

Statistical Issues in Assessing Forensic Evidence

Karen Kafadar^a

April 21, 2011

Technical Report 11-01
Department of Statistics
Indiana University
Bloomington, IN 47405

^aDepartments of Statistics, Indiana University. kkafadar@indiana.edu

Statistical Issues in Assessing Forensic Evidence

Karen Kafadar
Department of Statistics
Indiana University
Bloomington, Indiana 47408-3825

Abstract

The National Academy of Science released its report, *Stengthening Forensic Science in the United States: A Path Forward* (NRC 2009). An important finding was the increased need for scientific research in the evaluation of methods used in forensic science, such as bias quantification, validation, and estimates of accuracy and precision in different contexts. This article illustrates, using medical clinical trials and two fingerprint studies, the value of applying statistical methods in the design of studies that are needed to evaluate inferences from forensic evidence. Because many sources can affect both the accuracy and the consistency of decisions at each stage of the process, from specimen collection to final decision, this article discusses methods for identifying these sources, as well as the statistical principles involved in the quantification of, and uncertainty in, measured error rates. By contrasting the design of medical trials with two previous fingerprint studies and with bullet lead studies, this article emphasizes the need for reduced subjectivity, the types of measurements on physical evidence that can lead to more accurate and consistent decisions, and the importance of carefully designed studies in the evaluation of forensic evidence.

Key words: designed experiment, variability, bias, error rates, sensitivity, specificity, positive/negative predictive value, confidence interval, fingerprints, DNA evidence

1 Introduction

“With the exception of nuclear DNA analysis, however, no forensic method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source.”

— *Stengthening Forensic Science in the United States: A Path Forward*, p.7.

Probably no other sentence in the entire report released by the National Academy of Science (hereafter, “NAS report”; NRC 2009) has received as much attention, and as much of a reaction, as this one sentence describing the state of the scientific rigor for the methods currently in practice in forensic science. Fingerprint evidence, and, to a lesser extent, analyses of hair, pattern, or markings (e.g., on weapons), have been advocated as methods that can “uniquely identify” an individual. While some of these methods may be reasonably accurate in narrowing down the specific *class* of individuals that are consistent with the evidence, the inclusion of fingerprint evidence in that statement came as a huge shock to many people. Fingerprint analysis had been assumed for so long to be capable of identifying a unique *individual* that its validity and reliability had never been questioned, so the need for studies of its validity and reliability appeared to be unnecessary. Only because examiners claimed “zero error rate” in fingerprint analyses (Cole 2004, Zabell 2005) and when mistakes surfaced (e.g., Stoney 2004 on the Mayfield case) did scientists and others question the validity of fingerprint analysis. The Committee on Identifying the Needs of the Forensic Science System (hereafter, “Committee”) that authored the NAS Report found that the published studies of the accuracy and reliability of most forensic methods, other than nuclear DNA analysis, failed to meet the stringent criteria seen in other scientific disciplines, leading to the conclusion above.

As noted in Chapter 3 of the NAS Report, Courts’ attempts to apply scientific principles to the evaluation and consideration (admission or exclusion) of forensic evidence, such as the Daubert criteria (roughly, admission allowed if methods on which evidence was obtained are “testable”, “peer-reviewed,” have “known error rate[s],” and are “generally accepted in scientific community”) have failed. The judicial system is not the proper forum to establish scientific standards. Rather, the scientific laboratory is well positioned to design, execute, and evaluate the results from well-conceived scientific studies that will quantify accuracy (bias, validity) and precision (reliability, consistency) of forensic evidence, as well as reveal shortcomings that can be addressed, and ultimately strengthen, the value of the evidence.

This article describes considerations in the design of such studies, with a particular focus on latent fingerprint analysis. Section 2 defines the metrics for assessment (e.g., sensitivity, specificity), including a statistical calculation concerning numbers of features to assure identification with a high degree of confidence. Section 3 compares the situation between DNA and fingerprint analyses. Section 4 presents an example of a well designed study from the medical arena that was conducted to settle an important public health concern in 1954. Section 5 discusses this example in the context of latent fingerprint analysis, and Section 6 illustrates, via bullet lead evidence, the consequences of ignoring the principles of good study design. Section 7 concludes with some final comments encouraging research in these directions.

2 Statistical measures of performance

Numerous articles have appeared on various types of forensic evidence, including the process of collecting the evidence, the methodology that is applied, and, when possible, the description of drawing inferences from the evidence. (Chapter 5 of the NAS report reviewed the state of the scientific knowledge for many of these types of evidence.) Many fewer articles present the results of carefully designed studies to quantify accuracy, precision, sensitivity, and specificity in these methods. In this section, we review the definitions of key considerations in the evaluation of any scientific process, namely *consistency*, *reliability*, *sensitivity*, *specificity*, and *positive/negative predictive value*.

1. **Validity** (accuracy): Given a sample piece of evidence on which a measurement is made, is the measurement accurate? That is, if the measurement is “angle of bifurcation,” or “number of matching features,” does that measurement yield the correct answer? For example, if a bifurcation appears on an image with an angle of 30 degrees, does the measurement technology render a result of “30 degrees”, at least on average if several measurements are made? As another example, if a hair diameter is 153 micrometers, will the measurement, or average of several measurements, indicate “153”?
2. **Consistency** (reliability): Given the same sample, how consistent (or variable) are the results? If the measurement is repeated under different conditions (e.g., different fingers; different examiners; different analysis times, different measurement systems; different levels of quality in evidence), is the measurement the same? (Almost surely not; see Dror and Charlton 2006 for a small study of five examiners.) Under what conditions are the measurements most variable? That is, do measurements vary most with different levels of latent print quality? Or with different fingers of the same person? Or with different times of day for the same examiner? Or with different measurement (e.g., AFIS) systems? Or with different examiners? If it is determined that measurements are most consistent when the latent print quality is high and when AFIS system type A is used, but results vary greatly among examiners when the latent print quality is low or when other AFIS systems are used, then one would be in a good position to recommend the restriction of this particular type of forensic evidence under only those conditions when consistency can be assured. (Note that this consideration requires some objective measure of quality. For a measure of quality when imaging biological cells, see Peskin et al. 2010.) Notice that a measurement can be highly consistent around the wrong answer (consistent but inaccurate). For example, if an air traffic controller directs a pilot to “contact ground control at 121.7,” the pilot’s instrument must be both accurate (i.e., able to tune to 121.7, not 121.6 or 121.8) and precise (i.e., consistently reach 121.7).

3. **Well-determined error rates:** If the true condition is known, what is the probability that the measurement technology will return the correct answer? These error rates depend on whether the “condition” is “present” or “absent”, so two terms are used to describe error rates. For purposes of illustration, suppose that “condition is present” refers to the situation where an exact match is known; e.g., two different prints (latent and rolled, or two latent) are made on the same person. Conversely, suppose that “condition is absent” means that the two prints are known to have come from different individuals.

- *Sensitivity:* If, unbeknownst to the examiner, the two prints “match” (“condition is present”), what is the probability that the method of analysis yields a result of “match”? This probability is called *Sensitivity*:

$$\textit{Sensitivity} = P\{\text{analysis claims “match”} \mid \text{true match}\}$$

The opposite of *Sensitivity* is *False Negative Rate (FNR)*:

$$\textit{False Negative Rate (FNR)} = P\{\text{analysis claims “no match”} \mid \text{true match}\}.$$

- *Specificity:* If, again unbeknownst to the examiner, the latent print and the 10-print card do not come from the same individual (true non-match, or “condition is absent”), what is the probability that the analysis yields a correct result of “NO match”?

$$\textit{Specificity} = P\{\text{analysis claims “NO match”} \mid \text{true NON-match}\}$$

The opposite of *Specificity* is *False Positive Rate (FPR)*:

$$\textit{False Positive Rate (FPR)} = P\{\text{analysis claims “match”} \mid \text{true NON-match}\}.$$

The error rates *FPR* and *FNR* for a method can be estimated only from a designed experiment where the experiment designer (*not* the examiner nor the exam administrator) knows whether the presented prints match or do not match. The *uncertainty* in these estimates depends on the sample size; see point #6 below.

4. **Positive Predictive Value (PPV):** In the courtroom, one does not have the “true” answer; one has only the results of the forensic analysis. The question for the jury to decide is: *Given the results of the analysis, what is the probability that the condition is present or absent?* For fingerprint analysis, one might phrase this question as follows:

$$\textbf{Positive Predictive Value PPV} = P\{\text{true match} \mid \text{analysis claims “match”}\}$$

If PPV is high, and if the test result indicates “match,” then we have some reasonable confidence that the two prints really did come from the same person. But if PPV is low, then,

despite the test result (“match”), there may be an unacceptably high chance that in fact the prints did not come from the same person – i.e., we have made a serious “Type I error” in claiming a “match” when in fact the prints came from different persons. The opposite of the PPV is the probability of a false positive call: given that the analysis claimed “match,” what is the probability that in fact the specimens do **not** match?

$$\text{False Positive Call FPC} = P\{\text{true NON-match} \mid \text{analysis claims “match”}\}$$

(The quantity above is related to the “false discovery rate,” a term coined by Benjamini and Hochberg in 1995 in the context of multiple hypothesis tests; see Appendix.)

5. **Negative Predictive Value (NPV):** Conversely, the test should also correctly identify non-matches if in fact the two prints came from different sources. This aspect is called *Negative Predictive Value*:

$$\text{Negative Predictive value NPV} = P\{\text{true NON-match} \mid \text{analysis claims “NO match”}\}$$

If NPV is high, and if the analysis indicates “no match,” then we have some assurance (given by the probability) that the two prints really did come from different people. But if NPV is low, then, despite the analysis results (“no match”), the probability may be high that in fact the prints really arose from the same person; i.e., the analysis has resulted in a “Type II error” in claiming a “non-match” when in fact the prints came from the same person (freeing a potentially guilty person). The opposite of the NPV is the probability of a false negative call: given that the analysis claimed “no match,” what is the probability that in fact the specimens really matched?

$$\text{False Negative Call FNC} = P\{\text{true match} \mid \text{analysis claims “no match”}\}$$

(The quantity above is related to the “false non-discovery rate,” in analogy with Benjamini and Hochberg’s “false discovery rate”; see Genovese and Wasserman 2004.)

PPV and NPV, and hence the probabilities of false positive and false negative calls, cannot be determined on real-life cases because the “true” answer in such cases is unknown. Sensitivity and specificity can be *estimated* from *realistic* scenarios in which an administrator arranges test scenarios of print pairs that truly “match” and truly “do not match.” Bayes’ formula (e.g., Snedecor and Cochran 1972) provides the connection between PPV/NPV and sensitivity/specificity (see Appendix). PPV and NPV also depend upon the *prevalence* of the event; i.e., are we looking for 1 suspect out of 1 billion, or 1 suspect out of 10,000? The consequence of

this formula is that **low** prevalence, **high** sensitivity, and **high** specificity are needed for **high** PPV and NPV; i.e., for high probabilities of **correct** decisions; hence the need for sensible restriction of the suspect population, as well as highly reliable and accurate tests. Moreover, sensitivity and specificity, and hence probabilities of correct decisions given the evidence, *may well depend on various factors*, such as examiner, quality of evidence, measurement system, etc. For example, sensitivity and specificity may be lower for examiners with less experience. How do sensitivity and specificity vary with years of experience? with different levels of quality of the latent print? with different AFIS systems? Without this information, we cannot assess the probabilities of correct decisions (PPV, NPV).

6. **Uncertainties in estimates:** As indicated above, a well-designed study can provide *estimates* of sensitivity and specificity, from which estimates of PPV and NPV, and hence error rates, can be derived. But these estimates will be subject to uncertainty, because they will be based on a sample and hence the information is limited by the sample size. An unappreciated fact is that the upper 95% confidence limit on a proportion based on N tests that resulted in zero false positive calls is *not* zero but is roughly $3/N$. So, for example, if an analyst is presented with 50 print pairs, some of which are true matches and some true non-matches, and makes the correct calls on all of them (zero errors), the upper bound on the true probability of false calls is roughly $3/50$, or 6%; that is, probabilities of less than 6% are consistent with having observed 0 errors out of 50 trials, while probabilities greater than 6% would be inconsistent with having observed 0 errors out of 50 trials. Had 1 mistake out of 50 occurred, the upper 95% confidence bound would have been roughly $4.7/N$, or 9.4%. If 2 errors were called, the upper 95% confidence limit would have been roughly $6.2/N$, or 12.4%.

The committee was unable to find comprehensive studies of latent fingerprint analysis that addressed all of these issues with high levels of confidence.

3 DNA versus Latent Fingerprint Analysis

Latent fingerprint analysis has enjoyed the reputation of being reliable evidence of individualization; i.e., is capable of identifying not just a class of individuals, but in fact the specific individual, both with high reliability and with high confidence (in fact, some have stated that latent fingerprint analysis has essentially zero error rate, which is not possible). But the process is also acknowledged as subjective; many claim that it is not amenable to objective quantification. This subjectivity imposes great limitations on the demonstrated reliability of latent fingerprint analysis that would be reduced with more objective criteria. To better illustrate the disparity between DNA analysis

and fingerprint analysis, we consider both methods in this section.

Arguments put forth for the use of latent fingerprint analysis as forensic evidence of individualization have been offered most recently in *U.S. v. Titus Faison* (D.C. Superior Court 2008-CF2-16636, 12 April 2010). In that case, the United States argued to deny the defendant’s motion to exclude fingerprint evidence (i.e., argued to admit fingerprint evidence), because:

1. “The field of latent fingerprint identification is not a new science”
2. “the ACE-V method of fingerprint identification ... enjoys general acceptance throughout the relevant scientific community”
3. “arguments [to exclude latent fingerprint testimony] ignore nearly a century of forensic history”
4. “arguments already considered and rejected by at least five judges”
5. “other courts throughout the country have found this evidence highly probative and properly admissible”
6. “Instances of examiner error are exceedingly rare”

(see pages 3–5 of this motion). It hardly need be said that the mere existence of a technology for over a century is not equivalent to demonstrating its reliability, validity, and positive predictive value. Just because the technology has been used does not mean that the use has been proven to be appropriate, valid, and error-free under all conditions. (Certain operating systems have been in use for over 30 years, but nonetheless are known to contain hundreds of problems resulting in computer failures to perform as advertised.) The arguments above for fingerprints would be akin to those for using aspirin for a medical condition because:

1. Aspirin is not a new drug.
2. People have been using aspirin for over a century.
3. Administration of aspirin is generally accepted throughout the relevant scientific community.
4. The use of aspirin has been considered and accepted by at least five medical doctors.
5. Other doctors throughout the country have found this medication to be highly effective and properly prescribed.
6. Instances of prescription error are exceedingly rare.
7. There has been a lack of anecdotal evidence of improper administration.

These statements may all be true. But they hardly confirm the use of aspirin without the proper randomized studies confirming its benefit for the specific condition.

In contrast, consider the history of DNA evidence. First, DNA typing began in biochemistry labs, based on the seminal work of James Watson, Francis Crick, and Rosalind Franklin. They identified regions of great similarity among individuals, *but also identified specific regions of great differences*. The totality of these regions of difference leads to genetic uniqueness among individuals (except for identical twins at birth). Biochemists have identified 13 regions (loci) where individuals have been found to differ greatly. Specifically, multiple outcomes (alleles) are possible at each of the 13 loci. Suppose that there are 36 possible outcomes at locus 1, 231 at locus 2, 21 at locus 3, ..., 120 possible outcomes at locus 12, and 153 possible outcomes at locus 13. (See Table 1, taken from Table 1 in Budowle et al. 2009, p.62.) Then the number of possible unique DNA signatures is $36 \times 231 \times 21 \times \dots \times 120 \times 153$, making the number of possibilities huge (about 8×10^{21}). [The number may be larger if one includes low-frequency ($< 1\%$) alleles.] In addition, the loci appear on different chromosomes, and hence the outcomes at the loci are presumably *independent*; that is, knowing that locus 1 had allele number 3 provides no information on the allele that is present at any other locus. [It should be noted that both of these assumptions — number of low-frequency alleles, and independence of occurrence of alleles at two different loci — were based on DNA databases available at the time of the NRC (1996) report. Today, the U.S. Federal Bureau of Investigation (FBI) maintains a DNA database (CODIS: Combined DNA Indexing System), which now has over 7.2 million profiles, so these assumptions could be tested in greater depth; see Laurie and Weir 2003, and Weir 2004, 2007, 2009.] The frequencies of the outcomes (alleles) at the different loci have been estimated, so that, if an individual's DNA profile matches the suspect's profile at all 13 loci, the probability of a spurious match can be calculated to be very low indeed. Consequently, both PPV and NPV are extremely high. DNA analysis does not leave the selection of the 13 regions (loci) up to the examiner. The 13 regions are fixed; they have been designed into the DNA analysis process precisely because they are stable (i.e., the alleles at these loci are not likely to change over time) and because they provide a unique signature. One does not need to sequence an individual's entire DNA, but one also does not have the freedom to choose one's own signature. The signature has been carefully designed and evaluated to yield high PPV and NPV. So the science behind forensic DNA analysis has been carefully studied. In fact, DNA typing preceded its use in the courtroom, so, by the time it was proposed as a forensic tool, lab procedures for DNA analysis had already been well specified by biologists. Validation and proficiency tests for examiners have been established, and accredited DNA forensic laboratories are required to follow established standards for the analysis, interpretation, and reporting of DNA test results. (Even with standards in place, laboratory errors

can and do occur.)

Table 1: Number k of alleles (with frequencies $\geq 1\%$) and

		number of genotypes $k(k + 1)/2$ at 13 loci in the human genome						
Name	CSF1P0	FGA	TH01	TPOX	vWA	D3S1358	D5S818	
#alleles k	8	21	6	7	9	8	8	
#genotypes	36	231	21	28	45	36	36	
Name	D7S820	D8S1179	D13S317	D16S539	D18S51	D21S11		
#alleles k	8	10	7	7	15	17		
	36	55	28	28	120	153		

Now consider the situation with latent fingerprint analysis:

1. Fingerprint ridges are **presumed** to be unique, based on observation by Galton and Locard (cf. Stigler 1999).
2. Fingerprint ridge characteristics can change over time (e.g., become less distinct, with greater use, or absent altogether, with scarring).
3. The **Analysis** phase of the ACE-V method — assessment of print clarity/quality — is acknowledged to be subjective. At present, no objective measures of print quality have been proposed, as they have been for other images. For example, the methodology in Peskin et al. (2010) for assessing the quality of biological cell images (see Figure 1) could be adapted to derive a more objective measure of fingerprint image quality (see Figure 2).
4. The **Comparison** phase of the ACE-V method is highly subjective: the examiner **selects** regions for comparing a latent print with prints from a database
5. The **Evaluation** phase of the ACE-V method likewise is subjective: the examiner decides on a number of features (points or minutiae) needed to establish a “match.” The non-compulsory guidelines from the Scientific Working Group on Friction and Surface Technology (SWGFAST) recommend 6–12 points of agreement. But the measure of “agreement” is not made on the basis of measurements (e.g., distance between ridges; density of pores, etc.), but rather is subjective: “Source determination is made when the examiner concludes, based on his or her experience, that sufficient quantity and quality of friction ridge detail is in agreement” (NAS Report, p.138).
6. The **Verification** phase of the analysis is not conducted as an independent second review, but rather proceeds with “another qualified examiner [who] repeats the analysis and comes to same conclusion ... [this second examiner] may be aware of [first] conclusion” (NAS Report, p.138).

7. Unlike the probabilistic model for the frequency of alleles at the 13 loci in DNA analysis, from which one can calculate probabilities of false positive calls and false negative calls (from PPV and NPV), no reliable probabilistic or scientific model for the frequency of minutiae has been validated, and hence neither probabilistic estimates of error rates, nor the uncertainties in these estimates, can be made.

8. Finally, unlike DNA analysis which was being studied and evaluated across research laboratories all over the world, no extensive and comprehensive studies of performance of the latent fingerprint process or ACE-V methodology (that quantify the effects of multiple sources known to affect variability in performance), nor the error rates from the individual steps in the process, have been conducted. Langenburg (2011: Table 14-2, page 14-18) describes the various probability models that have been proposed for fingerprints, beginning with Galton in 1892 up to more recent attempts such as Champod and Evett (2001). He concludes that, while “fingerprint minutiae are highly discriminating features” and “the more minutiae that are shared between impressions, the less likely it becomes to randomly observe these features elsewhere in the population” (p.14-19, l.1-3), nonetheless “these models have not been validated” (p.14-19, l.10-11).

In short, then, the table below compares DNA analysis with latent fingerprint analysis. It is clear that the science underlying the former far outpaces that of the latter.

	DNA	Fingerprints
Scientific Basis	Before its use in Forensics	After its use in Forensics
Feature Selection	(13) specified markers (loci)	(6–12?) User-specified (loop, whorl, ridge...)
Feature Options	Many outcomes for each locus	Binary outcomes (Present/Absent)
Feature Independence	Loci outcomes are Independent	Dependence among minutiae is unknown
Error rate estimates	Based on allele frequencies	Unknown frequencies of minutiae
Calculations	Probability Models	No models
Potential Subjectivity	Identify allele presence/absence	Quality, minutiae, “match/no-match” call

A common question then arises for statisticians to answer: How many features are needed to provide a unique signature? For example, are 13 features (used for DNA typing) always sufficient for individualization? The answer depends on several factors:

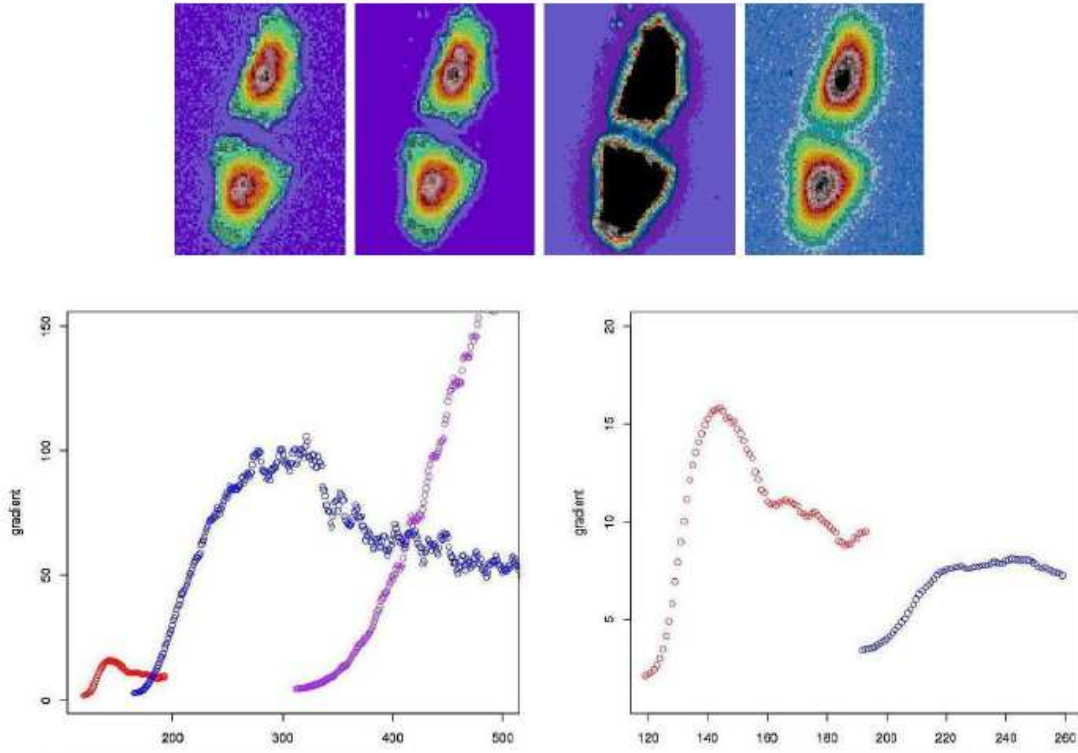


Figure 1: The top four pictures show two biological cells under four different imaging conditions. The bottom left plot shows gradient curves for images 1, 2, and 3: the high sharp peak for image 2 indicates a better image. The bottom right plot shows gradient curves for images 1 and 4: the higher sharp peak for image 1 indicates a better image. From Peskin et al. (2010).

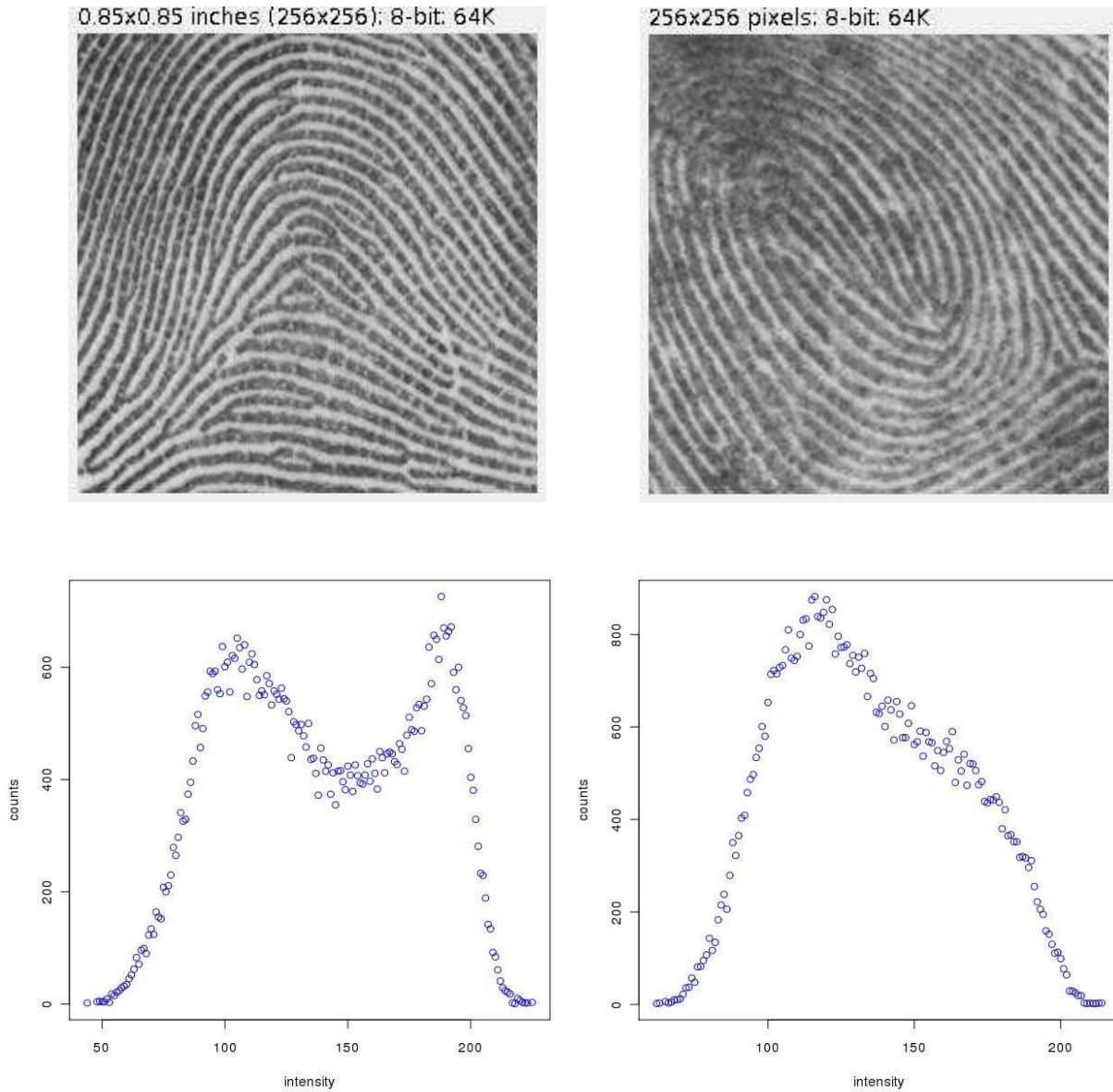


Figure 2: The top two pictures show two fingerprint images. The bottom plots show the gradient curves from them, using the methodology in Peskin et al. (2010) for biological images. The sharper peaks in the plot on the left indicates a better image. Clearly the curves would need to be region-specific, as different regions of the print would have different levels of clarity.

- The *sensitivity* of each feature; i.e., if the profiles really come from the same person, the probability that the features match in both profiles is high (e.g., above 0.90).
- The *specificity* of each feature; i.e., if the profiles really come from different persons, the probability that the features do not match in the profiles is high (e.g., above 0.90).
- The *independence* (or lack of strong dependence) among the features.
- The size of the population on which the signatures are being evaluated.

Under these conditions, 13 features (or even as few as 10) suffice to assure a positive predictive value of over 0.9995. When both the sensitivity and the specificity of each of k independent features exceeds 0.90, Figure 3 shows a plot of the PPV as a function of k = number of features, for various population sizes (e.g., the match occurs in 1 out of 10, 20, 50, 100 , ..., one million profiles), when both the sensitivity and the specificity of *each* of k independent features exceeds 0.90, However, when the sensitivity is only 0.80 and the specificity is only 0.50, then the number of *independent* features needed to assure $PPV = 0.90$ or higher can easily become very large; cf. Figure 4.

4 A well-designed study in public health

Polio was a greatly feared disease in the 1950s; thousands of people were afflicted or died. The Salk vaccine had been developed in the laboratory by scientist Jonas Salk, using killed polio virus. A serious public health issue arose: Should the vaccine be administered widely, as an effective means of curtailing a polio epidemic? Or might the vaccine be ineffective, or, worse, result in more cases of polio? The only way to answer this question was to conduct a large, well-designed clinical trial.

Many considerations had to be addressed in the design of this trial, as nicely presented by Meier (1957; 1980). Two studies were conducted. One was an observational community trial, in which the parents of all second-grade children in the community could volunteer their children to receive the vaccine, while the children in the community's first and third grades were left unvaccinated. Two possible grounds for criticism were raised with this trial. First, health professionals in the community would know whether a child aged 6 to 8 received, or did not receive, the vaccine; the study is not even blind (where the participant does not know the treatment being administered), much less double-blind (where neither participant nor administrator knows). Second, the children of parents who volunteer them for a study may not be representative of the general population of children. This "healthy-volunteer" bias is known to occur in other studies, and, in fact, occurred here as well: the rates of detected polio cases in this community trial were lower than those in the second trial, described below.

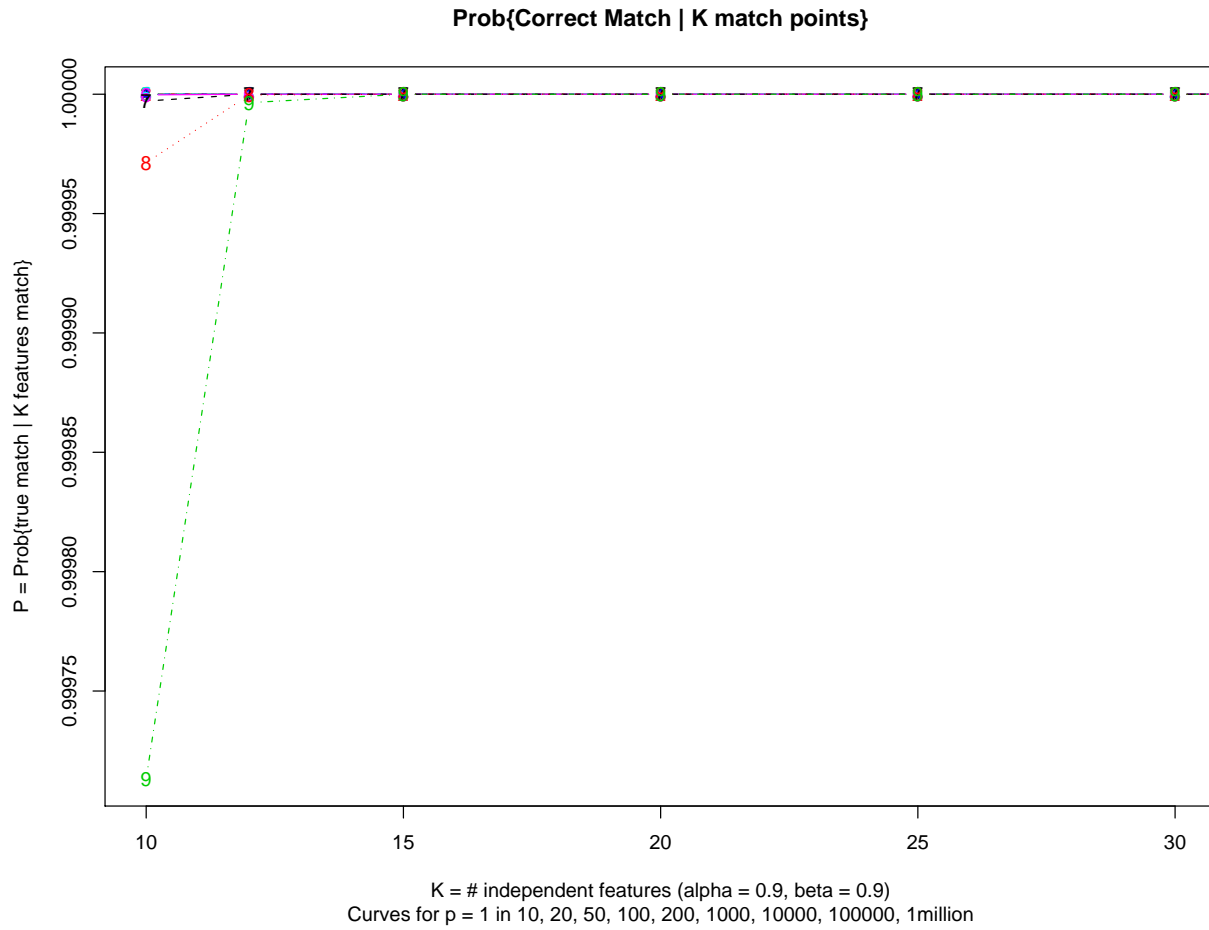


Figure 3: Plot of PPV = positive predictive value, as a function of number of independent features, when the sensitivity of each feature is 0.90 and the specificity of each feature is 0.90. Curves correspond to sizes of population in which match is believed to occur (100, 200, 500, 1000, ..., 1 million).

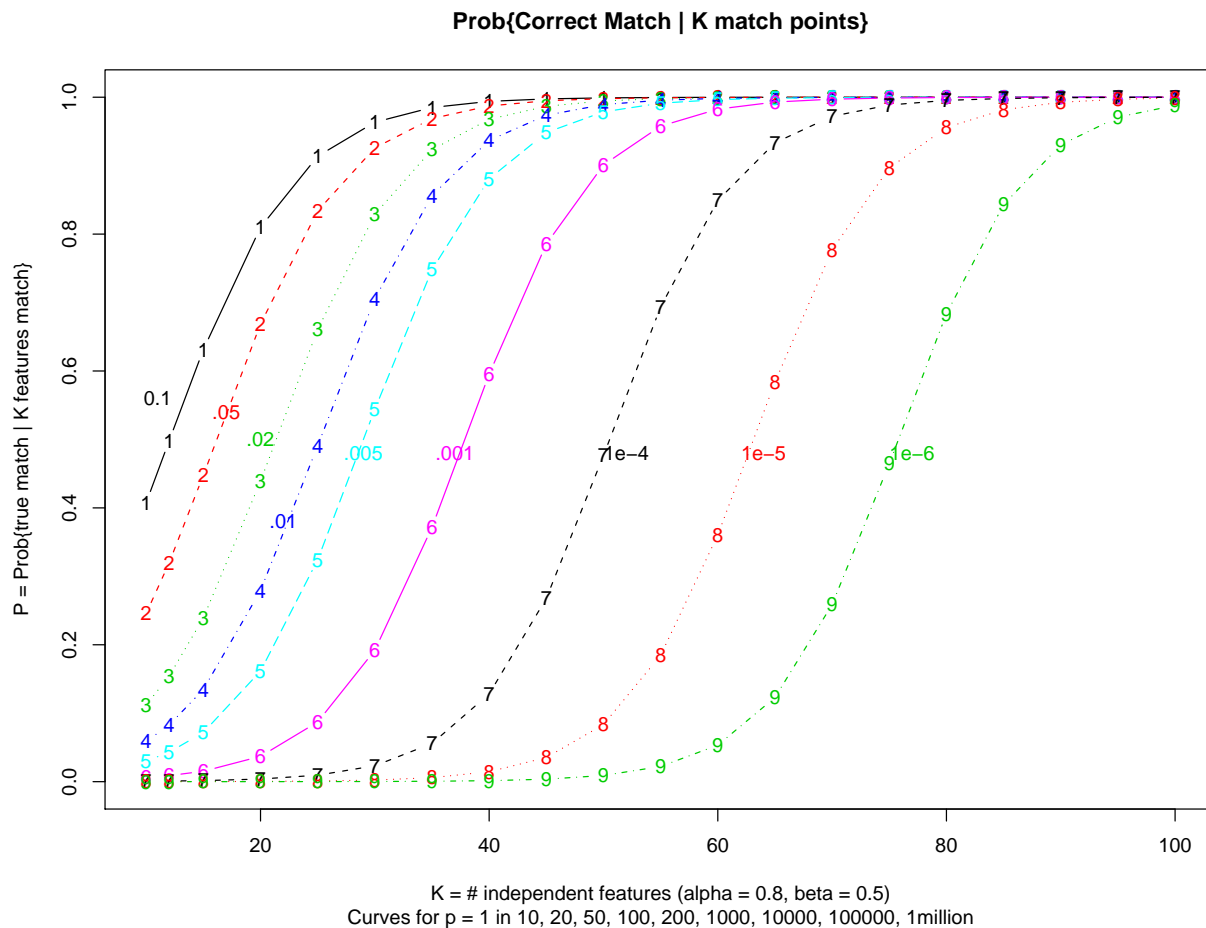


Figure 4: Plot of PPV = positive predictive value, as a function of number of independent features, when the sensitivity of each feature is 0.80 and the specificity of each feature is 0.50. Curves correspond to sizes of population in which match is believed to occur (100, 200, 500, 1000, ..., 1 million).

The second trial was designed to avoid the criticisms of the community trial. In this placebo-controlled randomized trial, the parents of approximately 402,000 children agreed to participate in the trial. Each child received an inoculation, which contained either the Salk vaccine or a placebo (no vaccine at all). The vials looked identical, so neither the child nor the administrator knew whether the vial contained vaccine or placebo. (In fact, even the physicians making the diagnoses of disease did not know whether the child received vaccine or placebo.) The choice of vaccine or placebo was decided on the basis of a coin flip; i.e., purely at random.

The relevant results of the placebo-controlled randomized trial were as follows. Among the 200,745 children that received the vaccine, 82 polio cases arose. Among the 201,229 children that received the placebo, 162 polio cases arose. Because the numbers of children in each group are roughly equal (about 201,000 in each group), we can make a direct comparison of these two numbers, 82 versus 162. *If the vaccine were not effective, then we would expect the same numbers of cases in each group.* Because the decision to receive or not receive vaccine depended on a coin flip, there was a 50-50 chance of ending up in one group or the other. So, if the vaccine were no more effective than the placebo, then the fact that the polio-stricken child ended up in the vaccine group was just the result of a coin-flip — it was just as likely that the child could have received placebo, and we would have expected the 244 cases to split evenly between the two groups, or 122 cases in each group. Under the “equal-effectiveness” hypothesis, a split of 120-124 cases is plausible, or even a 118-126 split. How extreme a split might we see, if the chances of getting polio really were equal in each group?

To answer this question, we can simulate the tossing of 244 coins, and count the number of “heads” versus “tails.” In the first run, the split was not 122–122, but was 114–130. In the second run, it was 125–119. In the third run, it was 131–113. In the fourth run, it was 117–127. Figure 5 shows a histogram of the “numbers of heads” (i.e., numbers of cases that fell into the vaccine group) in the 10,000 runs. Most (9500) of the splits were no further than 15 from 122–122, and in only 500 runs were the splits more extreme than 107–137 (i.e., 107 heads, or vaccine cases, and 137 tails, or placebo cases; or 137 heads and 107 tails). *None* of the splits was more extreme than 96–148. Consider the outcome of the trial: the split was 82–162. This is possible under the 50-50 model, but extremely unlikely: the probability of seeing a split as extreme as 82–162 (or even more extreme such as 81–163 or 80–164 or ...) is only 3.7×10^{-7} (less than one in a million). The data do not support the hypothesis that the vaccine is no more effective than the placebo; i.e., it is much more plausible that the effectiveness rates differ, by a factor of about 2 to 1; i.e., about twice as many cases in the placebo group as in the vaccinated group. (A 95% confidence interval for this ratio of effectiveness is (1.52, 2.63) – i.e., quite far from 1.00.)

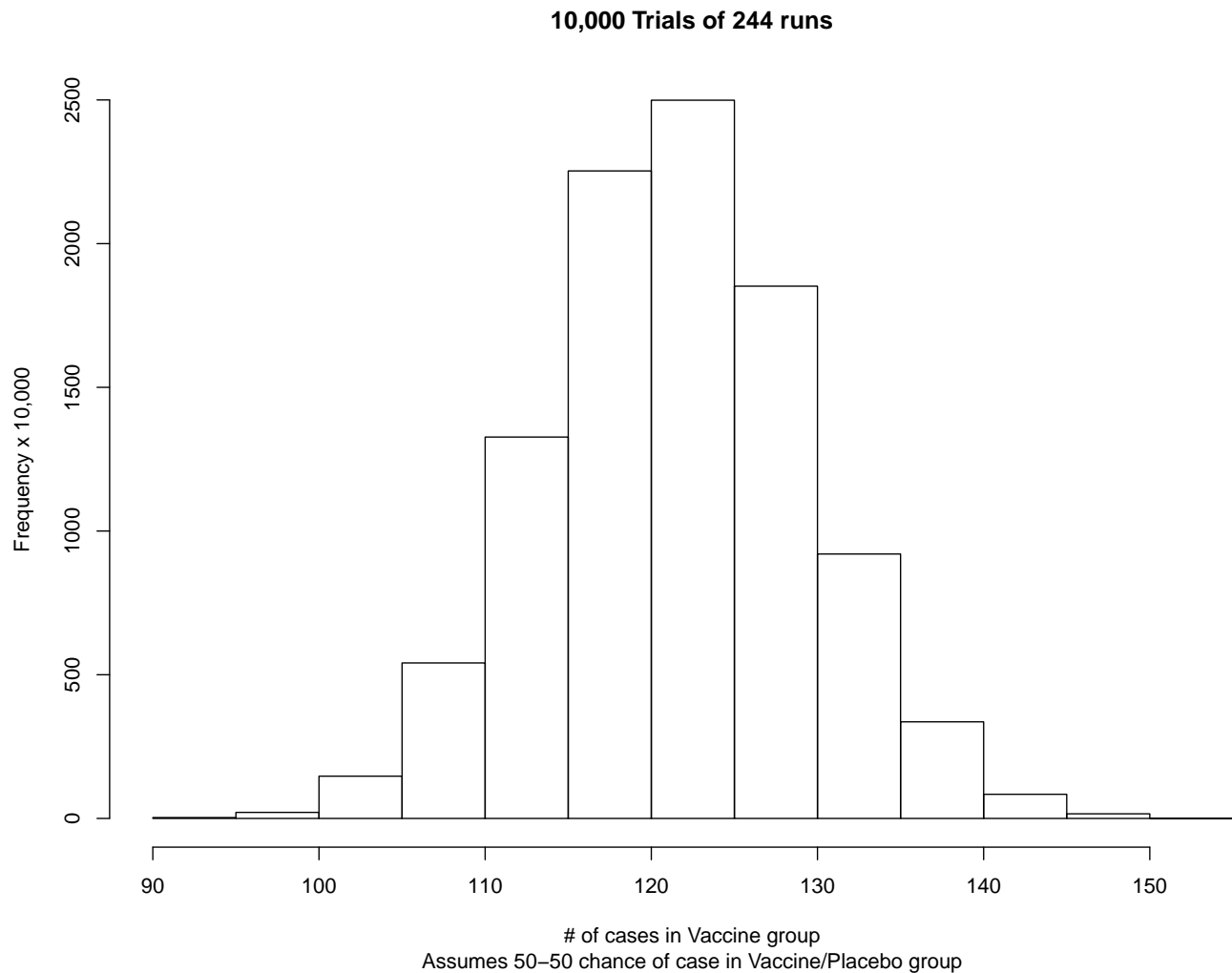


Figure 5: Histogram of 10,000 simulated runs of 244 coin flips. The outcome of each run is the simulated number of polio cases in the vaccinated group, assuming that the vaccine is equally effective as the placebo, so one expects each case to fall into the vaccine group with probability $1/2$. In most of the 10,000 runs, the fraction of the 244 cases in the polio group would be expected to be about $1/2$, or roughly 122 cases, if the placebo and vaccine were equally effective; in none of the 10,000 runs was the number of cases in the polio group expected to be more extreme than 96–148. Because the outcome of the real trial was a split of 82-162 cases, this result is far more extreme than would be predicted on the basis of the hypothesis of “equal effectiveness”, so the data do not support the hypothesis.

This trial presents an extremely important illustration of the value of a well-designed experiment. On the basis of it, polio vaccinations were institutionalized. (The Salk vaccine was replaced in 1962 by the Sabin vaccine using an attenuated virus.) One good study had a major effect on public health practices. Many poorly designed studies never have that kind of impact. The key features of this trial include both the design (two groups, one received “no” treatment, the other received “vaccine” treatment; double-blind, so neither participant nor health professional knew the treatment) as well as the analysis plan. The data analysis plan proceeded as follows:

1. Identify a metric (or set of metrics) that describes the essential features of the data. For this example, the metric was the fraction of cases that occurred in vaccine (vs placebo) group.
2. Determine a range on the metric(s) which is “likely to occur” (has a 95% chance of occurring) if “nothing interesting is happening.” In this example, “nothing interesting” means a 50-50 split; i.e., if N cases of polio arise in the trial, the “equal effectiveness” hypothesis is $N/2$ cases in each group, and 95% of possible likely outcomes (number of polio cases in the vaccinated group) can be expected to lie between $N/2 - \sqrt{N}$ and $N/2 + \sqrt{N}$.
3. Identify the “extreme range” = range of the metric(s) outside of the “95%” range. For this example, splits more extreme than $(N/2 - \sqrt{N})$ – $(N/2 + \sqrt{N})$ would be deemed “extreme” or “unlikely under the equal effectiveness hypothesis.”
4. Conduct experiment and calculate the metric(s). For this example, the vaccine/placebo was administered to 200,745/201,229 children (randomized to one group or the other), and a total of 244 cases were observed: 82 in the vaccine group and 162 in the placebo group, for a split of 82–162 (33.5% of the cases in the vaccine group).
5. If metric falls in “expected” range, then data are consistent with the hypothesis that “nothing interesting is happening”. If metric falls in “extreme” range, then data are not consistent with this hypothesis \Rightarrow hypothesis is not supported by the data. For this example, splits more extreme than 107–137 (e.g., 106–138, or 105–139, or ...) occur with probability less than 5%, and hence are deemed “extreme” or “unlikely under the equal effectiveness hypothesis.” Because 82–162 falls in “extreme” range, the data do not support the hypothesis of “equal effectiveness.”

(Above, “95%” could be replaced by a higher confidence level.) We next discuss how this process might apply to a study that evaluates the effectiveness of latent fingerprint analysis to correctly identify perpetrators.

5 Designing studies for assessing latent fingerprint analysis

When one compares the design of the Polio Salk Vaccine trial with typical studies of latent fingerprint analysis, many differences become clear. An often-cited illustration of the “success” of fingerprint analysis was the pairwise comparison of 50,000 high-quality prints with each other. The prints were not a random sample from any population; they were all “high-quality” (even the “partial” prints were obtained as 25% subsets of “high-quality” prints, a far cry from the typical quality of a latent fingerprint); the degree of similarity between two different prints compared with the *same* print was never evaluated; see Kaye (2005) for a more detailed criticism of this study.

A second example of a designed study was the Proficiency study conducted by Collaborative Testing Services (1995). In this study, test materials were sent to 228 participants in 94 crime laboratories. The materials included seven photographs of test prints with background information on the crime scene and related incidents, along with four 10-print cards; the instructions were to “find the best match.” The answers were: two of the prints (prints A and D) came from one of the four cards; three of the prints (prints B, E, G) came from another one of the four cards; and two of the prints (prints C and F) came from none of the four cards. Responses were received by only 156 (about 2/3) of the participants; Most (71) of the 94 labs provided responses from only one participant; 12 labs provided responses from two participants; 9 labs provided responses from 3,4,5, or 7 participants, and 11 respondents came from the lab denoted as “2483.” The test process did not specify the complete ACE-V method. The prints did not appear to have come from a random sample of a target population, and the method by which laboratories were selected for this proficiency test was not provided in the report.

The data from this proficiency test are summarized in the table below.

Test Print	A	B	C*	D	E	F*	G
Correct	137	126	150	108	115	127	146
Incorrect	2	2	6	2	3	29	1
Missed/NI	17	28	—	46	38	—	9

From these results:

- The mean proportion correct (“Sensitivity”) can be estimated as $126.4/156 = 0.81 = 81\%$, with a 95% confidence interval of $(0.763, 0.866) = (76.3\%, 86.6\%)$.
- The mean proportion incorrect (false positive rate) is estimated as $2/156 = 0.013 = 1.3\%$; a 95% confidence interval for the false positive rate is $(0.00, 0.04) = (0\%, 4\%)$.
- The mean proportion missed or not identified (NI) is estimated as $27.6/156 = 0.177 = 17.7\%$; a 95% confidence interval for the false negative rate is $(0.123, 0.242) = (12.3\%, 24.2\%)$.

- The same 2 people mis-identified prints A–F.

These results indicate that sensitivity and specificity of latent fingerprint analysis *under these idealized conditions* (not the complete ACE-V process; small population of prints for comparison) may be on the order of 96% and 80%, respectively, which suggests that the analysis under these favorable conditions has high PPV and NPV. But these results cannot translate into real practice.

The Polio Salk vaccine trial began by identifying the sources of greatest variability in the outcomes. Because that source was deemed to be individuals (versus geography, health professionals, doctor diagnosis, etc.), the trial focused on identifying large numbers of participants. Similarly, for assessing the performance of latent print analysis, one might begin by identifying the greatest sources of variability in the process, namely those aspects that are *subjective*.

The first step in the ACE-V methodology was noted as a *subjective* assessment of latent fingerprint quality. Because image analysis has faced this problem, presumably an objective measure of quality could be developed. We start by assuming that such a measure is available, though acknowledge that its creation may require additional research.

Assuming latent print quality passes acceptable threshold, consider again the steps used in the Polio Salk vaccine trial, but now applied to latent fingerprint analysis:

1. *Identify a metric (or set of metrics) that describes the essential features of the data.* For example, these metrics might consist of the numbers of certain types of features (minutiae), or average distances between the features (e.g., between ridges or bifurcations); or eccentricities of identified loops; or other characteristics that could in principle be measured.
2. *Determine a range on the metric(s) which is “likely to occur” (has a 95% chance of occurring) if “nothing interesting is happening”* (i.e., the two prints do not arise from the same source). For example, one could calculate these metrics on 10,000 randomly selected latent prints known to have come from different sources.
3. *Identify “extreme range” = range of the metric(s) outside of the “95%” range.* For example, one can calculate ranges in which 95% of the 10,000 values of each metric lie.
4. *Conduct the experiment and calculate the metric(s).* For example, from the “best match” that is identified, one can calculate the relevant metrics.
5. *If the metric falls in “expected” range, then data are deemed consistent with the hypothesis that “nothing interesting is happening”.* If the metric falls in the “extreme” range, the data are not consistent with this hypothesis and indicate instead an alternative hypothesis.

6 Bullet Lead as Forensic Evidence

In the 1960s, the U.S. Federal Bureau of Investigation (FBI) began performing Compositional Analysis of Bullet Lead (CABL), a forensic technique that compared the elemental composition of bullets found at a crime scene (CS) to that of bullets found in the possession of a potential suspect (PS). CABL was used when no gun could be recovered, or when bullets were too small or fragmented to compare striations on the casings with those on the gun barrel. (See Cork et al. 2011 for a discussion of the NRC report concerning the present infeasibility of a national ballistic database.) The National Academy of Sciences formed a Committee charged with the assessment of CABL’s scientific validity as a forensic technique in connecting a potential suspect with the crime scene bullets. Below is an assessment of the statistical procedures used in CABL, described in greater detail in the Committee’s report (National Research Council, 2004).

The “chemical signature” consisted of the measured concentrations of 3–7 elements found in the lead of the target bullet. (Before 1991, only three elements were measured by neutron activation analysis; eventually seven elements were measured by inductively coupled plasma optical emission spectroscopy (ICP-OES), so this discussion focuses on a seven-element signature.) Each concentration was measured three times, using three different standard reference materials from the U.S. National Institute of Standards and Technology (NIST), yielding sample means and standard deviations. The FBI test procedure for a “match” was as follows: If the “2-SD intervals” of all seven elements in the PS bullet “overlapped” with those from the CS bullet, the bullets were declared to be “analytically indistinguishable.” An agent often would testify that these two bullets “came from the same box of bullets,” implying guilt of the potential suspect. Occasionally the FBI replaced “2-SD intervals” with ranges (maximum – minimum of the 3 measurements), resulting in a “range test” but the former “2-SD interval overlap” procedure was more common. The FBI asserted that the measurement errors among the seven elements were uncorrelated. Based on a *selected* subset of 1837 bullets from the FBI’s log of bullet measurements over many years (approximately 70,000 measurements on about 17,000–18,000 bullets), the FBI found only 693 “matches” among all 1,686,366 ($1837 \times 1836 / 2$) pairwise comparisons and concluded a “false positive rate” of 0.04%.

Figure 6 shows an illustration comparing six “2-SD intervals” (left panel) and “range intervals” (right panel) from two bullets (data described in Peele et al. 1991). Because all six sets of 2-SD intervals overlap, the FBI would have called these two bullets “analytically indistinguishable.” One (Sb) of the six “range intervals” just barely fails to overlap, so the call from the “range overlap test” would have been “analytically indistinguishable on 5 elements.” [Note that the expected range in a Gaussian sample of size 3 is 1.69256σ (Harter 1961), considerably less than 4σ used in the “2-SD overlap test” so fewer “overlaps” would be expected with the former “test”.]

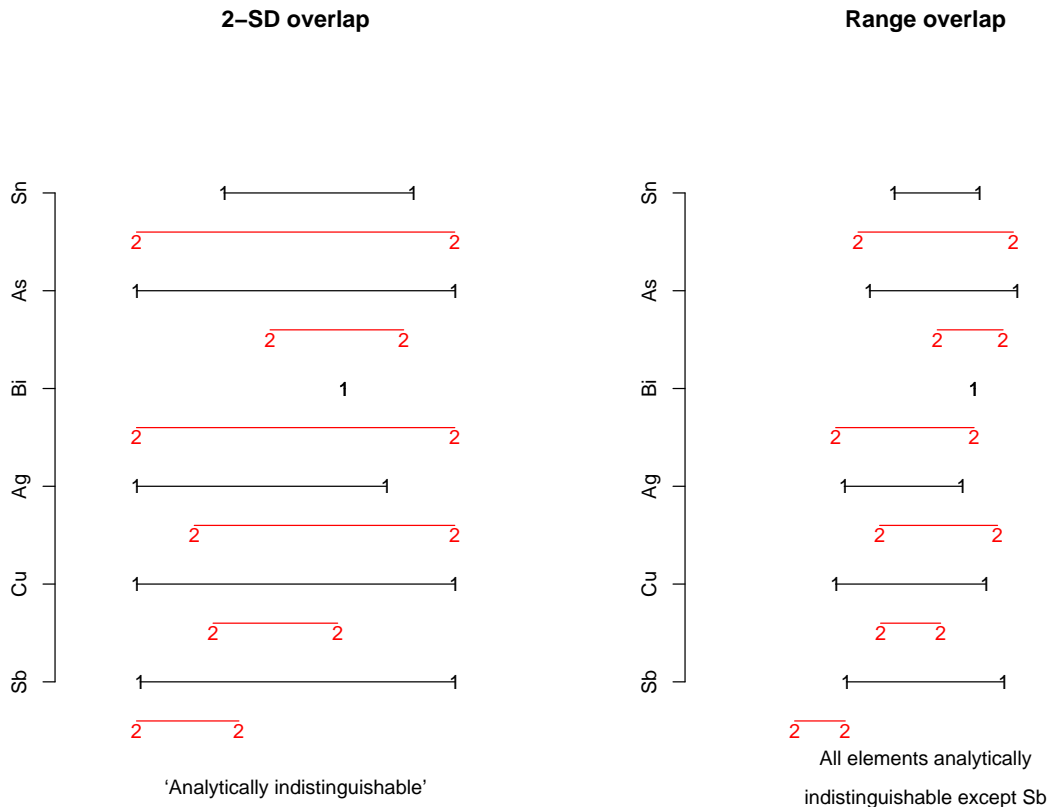


Figure 6: Comparing two bullets via “2-SD interval overlap” (left panel) and “range overlap” (right panel). All six pairs of “2-SD intervals” overlap, resulting in a claim that the bullets are “analytically indistinguishable.” All but one pair of the “range intervals” overlap, resulting in a claim that the bullets are “analytically indistinguishable on five elements.”

Formally, let X_{ij} and Y_{ij} , $i = 1, \dots, 7$ and $j = 1, 2, 3$, denote the j^{th} measured concentration of element i in the CS and PS bullets, respectively, where the elements are arsenic (As), silver (Ag), antimony (Sb), bismuth (Bi), copper (Cu), tin (Sn), and cadmium (Cd). (The copper measurement could be unreliable due to the presence of copper in the bullet casing, some of which might have been extracted along with the bullet lead.) The “2-SD intervals” are

$$(\bar{X}_i - 2 \cdot s_i^X, \bar{X}_i + 2 \cdot s_i^X), (\bar{Y}_i - 2 \cdot s_i^Y, \bar{Y}_i + 2 \cdot s_i^Y), \quad (1)$$

where

$$\bar{X}_i = \sum_{j=1}^3 X_{ij}/3; \quad s_i^X = \left[\sum_{j=1}^3 (X_{ij} - \bar{X}_i)^2 / 2 \right]^{1/2} = [((X_{i1} - X_{i2})^2 + (X_{i1} - X_{i3})^2 + (X_{i2} - X_{i3})^2) / 6]^{1/2} \quad (2)$$

$$\bar{Y}_i = \sum_{j=1}^3 Y_{ij}/3; \quad s_i^Y = \left[\sum_{j=1}^3 (Y_{ij} - \bar{Y}_i)^2 / 2 \right]^{1/2} = [((Y_{i1} - Y_{i2})^2 + (Y_{i1} - Y_{i3})^2 + (Y_{i2} - Y_{i3})^2) / 6]^{1/2}. \quad (3)$$

Thus, the “2-SD intervals” overlap if

$$|\bar{X}_i - \bar{Y}_i| < 2(s_i^X + s_i^Y). \quad (4)$$

In terms of our usual t-statistics, this equation (without the subscript i) can be written as

$$\begin{aligned} & \text{P}\{(\bar{X} - \bar{Y}) < 2(s^X + s^Y) \mid |\mu_X - \mu_Y| = \delta \} \\ &= \text{P}\{(\bar{X} - \bar{Y}) / (s_p \sqrt{2/3}) < 2(s^X + s^Y) / (s_p \sqrt{2/3}) \mid |\mu_x - \mu_y| = \delta \} \end{aligned}$$

where μ_X and μ_Y are the true mean concentrations and s_p is a pooled estimate of σ based on many degrees of freedom. If the measurements are normally distributed, then the sample standard deviations s^X and s^Y , each based on two degrees of freedom, have expectations 0.8812σ , while a pooled s_p , based on many degrees of freedom, is approximately σ . Hence,

$$\begin{aligned} & \text{P}\{(\bar{X} - \bar{Y}) < 2(s^X + s^Y) \mid |\mu_X - \mu_Y| = \delta \} \\ & \approx \text{P}\{(\bar{X} - \bar{Y}) / (s_p \sqrt{2/3}) < 4.317 \mid |\mu_x - \mu_y| = \delta \} \end{aligned}$$

The approximation comes from the fact that $\text{E}(\text{P}\{t < S\}) \neq \text{P}\{t < E(S)\}$. Nonetheless, the calculation indicates that the definition of “analytically indistinguishable” is highly generous.

The Committee’s report noted several problems with this procedure:

1. *The 1837-bullet subset was not a random sample.* Notes from the FBI that accompanied this data set included this paragraph (emphasis added):

“To assure independence of samples, the number of samples in the full database was reduced by removing multiple bullets from a given known source in each case. To do this, evidentiary submissions were considered one case at a time. For each

case, one specimen from each combination of bullet caliber, style, and nominal alloy class was **selected** and that data was placed into the test sample set. In instances where two or more bullets in a case had the same nominal alloy class, one sample was randomly selected from those containing the maximum number of elements measured. . . . The test set in this study, therefore, should represent an unbiased sample in the sense that each known production source of lead is represented by only one randomly selected specimen.”

(See also Koons and Basaglia 2005.) Obviously, the bullets were **selected** to be as **different** as possible, so the variability among the bullets in this sample is likely to be greater than that seen in the general population of bullets (which could vary over time). The data set provides information on likely ranges of concentrations that one might see in bullet lead but no formal inference can be derived from this biased data set.

2. *The estimates of the measurement error standard deviation are highly unstable.* Using only 3 observations to estimate s_i^X (and likewise for s_i^Y) yields highly unstable estimates of the measurement error standard deviation, as they each are based on only two degrees of freedom. Because the FBI had been conducting these ICP-OES analyses for many years on hundreds of bullets, presumably the lab had much data which could have been pooled to obtain more stable estimates. To guard against potential drifts in the measurement process, the report recommended plotting the s_i 's (from both CS and PS bullets) and monitoring the stability of these estimates over time using standard control charts. (Because chemical concentrations often are lognormally distributed, the report also suggested that a logarithmic transformation of the data may be more amenable to analysis.)
3. *The measurement errors appeared to be correlated.* The FBI provided triplicate measurements on 50 bullets in each of 4 boxes (200 bullets total) from each of 4 manufacturers (Peele et al. 1991). One of these four sets of 200 bullets included triplicate measurements (via ICP-OES) on six of the seven elements (all but Cd), enabling pooled estimates of the correlation among the measurement errors. The estimated correlation matrix was found to be:

	As	Sb	Sn	Bi	Cu	Ag	(Cd)
As	1.000	0.320	0.222	0.236	0.420	0.215	0.000
Sb	0.320	1.000	0.390	0.304	0.635	0.242	0.000
Sn	0.222	0.390	1.000	0.163	0.440	0.154	0.000
Bi	0.236	0.304	0.163	1.000	0.240	0.179	0.000
Cu	0.420	0.635	0.440	0.240	1.000	0.251	0.000
Ag	0.215	0.242	0.154	0.179	0.251	1.000	0.000
(Cd)	0.000	0.000	0.000	0.000	0.000	0.000	1.000

Notice in the above table that the most optimistic correlations (zero) were assumed for Cd, which was not measured in this data set. Nonetheless, all of the other estimated correlations appear to be rather different from zero, ranging from 0.154 (Ag and Sn) to 0.635 (Sb and Cu). As a consequence, the false positive rate of seven tests, say $\alpha_1, \dots, \alpha_7$, is not simply the product $\prod_{i=1}^7 \alpha_i$; in fact, simulations showed that, if all $\alpha_i = \alpha$, then the false positive rate among the seven correlated tests is closer to α^5 than to α^7 .

4. *The FBI's claimed false positive rate based on the 1837-bullet subset is unrealistic.* Figure 7 calculates the false positive rate in comparing bullets based on 1 element (left panel) and 7 elements (right panel). (The right panel was prepared under the assumption that the measurement errors are correlated as above.) This false positive probability is a function of the ratio of the true mean difference, say δ , to the measurement error standard deviation, say σ ; i.e., is a function of δ/σ . When this ratio is 2, the false positive probability is 0.43; at 3, it is 0.09. Clearly the chance of a false positive goes to zero as the true difference between the bullet lead concentrations increases relative to the variability in the measurements.
5. *Equivalence testing is more appropriate.* Given the “innocent until proven guilty” philosophy, the “null” hypothesis is more sensibly H_0 : *Concentrations differ by at least δ* , versus H_1 : *Concentrations differ by less than δ* . Using $\alpha = 0.30$ for each test, so that $\alpha^5 = 0.0025$ (1 in 400), the appropriate multiplier for each test is not 4.317, but rather would be 0.63 (if “analytically indistinguishable” means “ $\delta < 1.0\sigma$ ”) or 1.07 (if “analytically indistinguishable” means “ $\delta < 1.5\sigma$ ”). The allowance 0.63 or 1.07 arose from a noncentral t distribution used in the equivalence t test (Wellek 2003), assuming 3 measurements on each bullet and at least 400 degrees of freedom in s_p . Note that either allowance is considerably smaller than that used in the FBI procedure, which corresponded roughly to declaring “analytically indistinguishable” if “ $\delta < 4.3\sigma$.” Note also that one could use Hotelling’s T^2 statistic, but the robustness of multivariate equivalence testing with a misspecified covariance matrix has not been investigated (besides being considerably more complicated).

In the end, the complexity of the statistical issues, together with the fact that manufacturers’ batches of lead were sufficiently homogeneous that anywhere from 35,000 to 35 million bullets could be deemed “analytically indistinguishable”, the FBI chose to abandon CABL as a way of connecting potential suspects to crime scene bullets. However, the same issue continues to arise in other contexts, such as air compositions, glass, and other forensic evidence. Statisticians need to be involved in assessing the validity of the approaches to analyzing such forensic evidence.

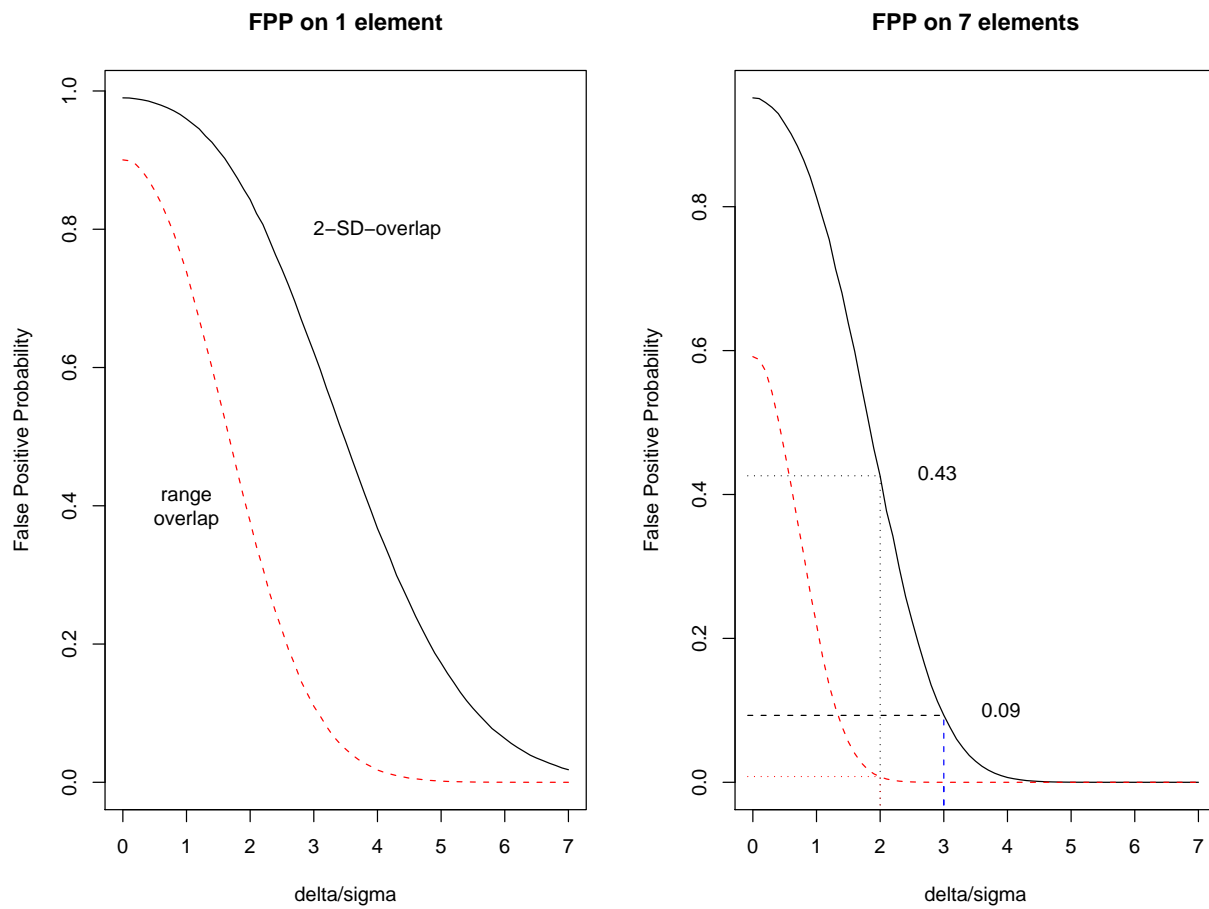


Figure 7: False positive probabilities of declaring “analytically indistinguishable” between two bullets, based on 1 element (left panel) and on 7 elements (right panel). Correlations among elements estimated from a data set on 200 bullets provided by the FBI.

7 Final comments

This article provides some guidance on scientific principles and statistical issues that should be considered when evaluating forensic evidence. Well-characterized, objective metrics need to be developed for each type of evidence, and the studies to evaluate its performance on realistic cases need to be designed and conducted that account for sources that can affect the results. Such studies can be beneficial not only in identifying conditions under which the evidence is valuable but also in raising issues which can be addressed and ultimately strengthen the value of the evidence.

It is important to emphasize that the Committee did not say that the methods are invalid, unreliable, or incapable of providing class evidence. On the contrary, the Committee simply described the state of the peer-reviewed published research on these methods, through hundreds of published articles, subjecting these articles to intensive scientific scrutiny, as would be done in any other area of science. The goal of this article is to provide some basis for understanding the ways in which the published studies failed to meet the high standard known as the *scientific method* (NAS report, Ch. 4), and the ways in which studies can be designed to better quantify the capabilities of the methods. Mearns (2010) emphasizes that better methods will ensure not only fewer false convictions but more proper convictions. **Only by better understanding both the strengths and the limitations of the methods can be totality of the evidence in a particular case be better evaluated.**

It is equally important to recognize that the NAS Report did *not* provide a blueprint for law reform, nor did it state how courts should now treat the admissibility of forensic evidence. The authors of the report certainly hope that its findings will be taken into consideration in the course of judicial proceedings. But as Judge Edwards stated in his testimony to the Senate Judiciary Committee (18 March 2009),

‘It will be no surprise if the report is cited authoritatively for its findings about the current status of the scientific foundation of particular areas of forensic science. And it is certainly possible that the courts will take the findings of the committee regarding the scientific foundation of particular types of forensic science evidence into account when considering the admissibility of such evidence in a particular case. However, **each case in the criminal justice system must be decided on the record before the court pursuant to the applicable law, controlling precedent, and governing rules of evidence.** The question whether forensic evidence in a particular case is admissible under applicable law is not coterminous with the question whether there are studies confirming the scientific validity and reliability of a forensic science discipline.’ (p.10; emphasis added)

One hopes that the relevant studies can be conducted so that forensic evidence can be recognized as a valuable tool in the search for truth.

Acknowledgements

The author thanks the members of the NAS Committee on Identifying the Needs of the Forensic Science Community (2007–2009), particularly the Honorable Harry T. Edwards, Professor Contantine Gatsonis, and Dr. Anne-Marie Mazza. While this article benefitted from their comments, the statements made herein (apart from attributable citations) remain the sole responsibility of the author and do not necessarily represent the views of individual committee members. This article was prepared in part with support from Grant Number W911NF0510490 from the Army Research Office, which is gratefully acknowledged.

Appendix: Estimating error rates

As indicated in Section 2, studies can be designed with fixed numbers of “true matches” and “true non-matches” which can then lead to estimates of sensitivity (number of “match” calls among the true matches) and specificity (number of “no match” calls among the true non-matches). Letting $Sens$ denote sensitivity, $Spec$ denote specificity, and p denote the probability of a “true match” in the population (e.g., $0.01 = 1$ of 100, or $0.001 = 1$ in 1000, or ... or $10^{-6} = 1$ in a million), the positive predictive value (PPV) and negative predictive value (NPV) are related to $Sens$, $Spec$, and p as follows:

$$PPV = \frac{Sens \cdot p}{Sens \cdot p + (1 - Spec) \cdot (1 - p)}$$

$$NPV = \frac{Spec \cdot (1 - p)}{Spec \cdot (1 - p) + (1 - Sens) \cdot p}$$

Consequently, the “false discovery rate” (false positive call probability) and the “false non-discovery rate” (false negative call probability) are

$$FDR = 1 - PPV = \frac{(1 - Spec) \cdot (1 - p)}{(1 - Spec) \cdot (1 - p) + Sens \cdot p}$$

$$FNDR = 1 - NPV = \frac{(1 - Sens) \cdot p}{(1 - Sens) \cdot p + Spec \cdot (1 - p)}$$

We calculate these quantities on the (fictitious) results of one examiner:

	‘Match’	‘No match’	Total
True match	87	13	100
Non-match	3	97	100
Total	90	110	200

- Estimate sensitivity: $87/100 = 0.87$; 95% confidence interval (CI) is (0.80, 0.93).

- Estimate specificity: $97/100 = 0.97$; 95% CI is (0.93, 1.00).
- Estimate PPV: $87/90 = 0.97$; 95% CI is (0.93, 1.00)*.
- Estimate NPV: $97/110 = 0.88$; 95% CI is (0.81, 0.93)*.
- Estimate FDR = 1-PPV: 3%; 95% CI is (0%, 7%).
- Estimate FNDR = 1-NPV: 88%; 95% CI is (7%, 19%).

* Confidence intervals for PPV and NPV are obtained via simulation, because standard formulas from the binomial probability distribution do not apply (the denominator is not fixed, as required for the binomial distribution).

References

1. Benjamini, Y.; Hochberg, Y. (1995), Controlling the false discover rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B* 57: 289–300.
2. Budowle, Bruce; Baechtel, F. Samuel; Chakraborty, Ranajit (2009), “Partial matches in heterogeneous offender data based do not call into question the validity of random match probability calculations,” *Int. J. Legal Medicine* 123: 59–63; DOI 10.1007/s00414-008-0239-1.
3. Champod, Christophe; Evett, Ian W. (2001), “A probabilistic approach to fingerprint evidence,” *Journal of Forensic Identification* 51: 101–122.
4. Cole, Simon (2004), “Grandfathering evidence: Fingerprint admissibility rulings from Jennings to Llera Plaza and back again,” *American Criminal Law Review* 41:1226–1231
5. Collaborative Testing Services, Inc. (1995), “Latent Prints Examination,” Report No. 9508, Herndon, VA.
6. Cork, D.; Nair, V.N.; Rolph, J. (2011), Some forensic aspects of ballistic imaging, *Fordham Urban Law Journal*, March 2011.
7. Dror I.E.; Charlton D. (2006), Why experts make errors, *J Forensic Identification* 56(4):600–616.
8. Genovese, Christopher R.; Wasserman, Larry (2004), Controlling the false discovery rate: Understanding and extending the Benjamini-Hochberg Method, <http://www.stat.cmu.edu/genovese/talks/pitt-11-01.pdf>.

9. Haber L.; Haber R.N.: Scientific validation of fingerprint evidence under Daubert, *Law, Probability, and Risk* 7(2):87–109 (2008)
10. Harter, H.L. (1961): Expected values of normal order statistics, *Biometrika* 48(1/2), 151–165 (Table 1 on pp 158–159).
11. Kaye D.H.: Questioning a Courtroom Proof of the Uniqueness of Fingerprints, *International Stat Rev* 71(3):521–33 (2003)
12. Koons, R.; Basaglia, J. (2004), Forensic significance of bullet lead compositions, *Journal of Forensic Science* 50(2), paper ID JFS2004212.
13. Langenburg, Glenn (2011), Scientific research supporting the foundations of friction ridge examinations, Chapter 14 in *The Fingerprint Sourcebook*, <http://www.ojp.usdoj.gov/nij/pubs-sum/225320.htm>.
14. Laurie, C.; Weir, B.S. (2003), “Dependency effects in multi-locus match probabilities,” *Theoretical Population Biology* 63:207–219.
15. Mearns, G.S. (2010): The NAS report: In pursuit of justice, *Fordham Urban Law Journal*, December 2010.
16. Meier, Paul (1957), “Safety testing of poliomyelitis vaccine,” *Science* 125(3257):1067–1071.
17. Meier, Paul (1984), “The biggest public health experiment ever: The 1954 Field Trial of the Salk Poliomyelitis Vaccine,” *Statistics: A Guide to the Unknown* (J. Tanur et al., eds.), 88–100.
18. National Research Council, *The Evaluation of DNA Evidence*, National Academies Press, 1996.
19. National Research Council, *Forensic Analysis: Weighing Bullet Lead Evidence*, National Academies Press, 2004.
20. National Research Council, *Strengthening Forensic Science in the United States: A Path Forward*, National Academies Press, 2009.
21. Peele, E.R.; Havekost, D.G.; Peters, C.A.; Riley, J.P.; Halberstam, R.C.; Koons, R.D. (1991), Comparison of bullets using the elemental composition of the lead component, *Proceedings of the International Symposium on the Forensic Aspects of Trace Evidence*, June 24–28 (ISBN 0-932115-12-8), 57.
22. Peskin, A.; Kafadar, K.; Dima, A. (2010), “A quality pre-processor for biological cell images,” submitted.

23. Snedecor G.; Cochran, W. (1972), *Statistical Methods*, Iowa State University Press.
24. Spiegelman C.S.; Kafadar K: Data Integrity and the Scientific Method: The Case of Bullet Lead Data as Forensic Evidence, *Chance* 19(2):17–25 (2006)
25. Stacey, Robert B. (2004), “A report on the erroneous fingerprint individualization in the Madrid train bombing case,” *J. Forensic Identification* 54: 706–710.
26. Stigler, Stephen M. (1999), *Statistics on the Table*, Harvard University Press.
27. Weir, B.S. (2004), “Matching and partially-matching DNA profiles,” *Journal of Forensic Sciences* 49:1009–1014.
28. Weir, B.S. (2007), “The rarity of DNA profiles,” *The Annals of Applied Statistics* 1: 358–370.
29. Weir, B.S. (2009), “A proposal to study the CODIS Database,” White paper, Department of Biostatistics, University of Washington, September 30, 2009.
30. Wellek, S. (2003), *Testing Statistical Hypotheses of Equivalence*, Chapman and Hall, New York.
31. Zabell, Sandy L. (2006), “Fingerprint Evidence”, *Journal of Law and Policy*, 143–179
(http://www.brooklaw.edu/students/journals/bjlp/jlp13i_zabell.pdf).