Statistical Evaluation of Firearms and Toolmark Evidence

Susan Vanderplas

March 11, 2020

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

Outline

- Legal challenges to forensic analysis
- Firearms and Toolmark Examination primer
- Issues
  - Scientific foundations
  - Subjectivity of comparisons
  - Error Rates

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

The Washington Post
Democracy Dies in Darkness

The Watch • Opinion

Incredibly, prosecutors are still defending bite mark evidence

(Getty Images)

Most Read Opinions

1 Opinion
The Trump administration's green ca
Catch-22

2 Opinion
The rest of the world is preparing for t
more years of Trump

3 Opinion
This presidential race is a New York t
throuple of white, 70-something male

4 Opinion
Bloomberg and Sanders are as wrong
they are self-serving

5 Opinion
Trump puts an unqualified loyalist in
charge of national intelligence

4 / 38

ABA JOURNAL

NEWS IN-DEPTH BLAWGS ABOUT

Home / In-Depth Reporting / Crime labs under the microscope after a string...

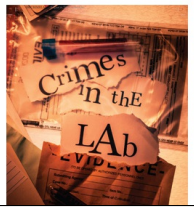
FEATURES

Crime labs under the microscope after a string of shoddy, suspect and fraudulent results

BY MARK HANSEN

SEPTEMBER 1, 2013, 10:20 AM CDT

Like 65 Share Tweet Share



In January, the New York City medical examiner's office confirmed that it was reviewing more than 800 rape cases from a 10-year period during which DNA evidence may have been mishandled by a lab technician who resigned in 2011 after an internal review uncovered problems with her work.

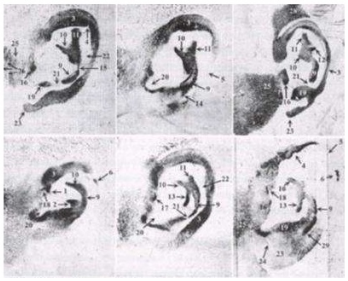
The review, then about half complete, had already turned up 26 cases in which the former technician failed to detect the presence of DNA evidence, including one in which the evidence has since led to an arrest in a 10-year-old rape case. The review uncovered 19 cases in which DNA evidence was commingled with DNA evidence from other cases.

A month earlier, a former chemist at a now-shuttered state drug lab in Boston was indicted on 27 counts of obstructing justice,

5 / 38

FEBRUARY 24, 2006

Earprints as evidence?



6 / 38

INNOCENCE PROJECT

About The Cases Get Involved Latest More

News 01.19.10

Three Freed, and FBI Continues to Review Ballistic Cases

It has been five years since the FBI stopped using an unreliable forensic test to determine the source of bullets, and a review of more than 2,500 cases involving the faulty evidence is still ongoing.

The Associated Press reports today

that at least three convictions have been overturned nationwide after bullet lead evidence was debunked, and the FBI has notified prosecutors in 187 cases that testimony offered by FBI experts "exceeds the limits of the science and cannot be supported by the FBI."

Challenges to Forensic Analysis

- 2009 National Academy of Sciences Report - Strengthening Forensic Science in the United States: A Path Forward
- 2016 President's Council of Advisors on Science and Technology (PCAST) report - Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature Comparison Methods

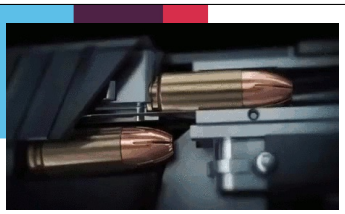
Fundamental Conclusions: Problems with

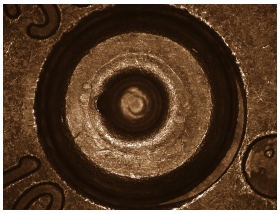
- Science - Poor or nonexistent scientific foundations for specific analyses
- Subjectivity - Conclusions are based off of subjective evaluations
- Screw ups - Estimates of error rates are nonexistent, not credible, or based on poorly designed studies

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

A Primer on Firearms Analysis

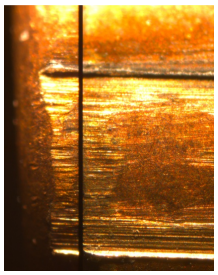
csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org





Locard's Exchange Principle:
Every contact leaves a trace

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org



csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

AFTE Theory of Identification

Option 1: Identification

> Agreement of a combination of individual characteristics and all discernible class characteristics where the extent of agreement exceeds that which can occur in the comparison of toolmarks made by different tools and is consistent with the agreement demonstrated by toolmarks known to have been produced by the same tool.

Option 2: Elimination

> Significant disagreement of discernible class characteristics and/or individual characteristics.

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

AFTE Theory of Identification

Option 3: Inconclusive

> (a) Some agreement of individual characteristics and all discernible class characteristics, but insufficient for an identification.
>
> (b) Agreement of all discernible class characteristics without agreement or disagreement of individual characteristics due to an absence, insufficiency, or lack of reproducibility.
>
> (c) Agreement of all discernible class characteristics and disagreement of individual characteristics, but insufficient for an elimination.

Under AFTE Theory of Identification, inconclusive results are not errors. An examiner could report nothing but inconclusive results for their entire career and testify that they have a 0% error rate.

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

AFTE Theory of Identification

Option 4: Unsuitable

> Unsuitable for examination.

Unsuitable evidence should be discarded before it is compared to known samples.

Issue 1: Scientific Foundations of Firearms Examination

Scientific Foundations

> Conclusions drawn in firearms identification should not be made to imply the presence of a firm statistical basis when none has been demonstrated

What would we need? According to Spiegelman & Tobin (2013)

- Every rifled firearm brand
  - different production settings, batches, tempering methods, barrel alloys
- Different ammunition types and sizes
- Different break-in periods for the guns
- Different maintenance procedures and lubrication types

Examine multiple fired bullets/cartridges from **each combination of factors** to determine if the markings are unique.

Compare features of the markings across caliber/material to differentiate class characteristic comparisons from individualizing marks

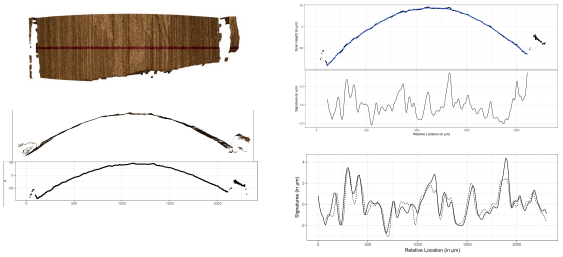
And even then, you can't generalize to new firearms

Issue 2: Examination is 100% Subjective

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

Subjectivity

Hare, Hofmann, and Carriquiry (2017) proposed a method for automated bullet matching



Numeric features derived from aligned signatures

Features used to train a random forest

Random forest votes used to assess similarity

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

Subjectivity

- Random Forest initially trained on data from the NIST Ballistics Toolmark Research Database
  - 2 sets of 35 bullets from the "Hamby" studies used for FTE training
  - Both sets use the same 10 consecutively rifled Ruger P-85 barrels
  - Digital scans with resolution of 1.5625 μm
- How well does the Hare, Hofmann, and Carriquiry algorithm generalize to other (similar) firearms?

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

Subjectivity

> Vanderplas, Nally, Klep, Cadevall, & Hofmann (2020) Comparison of three similarity scores for bullet LEA matching. Forensic Science International

- 3 different test sets
- different scan resolution (0.65 μm) than training data (1.5625 μm)
- different firearms (Ruger P-85, Ruger P-95, Ruger LCP)
  - Still Rugers - a brand known to mark well
- different types of ammunition

Subjectivity

Comparison of 3 different quantitative measures for bullet LEA matching:

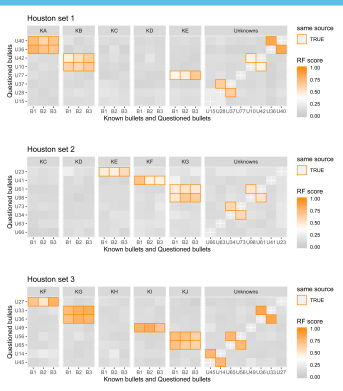
- Consecutive Matching Striae (CMS)
- Cross-correlation (CCF)
- Random forest score (RF)

Goals:

- Quantify error rates on external test sets
- Is the optimal cutoff (EER) stable across different firearms?

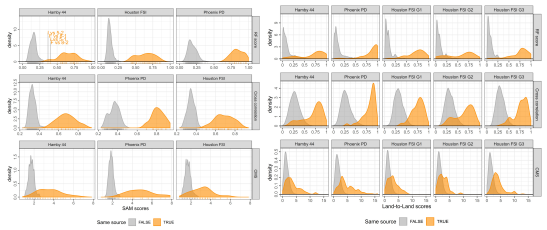
Hope to show that the RF score is better than other alternatives, and the same cutoff for RF score works across different sets

Subjectivity



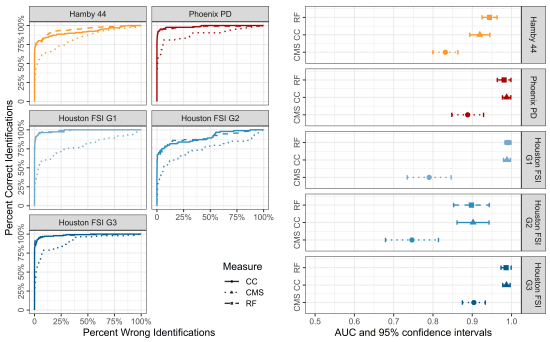
Future work: A paper comparing the matching algorithm's performance to examiner performance on the Houston FSC test sets.

Subjectivity



- Complete separation of bullet-to-bullet scores for both RF and CCF
- Consecutive Matching Striae are terrible at the land-to-land level and not great at the bullet-to-bullet level.
- The RF score has better separation on most land-to-land measures than CCF

Subjectivity



Forensic Software

- National Integrated Ballistic Information Network (NIBIN)
  - closed hardware
  - closed-source software
  - used by most forensics labs to identify matches in bullets and cartridges across jurisdictions

- x3p file format: OpenFMC (Open Forensic Metrology Consortium) ISO standard format
- x3ptools
  - R package for working with bullet and cartridge scans (or any other surface scan)
- bulletxtrctr
  - R package implementing the matching algorithm and feature extraction process for bullets

Issue 3: Screw-Ups (Error Rates)

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

Screw-Ups (Error Rate Estimates)

To be admitted in court, examiner testimony must pass the **Daubert standard** as codified in Rule 702 of Federal rules of evidence

- Relevance - the method is relevant to the evidence
- Reliability - the method rests on a reliable foundation
- Scientific Knowledge - the method is based in scientific methodology.

Important factors for scientific methodology:

- general acceptance by the community
- method has been through peer review and publication
- method can be tested
- the known or potential error rate is acceptable
- the research was conducted by unbiased individuals (e.g. the testing wasn't just for the specific court case)

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

Screw-Ups (Error Rate Estimates)

Ideal design:

- Varying numbers of same-source and different-source comparisons to prevent examiners from guessing or getting information from colleagues about test set composition
- Tests should have single pairs of evidence from one known source, with and one unknown for comparison
  - Ensures no additional information is available to examiners
  - Similar to most common casework scenario
- Examiners should use the same criteria for evaluating the evidence
- Examiners should not know they're being tested
- Sufficient comparisons and number of examiners to generalize well to the entire field

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

Inconclusives and Error Rates

	Examiner Decision		
Ground Truth	**Identification**	**Inconclusive**	**Elimination**
Same Source	a	b	c
Different Source	d	e	f

Options:

1. Condition on "Not Inconclusive" - c and d are errors, compare to a + d + c + f
2. Inconclusives are Correct (AFTE) - c and d are errors, compare to a + b + c + d + e + f
3. Inconclusives are Errors - b, c, d, e are errors, compare to a + b + c + d + e + f

Option 3 describes the error in the examination process; option 2 describes examiner error alone.

Error Rate Estimates The Good

Keisler et al. (2018) Isolated Pairs Research Study, AFTE Journal

- 9 Smith & Wessons
- 20 pairs of one known and one unknown cartridge
  - 12 same-source, 8 different-source
- 126 participants

	Examiner Decision		
Ground Truth	**Identification**	**Inconclusive**	**Elimination**
Same Source	1508	4	0
Different Source	0	203	805

- Fixed proportion of same-source comparisons
- Not blind
- Examiners used lab rules for classification (more variability)

Error Rate Estimates: The Good

Baldwin et al. (2014) A Study of False-Positive and False-Negative Error Rates in Cartridge Case Comparisons. Ames Laboratory report

- 25 Ruger SR9s
- Each participant evaluated 15 comparison sets of 3 knowns and 1 unknown
  - 5 same-source, 10 different-source
- 218 participants

	Examiner Decision		
Ground Truth	**Identification**	**Inconclusive**	**Elimination**
Same Source	1075	11	4
Different Source	22	737	1421

- Fixed proportion of same-source comparisons
- Not blind
- Examiners used lab rules for classification (more variability)

Error Rate Estimates: The Bad

Brundage-Hamby study: Hamby et al. (2019) A Worldwide Study of Bullets Fired From 10 Consecutively Rifled 9MM RUGER Pistol Barrels. Journal of Forensic Sciences.

- 10 consecutively manufactured Ruger P95 barrels
- Closed set study: 2 x 10 knowns, 15 unknowns
- 697 participants

	Examiner Decision		
Ground Truth	**Identification**	**Inconclusive**	**Elimination**
Same Source	10447	8	0
Different Source	0	?	?

- Focus is only on identification, not elimination
- Can't determine the total number of different source comparisons
- Closed set study
- Not blind

Error Rate Estimates: The Ugly

Lyons (2009) The Identification of Consecutively Manufactured Extractors, AFTE Journal.

	Examiner Decision		
Ground Truth	**Identification**	**Inconclusive**	**Elimination**
Same Source	174	1	3
Different Source	3	?	?

Results and Discussion

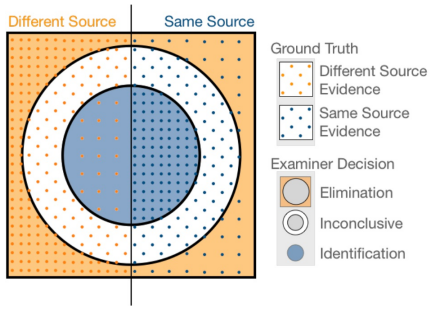
Some problems were discovered during the testing phase. These problems were not encountered during the pre-test phase. These problems ranged from the construction of the test set to the wording on the instruction sheet.

The first problem encountered was on test set number 4. The associated answer sheet was returned with ten identifications rather than the expected twelve. There was no indication of any inconclusive results on the answer sheet. When the submitting examiner was contacted it was explained that the examiner believed only ten identifications were requested. A review of the instruction sheet indicated that there was at least one unknown for each known, meaning that some knowns may have more than one unknown associated with it. The examiner understood this to mean there was only one unknown for each known and that the remaining two casings were from extractors other than the consecutively manufactured ones.

The answer sheet was returned to the examiner without the examiner knowing if the ten identifications submitted were correct. The examiner was asked to include the remaining two casings on the answer sheet, if in fact identifications could be made. The examiner completed the remaining examinations and re-submitted the answer sheet reflecting twelve identifications. These two additional identifications, as well as the first ten, were all correct.

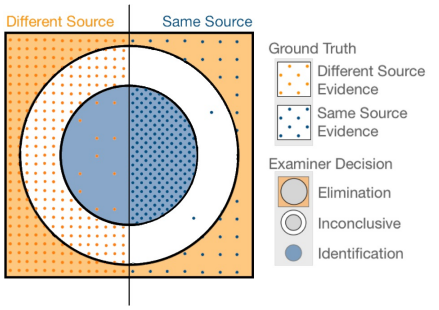
Inconclusive Problems

If inconclusives are a category that is necessary to describe situations where the bullet did not mark well, $P(SS|\text{Inconclusive}) \approx P(SS)$



Inconclusive Problems

Instead, what we see is more like this:



Ground Truth
- Different Source Evidence
- Same Source Evidence

Examiner Decision
- Elimination
- Inconclusive
- Identification

Across multiple studies,

- P(SS | Inconclusive) ≈ 0.02
- P(DS | Inconclusive) ≈ 0.98

Reasons:

- Lab rules
- Bias towards prosecution
- ???

Error Rates and Testimony

- Courts are disallowing identifications in jurisdictions across the country
- Language used for identifications is an issue because of the scientific foundations
  - we can't be sure that things came from the same gun
  - with good experiments, examiners can report P(SS|Identification), which is generally > 99%
- Reporting error rates as conditional on the examiner's decision provides more relevant information
  - $P(\text{Different Source}|\text{Identification})$, $P(\text{Same Source}|\text{Elimination})$

Error Rates and Testimony

- The "process error" that includes inconclusives is a better description of the overall error rate in court
  - Use examiner error for proficiency testing/evaluation
- Inconclusives (should) imply that there is no change from prior belief to posterior belief...
  instead, they're much more likely to be different source.
- Moving away from subjective evaluation removes this bias
  - Input features mimic the features examiners use
  - Numerical score provides jury with more information
  - Input features aren't biased towards the prosecution or the defense

Future Work

- Examine random forest scores compared with examiner conclusions for the same comparisons
- Refine the random forest model and refit with a more diverse set of data
- Work with examiners and laboratories to get 3D microscopes into labs and start using the random forest score in case work
- Conduct eye-tracking studies to see which features examiners use to make a conclusion; use this information to augment the feature space in the RF model

References

- Baldwin, D. P., Bajic, S. J., Morris, M., & Zamzow, D. (2014). A Study of False-Positive and False-Negative Error Rates in Cartridge Case Comparisons: Defense Technical Information Center. https://doi.org/10.21236/ADA611807
- Hamby, J. E., Brundage, D. J., Petraco, N. D. K., & Thorpe, J. W. (2019). A Worldwide Study of Bullets Fired From 10 Consecutively Rifled 9MM RUGER Pistol Barrels???Analysis of Examiner Error Rate. Journal of Forensic Sciences, 64(2), 551 - 557. https://doi.org/10.1111/1556-4029.13916
- Hare, E., Hofmann, H., Carriquiry, A., & others. (2017). Automatic matching of bullet land impressions. The Annals of Applied Statistics, 11(4), 2332 - 2356.
- Keisler, M. A., Hartman, S., & Kil, A. (2018). Isolated Pairs Research Study. AFTE Journal, 50(1), 56 - 58.
- Lyons, D. (2009). The Identification of Consecutively Manufactured Extractors. AFTE Journal, 41(3), 246??? 256.
- Spiegelman, C., & Tobin, W. A. (2013). Analysis of experiments in forensic firearms/toolmarks practice offered as support for low rates of practice error and claims of inferential certainty. Law, Probability and Risk, 12(2), 115 - 133. https://doi.org/10.1093/lpr/mgs028
- Vanderplas, S., Nally, M., Klep, T., Cadevall, C., & Hofmann, H. (2020). Comparison of three similarity scores for bullet LEA matching. Forensic Science International. https://doi.org/10.1016/j.forsciint.2020.110167