A PRIMER ON ERROR RATES IN FINGERPRINT EXAMINATION

Authors:
Heidi Eldridge, MS, RTI International and Professor Christophe Champod, PhD, Université de Lausanne, Lausanne, Switzerland

With a grant funded by the U.S. National Institute of Justice 2017-DN-BX-0170

Published May 2020

A Primer on Error Rates in Fingerprint Examination

The aim of this primer is to provide assistance to the Judiciary, legal professionals, and lay persons in understanding error rates in fingerprint examination. The document aims at defining the types of errors that may be presented in courts and explaining their meaning and limitations, both to assess the discipline as a whole or assess the proficiency of a specific expert in a given case. At the end of the document, it is illustrated by an example of a brief conversation between a member of the Judiciary and a fingerprint expert.

Is fingerprint comparison free from errors?

Fingerprint comparison (e.g. comparing a mark recovered from a crime scene to a set of prints from a person of interest) is carried out by fingerprint examiners mostly visually. They will ultimately make a decision regarding the identity of source based on their training and experience. The decision is entirely left to the examiner. Obviously, any human endeavour involving decisions is prone to potential errors.

Through fingerprint comparison, two types of error attract most of the attention:

(1) The decision to associate a mark to a person who was not, in fact, the source. This is an erroneous identification.

(2) The decision to exclude a person from being the source of a mark, whereas that person was, in fact, the source. This is an erroneous exclusion.

Scientists will often refer to the first type of error as a "false positive" and to the second as a "false negative".

There is anecdotal evidence of cases of wrong decisions that occurred during casework and were ultimately discovered. For erroneous identifications, fewer than 50 cases have been detected and documented worldwide.

It has also been reported following controlled studies involving fingerprint examiners, that both types of error can occur. These studies are said to be "controlled" in the sense that the participants were given marks and prints of known ground truth to compare. This means that the administrators of the study had selected marks and prints of known and undisputable sources.

The two most publicised controlled studies are the following:

(1) A study by Ulery and colleagues[1] published in the Proceedings of the National Academy of Sciences in 2011 on the comparison of fingerprints only. They reported a false positive rate of about 0.1% (1 in a thousand) and a false negative rate of about 7.5%.

(2) A 2014 study by Pacheco and colleagues from the Miami-Dade police funded by the National Institute of Justice[2] for which the false positive rate can be estimated[3] to be 0.9% with a false negative rate of 7.5%. This study included both finger and palm impressions and did not separate the two when calculating error rates.

For the rest of the discussion we will report on the error rates obtained in a controlled study devoted to palm prints and the meaning and range of application of the error rates obtained.

[1] Ulery BT, Hicklin RA, Buscaglia J, Roberts MA. Accuracy and Reliability of Forensic Latent Fingerprint Decisions. Proceedings of the National Academy of Sciences, USA. 2011;108(19):7733-8.

[2] Pacheco I, Cerchiai B, Stoiloff S. Miami-Dade Research Study for the Reliability of the ACE-V Process: Accuracy & Precision in Latent Fingerprint Examinations. Washington DC: National Institute of Justice; 2014, https://www.ncjrs.gov/pdffiles1/nij/grants/248534.pdf.

[3] Ausdemore MA, Hendricks JH, Neumann C. Review of Several False Positive Error Rate Estimates for Latent Fingerprint Examination Proposed Based on the 2014 Miami-Dade Police Department Study [with discussion]. Journal of Forensic Identification. 2019;69(1):59-119.

Palm prints: What is the error rate of the discipline?

The palm study error rates

The palm print study comprised of a set of 526 trials (palm marks and prints) of known sources that have been submitted to 226 trained examiners[4]. Each case was assessed by about 23 examiners, and each examiner assessed an average of 37 trials.

In total 7,620 decisions of identification or exclusion were made by these examiners with the following performance:

- In 1,785 submissions where the ground truth was that the mark **did not originate** from the person who provided the prints, 12 erroneous identifications were declared along with 1,773 correct decisions. This corresponds to **a false positive rate of 0.7%** (12/1,785).
- In 5,835 submissions where the ground truth was that the mark and the print under examination **shared the same** source, 552 wrong exclusions were declared along with 5,283 correct exclusions. This corresponds to a **false negative rate of 9.5%** (552/5,835).

Do these error rates reflect the performance of the whole fingerprint profession?

The study error rates are intrinsically linked with the study only. Do these rates allow us to infer the performance of the fingerprint profession at large? Yes, they will, but only if the study design can demonstrate that some assumptions have been fulfilled, such as the following:

- The study has used marks and prints that are of similar complexity and quality as the images used in casework.
- The study participants are representative of the examiners working in cases in operational laboratories.
- The participants carried out their examinations using the same protocol and dedication as in casework.

The study design can provide some indications on the above assumptions, but there will always some uncertainty. Typically, as with this study, examiners were willingly and knowingly participating, and it is difficult to assure that they carried out their activities as they would do in casework, without any bias caused by the study itself. This means that the error rates have to be taken as a ballpark figure. Even if we could process them with numerical accuracy, we should not have the illusion that they exactly reflect the practicing profession as a whole. For the palm study, these rates are adequate as ballpark figures; we could say for example that we have empirical data to suggest that it reasonable to claim that the profession has to allow for a **false positive rate on the order of 1% and a false negative rate on the order of 10%**. The use of the expression "on the order of" is made to clearly indicate that these orders of magnitude have been indeed informed by the palm study results, but some care should be exercised when generalising to the whole profession.

Shall we use confidence intervals associated with these rates?

The above error rates have been obtained by making a ratio of the trials with erroneous conclusions divided by the total number of trials. For example, the false positive rate of 0.7% is obtained by dividing 12 false identification over 1,785 trials. If the study had been more limited in size, we could have

[4] All results can be consulted at https://cchampod.shinyapps.io/Results_BBStudy/

obtained the same rate with 1 false identification over 142 trials. These are the same rates, but the confidence in them will vary. We would give more trust to the first rate because it is based on a larger study. The larger the sample size, the more confident we are that the observed rate is not the result of a sampling effect and truly reflects the rate of the event in the studied population.

For these reasons, statisticians have developed methods to qualify our trust in the observed rates, so that we can account for the difference in sample size. They will compute what are called confidence or credible intervals around them. The interval will give a lower and an upper bound. There is a range of techniques to compute these intervals, some requiring more statistical assumptions than others. The size of the intervals will depend on the confidence required (typically 95% or 99%). When the total number of trials is large, such as in this study, and the rates are not too small, all methods will give similar intervals[5].

In this palm study we can compute the following credible intervals:

	Number of Errors	Number of Trials	Rate (in %)	95% Lower Bound	95% Upper Bound
False Positive	12	1,785	0.0067 (0.67%)	0.004 (0.4%)	0.011 (1.1%)
False Negative	552	5,835	0.0946 (9.46%)	0.087 (8.7%)	0.102 (10.2%)

If we focused on the false positive, these credible intervals allow one to conclude that the false positive rate lies with 95% probability between 0.4% and 1.1%. Conversely for the false negative rate, there is a 95% probability for it to lie between 8.7% and 10.2%. These upper bounds are in the order of magnitude that we quoted before. Had these upper bounds been largely above our orders of magnitude, we should have reassessed our orders of magnitude.

Shall we quote these intervals in court?

Our position is that they should be computed as they reflect on the statistical robustness of the rates computed based on the study data. We don't think they should be quoted systematically as we would prefer adopting the order of magnitude approach we mentioned before. Computing credible intervals is part of a good scientific practice that enables us to define the orders of magnitude we would ultimately provide. Needless to say, the underpinning data and statistical treatment should be disclosed upon request, but to communicate about the rates themselves, they are not critical in our opinion.

What shall be the use of these error rates in casework?

We advise to use them as part of the debate regarding the admissibility of the field as an area of expertise. In that context, the error rates bring objective data regarding the general reliability of the method (here the comparison of palm prints). Alongside other factors that judges must weigh to decide on the admissibility of scientific evidence, these orders of magnitude provide tangible evidence of the reliability of the discipline.

As we will see in the next question, these error rates do not tell much about the probability of an error in a particular case. In other words, error rates help more in assessing admissibility and less in assessing the weight of evidence in a specific case.

[5] All computations can be carried out on https://cchampod.shinyapps.io/app_CI/

Palm prints: What is the error rate of the examiner in this case?

We shall take one of the best performing examiners in the study. She carried out 43 trials (37 from same source and 6 from different sources) without any errors.

It is important to know that this examiner committed no errors, but when it comes to error rates, we would not claim it to be zero. This is because we have only 43 trials for this examiner. Using the same logic as before, we can compute the credible intervals for that examiner:

	Number of Errors	Number of Trials	Rate (in %)	95% Lower Bound	95% Upper Bound
False Positive	0	37	0 (0%)	0 (0%)	0.076 (7.6%)
False Negative	0	6	0 (0%)	0 (0%)	0.348 (34.8%)

We can see the limit of the statistical exercise. Indeed, these numbers are correct, but the upper bound for the false positive rate of 7.6% has limited relevance in assessing the quality of the expert testifying. If such a rate were to be trusted it would mean that if this examiner were to conduct 100 comparison a week, we would expect on average 7 to 8 wrong identifications every week!

The factual information that she made no errors on that study is informative enough. The very large upper bounds are simply due to the limited number of trials (37 and 6).

Our position is that error rates and their upper bounds computed on a limited number of trials is not helpful to reflect the proficiency of an examiner who has carried out thousands of cases in her career including a fair number of annual proficiency tests[6].

Without dismissing the excellent performance of that examiner in the palm study (and the questions we may legitimately ask the examiners who made wrong identifications), the inquiry into the reliability of the expert should rely more on other information in relation to her training, experience, and demonstration of proficiency. For example, the information provided by this examiner in the palm study[7] indicates that she has been practicing for more than 10 years in latent print examination, she is a certified examiner of the International Association for Identification (IAI), she is annually subjected to a proficiency test, and she testified in court in the last year. She declared that she has never made a wrong identification in casework in her career. She had wrongly missed identifications in the past, but these cases were caught by the verification process in place in her agency.

The point we are trying to make here is that more than her good performance in the palm study, it is the disclosure of her qualifications and evidence of her good practice – including the quality management system in place in the agency – that testify to her credibility as an expert. Her personal upper bound false positive rate of 7.6% has limited bearing here.

6 There are ways to compute credible intervals allowing for some prior knowledge of the performance of the examiner, typically in past proficiency tests. This method is also available in https://cchampod.shinyapps.io/app_CI/. However, it transforms the endeavor into a statistical exercise, whereas we believe that indication of reliability should be searched elsewhere.

7 That examiner is user-0309 in the palm study.