Estimating error rates for bullet comparisons in forensic science

Yawei Ge[a,b], Heike Hofmann[a,b]

[a] *Department of Statistics, Iowa State University,*
[b] *Center for Statistics and Applications of Forensic Evidence, Iowa State University*

Abstract

This is the abstract.

It consists of two paragraphs.

Text based on elsarticle sample manuscript, see http://www.elsevier.com/author-schemas/latex-instructions#elsarticle

1. Introduction

Firearm examination is an important topic in forensic science to help reveal the underlying pattern of firearm evidence. XXX go into a bit more detail on what you mean by the underlying pattern Currently, those examinations are executed in forensic labs by Firearms and Toolmark Examiners (FTEs) under the regulation of AFTE (AFTE Criteria For Identification Committee, 1992). However, this process is in its in nature based on human decisions and therefore subjective. XXX Change from passive voice to active, i.e. The PCAST report and the NRC called out this subjectivity and ... The subjectivity in the conventional firearm evidence identification processes is called to be reduced or to be complemented by more objective procedures by the National Research Council (National Research Council, 2009). The President's Council of Advisers on Science and Technology (PCAST) also emphasized the importance to establish the scientific validity and reliability of subjective forensic feature-comparison methods by blind empirical test (President's Council of Advisors on Science and Technology (PCAST), 2016). The PCAST report also recognized the importance of developing of objective computer-based algorithms in its following addendum (President's Council of Advisors on Science and Technology (PCAST), 2017).

In response to this criticism, subsequent research on cartridge case comparisons introduced the method of congruent matching cells (CMC) (Song, 2015), which allows a quantitative assessment of the similarity of two breech face impressions. Error rates were reported by [@Song2018; @Zhang2019] on XXX which studies were used in these papers? based on distributional models XXX

Email addresses: yaweige@iastate.edu (Yawei Ge), hofmann@iastate.edu (Heike Hofmann)

could you expand on those distributional models?. XXX xiao-hue Tai's paper should be mentioned. XXX mention the data that all of these methods are based on.

XXX i'm not sure you need to go into bullet comparisons

In bullet comparisons, an automatic matching algorithm was proposed by Hare (Hare et al., 2016). The discussion about the degraded land comparison was also made in the following paper (Hare et al., 2017). The empirical error rates associated with this method based on three test sets are discussed by Vanderplas (Vanderplas et al., 2020). The most prominent single feature in the random forest proposed by Hare (Hare et al., 2016) is the cross-correlation function (Vorburger et al., 2011) according to the empirical studies (Hare et al., 2016; Vanderplas et al., 2020). Other automated comparison methods or improvement such as Robust LOESS in groove engraved area identification (Rice et al., 2020), Chumbley score method (Krishnan and Hofmann, 2019) are also proposed.

XXX What does this however refer to?

However, due to the limitation of sample sizes and designs of study, a sound distributional assumption about the resulting similarity scores was not well established in the bullet comparison field for automated matching algorithms. The error rates are estimated empirically from dozens of bullets which were originally designed for examiner proficiency tests. And many of those tests are also criticized for the closed-set design and potential information provided to the examiners as pointed out by PCAST (President's Council of Advisors on Science and Technology (PCAST), 2017, 2016). It's important to understand the underlying distributions of the similarity scores to estimate the theoretical error rates of any automatic algorithm. And that also helps later to choose proper thresholds for the methods instead of simply choosing 0.5 (in a 0-1 range of similarity scores) as a belief.

Generally, in forensic science, we consider the problem of distinguishing two pieces of evidence coming from a same source or different sources. Specifically, in bullet comparisons, we want to know if two bullets are coming from a same gun or not. When shooting a bullet, the barrel will leave markings on the surface of bullets due to the imperfectness of manufacturing. We distinguish land engraved areas (LEA) and groove engraved error areas. And the former is compared by examiners by inspecting the striations under confocal microscopes in labs (AFTE Criteria For Identification Committee, 1992). And the LEA is also used in the developed computer-based methods. A fired bullet has several LEAs according to the barrels' design. So, we are not doing one step to identify bullets, instead we consider each LEA and then combine the land level conclusions to a bullet level one. Accordingly, we start our discussion about underlying distributions and error rates from the land level.

There are two types of distributions we would consider in the bullet comparison cases. One is called known-match (KM) distributions for those actually matched lands. The other is called known-non-match distributions (KNM) for those actually not matched lands. Those two distributions are only available in experimental settings where we know the ground truth. They are the starting

point for any discussion about theoretical error rates. And particularly, we consider the cross-correlation function (CCF) for bullet signatures as a typical similarity score with range 0-1 used to distinguish between same source and different source comparisons. The cross-correlation function (CCF) is calculated after extracting the signatures of bullet lands by maximizing the cross correlation of two sequence of signatures by horizontally adjusting the signatures (Vorburger et al., 2011).

In section 2... In section 3...

2. The distributional forms of similarity scores

The quantitative methods used to objectively measure the similarity between LEAs report various quantities, such as counts, correlations, distances, probabilities and more general similarity scores (Hare et al., 2016; Ma et al., 2004; Nichols, 2003, 1997). To understand how those quantities reflect the strength of evidence and to study the underlying error rates in making decisions based on those quantities, distributional forms are usually set up (Song et al., 2018; Zhang, 2019). Particularly, we are focusing on the similarity scores which range from 0 to 1 and the probabilities reported as the likelihood of an actual match. The similarity scores reported in the forensic researches are classified into two categories as known matches (KM) and known non-matches (KNM), and the corresponding distributions are named as KM distributions and KNM distributions. When we make any decisions based on any quantitative measurement, we are actually making a distinguish between those two potential distributions. The strength of any identification process is also measured by the disparities of those distributions. However, in practice, we can hardly discriminate those two distributions entirely, thus, we are never 100% sure which distribution the observed score comes from. This is where the identification error raises.

The cross-correlation function (CCF) is the most prominent single feature in the automatic random forest algorithm (Hare et al., 2016) which has theoretical range from -1 to 1. But in real applications, this similarity score always has values in 0 to 1, e.g. in our case, none of 121992 comparisons has negative values. So, it is selected as a representative of similarity scores with range from 0 to 1 in this paper for further analysis. The similarity scores in 0 to 1 can be explained as probabilities that quantify the likelihood that a pair of LEAs are actually a match. Or we can think of them as general similarity measurement. As the name indicated, the higher the similarity score is, the stronger evidence is to support the same source assumption. For different combination of ammunition and firearms, the scores are distributed differently. It is expected that systematic differences exist there for different cases (Vanderplas et al., 2020). So, it is necessary to study the scores under controlled conditions.

We can see from the Figure ... (a figure of nonparametric fit of the CCF?), which is a typical one we usually have for the similarity scores, that the distributions of KM and KNM are apart for the majorities. In the bullet LEA comparison problems, we usually have a well separated bullet scores but for the land scores, there are usually some overlaps (Vanderplas et al., 2020). We

propose beta distributions for those scores. Because the beta distribution is a well-used distribution in statistics to describe a quantity from 0 to 1 which is usually a probability or proportion quantifying our knowledge for another distribution in Bayesian analysis. And it is very flexible to capture unimodal asymmetric shapes in 0 to 1. However, it may not be adequate to explain a heavy tail or even a second mode. Thus, we further consider the beta mixture distribution which is a more complex distribution than the beta distribution as a special case. In the beta mixture distribution, we introduce a hierarchical structure with a prior probability to combine a few beta distributions as one. The two-component beta mixture distributions are defined below.

The reported similarity scores are denoted as Y_{ij} for j^{th} LEA comparison within class i, where $i = 1$ is KM and $i = 2$ is KNM. Y_{ij}'s are considered independent and identically distributed within each class, i.e.

$$Y_{1j} \overset{iid}{\sim} Betamix(p_1, \mu_{11}, \phi_{11}, \mu_{12}, \phi_{12})$$
$$Y_{2j} \overset{iid}{\sim} Betamix(p_2, \mu_{21}, \phi_{21}, \mu_{22}, \phi_{22})$$

where μ_{ik} and ϕ_{ik} are distribution parameters for i^{th} class and k^{th} component, $k = 1$ or 2. And p_i is the prior probability for the first component in i^{th} class, thus, $1 - p_i$ is the prior probability for the second component in i^{th} class. Note that we are using the mean and precision parameterization of beta distributions which simplifies the math in calculation and is more intuitive (μ is the mean, and the variance is roughly proportional to the reciprocal of ϕ). It's equivalent to the usual α and β parameterization through the following transformation:

$$\mu = \frac{\alpha}{\alpha + \beta}$$
$$\phi = \alpha + \beta$$

3. LAPD data set and the estimated distributions

For the following sections of the paper, we will base our analysis on the LAPD data sets. This is a large data set of ... It is the first time such a large data set available to the researchers, which makes it possible for a statistical analysis for the distribution of similarity scores. (More on the data. Is there any formal documentation for LAPD? And number of land comparisons for different cases should be discussed here)

We consider cross correlation functions (CCF) for land comparisons, which is produced by calculating the maximized CCF between two signatures extracted from a pair of bullet LEAs.

The estimation was done using Nelder-Mead algorithm in R [reference]. This is a general purpose numeric method which works reasonably well for multidimensional optimization problems. In our case, the objective function is the log likelihood function of the beta mixture distribution. Therefore, we

finally found the maximum likelihood estimates (MLE). Instead of a simpler beta model, we start with the more complex two-component beta mixture model, see how well it fits the data and test if some components are necessary. We estimated beta mixture distributions for both KM and KNM as in Table 1 and Table 2 respectively. For the beta distributions, the KM distribution has a mean at 0.635 and the KNM distribution has a mean at 0.404. We can also see that for the two component beta mixture distributions for both KM and KNM, there are components with mean around 0.5 and the other components more separated from each other. This indicates the model successfully accounted for different situations of comparisons involving tank rash and random identification etc. This is an ideal property we would like to see and make use in explaining the similarity scores. Another point about the estimates that worth mentioning is that the ϕ of the KM beta distribution is much smaller than any of the two components of the two-component KM beta mixture distribution. This indicates that the two component distribution could be better because of smaller variances for components.

Full Data KM Distribution Estimation						
Model	Component Prior Probability	μ	ϕ	logLik	p-value	BIC
3-comp	0.247	0.406	24.097	6278		-12479
	0.521	0.664	13.342			
	0.232	0.819	33.823			
2-comp	0.419	0.466	15.812	6243	0	-12438
	0.581	0.759	18.476			
Beta		0.635	6.529	5712	0	-11404

Table 1: Parameter estimations for the Beta distribution (1-component beta mixture distribution), 2-component and 3-component distributions for the KM CCF. "2-comp" refers to the 2-component beta mixture distribution, and the same for "3-comp". Column "logLik" is the maximized log likelihood for each distribution. Column "p-value" is the p-value for asymptotic likelihood ratio tests between the current model and one-step more complex model, where 0 indicates that we would reject the hypothesis that the current one is sufficient to decribe the data

The estimated beta distributions are shown in Figure 1 and the estimated two-component beta mixture distributions are also shown in Figure 2. As expected, the estimated distributions show the properties we desire for the similarity scores. The majorities of the estimated distributions are apart from each other, while the minority part between the two distributions has some overlap. It's worth to note that both curves have heavy tails to the farther boundaries and the KM curve has the heavier tail compared to KNM. We can see the two-component distributions fit the data very well, while the single beta distributions are not as good as the two-component distributions for both KM and KNM. The single beta distribution for KM clearly failed to capture the potential second mode of the histogram indicating it is not sufficient. The single beta distribution for

Full Data KNM Distribution Estimation						
Model	Component Prior Probability	μ	ϕ	logLik	p-value	BIC
3-comp	0.650 0.314 0.032	0.345 0.490 0.667	35.017 24.979 18.454	78961		-157830
2-comp	0.674 0.336	0.358 0.494	39.908 13.324	78808	0	-157558
Beta		0.404	15.801	75349	0	-150675

Table 2: Parameter estimations for the Beta distribution (1-component beta mixture distribution), 2-component and 3-component distributions for the KNM CCF. "2-comp" refers to the 2-component beta mixture distribution, and the same for "3-comp". Column "logLik" is the maximized log likelihood for each distribution. Column "p-value" is the p-value for asymptotic likelihood ratio tests between the current model and one-step more complex model, where 0 indicates that we would reject the hypothesis that the current one is sufficient to decribe the data

KNM is not too bad but still failed to capture the distributional information for some parts. This indicates that even though beta distributions are flexible for variables in 0 to 1, they are still restricted too much to be able to describe the cases here. The two-component beta mixture distributions look more promising.

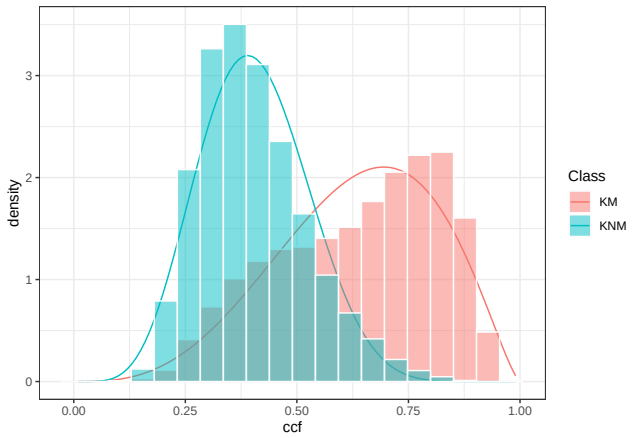

Figure 1: Estimated beta distributions for full data

Three candidate distributions for each of KM and KNM are considered. For the increasing of complexity, we have single beta, two-component beta mixture and three-component beta mixture distributions. And they are also nested in that order. Naturally, we first look at the maximized log-likelihood of each model, and we can do asymptotic log-likelihood ratio chi-square tests for single beta against two-component beta mixture distributions, and two-component against

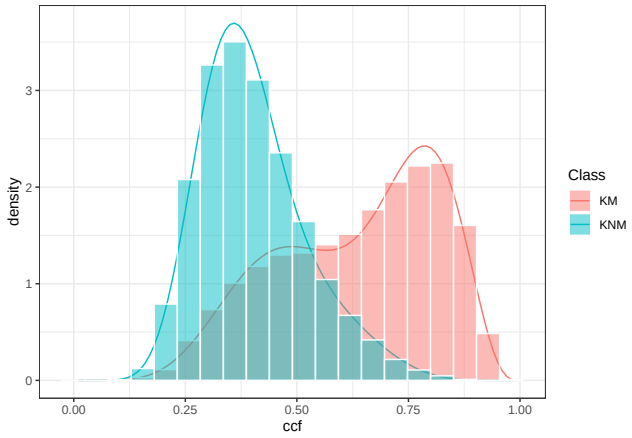

Figure 2: Estimated two-component beta mixture distributions for full data

three-component beta mixture distributions. The p-values of those test are shown in the Table 1 and Table 2 as the column "p-value". Surprisingly (or not), we found all those p-values are 0 which strongly suggests a more complex model when there is one. However, considering the size of the data we are using to fit these models, we can expect the statistical significance will be easily achieved since any small difference of the sufficiency will be detected. So we have to take the sample size effect into consideration. As Bayesian information criterion (BIC) is a well used criterion which takes the model complexity, sufficiency and the sample size (by a log function) into account. As shown in the column "BIC" in the tables, the BICs for the single beta distributions are larger than that of the two-component beta mixtures by relatively large proportions. And the BICs for the two-component beta mixture distributions are a little larger than that of the three-component beta mixture distributions for both KM and KNM. Obviously, we would prefer the two-component beta mixture distributions, but we would cast a doubt when it comes to the three-component beta mixture distributions. We still prefer the two-component one instead of the three-component one. The reasons are: 1) these differences of BICs between two and three component distributions are really small in proportion (by 0.3% for KM and 0.017% for KNM), 2) the estimation cost is not accounted, which will be higher for a more complex model, 3) the BICs still don't take the sample size effect fully into account since the log function for sample size goes to flat when the sample sizes are large. By simple calculation, we can see that the log-likelihood increased by a multiplicative factor more than 10 when the sample sizes increased from KM case to KNM case, however, at the same time, the log of sample size only roughly increased from 9 to 11. So we choose the two-component beta mixture distributions for both KM and KNM. And also as we have seen, these distributions have potential good forensic interpretations.

It's also helpful to see how the individual components look like in the beta

mixture distributions as in Figure 3. The KNM and KM distributions seem to share a common component while keeping the other components far apart from each other. The separated components represent the ideal cases where the bullet land engraved areas preserved the information of source well and result in clearly distinct separation. The shared components represent the cases where the KM results in lower scores because of some degree of tank rash, pitting, breakoff or other damages on the bullets and the KNM results in higher score because of the random identification effect. According to the estimated prior probabilities, both distributions put less weight on the common component while putting larger weight on the the components characterizing the differences of KM and KNM respectively, which agrees on our expectation that majorities of the distributions are separated while the minorities overlapped. These properties together well explained the observed empirical distribution of similarity scores in our cases. Particularly, the heavier tail of the KM comparisons is explicitly included in the form of the model by one of the components.

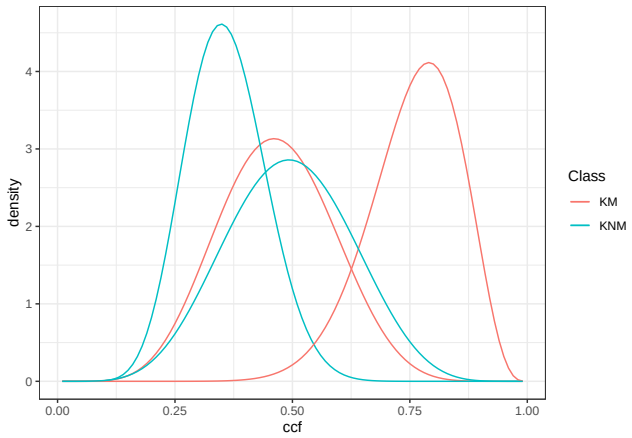

Figure 3: Estimated-components

4. Evaluate the error rates

- Given the iid assumption, could we develop a further model for the bullet level distribution and estimate the corresponding error rates?

Estimating the underlying theoretical error rates is a fundamental challenge in the forensic science (President's Council of Advisors on Science and Technology (PCAST), 2017, 2016). The estimation process becomes more straightforward after we finished the distribution estimation. There are many types of error rates can be defined. We focused on the four most used ones, i.e. 1) False positive rate (FPR), 2) False negative rate (FNR), 3) False identification rate (FIR), 4) False exclusion rate (FER), as defined in Table 3. FPR and FNR

are considered source-specific assessment. They are both probabilities given the ground truth to make false conclusions, which is useful especially when we evaluate an algorithm. FIR and FER are considered decision-specific assessment. They are both probabilities given the algorithm decisions to false claim the actual source, which is useful in real cases when jury is presented with the forensic conclusions to make judgment.

	Ground Truth		
	Match	Non-Match	Error Rate
Identification	Correctly judged as matches, True Positives (TP)	Incorrectly judged as matches, False Positive (FP)	False Identification Rate (FP/(TP + FP))
Elimination	Incorrectly judged as non-matches, False Negative (FN)	Correctly judged as non-matches, True Negative (TN)	False Exclusion Rate (FN/(TN + FN))
Error Rate	False Negative Rate (FN/(FN+TP))	False Positive Rate (FP/(FP + TN))	

Table 3: xxxx

The error rates are also affected by the threshold selected to make decisions. In an extreme example, if we are very conservative in making identifications, we can set the threshold to be large enough, say CCF >= 0.99, then the FPR will be 0. But at the same time, the FNR will be 1. It's clear that there are trade-offs between the error rates. So, to make these error rates comparable and consistent, for the current paper, we refer the threshold as the value where the probability densities of KM and KNM distributions equal with each other. This is not a bad choice, since it has good statistical meaning that at this value, the likelihood ratio equals 1 and any deviation from 1 would give preference to a certain decision (as if a threshold). For the two component beta mixture distributions we estimated in last section, this threshold is 0.529. In the application of the algorithms, the threshold need to be carefully chosen where the study of associated error rates will in turn be essential guiding that process.

The four theoretical error rates are directly associated with the distributions we estimated as shown in the Figure 4 and Figure 5. The estimated values of the error rates are reported in Table 4. The FPR is 0.148, while the FNR is 0.301 which is much larger. This means for the given threshold value 0.529, the algorithm has probability 0.148 to misclssify an actual different-sourced pair of land engraved areas (LEAs) into identifications. And the algorithm has a probability 0.301 to misclassify an actual same-sourced pair of land engraved areas (LEAs) into eliminations. The fact that FNR is larger than FPR implies that for the current threshold, we are more likely to misclassify an pair of KM LEAs instead of KNM LEAs. This is partly due to the fact the KM distribution is more spread than the KNM distribution. The FPR will decrease rapidly when increasing the threshold, while the FNR will be small only when the threshold is

very low at great cost of FPR. And similar pattern can be observed from FIR and FER. However, these error rates are explained in a totally different view. FIR is 0.212 means when the algorithm reports an identification (to the jury in court potentially), the algorithm could have made wrong claim at probability 0.212. As we aforementioned, the FPR and FNR are used to evaluate algorithms overall while the FIR and FER are used to evaluate a certain result reported.

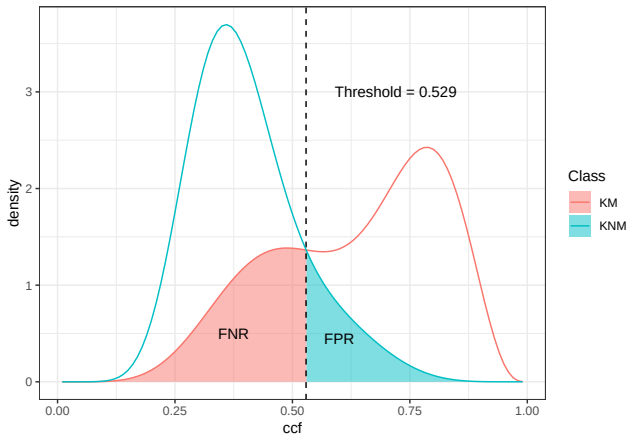

Figure 4: False Negative Rate(FNR) and False Positive Rate(FPR)

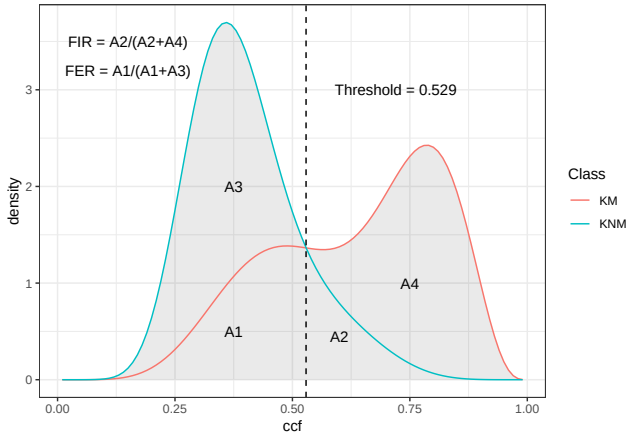

Figure 5: False Indentification Rate(FIR) and False Exclusion Rate(FER)

Overall, the error rates are not small enough to give confidence to any conclusions. But remember the current comparison discussed is in the land level. In real cases, we will need to combine the land level conclusions to bullet level ones. And this process of combining multiple pieces of evidence will improve the performance greatly. One of the common ways in computer-based bullet analysis

Type of Error Rates	Estimated Values
False positive rate (FPR)	0.148
False negative rate (FNR)	0.301
False identification rate (FIR)	0.212
False exclusion rate (FER)	0.354

Table 4: xxxx

is to calculate the sequential average maximum (SAM) [susan's paper and the original paper of SAM?] of land scores. Next, we are going to making use of the KM and KNM distributions in the land level to explore the error rates of the bullet level.

5. Estimations with changing sample size

(quantify the variation, reproduce some results in the previous sections with changing sample size)

6. Conclusion

References

AFTE Criteria For Identification Committee, 1992. Theory of identification, range striae comparison reports and modified glossary definition afte criteria for identification committee report. AFTE Journal 24, 336–340.

Hare, E., Hofmann, H., Carriquiry, A., 2016. Automatic matching of bullet land impressions. Annals of Applied Statistics 11, 2332–2356.

Hare, E., Hofmann, H., Carriquiry, A., 2017. Algorithmic approaches to match degraded land impressions. Law, Probability and Risk 16, 203–221. doi:10.1093/lpr/mgx018

Krishnan, G., Hofmann, H., 2019. Adapting the chumbley score to match striae on land engraved areas of bullets. Journal of Forensic Sciences 64, 728–740. doi:10.1111/1556-4029.13950

Ma, L., Song, J.-F., Whitenton, E.P., Vorburger, T.V., Zhou, J., Zheng, A., 2004. NIST Bullet Signature Measurement System for RM (Reference Material) 8240 Standard Bullets 49(4).

National Research Council, 2009. Strengthening forensic science in the united states: A path forward, Strengthening Forensic Science in the United States: A Path Forward. National Academies Press. doi:10.17226/12589

Nichols, R., 2003. Firearm and toolmark identification criteria: A review of the literature, part II. Journal of forensic sciences 48, 318–27. doi:10.1520/JFS14149J

Nichols, R.G., 1997. Firearm and Toolmark Identification Criteria: A Review of the Literature. Journal of Forensic Science 42, 466–474. doi:10.1520/JFS14149J

President's Council of Advisors on Science and Technology (PCAST), 2016. Report to the president, forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods. Executive Office of the President of the United States, President's Council ....

President's Council of Advisors on Science and Technology (PCAST), 2017. An Addendum To The PCAST Report On Forensic Science In Criminal Courts.

Rice, K., Genschel, U., Hofmann, H., 2020. A robust approach to automatically locating grooves in 3D bullet land scans. Journal of Forensic Sciences 65, 775–783. doi:10.1111/1556-4029.14263

Song, J., 2015. Proposed "Congruent Matching Cells (CMC)" Method for Ballistic Identification and Error Rate Estimation 47, 9.

Song, J., Vorburger, T.V., Chu, W., Yen, J., Soons, J.A., Ott, D.B., Zhang, N.F., 2018. Estimating error rates for firearm evidence identifications in forensic science. Forensic Science International 284, 15–32. doi:10.1016/j.forsciint.2017.12.013

Vanderplas, S., Nally, M., Klep, T., Cadevall, C., Hofmann, H., 2020. Comparison of three similarity scores for bullet lea matching. Forensic Science International 308. doi:10.1016/j.forsciint.2020.110167

Vorburger, T.V., Song, J.F., Chu, W., Ma, L., Bui, S.H., Zheng, A., Renegar, T.B., 2011. Applications of cross-correlation functions. Wear, The 12th International Conference on Metrology and Properties of Engineering Surfaces 271, 529–533. doi:10.1016/j.wear.2010.03.030

Zhang, N.F., 2019. The use of correlated binomial distribution in estimating error rates for firearm evidence identification. Journal of Research of the National

Institute of Standards and Technology 124. doi:10.6028/jres.124.026