BRIEF REPORT

Hidden Multiple Comparisons Increase Forensic Error Rates

Susan Vanderplas[a,1,2], **Alicia Carriquiry**[b, c], **and Heike Hofmann**[b, c, 1]

This manuscript was compiled on April 17, 2024

When wires are cut, the tool produces striations on the cut surface; as in other forms of forensic analysis, these striation marks are used to connect the evidence to the source that created them. Here, we argue that the practice of comparing two wire cut surfaces introduces complexities not present in better-investigated forensic examination of toolmarks such as those observed on bullets, as wire comparisons inherently require multiple distinct comparisons, increasing the expected false discovery rate. We call attention to the multiple comparison problem in wire examination and relate it to other situations in forensics that involve multiple comparisons, such as database searches.

Forensic Evidence | Statistics | Wire cuts | Toolmark analysis

In forensic evaluations, a single conclusion often relies on many comparisons, either implicitly or explicitly. Multiple comparisons arise persistently when developing statistical methods to address scientific problems (1), and greatly increase the probability of false discoveries. Now that vast databases and efficient algorithms are routinely used in forensic evaluations to propose matches to crime scene items, the problem of close non-matches (2) due to multiple comparisons becomes critically important. This often ignored issue increases the false discovery rate, and can contribute to the erosion of public trust in the justice system through conviction of innocent individuals. The multiple comparison problem is not new: it has been raised in the past with regard to DNA (3) and latent print evaluations (4). One of the root causes (5) leading to the wrongful accusation of Brandon Mayfield in the 2004 Madrid train bombing case was that the large size of the IAFIS database used to search for similar prints made it possible to locate ‘unusually’ close non-matches. As database size increases, so does the probability of finding a close non-match.

Compounding this issue, the use of algorithms also results in a large number of comparisons that are not obvious to the user. For example, the cross-correlation function (6), which computes the correlation for each alignment of two sequences, was one of the first measures proposed to quantify the similarity between two patterns in response to the 2009 NRC report (7), and continues to be used in many pattern searching algorithms to find the best alignment between two images and to quantify their overall similarity. Finding the best alignment often consists in sliding one surface across the whole length (for one-dimensional patterns, such as striations) or area (for two dimensional sources, such as impression marks) of the other item while keeping track of the value of a similarity measure. This mirrors the forensic examination process: the examiner visually rotates and shifts items under a comparison microscope to align two surfaces. In order to avoid false accusations and the corresponding impact on public perception of forensics, we must address the problem of multiple comparisons in database and alignment searches and control their effect on false discovery rates.

Here, we consider the multiple comparisons problem that arises from a relatively simple toolmark examination: matching a cut wire to a wire-cutting tool. We describe the comparison approach, estimate the (minimal) number of comparisons that are needed to carry out the examination, and discuss how the false discovery rate changes with the number of comparisons involved, using error rates derived from published black-box studies.

Examination Process

A forensics examiner tasked with determining whether a wire in evidence was cut by a recovered tool will create one or more blade cuts, which are then compared to the cut surface of the wire recovered from the scene. These cuts are made in

Author affiliations: [a]Statistics Department, University of Nebraska Lincoln. 350 Hardin Hall, 3310 Holdrege North Wing, Lincoln, NE 68503; [b]Department of Statistics, Iowa State University. 1121 Snedecor Hall, 2438 Osborn Dr, Ames, IA 50011; [c]Center for Statistics and Applications in Forensic Evidence. 195 Durham Center, 613 Morrill Road, Ames, Iowa 50011

Please provide details of author contributions here.

The authors have no competing interests to declare.

[1]SVP (Author One) contributed equally to this work with HH (Author Two).

[2]To whom correspondence should be addressed. E-mail: susan.vanderplas@unl.edu

a sheet of material matching the wire composition, and may be performed at multiple angles, as the angle of the tool to the substrate can affect which striations are recorded on the substrate surface. The blade cuts will then be compared to the wire under a comparison microscope, though eventually, automatic comparison algorithms may also be validated for lab use. Each side of each blade cut will be compared to each side of the wire; different tool designs have between 2 and 4 cutting surfaces in contact with the substrate.

Calculating the Number of Comparisons

In order to calculate the number of comparisons carried out in the course of one examination, we define b to be the length of the blade cut, and d to be the diameter of the wire. We assume that the wire is covered with striations suitable for comparison across its full diameter d. If this is not the case, we reduce the value d. Both the blade and the wire are either digitally scanned at resolution r mm per pixel, or visually examined using a microscope with a digital resolution that can be expressed as r equivalent to the digital scan. An illustration of the sliding comparison process is shown in Figure 1. Imagine that we move the cut wire along the blade cut in order to assess whether striations on the blade cut match the striations on the wire. We can move the wire unit-by-unit, or we can move the wire by its full length, with no overlap to the previous comparison.

The first option gives us the maximum number of comparisons $(b/r - d/r + 1)$, while the second option gives us the minimum number of comparisons b/d. In the first case, sequential comparisons share much of the same physical data and are highly related; in the second case, no data are shared between physical comparisons and we can expect that they are statistically independent, though empirically there will be nonzero correlations due to physical similarities between striations. For simplicity, let us consider the number of comparisons to lie somewhere between these two estimates. Note that when $b/d \approx 1$, as in some toolmark comparisons, the upper number of comparisons goes to 1. Finally, we must consider the number of surfaces which must be compared: the wire may have one or two sets of striae and there may be two to four blade cut surfaces to examine, depending on the tool. This results in a multiplier of as much as 8.

A concrete example. Let us consider a wire-cutting tool with a 1.5 cm razor blade that meets a cast surface (one such tool is shown in Figure 1); the wire is held against this rectangular cast surface as the blade is pushed into the wire, splitting it in two. This is a minimal scenario - the wire will acquire striations from one side of the blade, while the blade itself has two cutting edges, which we will call side A and side B. A blade cut of a sheet of aluminum will thus produce two striated edges corresponding to side A and side B which are compared to cut wires to assess similarity. We also have a 12 gauge aluminum wire (2 mm diameter) which may have been cut by the wire-cutting tool described above. Class characteristics, which are shared by all tools of similar manufacture, appear to match: there is a flat impression on one side of the wire corresponding to the cast metal backstop of the tool, and the wire is cut such that the blade and the backstop appear to be perpendicular (that is, the wire appears to have been cut with a tool of similar configuration).

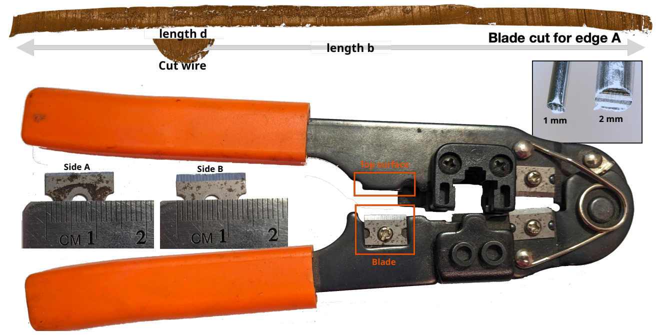

Fig. 1. (Top) A comparison between a wire and a blade cut requires sliding the wire along the entire blade cut length to determine the best match (or whether there is a match). Surfaces shown are rendered 2D topographical scans of a wire and blade cut taken with a confocal light microscope. (Bottom) RJ45 Crimp tool with a 1.5 cm razor blade used for cutting. 1 mm and 2 mm diameter aluminum wires cut with the pliers are shown in a box in the top right corner.

In this example, $b = 15$ mm, $d = 2$ mm, and there are at least $b/d = 7.5$ comparisons between a wire cut and a blade cut. As there are two blade cuts (side A and side B), the minimal number of comparisons is 15, as these comparisons are non-overlapping and independent (on average).

Assuming a resolution of 0.645μm per pixel, the maximum number of comparisons per blade cut is around $20,000$; thus, we need $40,000$ comparisons in order to find the optimal alignment between the wire and the blade cut. These comparisons are implicit in the calculation of cross-correlation, which is the first and often the only step used to quantitatively assess the similarity between striated evidence such as bullets, aperture shear, and firing pin impressions. Implicit comparisons are not unique to algorithms; an examiner would need to physically align the wire and the blade cut by searching along the length of the cut to visually match striations, performing the same process physically that the algorithm performs computationally. While these sequential comparisons are highly auto correlated, and we cannot assume sequential independence when calculating the probability of an error, they serve as an upper bound on the number of comparisons which could be performed. As the number of comparisons increases, the probability of encountering a coincidental match increases. Statisticians call this the *family-wise error rate* E; it is an important quantity to control when conducting a series ("family") of tests(8).

Probability of False Discoveries

There are at least two components of the false discovery rate (FDR): identifying two pieces of evidence that have similar characteristics but are from different sources (a coincidental match) and procedural failures (e.g. lab process errors) (2, p 50). In objective disciplines with standardized evaluation rules (e.g. DNA), these sources can be distinguished. However, in toolmark examination, no objective evaluation rules are used; examiners testify based on subjective rules for how much similarity is sufficient for an identification. Assuming that lab procedure errors are not a factor in studies, we use reported error rates from three open-set studies of striated evidence(9–11) to obtain a ballpark estimate of the coincidental match rate of a single wire-cut

Study	FDR e	False Discoveries (%) in N comparisons E_{10}	E_{100}	$E_{1,000}$	$E_N < 0.1$
Mattijssen (2021)	7.24%	52.8	99.9	100.0	1
Pooled Error	2.00%	18.3	86.7	100.0	5
Bajic (2020)	0.70%	6.8	50.7	99.9	14
Best (2022)	0.45%	4.5	36.6	98.9	23
	1 in 1,000	1.0	9.5	63.2	105
	1 in 10,000	0.1	1.0	9.5	$1,053$
	1 in 100,000	10^{-4}	0.1	1.0	$10,535$

Table 1. Table showing the relationship between false discovery rates and the chance of a false discovery in N comparisons for a set of different FDRs and different number of comparisons. The last column gives the number of comparisons allowed while ensuring a familywise false discovery percentage of at most 10%.

comparison. These studies have FDRs between 0.0045 (11) and 0.072 (10); pooling data from these studies weighted by sample size yields an FDR of 0.02. For a single-comparison FDR of e, the family-wise FDR for n comparisons, E_n is $1 - [1 - e]^n$. Table 1 shows the impact the number of comparisons has on these published error rates. With an error rate of 0.07, as suggested by Bajic (2020), examiners can make up to 14 comparisons, i.e. even the simple example in this paper exceeds an upper bound of 10% for the family wise false discovery error. To conduct a search of a modestly sized database with 1000 entries, the initial FDR cannot exceed 1 in 10,000 to guarantee a family-wise total false discovery error of at most 10%.

Under these constraints, the accuracy of an examination involving multiple comparisons between a wire and a tool will be low, as the number of candidate alignments that must be examined is high. Even the most innocuous example (small blade, only 2 cutting surfaces, and a relatively large wire) involves a minimum of 15 comparisons. Examiners would make cuts under multiple angles (12), increasing the number of comparisons and making a false discovery even more probable. As a result, it is questionable whether wire comparisons made under current protocols are reliable enough to be presented at trial.

Clearly, studies for wire evidence, and larger studies for striated evidence in general, are necessary. Moving away from binary assessments toward quantification of striation similarity and observed pattern frequency will also reduce the severity of this issue and allow examiners to assign unusual striation patterns more weight in the process.

Discussion & Conclusions

Forensic practitioners often report the findings from their examinations in the form of a categorical conclusion reflecting a single decision. This is misleading when the decision relies on multiple comparisons which are not individually presented in reports or testimony. In this short contribution, we have shown that the implicit comparisons performed during forensic analysis of wire cuts increase the family-wise error rate.

We describe a simple scenario where a wire is cut using a two-sided blade, but findings apply to any situation where a forensic evaluation involves multiple comparisons, including, e.g., database searches. Forensic practitioners should understand how the number of comparisons can affect the accuracy of their final conclusion. We propose three strategies to enhance transparency and enable more reliable estimates of examination-specific error rates.

First, examiners should report (or defense attorneys should request) the overall length or area of surfaces generated during the examination process, along with the total consecutive length or area of the recovered evidence. These pieces of information will take the place of b and d and facilitate calculation of examination-wide error rates.

Second, researchers should conduct studies relating the length/area of comparison surface to the error rate. For instance, we have pooled studies looking at bullet striations and firing pin shear marks because we could not find black-box error rate studies of wire cuts. The striated surfaces are of orders of magnitude different lengths, but represent the best estimate of the error rate for striated materials. New studies should be designed to assess error rates (false discovery and false elimination) when examiners are making difficult comparisons.

Finally, when databases are used at any stage of the forensic evidence evaluation process (from suitability assessment and triage to reports which will be used at trial), the number of database items searched (or comparisons made) and the number of results returned must be reported. Additionally, the number of results used for further manual comparison should also be reported. For example, if a firearms examiner searches a local NIBIN database with 1000 entries, requests the 20 closest matches to her evidence, and then carries out a physical examination of five exemplars from the list of 20, all of those values should be clearly reported to enable estimation of the familywise error rate. This will help make the multiple comparison issue accessible to everyone involved in evaluating the value of forensic evidence: examiners, lawyers, jurors, and judges.

ACKNOWLEDGMENTS. This work was partially funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreements 70NANB15H176 and 70NANB20H019 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, Duke University, University of California Irvine, University of Virginia, West Virginia University, University of Pennsylvania, Swarthmore College and University of Nebraska, Lincoln.

1. Y Benjamini, Y Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. Royal Stat. Soc. Ser. B (Methodological)* **57**, 289–300 (1995).
2. President's council of advisors on science and technology, *Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods.* (Executive Office of the President of the United States, President's Council), (2016).
3. WC Thompson, F Taroni, CGG Aitken, How the probability of a false positive affects the value of dna evidence. *J. Forensic Sci.* **48**, 2001171 (2003).
4. JJ Koehler, S Liu, Fingerprint error rate on close non-matches. *J. Forensic Sci.* **66**, 129–134 (2021).
5. GA Fine, A review of the fbi's handling of the brandon mayfield case unclassified executive summary, (Washington DC), Technical report (2006).
6. T Vorburger, et al., Applications of cross-correlation functions. *Wear* **271**, 529–533 (2011).
7. NRC, *National Research Council: Strengthening Forensic Science in the United States: A Path Forward.* (National Academies Press), (2009).
8. J Tukey, Multiple comparisons. *J. Am. Stat. Assoc.* **48**, 624–625 (1953).
9. S Bajic, LS Chumbley, M Morris, D Zamzow, Validation study of the accuracy, repeatability, and reproducibility of firearm comparisons, (Ames, IA), Technical report (2020).
10. EJAT Mattijssen, et al., Firearm examination: Examiner judgments and computer-based comparisons. *J. Forensic Sci.* **66**, 96–111 (2021) _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1556-4029.14557.
11. BA Best, EA Gardner, An assessment of the foundational validity of firearms identification using ten consecutively button-rifled barrels. *AFTE J.* **54**, 28–37 (2022).
12. M Baiker, Toolmark variability and quality depending on the fundamental parameters: Angle of attack, toolmark depth and substrate material. *Forensic Sci. Int.* p. 10 (2015).