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, 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.”

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. However, 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 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*:

  \[
  \text{Sensitivity} = P\{\text{analysis claims “match” } | \text{ true match}\}
  \]

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

  \[
  \text{False Negative Rate (FNR)} = P\{\text{analysis claims “no match” } | \text{ true match}\}.
  \]

- **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”? *Specificity* is defined as:

  \[
  \text{Specificity} = P\{\text{analysis claims “NO match” } | \text{ true NON-match}\}
  \]

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

  \[
  \text{False Positive Rate (FPR)} = P\{\text{analysis claims “match” } | \text{ true NON-match}\}.
  \]

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:

  \[
  \text{Positive Predictive Value PPV} = P\{\text{true match } | \text{ analysis claims “match”}\}
  \]

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?

**False Positive Call FPC = P\{true NON-match | 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:*

**Negative Predictive value NPV = P\{true NON-match | 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 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?

**False Negative Call FNC = P\{true match | 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). The consequence of this formula is that **high** sensitivity and specificity are needed for **high** PPV and NPV; i.e., for high probabilities of **correct** decisions.
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 mistakes out of 50 occurred, the upper 95% confidence limit 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.

Suppose that a diagnosis on a person is made: multiple sclerosis. The health professional tells the person to take aspirin. “Why?” the afflicted individuals sensibly asks. The health professional
offers these reasons:

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 for multiple sclerosis. Aspirin has been available for over a century and has been used in many cases. But that does not imply that prescription of aspirin provides the desired results in all cases. Moreover, the fact that five other medical doctors have prescribed aspirin would hardly provide the patient much assurance that aspirin is the “treatment of choice” for managing this extremely serious condition. In fact, most people are not likely to take a drug for a condition unless a well-designed randomized study confirmed that use of the drug resulted in significant improvements over the use of another medicine — regardless of how long that drug has been available.

Now, consider the arguments put forth for the use of latent fingerprint analysis as forensic evidence of individualization, 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.)

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 in Budowle et al. 2009, p.62.) Then the number of possible unique DNA signatures is \[36 \times 231 \times 21 \times \ldots \times 120 \times 153,\] making the number of possibilities huge (about \(8 \times 10^{21}\)). [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. 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.)

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; cf. Peskin et al. 2010).

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) are 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 studies of performance of the latent fingerprint process or ACE-V methodology, nor the error rates from the individual steps in the process, have been conducted.

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.
A common question to statisticians is: How many features are needed to provide a unique signature? For example, are 13 features always sufficient for individualization? The answer depends on several factors:

- 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 1 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 2.

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
Figure 1: 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 2: 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).
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 participate in the 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 in the 2nd trial, described below.

As a result of these concerns, a 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. So, if there were no difference in the effectiveness of the vaccine over placebo, 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 3 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 were 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 $3.7 \times 10^{-7}$ (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.)

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) are 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
Figure 3: 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 82 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.
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 ⇒ 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.”

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.

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

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 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. \textit{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. \textit{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. \textit{Conduct the experiment and calculate the metric(s).} For example, from the “best match” that is identified, one can calculate the relevant metrics.

5. \textit{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 \textbf{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 \textit{scientific method} (NAS report, Ch. 4), and the ways in which studies can be designed to better quantify the capabilities of the methods. \textbf{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 \textit{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., $0.01 = 1$ in 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:

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

$$NPV = \frac{\text{Spec} \cdot (1 - p)}{\text{Spec} \cdot (1 - p) + (1 - \text{Sens}) \cdot p}$$
Consequently, the “false discovery rate” (false positive call probability) and the “false non-discovery rate” (false negative call probability) are

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

\[
FNDR = 1 - NPV = \frac{(1 - \text{Sens}) p}{(1 - \text{Sens}) p + \text{Spec} (1 - p)}
\]

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

<table>
<thead>
<tr>
<th></th>
<th>‘Match’</th>
<th>‘No match’</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>True match</td>
<td>87</td>
<td>13</td>
<td>100</td>
</tr>
<tr>
<td>Non-match</td>
<td>3</td>
<td>97</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>90</td>
<td>110</td>
<td>200</td>
</tr>
</tbody>
</table>

- 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


