Toward an objective algorithm to compare bullets and cartridge cases

Alicia Carriquiry, Heike Hofmann

csafe
Center for Statistics and Applications in Forensic Evidence

Outline

- Uncertainty in forensic evaluations.
- The prosecutor's fallacy and the likelihood ratio.
- Unique challenges associated with pattern evidence.
- Current practice in toolmark and firearm comparisons.
- From subjective to objective comparisons.
- What still needs to be resolved.

Justice has failed an unknown number of individuals

- In the last two decades, the cases of hundreds of persons who spent years in jail for crimes they did not commit have come to light.
- The Innocence Project estimates that lack of scientific validity of many forensic tools, and exaggerated claims by forensic examiners is the second cause for this miscarriage of justice. (First cause is mistaken eye witness identifications.)
- **Lack of scientific validity**: Many forensic tools have no scientific underpinning or have not been evaluated in controlled, well-designed experiments.
- **Insufficient data**: Datasets for evaluation have been small, non-representative, or both.
- **Over-statement of significance of a "match"**: Forensic experts have routinely over-stated the significance of their findings, using terms such as *scientific reliability*, *practical certainty*, etc., that convey to juries a degree of precision that is not justified by the available science.

Is there science in forensic *science*?

- With the exception of DNA, a long list of National Academy of Sciences reports have found that science is missing in forensics.
- A 2009 report entitled *Strengthening Forensic Sciences in the United States: A Path Forward* found that most forensic tools lack scientific foundation, are plagued with subjectivity and are subject to unknown error rates.
- Junk science is routinely admitted in Court as "expert witnesses" are unchallenged by lawyers and even judges who do not have the background to decide what is good science.
- Much of this unsupported "science" is offered by the forensic practitioners themselves who often view their role as one of support for the prosecution.
- Defendants without the means to hire their own experts are at a clear disadvantage in the US criminal justice system.

Uncertainty in forensic evaluations

- Uncertainty is inevitable in science, the interpretation of results of scientific inquiry, and in any decision-making following scientific activity.
- **Measurement uncertainty:** Uncertainty creeps into measurement processes in different ways.
    - Measurement accuracy of the instrument used to collect the data. Typically well understood.
    - Variability due to operators, instrument, measurement occasion. Understood less well and rarely taken into account in reporting.

Uncertainty in forensic evaluations (cont'd)

- **Uncertainties around conclusions of forensic analyses:** Forensic scientists are often asked to compare two or more items, estimate the time of an event, decide whether two samples are *indistinguishable*.
- What is the chance that we conclude that two items "match" when in reality they have a different origin? E.g.,
  - Was a bullet fired from a putative gun (specific source question).
  - Were two bullets fired from the same gun (common source question).
  - What is the "confidence interval" around an estimated time of death.
- These are (arguably) purely statistical questions.
- When people talk about *error rates*, these are the questions they have in mind.

Uncertainty in forensic evaluations (cont'd)

- **Uncertainty about the probative value of evidence:** If two pieces of evidence are *indistinguishable*, what does it mean?
- This question is in the fact finder's portfolio, but forensic scientists often address it, as in "this gun, and no other, fired this bullet".
- The most important and hardest question.
- A related question: what is the chance that these two samples would be indistinguishable *even if they have a different source*?
- This is what is known as a *random match probability* in the DNA world and *probability of a coincidental match* everywhere else.

What is current practice in forensics?

- During trial, a competent forensic examiner will:
  - Describe the analytical techniques used to process the evidence.
  - Provide an assessment of the error rate associated with the instrument or test.
  - Interpret the results of the analyses for the jury.
- Depending on type of evidence, the interpretation step roughly consists in declaring that the crime scene and suspect's samples **match** or **do not match**, or that the tests were **inconclusive**.
- There is a need to push the forensics community to talk about
  - *Degree* of the strength of the match.
  - *Probative value* of a match.

Current practice (cont'd)

- Most lay jurors equate a match with *same source*.
- **Except in the case of single donor DNA among non-relatives, a match does not imply same source.**
- This is a critically important concept, and one that has escaped the attention of the forensic and legal professions until recently.
- The concept of a **coincidental match** in evidence other than DNA has emerged as important only in the last 20 years or so.
- For the vast majority of evidence types, we do not know the probability of a coincidental match.

Probability of a coincidental match

- Loosely, defined as the probability of observing that two samples are indistinguishable even though they have a different source.
- Consider blood types (A,B,O) in the US. Blood found at the crime scene is type A+ and so is the suspect. We consider two scenarios:
    - The suspect left the sample at the crime scene and if so, we would expect a match.
    - The suspect was not the donor. What is the probability some some other random person might have left an A+ sample at the crime scene?
- The latter is the probability of a coincidental match and for this example, we know that this probability is about 0.35.
- The 0.35 value is derived from the known frequency of A+ persons in the US, which is $\sim$ 35.7%.

Prosecutor's fallacy

- Consider the following statement: if the suspect is the source of the sample at the crime scene, then the probability that we would observe a "match" is very high.
- For kicks, suppose that the probability of a match in that case is 0.90.
- In statisticalese, we write $\Pr(E|S) = 0.90$.
- The **prosecutor's fallacy** consists in INCORRECTLY reversing the probability and concluding that therefore, the probability that the samples have the same source given that they match is also 0.90.
- This is almost never true! Again in statisticalese,

$$\Pr(E|S) \neq \Pr(S|E)!!!$$

- To compute $\Pr(S|E)$ we need an additional probability: $\Pr(E|\bar{S})$, where $\bar{S}$ means "different source"'. This is the *probability of a coincidental match*.

The “weight of evidence” paradigm

- To move beyond the binary “match/non-match” framework, statisticians have proposed a one-step approach, that consists in evaluating the evidence under two competing hypotheses:

$$\begin{aligned} H_p &: \quad \text{The suspect is the source of the evidence} \\ H_d &: \quad \text{Someone else is the source of the evidence} \end{aligned}$$

- A **likelihood ratio** statistic (LR)

$$LR = \frac{\Pr(E|H_p)}{\Pr(E|H_d)}$$

can be computed to decide whether the evidence E supports the prosecutor’s or the defense’s hypothesis.

A simple example

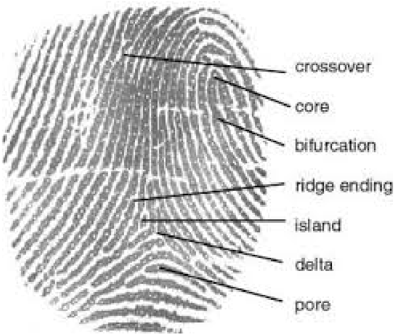
- A latent print is found at the crime scene.
- The arrangement of minutiae in the latent is indistinguishable from the suspect's print.
- How do we interpret this evidence?
  - If the suspect left the latent, we expect that $\Pr(E|S)$ will be close to 1.
  - If the suspect did not leave the latent, what do we know about $\Pr(E|\bar{S})$? Not much!
- We *assume* that fingerprints are unique to the individual, but in practice, smudged, partial latents may be indistinguishable from more than one reference print.
- If $\Pr(E|\bar{S})$ is, for example, 0.25, then the LR would be 4, meaning that it is only 4 times more likely that we would observe the same minutiae if the suspect happened to be the source of the latent.

Computing a LR

- In the absence of contamination or lab error, the numerator in the LR will be close to 1.
- Computation of the numerator only requires comparison of known and questioned samples.
- The denominator is much more challenging:
  - We need to define H_d.
  - For each potential H_d we need a statistical model to compute $P(E|H_d)$.
  - We have to have reference databases that are relevant for each plausible H_d.
- At this time, we can compute $P(E|H_d)$ only for single-donor DNA samples or for simple DNA mixtures.

The challenging area of pattern evidence

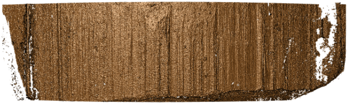
- Pattern evidence includes fingerprints, ballistics and other toolmarks, handwriting, shoe prints....
- The evidence typically consists of a 2D or a 3D image of the sample.



- No statistical models, no (or questionable) reference databases.

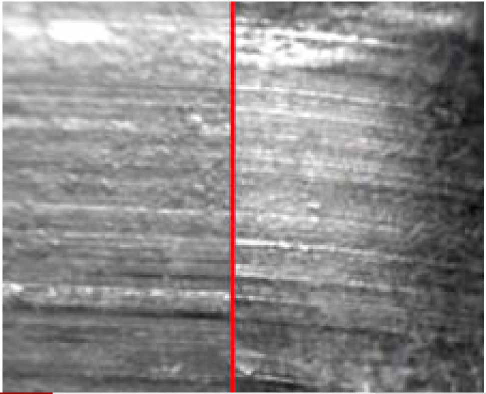
The case of bullet striae

- When bullets travel down a barrel after they are fired, the “rifling” of the barrel and manufacturing imperfections leave marks or striation on the bullet surface.



The art of ballistics

- By comparing striation in two bullets, firearms examiners attempt to determine whether a specific gun fired both bullets (identification) or at least whether the two bullets were fired by the same gun (same source).
- Current practice: “I know a match when I see one”.

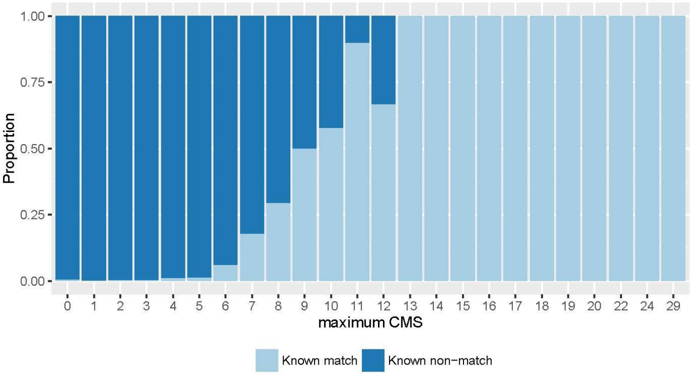


The AFTE Theory of Identification

...opinions of common origin to be made when the unique surface contours of two toolmarks are in *sufficient agreement*....Agreement is significant when it exceeds the best agreement demonstrated between toolmarks known to have been produced by different tools and is consistent with agreement demonstrated by toolmarks known to have been produced by the same tool................

An attempt at objectivity - CMS

- CMS: consecutively matching striae.
- Idea: many consecutively matching striae are indication of a "match"; a match suggests that the two bullets were fired by the same gun.
- But what is the threshold?

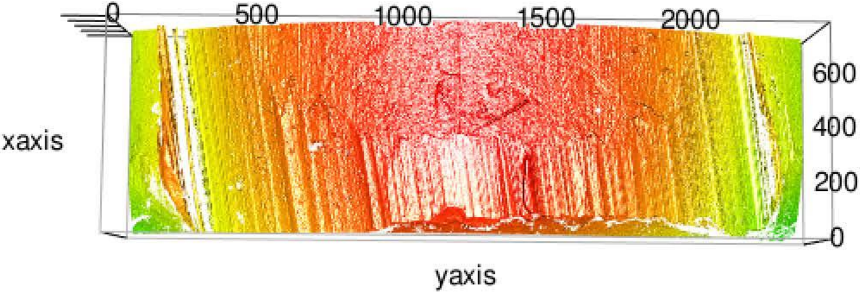


CSAFE's contribution

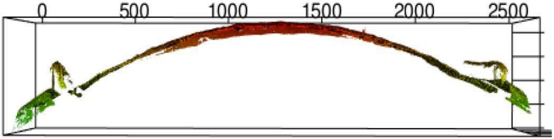
- The Center for Statistics and Applications in Forensic Evidence (CSAFE) is a NIST Center of Excellence established in July of 2015.
- We are a consortium of four universities: Iowa State (lead), Carnegie Mellon, Univ of Virginia, and Univ of California Irvine.
- Initial funding provided for five years – potential to request funding for a second five-year term.
- Three missions: research, outreach and training, with a focus on pattern and digital evidence.
- Main areas of research: firearms and toolmarks, shoe prints, hand writing, steganalysis, blood spatter, and also human factors.
- Training: for lawyers, judges, forensic practitioners, students in statistics and forensic sciences.
- Outreach: transfer knowledge to practitioners, receive feedback in terms of implementation issues.

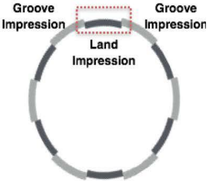
A more objective match criterion

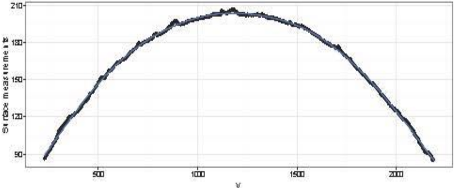
- Cutting edge: confocal 3D microscopy to capture surface topography of bullets.

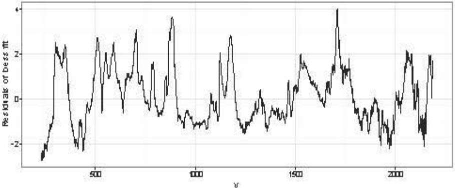


Extracting a signature



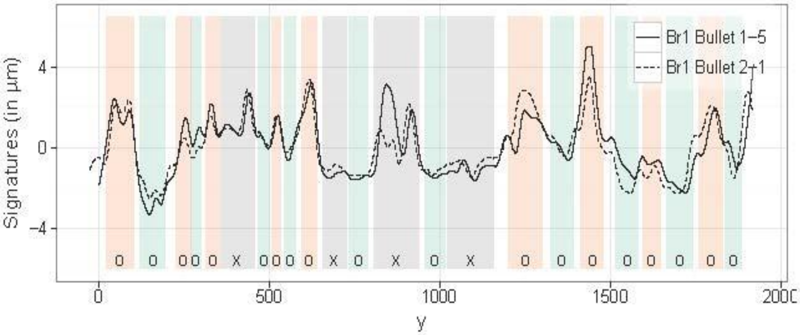






Comparing signatures

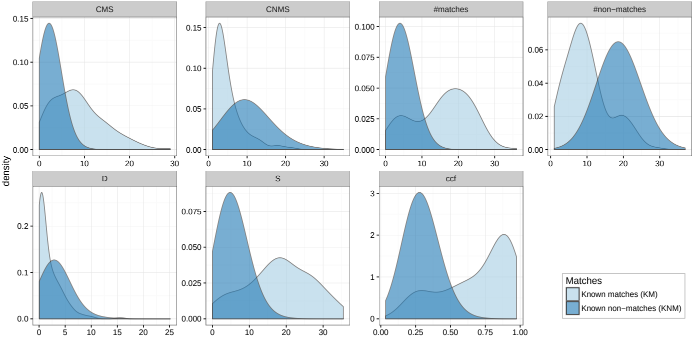
- Given two signatures, we can construct a “score” for the difference using machine learning methods.



Properties of a match criterion

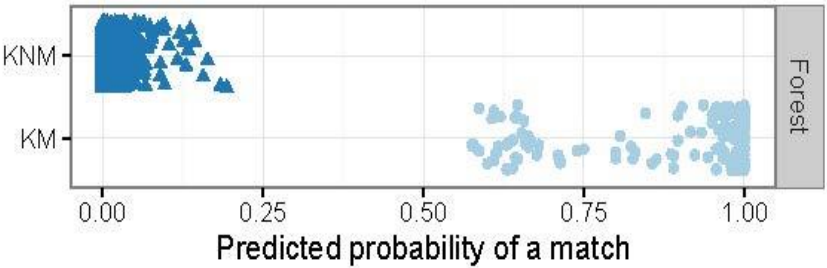
- Ideally:
    - High sensitivity or low probability of false positives
    - High specificity or low probability of false negatives.
- Need an experimental dataset with known matching pairs and known non-matching pairs to determine behavior of score.

Extracted features



Performance of score

- Features combined into score using random forests.
- Applied method to small (but challenging) dataset:

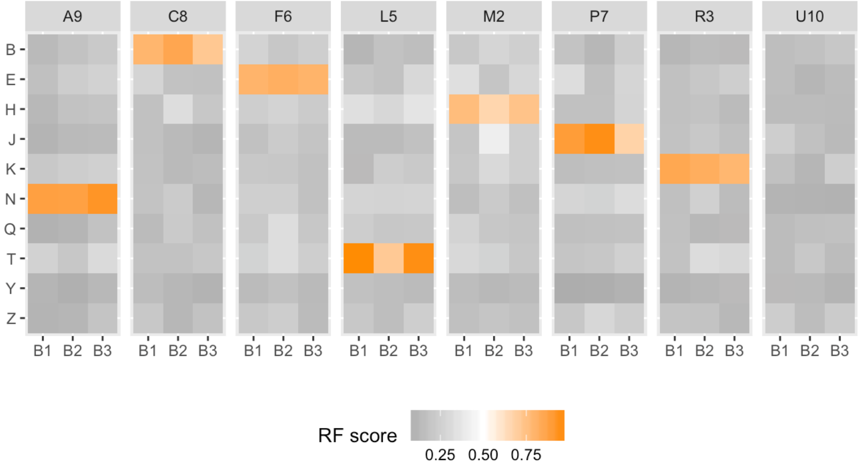


- Tested in several other datasets: no classification errors.

The Phoenix study

- Eight guns, three test shots per gun.
- Ten questioned bullets.
- Open set:
  - Some bullets may not have been fired by any of the 8 guns.
  - Some guns may not have fired any of the 10 questioned bullets.
- We used the model developed on the Hamby bullets to classify the Phoenix bullets.

Results from the Phoenix study



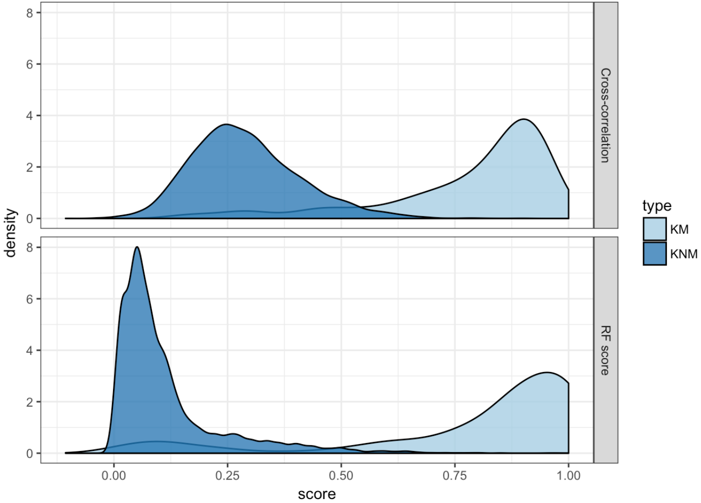
Very nice, but...

We still know nothing about the probative value of a match!

Distribution of scores under two hypotheses

- To decide whether a match between two bullets has probative value, we need to know what values of the score we can expect when comparing bullets fired from the same gun and bullets fired from different guns.
- To do so, we:
    - Assemble a LARGE database of pairs of bullets known to be fired by the same gun;
    - Assemble an EVEN LARGER database of pairs of **relevant** bullets known to be fired from different guns;
    - Compute the random forest score for each pair of known matches and known non-matches;
    - Build the distributions of RF scores among pairs of known matches and among pairs of known non-matches.

Scores for Phoenix dataset



Score-based LRs

- In principle, we could:
  - Fit densities to the empirical distributions of scores for matching pairs and for non-matching pairs.
  - Construct a likelihood ratio using the fitted densities $L(f(y)|H_p, \theta_p)$ and $L(f(y)|H_d, \theta_d)$, for y the vector of features, $f(.)$ the function that maps features into scores, and θ_p, θ_d the vectors indexing the likelihoods under the two competing hypothesis.
- **Problem:**

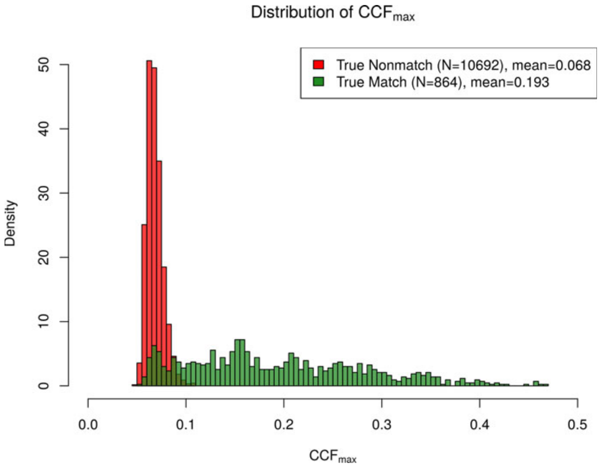
  $$LR = \frac{L(y)|H_p)}{L(y|H_d)} \neq LR_f = \frac{L(f(y)|H_p)}{L(f(y)|H_d)},$$

  except for trivial $f(.)$. In fact, LR and LR_f are typically not even proportional to each other.
- Still and open problem.

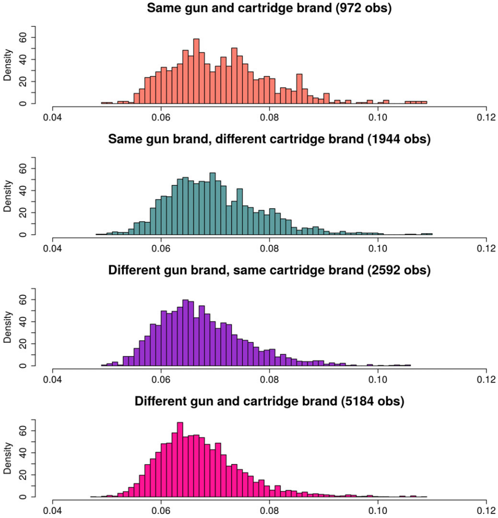
An algorithm for breech face comparison

- Our partners at CMU have developed an algorithm to compare breech face impressions using 2D images.
- Different “mathematics”, same idea.
- Promising results: their algorithm does better than anything else available (ours for bullets does as well!),

Distribution of breech face comparison scores



Choices for non-matching distribution



The work ahead of us

- The Obama administration was aware that the scientific and statistical bases of most forensic disciplines must be shored up.
- The new administration has radically different priorities, so we do not know whether federal support will be continued.
- We are confident that we have an objective, fully automated approach for firearm examination.
- But we still need to:
  - Collect ENORMOUS amounts of 3D images of bullets – and also of breech face impressions.
  - Validate our methods on large datasets.
  - Write user-friendly software.
  - Engage practitioners and transfer knowledge.
- We are hopeful about achieving similar progress in other forensic disciplines.

THANK YOU FOR YOUR ATTENTION

csafe

Center for Statistics and
Applications in Forensic Evidence

alicia@iastate.edu