the total number of specimens to construct the *within* distributions ranging from 5 specimens (corresponding to 10 data points) up to 60 (1770 data points).

The resampling scheme is as follows: for each number of specimens (5 to 60), a sample is drawn (with replacement) from the available 1770 data points. This sample (of 10 to 1770 data points) is used to compute the *within* distribution (bi-dimensional normal distribution density estimation). The LRs associated with all the simulated cases (considering either H_p or H_d) are computed using this *within* distribution and the *between* distribution conditioned on the appropriate brand of ammunition. From all

Table 1
Details of the samples used in this study. Four different ammunition brands have been used: Geco, Geco Sintox, Fiocchi and Winchester.

Distributions	Firearms	Ammunition brand	Quantity of cartridge cases
8 × Within distributions (2 firearms – W1 and W2*)	SIG Sauer P228 / called W1	Geco Sintox	60 × 4 from the same firearm
	SIG Sauer P226 / called W2	Geco	60 × 4 from the same firearm
4 × Between distributions	SIG Sauer** / 79 firearms	Fiocchi	79 × 4 from 79 different firearms
		Winchester	

* The abbreviation W is used to denote the within distribution W1 and W2.
** P226 (42 firearms), P228 (14 firearms) and Sig Pro (23 firearms).

these LRs, RMEP and RMED are computed. The process is repeated 500 times, recording for each iteration the rates of misleading evidence. Then, we explore the behaviour of the means of the rates of misleading evidence against the number of specimens in order to make an informed decision as to the appropriate minimum number of specimens.

2.4.4. Review procedure of the computed LRs

For the system to be applied in casework there is a need to define a procedure whereby the computed LRs are critically and transparently reviewed by an expert in order to detect cases where the system did not perform properly. In such cases, the computed LRs should be investigated, discussed and can be dismissed if necessary according to pre-defined criteria; otherwise the LRs obtained for the marks will be considered in the assessment. Note that it is expected that all computed LRs will be included in the casefile, including the discussion and the outcomes thereof regarding their ability to be used in the case or not. The following criteria to be fulfilled have been defined:

1. This criterion is related to a judgment with regards to the source and quality of the marks and their measurements. If the primer cup surface shows striated features that don't originate from the firearm (for example because they are due to the manufacturing process of the cartridge case), then it is not relevant to compute a LR based on these features. The system is not able to make that distinction. Hence there is a first decision by the expert even before considering using the system. That judgment should be carefully monitored with appropriate quality assurance procedures.
2. Once an LR is obtained, the numerator and the denominator of the likelihood ratio have to be scrutinised. If they are both less than $10e^{-4}$ (a threshold which we chose based on the results in [22]), it means that the probability densities are very low under both propositions. The obtained score lies on the tail of both distributions (*within* and *between*) as shown for example in Fig. 5. In such cases, it means that the LR is assessed regarding a pair of propositions that neither explain the results well. As a matter of policy, we propose to not consider such a likelihood ratio. In addition, when such scores occur on the individual marks (respectively from the breechface or from the firing pin), they should trigger an alarm when considering the LR obtained following the fusion of both marks (PCA).
3. The LR of interest is primarily obtained following the fusion between information derived from the breechface mark and from the firing pin mark. If one of these marks cannot be exploited by the system (either due to criterion 1 or 2 above or a failure to align by the algorithm resulting in a "Not a Number" abbreviated NaN), the LR retained will be based on one mark only and will only be used for that specific mark (during the experiment it turned out that most of time the firing pin mark resulted in the highest number of not exploitable cases).

2.4.5. Blind test experiments

The previous sections allowed designing a procedure that could be used in casework. To assess its operational feasibility, blind test experiments have been conducted as described below.

Seven tests have been prepared by one specialist of a large European forensic laboratory using samples fired by firearms coming from their reference collection. These firearms were not used for the development of the system and its underpinning distributions. Each set was composed of one questioned cartridge case and seven specimens. For each test set, only one ammunition brand has been used. This was done to mimic operational practice in which, when possible, the same ammunition brand is used for the test specimens as was used in the shooting incident (the questioned cartridge case(s)). The firearms and ammunition brands used to prepare the tests are summarised in Table 2.

These cases have been submitted to eight qualified firearm experts from the same European institute without any knowledge of the ground truth. They were asked to conduct the examination using their standard comparison procedures including case notes detailing the features they relied upon to reach a conclusion. For each test set, the experts were asked to perform the comparisons and formulate three conclusions based on the comparison of the firing pin mark, the breechface mark and of both marks combined.

The conclusions had to be expressed using the verbal scale which the firearms section of the considered forensic laboratory used at the time. This verbal scale was composed of 5 steps: two levels in support of the prosecutor's proposition (same source), two levels in support of the defence's proposition (different sources) and one without support for either the prosecutor's or defence's proposition. To allow a comparison with the calculated

Table 2
Information about the firearms and the ammunition brands used to establish the blind tests.

Test N°	Firearm - Questioned	Firearm – specimens	Ammunition
1	SIG P210	The same as questioned	Geco Sintox
2	SIG P228	Another SIG P228	Geco Sintox
3	SIG P226	The same as questioned	Geco
4	SIG P220	The same as questioned	Geco Sintox
5	SIG P210	Another SIG	Geco Sintox
6	SIG P228	The same as questioned	Geco
7	SIG model unknown	Another SIG	Geco

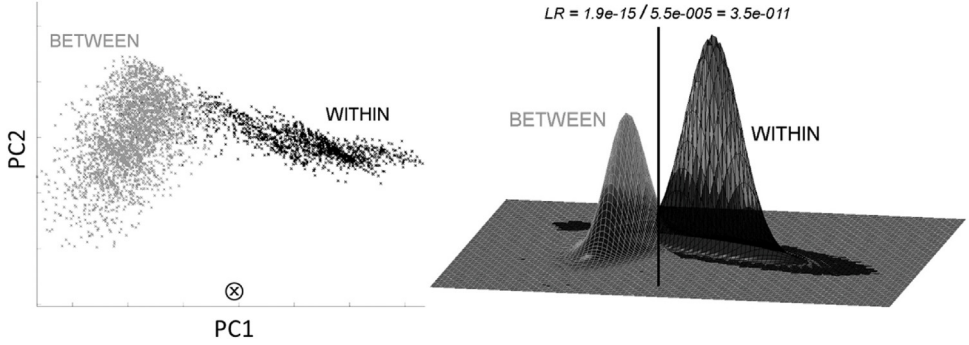

Fig. 5. The left graph shows the *within* and *between* distributions under the shape of a bi-dimensional plot. The black cross represents the scores of a new comparison between two cartridge cases. The right figure shows the same situation of the left graph but after modelling the data using KDE. The black cross is now represented by the vertical line. The coordinates of the line are very far from the two distributions. This explains the extremely low numerator and denominator of the LR (numerator = 1.9e–15 and denominator = 5.5e–005).

LRs by the automaticd system, as will be discussed later, a numerical equivalent between -2 and +2 has been assigned to each level of the verbal scale (Table 3).

In parallel and independently, an operator who had no prior knowledge of the true source of the cartridge cases has submitted them to the automatic system. The seven specimens for each test set were scanned and compared to obtain the *within* distribution. Then, the questioned cartridge case was compared to each of the test specimens, leading potentially to 21 LRs for each questioned cartridge case when considering both marks separately (7 LRs for the breechface mark and 7 LRs for the firing pin) and then jointly (7 combined LRs).

The mean, maximum and minimum values of the results obtained by these two distinct procedures (conclusions reached by the experts / results obtained by the system) for each blind test set were compared.

3. Results

3.1. The influence of the brand of ammunition on the LRs

The results show that the brand of ammunition can have a significant influence on the *within* and the *between* distributions and thus on the LRs calculated and the corresponding rates of misleading evidence. This is observed regardless of the type of impressions considered (firing pin marks "FPM", breechface marks "BFM" or both together "FPM + BFM"). Examples of the differences in distributions between two brands of ammunition are shown in Figs. 6–8.

Moreover, some brands of ammunition lead to a better separation between the two distributions (*within* and *between*) for the breechface mark and a lower separation for the firing pin mark or vice-versa. This is for example illustrated in Fig. 9, where the firing pin mark distributions for the Geco ammunition are compared to those of the Fiocchi ammunition left by the same firearm.

These results show that to derive robust LRs, it is essential to properly condition on the brand of ammunition. Indeed, as we will show later, the rates of misleading evidence will change significantly if data of one particular brand of ammunition is used to evaluate a case involving another brand of ammunition. Hence, if a *between* distribution corresponding to the brand of ammunition involved in the case is available, it should be used. To assess whether it would be possible to use a *general between* distribution (obtained by merging Geco, Geco Sintox, Fiocchi and Winchester data) to solve the problem when the above-described ideal dataset is not available we will explore Fig. 10. This figure illustrates the distributions obtained under both scenarios: (A) a *between* distribution built using only one brand of ammunition and (B) the *general between* distribution.

The LRs resulting from these two distinct procedures have been compared in Fig. 11 for the Geco Sintox. A linearity between the LRs ($R^2 = 0.97$) is observed.

However, that behaviour is not constant across all brands of ammunition considered. For example, in the case of the Winchester ammunition, the linearity ($R^2 = 0.73$) is lost (Fig. 12).

In the absence of data matching the brand of ammunition involved in the case, defaulting to the use of a *general between* distribution is thus not advised, because the amount of deviation from the linearity is hard to predict.

For each brand of ammunition, the rates of misleading evidence have been calculated for both firearms (W1 and W2). The evolution of such error rates (RMED and RMEP) in function of the increase of the LRs are presented in the Table 4.

3.2. The effect of the number of specimens on the LRs

As shown in the Table 5, when the number of specimens used to build the *within* distribution increases from 7 to 60, the average error rates (on 500 repetitions) do not change drastically for all brands of ammunition considered (same firearm).

These results allow us to conclude that a small number of specimens (7) can be used to compute the *within* distribution without significantly affecting the rates of misleading evidence. Acquiring 7 specimens is viewed as an acceptable burden from an operational perspective.

3.3. Results from the blind test experiments

3.3.1. Using the automatic system leading to an LR

The LRs obtained using the automatic system while applying the procedure proposed above are summarised in Table 6. For each test, the minimum, mean and maximum of the seven LRs calculated for each mark comparison (firing pin mark, breechface mark and the fusion of both marks) are shown in relation to the ground truth information.

Based on the results of the fusion (except for Case F where only the LR associated with the firing pin mark is available), the retained LRs provided support according to the state of the ground truth in 5 cases. Case B showed an inconsistency between the results obtained when the marks are considered separately or jointly. Indeed, the breechface comparisons support the hypothesis of different sources (which is in line with ground truth), whereas the fusion of the marks supports the hypothesis of a common source. The probability densities calculated for the firing pin mark comparison were very low (hence declared as NA* in the Table 6 as per our procedure). When the first two scores associated with the firing pin are shown in relation to the values obtained for the *within* and *between* transactions (Fig. 13), it can be seen that the first score does not allow a clear discrimination between the questioned samples and the *within* distribution, and so it explains why the fused results are pointing in a misleading direction.

For the fusion of both marks, this first score associated with the firing pin comparisons has the highest weight in the PCA (the 6 scores being considered together without any distinction). It means that the seven LRs for the fusion are dominated by this first score. In case B, the breechface scores have a limited contribution in the fused results and cannot compensate for the misleading firing pin information.

The first score (correlation index) is more influenced by the global shape of the marks which in this case are in correspondence (although the exhibits have been fired by different firearms) as

Table 3
Correspondence between the verbal scale used by the experts and the numerical equivalence adopted to quantify the results.

Verbal scale	Numerical equivalent
The findings are far more probable when H_p is true then when H_d is true	+2
The findings are more probable when H_p is true then when H_d is true	+1
The findings are approximately equally probable under both hypotheses	0
The findings are more probable when H_d is true then when H_p is true	−1
The findings are far more probable when H_d is true then when H_p is true	−2

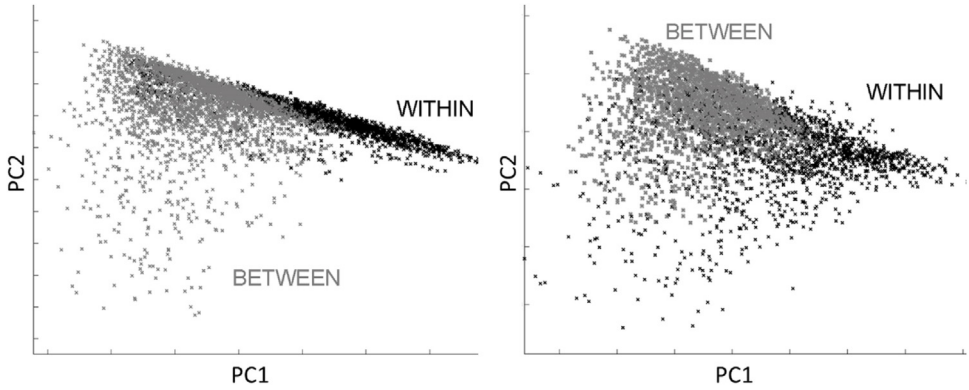

Fig. 6. *Within* and *between* distributions for the Geco Sintox (left) and Winchester ammunition (right). The data represent the scores for the firing pin marks (FPM).

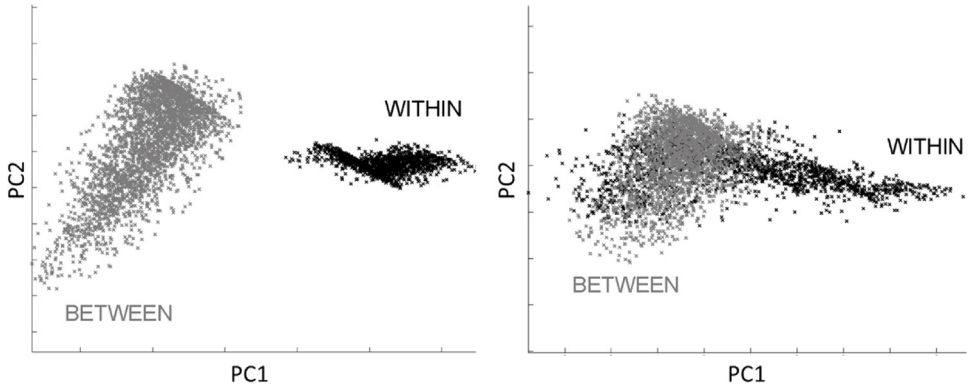

Fig. 7. *Within* and *between* distributions for the Geco Sintox (left) and Winchester ammunition (right). The data represent the scores for the breechface marks (BFM).

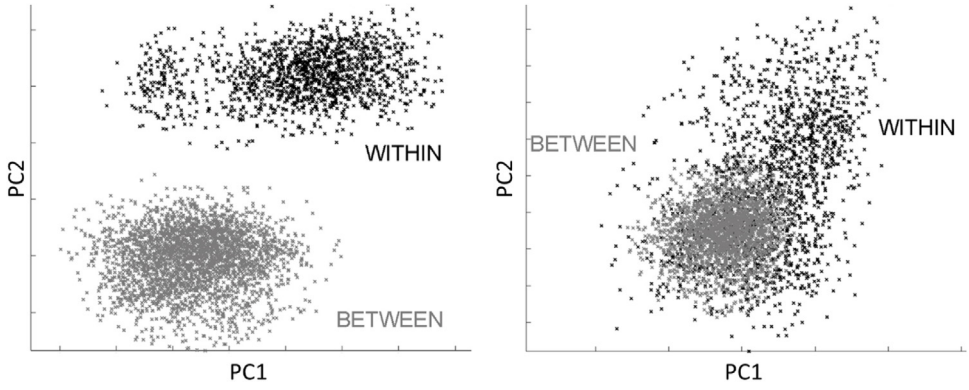

Fig. 8. *Within* and *between* distributions for the Geco Sintox (left) and Winchester ammunition (right). The data represent the scores for the fusion of the breechface and the firing pin marks (FPM + BFM).

shown in Fig. 14. The second score, based on the "Euclidean" distance, is more sensitive to the difference of finer characteristics, which in this case would allow for more discrimination. That potential discrimination however is lost when both scores are used jointly such as here.

The above detailed analysis requires knowledge of the ground truth. Without that knowledge (as in the blind test environment), conflicting results (here fused LRs providing support for another proposition than the breechface LRs), as seen here, should lead to a cautious approach when considering the computed LRs.

In Case G the densities obtained for the firing pin mark are low for both the numerator and the denominator, hence they have been dismissed following our review procedure. The breechface mark is almost non-informative and the fusion of the marks gave LRs below 1. In this case the firing pin marks of the questioned and known cartridges are so different (Fig. 15) in terms of shape and size that the resulting scores were not represented in the available *between* distribution. Such differences are rare among the comparisons between the different firearms available in the database. This means that the *between* distribution does not fully represent the population of interest. An increase of the size of the database should allow to solve this issue.

3.3.2. Results from the experts and comparison with the system

The conclusions provided by the eight experts have been converted to a numerical form according to Table 3. The results are

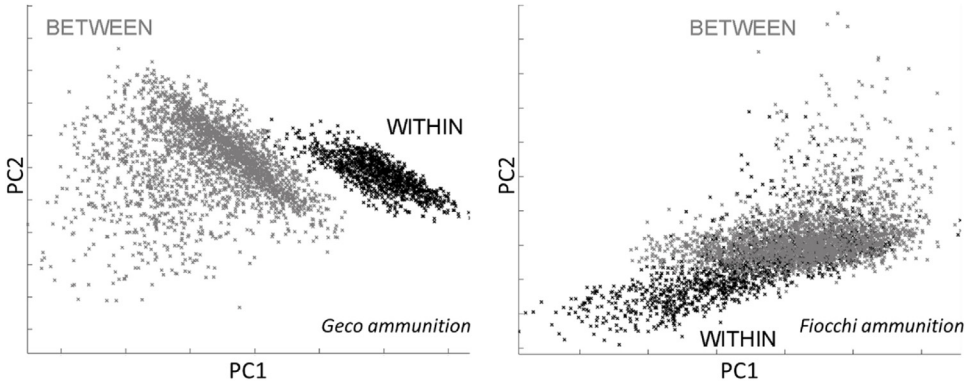

Fig. 9. *Within* and *between* distributions for the Geco (left) and Fiocchi ammunition (right). The data represent the scores for the firing pin marks left by the same firearm.

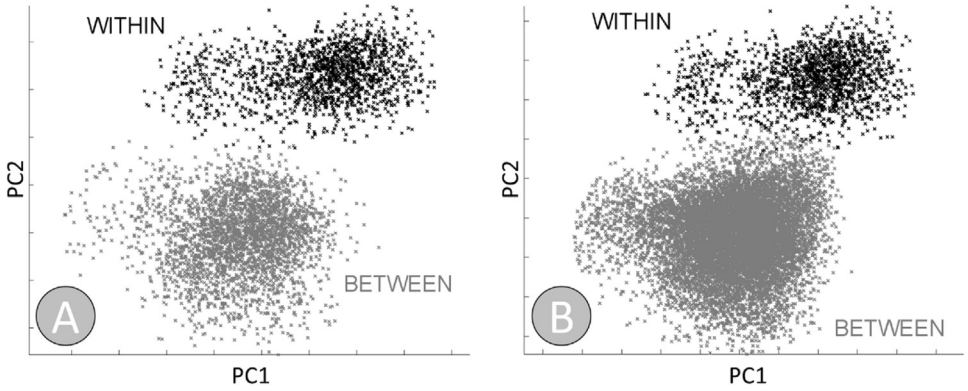

Fig. 10. *Within* and *between* distributions for the Geco Sintox ammunition (A) and the same *within* distribution with the *general between* distribution (B).

presented in Table 7 using the minimum, the mean and the maximum of the converted values.

In all cases, the experts concluded according to the ground truth and they were able to assign weights to their findings in cases where the automatic system didn't provide any numerical output (cases with NA* and NaN).

Two cases involving firing pin marks allow to assess the difference between the features considered by the automatic system and the experts, respectively. In case F, the system provided a $\log_{10}$(LR) of 8, whereas, 7 out of the 8 experts concluded this comparison as "approximately equally probable" (0 on the transformed scale, i.e., an LR of 1). The 3D scans of the firing pin marks show that the firearm reproduces the general shape and also some of the fine characteristics of the firing pin. The alignment based on these fine characteristics has been performed with success for all the firing pin marks hence resulting in a very concentrated *within* distribution without overlap with the *between* distribution. Comparatively, the majority of firearm experts, during subsequent interviews, indicated that the shape and size of the firing pin marks (questioned and test fires) were compatible but that the absence of fine details visible under the comparison microscope did not allow them to move more strongly in one direction or the other. The difference between both conclusions is mainly due to the considered characteristics of the firing pin marks: finer details with the 3D acquisition which were difficult to visualise with the optical microscope. Hence, in this case, the "information" available to the system and the expert respectively was different and cannot be compared in a straightforward fashion.

In case D, the firearm experts indicated that they used some fine striated characteristics situated on the slope of the firing pin mark (Fig. 16 – left) whereas the system did not take into account that information due to the applied segmentation and only used features further down in the firing pin mark (Fig. 16 – right).

The following graph illustrates the relationship between the mean of the expert conclusions and the collected data for the joint consideration of both marks (Fig. 17). Overall, we cannot fit a simple linear relationship between the answers on both scales, but it shows some logical consistency between LRs and evidential strength levels reached by the experts.

The contrast between case A and D is worth noting: all firearm experts reached the same conclusion: level +2, meaning that the "findings are far more probable when H_p is true then when H_d is true". The system however provided an average LR of different magnitude for each case ($\log_{10}$ = 2.8 for the case A and 23.1 for case D). This might be explained by the difference in applied conclusion scales. The experts use a discrete scale, where they will need to pass a certain threshold to reach the highest conclusion. All comparisons resulting in a conclusion above that threshold will still receive the same conclusion, while the automatic system provides conclusions on a continuous scale being able to provide more nuances. It is important to also note that our choice of a Kernel density estimation (KDE) has some limitations too and can affect the robustness of the LRs. Indeed, with KDE, there is a risk of underestimating the densities in the tail of the distribution, which will lead to overestimations of LRs. To minimise such effects, the robustness of the calculated LRs could be further investigated and can potentially be improved by choosing other types of distributional assumptions, e.g. adopting log-normal or Generalized Extreme Value (GEV) distributions or be limiting the minimum and maximum LRs based on the size of the *within* and *between* datasets [14,27].

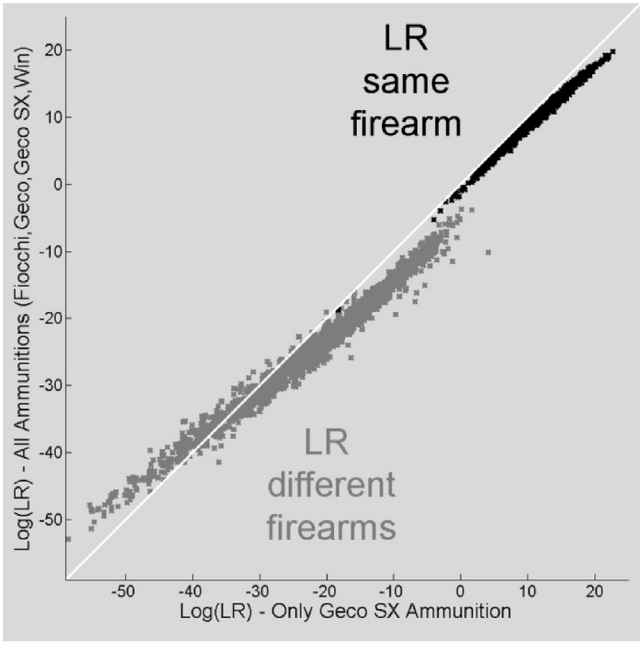

Fig. 11. The x-axis (in logarithmical scale) of this plot represents the LRs calculated using the *between* distribution of the Winchester ammunition. The same comparisons have been used to calculate the LRs using a *general between distribution* composed by data coming from four different ammunition brands (Fiocchi, Geco, Geco Sintox and Winchester); these LRs are represented on the y-axis. The grey and black dots are respectively the LRs which come from comparisons between cartridge cases fired by different firearms and by the same firearm. The white line represents the perfect linearity between the groups of LRs calculated in two different conditions. The linear regression fitted on this data is characterised by $R^2 = 0.97$.

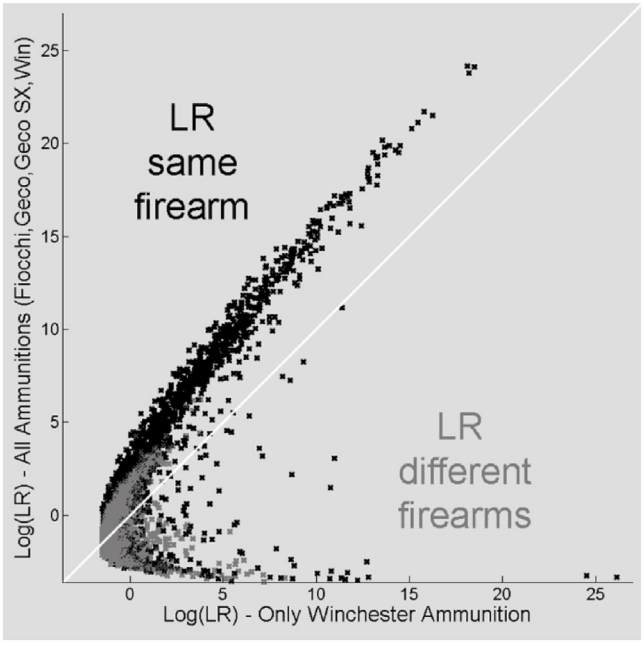

Fig. 12. The x-axis (in logarithmical scale) of this plot represents the LRs calculated using the *between* distribution of the Winchester ammunition. The same comparisons have been used to calculate the LRs using a *general between distribution* composed by data coming from four different ammunition brands (Fiocchi, Geco, Geco Sintox and Winchester); these LRs are represented on the y-axis. The grey and black dots are respectively the LRs which come from comparisons between cartridge cases fired by different firearms and by the same firearm. The white line represents the perfect linearity between the groups of LRs calculated in two different conditions. The linear regression fitted on this data is characterized by $R^2 = 0.73$.

4. Discussion

The system deployed in this study has been developed to support the firearm expert during the evaluative phase. The main aim of this study was to evaluate the constraints encountered during the application of this methodology from an operational perspective. They are summarised and discussed below.

4.1. Nature of the data to support the within *and* between *distributions in casework*

The results obtained show that the *within* distribution can be reasonably approached with a limited number of test fires without seriously affecting the rates of misleading evidence of the system. Obtaining seven test fires is viewed as operationally feasible. However, a critical constraint is the brand of ammunition that needs to be shared between the questioned cartridge, the test fires and the data used to establish the *between* distribution. We failed to identify any mechanism to bypass this constraint. Failure to comply with this requirement may lead to an under- or overestimation of the LRs. From an operational point of view this limits the applicability of the system, because only cases sharing the same brand of ammunition can be handled.

4.2. Review procedure of the computed LRs

The results obtained from the blind tests show that only one case leads to misleading information by the automatic system (Case B). Note however that this case was flagged as problematic for the system for its firing pin mark. The review procedure put in place seems to offer ways to identify problematic cases and alert the user when more care must be applied with the output.

4.3. Comparison between computed LRs and assessments by firearm experts

There are several aspects on which the approach used by the system showed differences with the traditional methodology.

4.3.1. The use of the within source variation

The use of information associated with the *within* source variability is not the same for the system and the experts. Data associated with the *within* source variation is used by the experts to explain potential differences, but once differences have been resolved, the *within* source variation does not have a huge explicit influence on the assignment of the weight of evidence. The system, however, will account for such variation in the numerator of the likelihood ratio. Our experience with experts has shown that this "explaining away" mechanism is even more salient when some marks within the complete impression mark are observed to be in agreement. The potential poor level of reproducibility is simply disregarded by virtue of the selectivity of the features in agreement and the numerator is subjectively set to approximately 1 when agreement is observed. In contrast, the system will always account for the numerator in the same way, regardless of the value of the denominator. In terms of transparency in the way the *within* variability is accounted for, the automatic system has a competitive edge compared to the expert.

4.3.2. Data associated with between sources variation

The concept of the *between* source variation may vary between the automatic system and the expert. On one hand, the expert takes advantage of a virtual representation of the *between* source

Table 4
The values in the table represent the percent of misleading evidence (RMED and RMEP) when LR > abs($\log_{10}$(X)) when X is a threshold value increasing from 0 to 10. The values have been calculated respectively for each firearm (W1 and W2) and ammunition combination (Fiocchi, Geco, Geco Sintox and Winchester). The NaN (Not a Number) value means that the error of misleading evidence is "zero" also when the threshold value X = 0.

			% of error rate when LR > abs($\log_{10}$(X))										
Error type	Firearm	Ammunition	X = 0	X = 1	X = 2	X = 3	X = 4	X = 5	X = 6	X = 7	X = 8	X = 9	X = 10
RMED	W1	Fiocchi	100.0	71.0	16.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Geco	100.0	30.2	7.2	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Geco Sintox	100.0	40.0	20.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Winchester	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	W2	Fiocchi	100.0	58.8	13.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Geco	100.0	72.4	48.3	37.9	34.5	24.1	17.2	6.9	6.9	6.9	6.9
		Geco Sintox	100.0	66.7	36.8	28.1	14.0	7.0	5.3	5.3	5.3	3.5	1.8
		Winchester	100.0	53.3	40.0	26.7	20.0	13.3	8.3	6.7	4.2	2.5	2.5
RMEP	W1	Fiocchi	100.0	30.9	9.0	2.7	0.5	0.0	0.0	0.0	0.0	0.0	0.0
		Geco	100.0	9.4	1.7	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Geco Sintox	100.0	15.6	2.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Winchester	100.0	15.0	6.9	2.9	1.5	0.2	0.0	0.0	0.0	0.0	0.0
	W2	Fiocchi	100.0	36.1	11.8	3.5	0.7	0.0	0.0	0.0	0.0	0.0	0.0
		Geco	100.0	50.0	25.0	25.0	25.0	0.0	0.0	0.0	0.0	0.0	0.0
		Geco Sintox	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
		Winchester	100.0	9.1	1.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 5
Mean values of RMED and RMEP for four ammunitions types (Fiocchi, Winchester, Geco and Geco Sintox) calculated by bootstrapping considering 7 and 60 specimens respectively.

Ammunition	$RMED_7$	$RMED_{60}$	$RMEP_7$	$RMEP_{60}$
Fiocchi	12.03%	12.88%,	4.96%	5.72%
Winchester	30.70%	34.72%	7.09%	9.69%
Geco	10.00%	14.86%	3.07%	4.73%
Geco Sintox	0%	0.34%	0.12%	0.09%

variability based on their training and experience. On the other hand, the "experience" of the system is entirely based on the data available to the system. These data can be disclosed and structured according to the brand of ammunition and the specificity of the relevant population. The system can adapt its outcomes as a function of these choices, whereas the expert is asked to carry out such tasks holistically. Again, as far as transparency is concerned, the automatic system has the advantage.

4.3.3. The combination of the distinct marks

The system considers and assesses marks jointly when scores associated with both marks are "fused" together. In terms of weight of association, it is not an additive process. For example, if the firing pin mark shows some dissimilarities and the breechface marks show similarities, the overall LR may be lower than the LR assigned for the breechface mark only. The experts use the mark with the strongest conclusion and consider the other mark as "inconclusive", hence not contributing.

The system operates differently as shown in case A for example: the results obtained for the firing pin mark (LR = 1) have been combined with the breechface mark results (LR of the order of 10^5) to give an LR for the fusion of the marks of the order of 10^3.

The cases illustrated something that was expected at the outset of this study: the observed features and their representations are different between the automatic system and the experts. Case D provided evidence for that (Fig. 16). Hence, any process to reconcile the respective conclusions of the expert and the automatic system will have to account for that potential difference in the considered features [28].

4.3.4. The indicators of reliability

The automatic system can be subject to an assessment of the rates of misleading evidence and the robustness of these rates as a function of the brand of ammunition and the type of firearm. These rates represent quality measures associated with the system that can be disclosed alongside the reported likelihood ratio. It shows all case-specific data that can then be submitted to review and critical assessment. By comparison, the experts provide assessments that can only be tested through controlled experiments such as proficiency testing [29] or part-declared blind tests [30,31]. Such testing regimes are expensive and resource intensive as, ideally, experts should be tested individually and in a double-blind procedure.

4.4. Who shall we trust, the system or the expert?

It is fair to recognize that this question is rather new as in many other areas of forensic science where holistic judgments on features have been used for many years. Bringing to the table a "new expert" offering automatic capabilities and asking how this new expert will cooperate with the traditional expertise is not a

Table 6
The results obtained using the automatic comparison system.

Test N°	Ground truth state (Questioned vs. Test fires)	Log10 LRs for the firing pin comparisons [min, **mean**, max]	Log10 LRs for the breechface comparisons [min, **mean**, max]	Log10 LRs obtained following the fusion of the marks [min, **mean**, max]
A	Same firearm	[−1, **0.1**, 0.5]	[1.9, **4.6**, 11.3]	[1.4, **2.8**, 6.7]
B	Different firearms	NA*	[−0.9, −**0.5**, 0.0]	[1.7, **3.4**, 3.9]
C	Same firearm	[7.9, **11.9**, 14.2]	[0.0, **2.4**, 4.3]	[6.1, **10.6**, 14.8]
D	Same firearm	[4.1, **8.6**, 10.8]	[7.1, **19.8**, 34.2]	[16.6, **23.1**, 33.2]
E	Different firearms	NA*	[−10.8, −**8.1**, −2.7]	[−5.4, −**2.0**, −0.8]
F	Same firearm	[8.0, **9.0**, 10.0]	NaN	NaN
G	Different firearms	NA*	[−1.0, **0.3**, 1.2]	[−4.2, −**2.4**, −1.0]

NA*: in this case the system provided an LR value, but their numerator and denominator are both less than 10e-4, the LR is thus dismissed. NaN: The poor quality of the breechface marks didn't allow the alignment of the marks. No value is thus provided by the system.

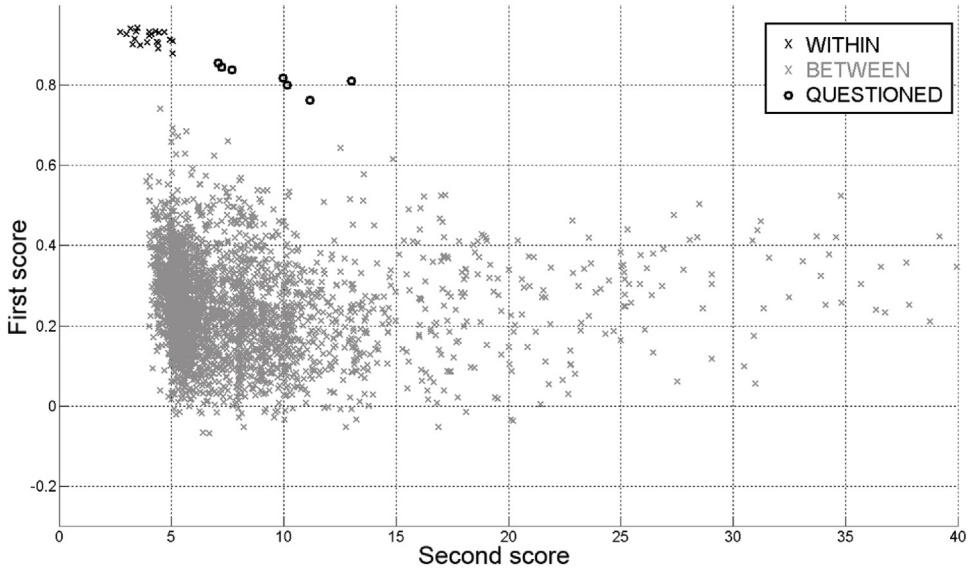

Fig. 13. Bi-dimensional plot of the *within* distribution (black cross), the *between* distribution (grey cross), and the comparisons between the test fires and the questioned cartridge case (7 black circles). The data represent two of the three scores generated for the firing pin comparisons.

Fig. 14. Firing pin marks of the blind test set B. Two test fires (above) and the questioned firing pin mark (below).

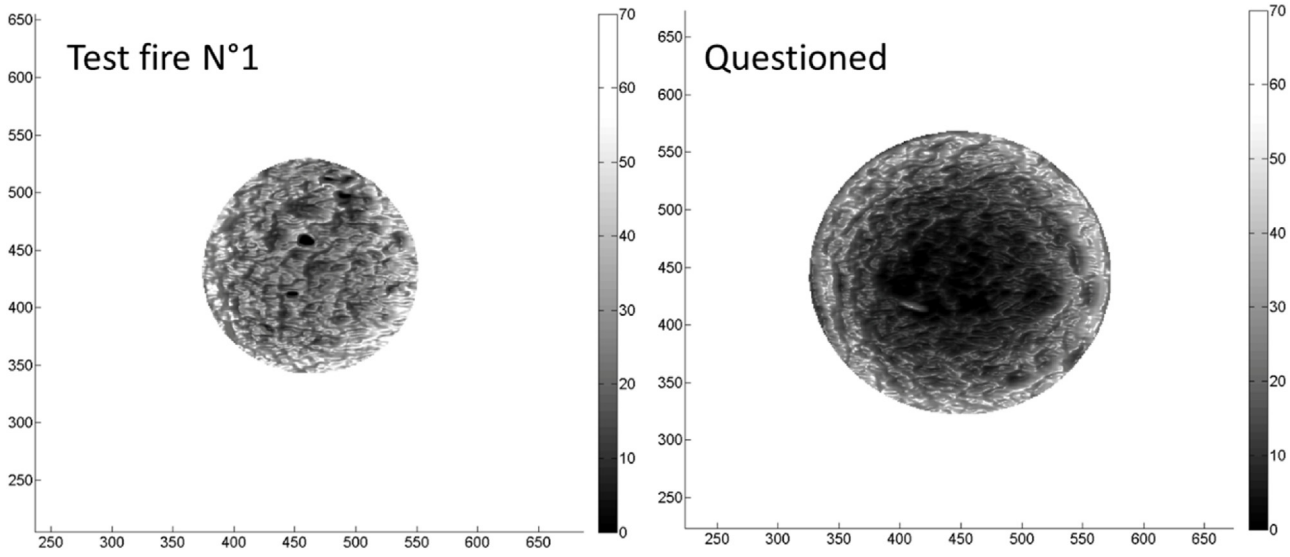

Fig. 15. Firing pin marks of the blind test set G. The test fires (left) and the questioned firing pin mark (right).

Table 7
The results provided by the experts converted to the numerical form in according with Table 3.

Test	State of the ground truth (Questioned vs. Test fires)	Firing pin [min, **mean**, max]	Breechface [min, **mean**, max]	Fusion of the marks [min, **mean**, max]
A	Same firearm	[0, **0.6**, 1]	[2, **2**, 2]	[2, **2**, 2]
B	Different firearms	[0, **0**, 0]	[−1, −**0.3**, 0]	[−1, −**0.3**, 0]
C	Same firearm	[1, **1.5**, 2]	[1, **1.6**, 2]	[2, **2**, 2]
D	Same firearm	[1, **1.9**, 2]	[2, **2**, 2]	[2, **2**, 2]
E	Different firearms	[−2, −**2**, −2]	[−2, −**1.5**, −1]	[−2, −**2**, −2]
F	Same firearm	[0, **0.1**, 1]	[1, **1.4**, 2]	[1, **1.5**, 2]
G	Different firearms	[−2, −**1.3**, −1]	[−2, −**0.4**, 0]	[−2, −**1.1**, −1]

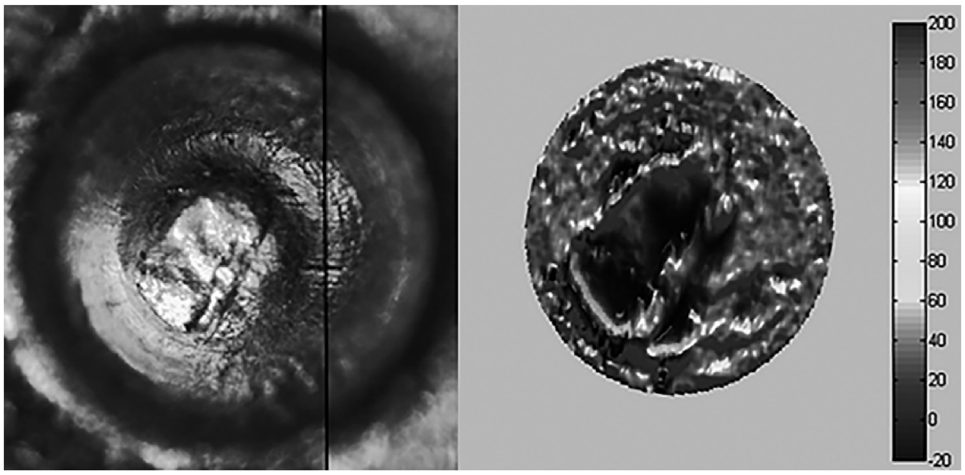

Fig. 16. Firing pin marks of the blind test set D.

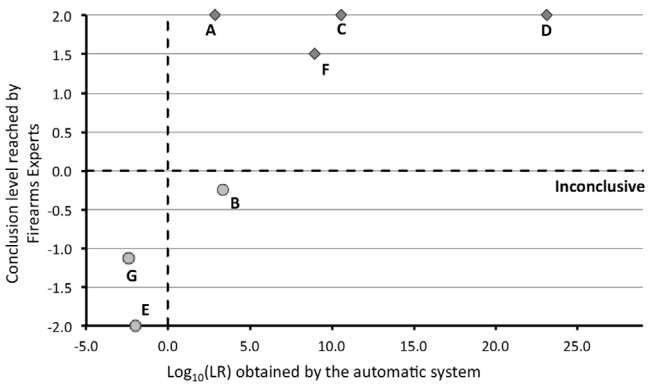

Fig. 17. Graph showing the correlation between the results obtained by the system (x-axis) and those obtained by the firearm experts following the scale presented in Table 3 (y-axis).

trivial question. We don't think that based on this study, all possible avenues have been explored. The results provided so far show that there is a potential to deploy such an automatic system in casework operation alongside experts and that these activities should still be monitored over an extensive trial period. Additional studies (such as e.g., 14), with known ground truths, will allow understanding the operational constraints of both systems, identifying gaps both in data supporting the system or expert training. Overall, we foresee a period of parallel use with full documentation while leaving the system under the direction of the expert. Based on such a trial period, it may be decided that the automatic system is granted the status of a trusted party. In that case, it might gradually become a systematic second expert in casework. Any interpretation conflicts arising, could then be handled through usual conflict resolution procedures as in the case involving two human experts [28].

5. Conclusion

A complete procedure to objectively support the firearm expert in the interpretation phase using LRs calculated by a 3D comparison system coupled to a bi-dimensional interpretative model has been developed and its performance analysed in different situations and configurations.

Although the automatic system worked for practically feasible samples sizes for case specific *within distributions,* the obtained results highlighted the difficulties related to the application of this technique in casework. In particular, it has been demonstrated that there is a dependency between the calculated LRs and the data used to establish the *between* distribution. This aspect raises questions about the generalisation of collected data affecting the versatility of such an approach. Indeed, in real cases, the same ammunition brand should be used to perform comparisons and collect data for the establishment of the *within* and the *between* distributions. To use data collected of ammunition brands that differ from those involved in the actual case can result in an under- or overestimation of the LRs. This doesn't preclude the use of these data but in such cases the firearm expert has to be conscious of the possible consequences on the reliability of the resulting LRs.

Despite this aspect, the adopted procedure provided results for the blind test allowing the presentation of the results under the form of an LR for marks taken into account separately or jointly which can easily be related to a case specific error rate.

CRediT authorship contribution statement

Fabiano Riva: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration. **Erwin J.A.T. Mattijssen:** Validation, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration. **Rob Hermsen:** Validation, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration, Funding acquisition. **Pascal Pieper:** Validation, Investigation, Resources. **W. Kerkhoff:** Validation, Resources, Writing - original draft, Writing - review & editing, Visualization. **Christophe Champod:** Conceptualization, Methodology, Software, Validation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Acknowledgments

We would like to thank Jean-Michel Carrier of the School of Criminal Justice, University of Lausanne, Switzerland, who provided part of the samples used for this research. Many thanks are also extended to all the firearms examiners, who enormously contributed to this research.

References

[1] A. Schwartz, A systemic challenge to the reliability and admissibility of firearms and toolmark identification, Columbia Sci. Technol. Law Rev. 6 (2005) 1–42.
[2] National Research Council (NRC) – Committee, D.L. Cork, J.E. Rolph, E.S. Meieran, C.V. Petrie, Ballistic Imaging – Committee to Assess the Feasibility, Accuracy and Technical Capability of a National Ballistics Database, National Academies Press, Washington, DC, 2008.
[3] National Research Council (NRC) – Committee on identifying the needs of the forensic sciences Community, Strengthening Forensic Science in the United States: a Path Forward, National Academies Press, Washington, DC, 2009.
[4] President's Council of Advisors on Science and Technology, Report to the President Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-comparison Methods, (2016) Retrieved from Washington DC: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.
[5] I.E. Dror, D. Charlton, A.E. Peron, Contextual information renders experts vulnerable to making erroneous identifications, Forensic Sci. Int. 156 (2006) 4–8.
[6] S.M. Kassin, I.E. Dror, J. Kukucka, The forensic confirmation bias: problems, perspectives, and proposed solutions, J. Appl. Res. Mem. Cogn. 2 (2013) 42–52.
[7] R.D. Stoel, C.E.H. Berger, W. Kerkhoff, E.J.A.T. Mattijssen, I.E. Dror, Minimizing contextual bias in forensic casework, in: K.J. Strom, M.J. Hickman (Eds.), Forensic Science and the Administration of Justice: Critical Issues and Directions, SAGE Publications, Inc., Thousand Oaks, California, 2014.
[8] S.A. Cole, Implementing counter-measures against confirmation bias in forensic science, J. Appl. Res. Mem. Cogn. 2 (2013) 61–62.
[9] W.C. Thompson, Painting the target around the matching profile: the Texas sharpshooter fallacy in forensic DNA interpretation, Law Probab. Risk 8 (2009) 257–276.
[10] D.E. Krane, S. Ford, J.R. Gilder, L. Inman, A. Jamieson, R. Koppl, I.L. Kornfield, D. M. Risinger, N. Rudin, M.S. Taylor, W.C. Thompson, Sequential unmasking: a means of minimizing observer effects in forensic DNA interpretation, J. Forensic Sci. 53 (2008) 1006–1007.
[11] E.J.A.T. Mattijssen, W. Kerkhoff, C.E.H. Berger, I.E. Dror, R.D. Stoel, Implementing context information management in forensic casework: minimizing contextual bias in firearms examination, Sci. Justice 56 (2016) 113–122.
[12] 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 (2003) 77–90.
[13] E.J.A.T. Mattijssen, C.L.M. Witteman, C.E.H. Berger, R.D. Stoel, Cognitive biases in the peer review of bullet and cartridge case comparison casework: a field study, Sci. Justice (2020) (in press).
[14] E.J.A.T. 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.
[15] A. Banno, T. Masuda, K. Ikeuchi, Three-dimensional visualization and comparison of impressions on fired bullets, Forensic Sci. Int. 140 (2004) 233–240.
[16] B. Bachrach, A Statistical Validation of the Individuality of Guns Using 3D Images of Bullets, National Institute of Justice, Washington, DC, 2006 March; Document No.: 213674.
[17] N. Senin, R. Groppetti, L. Garofano, P. Fratini, M. Pierni, Three-dimensional surface topography acquisition and analysis for firearm identification, J. Forensic Sci. 51 (2) (2006) 282–295.
[18] T.V. Vorburger, J.H. Yen, B. Bachrach, T.B. Renegar, J.J. Filliben, L. Ma, et al., Surface topography analysis for a feasibility assessment of a national ballistics imaging database, NISTIR 7362: a Report Prepared for the National Academies Committee to Assess the Feasibility, Accuracy and Technical Capability of a National Ballistics Database Under National Institute of Justice Grant 2003-IJ-R-029 With the NIST Office of Law Enforcement Standards (2007) http://www.nist.gov/manuscript-publication-search.cfm?pub_id=822733 (accessed March 18, 2020).
[19] U. Sakarya, U.M. Leloglu, E. Tunali, Three-dimensional surface reconstruction for cartridge cases using photometric stereo, Forensic Sci. Int. 175 (2) (2008) 209–217.
[20] B. Bachrach, A. Jain, S. Jung, R.D. Koons, A statistical validation and repeatability of striated tool marks: screwdrivers and tongue and groove pliers, J. Forensic Sci. 55 (2) (2010) 348–357.
[21] C. Gambino, P. McLaughlin, L. Kuo, F. Kammerman, P. Shenkin, P. Diaczuk, et al., Forensic surface metrology: tool mark evidence, Scanning 33 (2011) 272–278.
[22] F. Riva, Etude sur la valeur indicielle des traces présentes sur les douilles. [Thèse de doctorat], Université de Lausanne, Lausanne (Switzerland), 2011.
[23] 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.
[24] F. Riva, R. Hermsen, E.J.A.T. Mattijssen, P. Pieper, C. Champod, Objective evaluation of subclass characteristics on breech face marks, J. Forensic Sci. 62 (2) (2017) 417–422.
[25] J. Song, T.V. Vorburger, W. Chu, J. Yen, J.A. Soons, D.B. Ott, N.F. Zhang, Estimating error rates for firearm evidence identifications in forensic science, Forensic Sci. Int. 284 (2018) 15–32.
[26] D. Roberge, A. Beauchamp, S. Lévesque, Objective identification of bullets based on 3D pattern matching and line counting scores, Intern. J. Pattern Recognit. Artif. Intell. 33 (11) (2019) 1–34.
[27] P. Vergeer, A. van Es, A. de Jongh, I. Alberink, R. Stoel, Numerical likelihood ratios outputted by LR systems are often based on extrapolation: When to stop extrapolating? Sci. Justice 56 (6) (2016) 482–491, doi:http://dx.doi.org/10.1016/j.scijus.2016.06.003.
[28] I. Montani, R. Marquis, N. Egli Anthonioz, C. Champod, Resolving differing expert opinions, Sci. Justice 59 (1) (2019) 1–8.
[29] P. Pauw-Vugts, A. Walters, L. Øren, L. Pfoser, FAID2009: proficiency test and workshop, AFTE Journal 45 (2013) 115–127.
[30] W. Kerkhoff, R.D. Stoel, C.E.H. Berger, E.J.A.T. Mattijssen, R. Hermsen, N. Smits, H.J.J. Hardy, Design and results of an exploratory double blind testing program in firearms examination, Sci. Justice 55 (6) (2015) 514–519.
[31] W. Kerkhoff, R.D. Stoel, E.J.A.T. Mattijssen, C.E.H. Berger, F.W. Didden, J.H. Kerstholt, A part-declared blind testing program in firearms examination, Sci. Justice 58 (4) (2018) 258–263.