Statistical Issues in Forensic Science

Hal S. Stern

Department of Statistics, University of California, Irvine, California 92697-1250; email: sternh@uci.edu

Annu. Rev. Stat. Appl. 2017. 4:225–44

First published online as a Review in Advance on December 23, 2016

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi: 10.1146/annurev-statistics-041715-033554

Keywords

Bayes factor, likelihood ratio, pattern evidence, score-based inference

Abstract

Forensic science refers to the use of scientific methods in a legal context. Several recent events, especially the release in 2009 of the National Research Council (NRC) report *Strengthening Forensic Science in the United States: A Path Forward*, have raised concerns about the methods used to analyze forensic evidence and the ways in which forensic evidence is interpreted and reported on in court. The NRC report identified challenges including the lack of resources in many jurisdictions compared with the amount of evidence requiring processing, the lack of standardization across laboratories and practitioners, and questions about the analysis, interpretation and presentation of evidence. With respect to the last, the NRC report raises questions about the underlying scientific foundation for forensic examinations on some evidence types. Statistics has emerged as a key discipline for helping the forensic science community address these challenges. The standard elements of statistical analysis—study design, data collection, data analysis, statistical inference, and summarizing and reporting inferences—are all relevant. This article reviews the role of forensic evidence, the heterogeneity of forensic domains, current practices and their limitations, and the potential contributions of more rigorous statistical methods, especially Bayesian approaches and the likelihood ratio, in the analysis, interpretation, and reporting of forensic evidence.

1. INTRODUCTION

When a crime is committed, the investigation may identify a variety of types of evidence or information that can be used to help identify the criminal and potentially reconstruct the crime (e.g., by identifying the time of death or the sequence of events). This can include physical evidence at the crime scene (e.g., broken glass, hair/fiber, fingerprints, shoeprints, toolmarks, blood spatter, measurements of the victim) or digital evidence (e.g., video or voice recording). The analysis and interpretation of such evidence are the domain of forensic examiners. Examiners summarize their findings in a report that is used in the investigation and may be presented at a trial. Forensic evidence can be very powerful in courtroom settings, but several recent events have raised questions about the scientific foundation underlying the analysis and interpretation of forensic evidence. One example is a very public case in which fingerprint examiners from the Federal Bureau of Investigation (FBI) mistakenly identified Brandon Mayfield as the source of a latent fingerprint found at the scene of a 2004 train bombing in Spain. Other examples can be found in the work of the Innocence Project, a nonprofit legal organization founded in 1992, which has led to the freeing of more than 300 wrongfully convicted individuals through the beginning of 2016. Unreliable or improper forensic science was a contributing factor in roughly half of these cases. The 2009 National Research Council (NRC) report (NRC 2009) *Strengthening Forensic Science in the United States: A Path Forward*, prepared in response to a request by the US Congress, identified a number of challenges associated with the practice of forensic science. Similar concerns were expressed more recently in a 2016 report of the President's Council of Advisors on Science and Technology (PCAST 2016). Though addressing the concerns expressed in the NRC and PCAST reports is a multidisciplinary challenge, it is clear that the field of statistics has a significant role to play in improving the practice of forensic science.

1.1. Forensic Evidence and Forensic Examinations

Forensic examiners carry out a wide range of tasks including identifying the time of death when a dead body has been found, reconstructing a crime from blood spatter patterns at the scene, and comparing physical evidence at the crime scene with samples from a suspect. This article focuses on the last of these tasks, assessing evidence at the crime scene and determining whether it is consistent with the suspect (or an object in the suspect's possession). The assessment of such evidence is one place where statistical issues abound. It is important to note that in examining a single piece of evidence, we do not try to determine the guilt or innocence of the suspect, we merely try to identify whether the evidence is associated with the suspect. There may be any number of legitimate reasons for evidence from the suspect to be found at the crime scene (e.g., the suspect may have visited the crime scene at a different time).

There are a wide range of evidence types, which are briefly reviewed here along with the contexts in which they are often applied. In Sections 2 and 3, we provide a more detailed discussion of the role of probability and statistics in assessing the evidence. Among the most powerful forms of forensic evidence at the present time is DNA evidence obtained from a biological sample at the crime scene. The forensic examiner's task in this case is to determine if the DNA profile of a suspect matches the DNA profile found in the crime scene sample and to assess the significance of this agreement. The term trace evidence is used to refer to evidence types that can be characterized as a fragment or sample of a larger object that is left behind during the commission of a crime. This could include glass fragments from a broken window, hairs from an individual, or fibers from clothing or a carpet. The challenge here is to determine if a sample of trace evidence from the crime scene matches another sample obtained from a suspect (or perhaps from an object in the

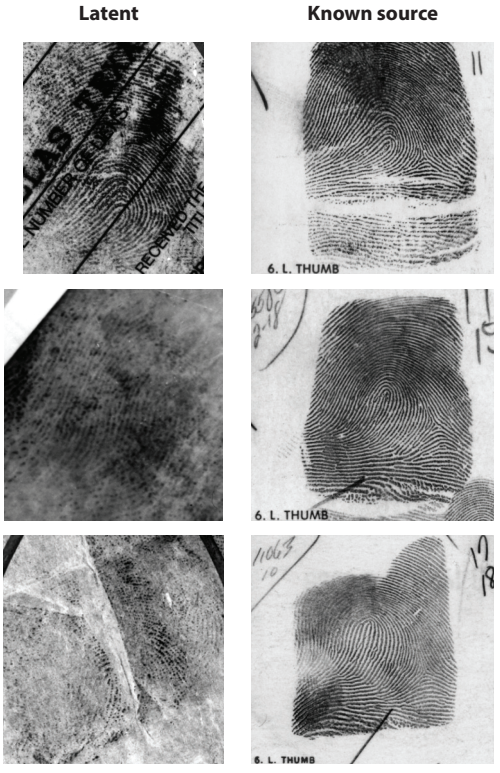

Figure 1

Latent fingerprints (*left column*) and matching images from known sources (*right column*) made available as part of National Institute of Standards and Technology's Special Database 27 (Garris & McCabe 2000). The top row is an example identified as a good latent print, the middle row an example identified as bad, and the bottom row an example identified as ugly.

suspect's possession). Pattern evidence refers to evidence left at the crime scene that is the result of an impression left by a person or object. The forensic examiner must attempt to determine if the pattern of the evidence found at the crime scene matches the pattern of an analogous sample obtained from the suspect or an object known to belong to the suspect. Types of pattern evidence include fingerprints, shoeprints, documents/handwriting, toolmarks, and firearm impressions. **Figure 1** presents examples of fingerprint evidence from the National Institute of Standards and Technology (NIST) Special Database 27 (Garris & McCabe 2000). Three example latent prints are shown along with the matching exemplar from a tenprint card taken under ideal conditions. The examples illustrate the wide range of latent prints that examiners encounter. Increasingly, digital evidence (e.g., video images from a security camera, audio recordings) is becoming relevant. In many ways the analyses of these types of digital evidence resemble analyses of pattern evidence. The evidence types vary considerably in the data that they yield and in the strength of the scientific literature that underlies the methods used to analyze and interpret the evidence.

The use of expert testimony regarding forensic evidence in federal courts is governed by guidelines established in the court case *Daubert v. Merrill Dow* (1993) and known as the Daubert standard. The Daubert decision identified several factors that should be considered by a judge in deciding whether to allow expert scientific testimony about a given form of scientific evidence. The factors include determining whether the forensic approach follows the scientific method,

is generally accepted in the scientific literature, and provides information about the error rate associated with the approach. Though none of these characteristics is explicitly required for expert testimony to be admitted, it is clear that the ruling asks the judge to assess the underlying scientific basis for the testimony. As an extreme example, there appears to be little evidence at the present time that forensic examiners can reliably assess whether a particular suspect is the source of a bite mark left on a victim (Saks et al. 2016). Though such testimony has been admitted for more than 30 years, jurisdictions are increasingly concerned that bite-mark analysis does not satisfy the Daubert standard (see, e.g., Eckholm 2016). Though many states have adopted the Daubert standard, several states, including California and New York, have not. Expert testimony in state and local courts in such states continues to be governed by the Frye standard (*Frye v. United States* 1923), which allows expert opinion based on a scientific technique only if the technique is generally accepted in the relevant scientific community. The Frye standard does not provide explicit standards for a judge to consider in determining whether to allow expert testimony beyond the notion of general acceptance.

Forensic practice for reporting the results of forensic examinations varies considerably across the many forensic disciplines. Thus, although quantitative assessments of the strength of the evidence (e.g., using a likelihood ratio) are common for DNA (at least for single-source samples), they are rarely utilized in other disciplines. Common practice for a wide range of trace and pattern evidence disciplines is for examiners to report their conclusion as an expert opinion regarding the evidence. For many disciplines, such as firearms and toolmarks, examiners will report one of three outcomes after their analysis: They will report that it is their expert opinion that the suspect is the source of the crime scene evidence (known as an identification), that in their opinion the suspect is definitely not the source of the crime scene evidence (known as an exclusion), or that their analysis is inconclusive in that it does not reach either of the other two opinions. Some disciplines allow for intermediate conclusions; for example, a forensic document examiner may report in a questioned document examination that the suspect was probably the source of the questioned evidence. Understanding the scientific basis for these expert opinions is a key issue raised by the NRC report.

1.2. Controversies and Errors

The NRC (2009) report identified several challenges facing the forensic science community. It cited the lack of standards for forensic examiners, which allows for considerable heterogeneity in the practices of examiners; concerns about the literature supporting the conclusions and interpretations being drawn; the need for additional resources to address the large number of samples that need to be processed; and considerable heterogeneity across jurisdictions in the resources provided. The report raised important questions about the science underlying a number of forensic science disciplines. In this respect it built on a number of earlier reports related to assessing evidence including a report on polygraphs and lie detection (NRC 2003), a report on bullet lead evidence (NRC 2004), a report on ballistic imaging (NRC 2008), and a report on the role of statistical assessments in evidence (Fienberg 1989). The 2009 NRC report and some of the highly visible forensic science errors mentioned earlier have led to increased attention on forensic science in recent years. The National Commission on Forensic Science (NCFS) was created in 2013 to provide advice to the US Attorney General regarding issues in forensic science. The Organization of Scientific Area Committees for Forensic Science (OSAC), a set of boards, committees, and subcommittees focused on different disciplines and other aspects of forensic science, was established in 2014 to produce standards and guidelines that define the best practices in each forensic science discipline. The various committees and subcommittees include practitioners representing

local, state, and federal agencies, as well as academic and other researchers. A cross-disciplinary statistical task group within the OSAC reviews all proposed standards and guidelines, and in that role, it is considering many of the issues reported on here. The National Institute of Justice has provided research funding focused on improving the methods and technology used for forensic science. The NIST has also increased its focus on forensic research, both in its intramural research program and through the funding of a National Center of Excellence.

The work of the Innocence Project, described briefly above, indicates that faulty interpretation or testimony regarding forensic evidence definitely occurs and can play a role in wrongful convictions. There do not exist published data on how often such problems occur, but there is increasing attention to the question. A recently completed study by the FBI in conjunction with the National Association of Criminal Defense Lawyers and the Innocence Project (FBI 2015) reviewed FBI testimony regarding microscopic hair comparisons in cases that led to a conviction prior to the year 2000, and for which a positive hair association was provided by the examiner. The review identified three different types of examiner statements that were judged to exceed the limits of the underlying science and then reviewed examiner testimony to assess how often such statements were made. They found such testimony in more than 95% of the initial set of 268 trials reviewed. The Department of Justice is currently developing a plan to assess forensic examiner testimony for a range of disciplines in more recent years. The proposed review is aimed at assessing whether testimony complies with best practices and learning how to improve the treatment of forensic evidence.

Another concern expressed in the 2009 NRC report is whether the forensic practitioner community has a full appreciation of the role of uncertainty in forensic examinations. For many years (through the early 2000s), it was common for latent print examiners to support a claimed identification by noting that the process they followed had zero error rate. Another popular claim was that the source of a print was identified to the exclusion of all other people that had ever lived or ever would live. Those who work in scientific disciplines and appreciate the role of uncertainty know that such claims are not credible. Recent studies (Ulery at al. 2011) have demonstrated a low but nonzero misidentification error rate for latent fingerprint examiners. In other forensic disciplines, it is common to have examiners testify to a "reasonable degree of scientific certainty." This language was recently criticized by the NCFS because it does not have a standard definition and might confuse or mislead jurors (NCFS 2016). Recent discussion within many of the OSAC subcommittees is focused on developing appropriate ways to testify about expert opinions regarding forensic evidence without going beyond the results available in the scientific literature.

More recently, while this article was being prepared for publication, the President's Council of Advisors on Science and Technology (PCAST) issued a report to the President (PCAST 2016) in response to the President's question about whether there were scientific steps that could be taken to enhance the state of forensic science in the US legal system. PCAST reviewed published studies and consulted with a wide range of stakeholders of the forensic science community. The report presents their conclusions about what is required to assess the scientific validity of forensic science pattern comparison methods and applies their criteria to several disciplines including fingerprint analysis, shoeprint/footwear analysis, bitemark analysis, firearms analysis, and DNA analysis. The report finds that the existing literature does not provide support for the scientific validity of shoeprint analysis, bitemark analysis, firearms analysis, and DNA analysis of complex mixtures, and it encourages additional peer-reviewed studies in these areas.

1.3. The Role of Statistics in Forensic Science

The concerns expressed in the NRC report are quite broad and require the attention of a range of scientific communities, including chemists, physicists, biologists, physicians, psychologists,

anthropologists, and others. The report also makes clear that when it comes to analysis and interpretation of forensic evidence, there is the need for increased use of statistical methods. The bulk of this article is focused on the technical ways in which statistical methods can be used to quantitatively assess the strength of particular types of forensic evidence. Before focusing on that, however, it is important to briefly address some of the other ways in which statistical concepts and methods can play a role in improving forensic science.

1.3.1. Reliability and validity. At present it is only possible to quantify the strength of evidence in very limited settings, with DNA being the evidence type for which statistical methods are most well developed. The reasons for this are discussed further in Section 3. In the disciplines for which formal quantitative measures of the strength of evidence do not exist, assessing the reliability and validity of forensic examinations is a critical need. Validity refers to the accuracy of the conclusions drawn by a forensic examiner. If a fingerprint examiner concludes that a latent print at the crime scene comes from the same source as a test impression made by the suspect (this conclusion is sometimes known as an identification), then it is natural to want to know the validity or accuracy of that conclusion. It is also natural to wonder about the reliability of the process: Would the same examiner draw the same conclusion if presented with the same evidence at a different time? Would a different examiner draw the same conclusion if presented with the same evidence? Studies to address such questions require that statisticians work with practitioners and other experts on study design and analysis. This work is only now beginning with the recent studies of latent print examiners providing noteworthy examples (Ulery et al. 2011, 2012).

1.3.2. Task-relevant information and cognitive biases. It has been demonstrated in a wide range of settings that human judges are prone to cognitive biases (Dror 2015). The term bias in this instance refers to a difference between observed behavior and what one might think of as rational decision making. Examples include framing effects, where the same question may receive different answers from respondents depending on other, often irrelevant, information that is provided, and confirmation bias, where individuals tend to favor interpretations that confirm their own preconceptions about an issue. Statisticians can play a role in teams that study cognitive bias and in teams that try to determine what information should be deemed relevant for a particular task.

1.3.3. Causal inference. Forensic examiners in disciplines like crime scene investigation, arson, and blood spatter analysis attempt to reconstruct a crime based on evidence found at the crime scene, which can be viewed as attempts to infer the causes of observed effects. This can be a challenging task because it is difficult to carry out realistic controlled experiments that would allow one to reliably distinguish between competing explanations (e.g., fires that develop naturally versus those that use an accelerant). Statistical collaboration with practitioners in relevant disciplines will be valuable in strengthening inferences in these settings.

1.3.4. Case processing and procedures. Any process can benefit from careful analysis to understand potential limitations, bottlenecks, or sources of error, and the forensic evidence analysis process is no exception. Indeed, the large workload in many crime laboratories leads to backlogs that can slow the treatment of evidence. At the same time, maintaining quality forensic examinations requires a quality assurance program that incorporates reanalysis or verification of some conclusions. Statistical methods can play a role in designing quality assurance programs that can improve efficiency of lab operations while simultaneously insuring the accuracy of conclusions. For example, statistical studies of reliability and validity may show that relatively easy comparisons

require only occasional random verifications, whereas more complex comparisons require blind verification by another examiner.

1.3.5. Testifying on forensic evidence. The remainder of the article focuses on the potential to generate rigorous quantitative evaluations of forensic evidence. A related area of research is how such analyses of forensic evidence should be presented in court. There have been studies demonstrating the difficulty of understanding likelihood ratios and Bayes factors for jurors (Martire et al. 2013, Thompson & Newman 2015). Statisticians have an important role to play in developing approaches to presenting quantitative evaluations of evidence and in the design and analysis of juror studies.

2. A STATISTICAL FRAMEWORK FOR ASSESSING FORENSIC EVIDENCE

2.1. Evidence

The primary focus of this article is on the likelihood ratio and its role in the assessment of forensic evidence. Let E denote the evidence being considered by an examiner. As one example, E might be the DNA profile obtained from a biological sample obtained at the crime scene and the DNA profile of a suspect. In another case, E might refer to measurements of the chemical composition of glass fragments found at the crime scene and similar measurements from glass fragments found on the clothing of a suspect. In the case of pattern evidence, E usually refers to an image or impression (e.g., a fingerprint, a shoeprint) found at the crime scene and an analogous image or impression gathered from a potential source (e.g., the suspect's finger, the suspect's shoe). In the next section, it will be convenient at times to think separately of the evidence from the crime scene (E_x) and the evidence from the suspect (E_y), but in this preliminary discussion we will use E to represent both. Of course, there can be more than a single piece of evidence in a case. This article focuses on evaluating the evidence for a single evidence type; we return to the question of how multiple types of evidence might be handled in Section 4.

2.2. Hypotheses

The evidence E needs to be evaluated in the context of two alternative hypotheses or propositions regarding the source of the crime scene evidence. The first hypothesis, commonly called the "same source hypothesis," is that the suspect is the source of the evidence found at the crime scene. For example, it may mean that the shoeprint found at the crime scene and an image taken of the suspect's shoe both come from the same source (the shoe in question). "Same source" is not always precisely the correct terminology. For example, the question in a ballistics exam may be whether the markings on a bullet casing found at the crime scene match those that would be expected from a gun found in the possession of the suspect. Thus it is more appropriate to ask if the gun is the source of the bullet casing. Even here, however, that determination would be made by comparing the casing of unknown origin (found at the crime scene) with a casing from a test fire of the suspect's gun and trying to determine if these two casings have the same source. The same source hypothesis is often also referred to as the prosecution hypothesis, and it is denoted as H_p below. The second hypothesis is the "different source" hypothesis, sometimes known as the defense hypothesis H_d. Under this hypothesis, the evidence found at the crime scene does not arise from the suspect (or from the same source as the evidence provided by the suspect). Any observed similarities in this case must be due to chance. For purposes of this article, we assume that H_p

and H_d are mutually exclusive and exhaustive hypotheses. This is generally the case considered in forensic examinations, but there are examples in which it is not the case.

It is important to emphasize that the hypotheses H_p and H_d do not refer to the guilt or innocence of the suspect. That determination is ultimately made by the jury based on a complete evaluation of the entire collection of evidence, and the terms guilt and innocence are not relevant to the assessment of a single type of evidence. It is also important to recognize that the discussion above considers the hypotheses in their most general form—the specific hypotheses in a given case may in fact be different than those described above. One possible situation is that the suspect may be known to come from a limited pool of suspects (i.e., only those individuals present at a party), in which case the defense hypothesis can be more specific than just "different source" and would likely refer to the remaining candidates.

2.3. Contextual Information

In addition to the evidence E, there may also be additional information I available that should be considered in evaluating the evidence. This can include details about the way the evidence was collected (e.g., the chemical substrate used to obtain a latent print at the crime scene) or information about the population distribution of various characteristics (e.g., allele frequencies in a DNA analysis). What can and should be included in I is itself a source of discussion in the forensic science community. If I were to include information about other evidence in the case or other local circumstances, then this could open up the possibility that such task-irrelevant information might influence the examiner (even subconsciously). The recently established NCFS has issued a guidance document that addresses this issue (NCFS 2015).

2.4. The Bayesian Framework and the Likelihood Ratio

Bayes' Theorem provides a logical framework for evaluating the evidence regarding the two competing hypotheses described in Section 2.2. In forensic science this idea dates back at least to Lindley (1977). Bayes' Theorem can be most conveniently written for forensic evidence in the form

$$\frac{\Pr(H_p|E,I)}{\Pr(H_d|E,I)} = \frac{\Pr(E|H_p,I)}{\Pr(E|H_d,I)} \frac{\Pr(H_p|I)}{\Pr(H_d|I)}. \tag{1}$$

This expression indicates that the posterior odds of the two hypotheses are determined as the product of the a priori (pre-evidence) odds and the term $\Pr(E|H_p,I)/\Pr(E|H_d,I)$, which is known as the likelihood ratio or Bayes factor (see below for more on the terminology). Throughout this article we write the likelihood ratio in terms of probabilities and use the notation Pr; it should be understood that this is intended to refer to an appropriate probability distribution for the relevant evidence (i.e., either a probability mass function or a probability density function). The likelihood ratio measures the relative probability of obtaining the evidence E under the two hypotheses. A large likelihood ratio means that the observed evidence is much more likely under the prosecution hypothesis than under the defense hypothesis. An extremely small likelihood ratio means that the observed evidence is much less likely under the prosecution hypothesis. The equation above tells an evaluator of the evidence how to modify her prior odds in the face of the evidence E to obtain the posterior odds of the two hypotheses. The forensic examination can ideally provide enough information to supply the likelihood ratio, and that information would be shared with the individual or individuals charged with assessing the evidence and forming a conclusion about the evidence (and about the ultimate guilt or innocence of the suspect). There

is considerable uncertainty about whether jurors do—or even can—process information in the way described by the formula above (see, e.g., Martire et al. 2013, Thompson & Newman 2015).

Using the likelihood ratio as a quantitative summary of the evidence is sometimes known as a Bayesian approach to the analysis of forensic evidence because of the likelihood ratio's appearance in Bayes' theorem. This has caused some confusion for several reasons. One reason is because there are many ways in which the term Bayesian is used. The term Bayesian can reasonably be used to refer to any inferential procedure that relies on Bayes' theorem to draw its inference by combining prior information with observed data. This justifies thinking about the likelihood ratio as being central to a Bayesian approach to forensic inference. However, this is confusing for some, because the likelihood ratio by itself is not a sufficient summary of a fully Bayesian inference about the hypotheses under evaluation, as it does not address the role of the prior probabilities assigned to the two hypotheses. Fenton et al. (2016) discussed some of the reasons why fully Bayesian inference has not played a larger role in the legal system. Another source of confusion in referring to the likelihood ratio as a Bayesian approach is that the likelihood ratio is definitely not an exclusively Bayesian concept. The likelihood ratio features prominently in traditional approaches to inference, for example, in the likelihood ratio test statistic often used in developing hypothesis tests for nested hypotheses (Lehmann & Romano 2008). One final potential source of confusion relates to a bit of subtlety in the terminology that is related to the treatment of nuisance parameters in the likelihood ratio. For many evidence types, the likelihood ratio will depend on unknown parameters such as the mean and variance of the distribution of chemical concentration measurements for a sample of glass fragments, or the population frequency of a specific allele in a DNA analysis. Such parameters can be estimated using traditional frequentist-justified methods (in which case the ratio is typically referred to as the likelihood ratio) or by averaging over the distribution of the unknown parameters in a Bayesian analysis (in which case the ratio is often referred to as the Bayes factor). Thus, the likelihood ratio itself can be estimated using Bayesian or frequentist approaches. For the remainder of this article, we discuss the likelihood ratio in the context of different types of forensic evidence. We consider different approaches (Bayesian or non-Bayesian) to estimating the likelihood ratio as relevant for a particular evidence type.

2.5. Alternatives to the Likelihood Ratio

Though this article focuses primarily on the likelihood ratio approach to assessing forensic evidence, it is important to note that there are alternatives. It is especially important at the present time, when the probability models and reference databases required to develop formal likelihood ratios are not available for many forensic disciplines, to understand what alternative approaches are possible. One alternative approach is to consider forensic procedures that are used by examiners to reach decisions and assess the performance of such procedures. Typically the decisions considered are an identification (a decision of same source), an exclusion (a decision that the crime scene evidence does not come from the suspected source) or a finding that the comparison is inconclusive. The decision in a specific case can be based solely on expert opinion expressed by the examiner, as is common now, or it may in the future depend on a categorization determined by a similarity score. For such procedures it is possible to estimate error rates through careful validation studies. The emphasis is often on the false positive rate (incorrect identifications) and the false negative rate (incorrect exclusions). A study of this type in the fingerprint domain was published by Ulery et al. (2011). When forensic conclusions are based on a similarity score (assessing the similarity between the crime scene sample and the known sample from the suspect) it is possible to consider a range of decision procedures, as the threshold for declaring an identification is varied. Such procedures can be evaluated or assessed using receiver operating characteristic (ROC) curves that

measure the true positive (correct identification) rate (also known as the sensitivity) and the false positive rate as the decision threshold is varied. ROC curves are well established in the context of medical diagnostic procedures and are supported by an extensive peer-reviewed literature (see, e.g., Zweig & Campbell 1993, Pepe 2003).

3. LIKELIHOOD RATIOS FOR FORENSIC EVIDENCE

There are several issues that must be addressed before the likelihood ratio can be developed for a given type of forensic evidence. It must be possible to probabilistically describe the likelihood of observing the evidence under each hypothesis. For both the numerator and the denominator, this requires an understanding of the process by which the data are produced. This can be a biological process in the case of DNA evidence, or a manufacturing process in the case of glass fragments or shoeprints. In addition, the denominator requires some knowledge of the distribution of evidence measurements in a broader population. Finally, as indicated by the Daubert standard, there is a need for some evidence that the scientific community accepts the resulting models and calculations, for example through peer-reviewed publications. The plausibility of using likelihood ratios for forensic evidence varies considerably across the forensic disciplines. This section considers three distinct evidence types that cover the current range of support for the use of likelihood ratios: DNA evidence, for which likelihood ratios are currently computed and accepted in many common scenarios; glass fragments and other types of trace evidence, for which there is a peer-reviewed literature indicating how likelihood ratios can work but for which there has not been widespread adoption in practice; and fingerprints and other types of pattern evidence, for which there are initial efforts underway to develop likelihood ratios but for which there remain significant challenges.

3.1. Likelihood Ratios for DNA Evidence

Likelihood ratios are best established in the forensic analysis of DNA profiles. There is an extensive literature that describes the approach (NRC 1996, Steele & Balding 2014). A brief review is provided here to show clearly what is needed to effectively apply the likelihood ratio framework—this can help to identify the elements that need to be addressed to apply the framework in other disciplines. The evidence E in this case comprises the DNA profile of the biological sample found at the crime scene (and presumed to belong to a single source, the guilty party) and the DNA profile of the suspect, where a DNA profile identifies the alleles at a set of specific locations along the genome where there is known to be heterogeneity in the population. There are a few different types of DNA evidence (short tandem repeat analysis, mitochondrial analysis, etc.). We do not delve into the details of the different approaches because the focus here is on the statistical issues involved in analyzing the resulting evidence. It is convenient to write the evidence as $E = (E_x, E_y)$, where E_x is the evidence obtained from the crime scene sample (the sample with an unknown source) and E_y is the evidence obtained from a suspect (the sample with a known source). Each profile is composed of the alleles identified at p locations along the genome, so we can write $E_x = (E_{x1}, E_{x2}, \ldots, E_{xp})$ and $E_y = (E_{y1}, E_{y2}, \ldots, E_{yp})$ where E_{xi} denotes the alleles identified at the ith location for sample x. Each individual has two alleles at each location, one inherited from each biological parent. In short tandem repeat analyses carried out using the FBI's Combined DNA Index System (CODIS), the evidence comprises the alleles identified at $p = 13$ different locations. The likelihood ratio (LR) can be expressed as

$$\mathrm{LR} = \frac{\Pr(E_x, E_y|H_p)}{\Pr(E_x, E_y|H_d)} = \frac{\Pr(E_x|H_p)}{\Pr(E_x|H_d)} \frac{\Pr(E_y|E_x, H_p)}{\Pr(E_y|E_x, H_d)} = \frac{\Pr(E_y|E_x, H_p)}{\Pr(E_y|E_x, H_d)} \quad (2)$$

where the probability of observing the crime scene evidence E_x is assumed to be the same under H_p and H_d. Assuming the information at the locations along the genome is independent, a plausible assumption for CODIS profiles, one obtains

$$\mathrm{LR} = \prod_{j=1}^{p} \frac{\Pr(E_{yj} | E_{xj}, H_p)}{\Pr(E_{yj} | E_{xj}, H_d)}. \tag{3}$$

Evaluation of the value of the likelihood ratio for a single location can be fairly complex because, for maximum precision, the relevant probabilities should be adjusted for the possibilities of shared ancestries in the database used to identify allele frequencies and the samples under evaluation. Detailed descriptions can be found in Steele & Balding (2014) or Butler (2015). We focus on a simple version of the calculation here, as this makes the likelihood ratio concept most clear. The numerator of the likelihood ratio is the probability of observing a given pair of alleles at location j in the suspect sample given that (*a*) alleles E_{xj} are observed in the crime scene sample and (*b*) the suspect is the source of the crime scene sample. This probability (again ignoring some complications) is one if $E_{yj} = E_{xj}$ and zero otherwise. The likelihood ratio is thus only relevant if the crime scene and suspect samples share the same alleles. The denominator of the likelihood ratio is the probability that a random person from the relevant population will have a genotype that matches the suspect at location j. This is easily calculated given data regarding the population frequencies of the different possible alleles. If the sample is homozygous (two copies of the same allele) with an allele having population frequency p, then the probability of a random match is p^2. If the sample is heterozygous (two different alleles) with allele frequencies p_1 and p_2, then the probability of a random match is $2p_1p_2$, with the factor of two needed to account for the two ways these can be inherited from the two parents. The likelihood ratio thus depends on unknown parameters, the frequencies of different alleles in the population. Fortunately these are reasonably well determined from a variety of population studies.

The likelihood ratio for the full DNA profile is then obtained as the product of a series of terms, one for each location along the genome that is being considered. In our simple description, the numerator is one if the two samples match on every set of alleles and zero otherwise. The likelihood ratio is thus determined by the denominator, which is sometimes known as the random match probability. If the random match probability is low, then the likelihood ratio is high. In fact, for DNA evidence, extremely large likelihood ratios (e.g., in the billions or even larger) are not uncommon.

Likelihood ratios are well accepted in the analysis of single-source DNA evidence for several reasons. The underlying biology of genetic inheritance is well understood and leads to a well-understood probability model for the likelihood of matching DNA profiles. The unknown elements of the model (i.e., the population frequencies of different alleles) are available from published databases. There is a rich peer-reviewed scientific literature that supports the analysis [see, e.g., Steele & Balding (2014) and the references therein]. This should not be taken as a blanket endorsement of likelihood ratios for DNA evidence in all situations.

There are a few challenges associated with the analysis of DNA evidence. First, the technology that amplifies and records the length of DNA fragments must be read to identify the alleles present in a sample, and this step can require subjective decisions. Second, DNA technology has improved to the point that it is possible to reliably identify DNA profiles from smaller samples of biological substrate. This is a positive development for the field but it also increases the chance that so-called touch DNA resulting from the transfer of a small number of cells from one surface to another may contaminate a sample. Finally, though samples known (or believed) to have resulted from a single source can be analyzed effectively using the approach described above, it is still challenging to determine whether a given suspect is a contributor to a sample that is known (or believed) to represent

a mixture of genetic material from multiple individuals. There remains considerable uncertainty about how to analyze such mixtures, and interlaboratory studies have demonstrated considerable variability in the conclusions reached by different labs for a given sample. One recent comparison of different statistical approaches in the mixture setting is that of Kelly et al. (2014). A contributing factor to the continuing confusion in this arena is the availability of competing computer programs for which source code is unavailable. DNA mixture analysis is an area of ongoing research.

3.2. Likelihood Ratios for Trace Evidence

Trace evidence refers to materials that can be transferred during the commission of a crime. This is a broad class of evidence type, as it includes hair, fibers, soil and glass. The analysis of these types of evidence varies quite a bit depending on the specific type of trace evidence. Here, we use the example of glass, for which the actual evidence comprises measurements of the concentrations of various chemical elements in the sample (or ratios of chemical concentrations). For other types of trace evidence like hair and fibers, the forensic examination more closely resembles the analysis of pattern evidence (which is described in more detail below). Glass evidence E comprises measurements on multiple glass fragments found at the crime scene (E_x) and multiple glass fragments found on a suspect (E_y). We can write $E_x = (x_1, \ldots, x_m)$ where x_i is a vector of element concentrations for the ith crime scene sample and $E_y = (y_1, \ldots, y_n)$ where y_j is a vector of element concentrations for the jth suspect sample.

The standard approach for assessing glass evidence is to compare the population mean concentrations for the population of glass fragments that provided the crime scene evidence E_x and the population mean for the population of glass fragments that provided the suspect evidence E_y using a standard statistical significance test (or related procedure) (Almirall & Trejos 2006). Failure to reject the hypothesis of equal population means is often said to indicate that the two sets of glass fragments are indistinguishable. The number of samples and the threshold required for declaring a difference to be significant can be set to achieve desired performance in terms of false positive and false negative decisions regarding the hypothesis of equal means. If the population means can be distinguished, then this provides reasonably clear evidence that the two samples do not have the same source. The interpretation of results when the population means cannot be distinguished is more nuanced. If the hypothesis of equal population means cannot be rejected, then the two samples cannot be distinguished statistically, but this does not directly address the two hypotheses (H_p and H_d) described above in talking about the likelihood ratio approach. Samples that fail to reject the hypothesis of equal means may be obtained (and are to be expected) when the two samples have the same source, but such results may also be obtained when the two samples have different but similar sources. A complicating factor is that more substantial measurement error or more heterogeneity among fragments from the same source, neither of which should favor one hypothesis over the other, will make it more likely to find two samples indistinguishable. For this reason, the way that a failure to reject the null hypothesis (i.e., a finding of indistinguishable populations) is presented in reports and testimony is critical. If the significance of the finding is overstated, then there is a chance that the burden of proof may shift from the prosecution to the defense.

The use of likelihood ratios for trace evidence is not as well established as for DNA evidence, but there has been progress in the case of glass fragments. The likelihood ratio is defined in the usual manner,

$$\mathrm{LR} = \frac{\Pr(E_x, E_y | H_p)}{\Pr(E_x, E_y | H_d)}. \tag{4}$$

The key to developing the likelihood ratio in this case is some knowledge about the variability among chemical concentration measurements obtained from a sample of glass fragments obtained from a single source (i.e., the same window) and some knowledge about the variability among chemical concentration means for samples obtained from different sources. The former describes the variation under H_p and the latter contributes to the variation to be expected under H_d.

Though a careful treatment of glass evidence is beyond the scope of this article, we can learn a great deal by considering a simple plausible starting point that is described by Aitken & Lucy (2004). Initially, suppose that the $E_x = (x_1, \ldots, x_m)$ are a sample of exchangeable measurements from the crime scene window, and assume that these can be plausibly modeled as Gaussian (perhaps after some transformation) with mean μ_x and variance matrix Σ_w, with the subscript w denoting this as describing the variation among fragments from within the same piece of glass. The description here does not explicitly separate out variation due to the measurement process and variation due to heterogeneity within a single glass source. The sample mean $\bar{x}$ is a sufficient statistic for the evidence, and under the Gaussian model $\bar{x}|\mu_x, \Sigma_w \sim N(\mu_x, \Sigma_w/m)$ where $N(\alpha, V)$ denotes the Gaussian distribution with mean α and variance matrix V. As different glass sources will have different chemical concentrations, we can model that variation by considering μ_x as a draw from a normal population of glass sources with $\mu_x \sim N(\theta, \Sigma_b)$, where the subscript b denotes this as describing variation between different sources. Integrating over the population distribution yields a normal distribution for $\bar{x}$ that reflects both the within and between variance, $\bar{x}|\theta, \Sigma_w, \Sigma_b \sim N(\theta, \Sigma_w/m + \Sigma_b)$. The marginal distribution for the evidence E_y is defined in analogous fashion, with the result that $\bar{y}|\theta, \Sigma_w, \Sigma_b \sim N(\theta, \Sigma_w/n + \Sigma_b)$. There is, however, a key difference in the joint distribution of E_x, E_y under the two hypotheses. In the numerator, the two sample means are not independent, because they share a common source so that the joint distribution must account for the covariance of $\bar{x}$ and $\bar{y}$, which is equal to Σ_b. In the denominator, the two sample means are independent because they come from different sources. As in the DNA example, the likelihood ratio depends on unknown parameters: θ and Σ_b describe the variation among chemical concentration means for glass sources from the relevant population, and Σ_w describes the variation among repeated measurements from a single source. As mentioned at the outset, the uncertainty regarding these parameters can be treated in a Bayesian fashion by assigning prior distributions (perhaps developed through the analysis of earlier data sets) or through more standard point estimates for mean and variance parameters. In either case, however, there is some subtlety to how the inference is carried out—for example, whether one should sample glass fragments from a set of distinct glass samples (e.g., windows) obtained from a given neighborhood, a given city, the entire country, et cetera. The population of interest is not as well defined in this context as it is for the DNA example described in the previous section. Models similar to the one described here for glass evidence have been applied in the context of copper wire (Dettman et al. 2014) and bullet lead (Carriquiry et al. 2000).

As a summary, we note that for trace evidence characterized by chemical concentrations or other continuous measurements, it is possible to develop likelihood ratios. A key requirement is to develop plausible probability models for the variation observed within a single source and across multiple sources. The discussion above assumes a simple form for these distributions, but as demonstrated by Aitken & Lucy (2004), it is possible to make less restrictive assumptions. A limiting factor at present is the challenge of finding reliable representative data for characterizing population variability.

3.3. Likelihood Ratios for Pattern Evidence

Considerable public attention after the publication of the NRC (2009) report has focused on the forensic examination of pattern evidence. Pattern evidence refers to evidence produced when

one object comes in contact with another and leaves an impression. This includes fingerprints, shoeprints, toolmarks, tire treads, bitemarks, and impressions left on bullet casings or bullet fragments. It also is generally considered to include handwritten or typewritten documents. The key feature of all of these types of evidence is that a pattern or impression left at a crime scene whose source is unknown must be compared with another impression obtained from a known source (often a suspect). When viewed in those terms, the set of evidence types that might be considered pattern evidence grows even further, as some types of digital evidence (e.g., voice recordings, photos of a suspect) can also involve this kind of comparison. The analysis of pattern evidence has generally not been amenable to quantification and evaluation through the likelihood ratio approach, although this is beginning to change.

The standard approach for analyzing pattern evidence is perhaps best described by considering the case of fingerprint examination (see, e.g., Expert Work. Group Hum. Factors Latent Print Anal. 2012, Neumann & Stern 2016). We provide a short summary here. The analysis begins with an examination of the crime scene impression, E_x, which is typically a partial fingerprint, also known as a latent print, often of relatively low quality. Forensic analysts examine the impression to assess if they can identify a sufficient number of features (e.g., ridge bifurcations, ridge endings) of high enough quality to proceed with the analysis. The identified features are marked on the latent print. The examiner then compares the crime scene impression with the impression obtained directly from the suspect, E_y, with the aim of identifying whether the same features are present in roughly the same constellation on the suspect impression as they were found on the crime scene sample. The examiner then evaluates whether the observed correspondences are sufficient in number and quality to identify the two impressions as having come from a common source. In this respect, the ultimate conclusion is the judgment of the expert rather than a statistical measure or the result of a statistical test. This approach to fingerprint examination is known as ACE-V with the letters identifying the phases of the examination: analysis, comparison, evaluation, and verification. The final stage refers to the practice of having another examiner in the lab confirm the first examiner's conclusion.

There are several issues associated with the ACE-V approach that are worthy of mention. One is that there are examples of examiners cycling back and forth between the two prints and looking for features to match, which can increase the chance of an incorrect decision. Indeed, a report by the US Office of the Inspector General (OIG) (Fine 2006) identifies this as a contributing factor in the mistaken identification of Brandon Mayfield as the source of a latent print found at the scene of the 2004 Madrid train bombing. A second issue relevant to latent print analysis, though not specific to the ACE-V approach, concerns the potential for finding nonmatching fingerprints that resemble a given print when large database searches (using an automated fingerprint identification system) are used to identify potential sources of a latent print. This too was found by the OIG report to be a contributing factor in the erroneous Mayfield identification. **Figure 2** presents images from the OIG report showing the correspondence between features identified on the latent print and analogous features on the exemplar prints of Brandon Mayfield and the believed true source, Ouhnane Daoud. Finally, it is worth noting that there are studies suggesting the potential for latent print examiners to be affected by cognitive biases (Dror et al. 2006). This is not terribly surprising, as such biases are common when humans make judgments.

The likelihood ratio approach is an attractive one for quantifying the probative value of pattern evidence and for ensuring that the plausibility of the two impressions under each of the relevant hypotheses is considered carefully. In the standard approach to pattern evidence, the examiner implicitly addresses the relative likelihood of matching features under each hypothesis, but this is assessed primarily through the lens of the examiner's experience rather than through a formal

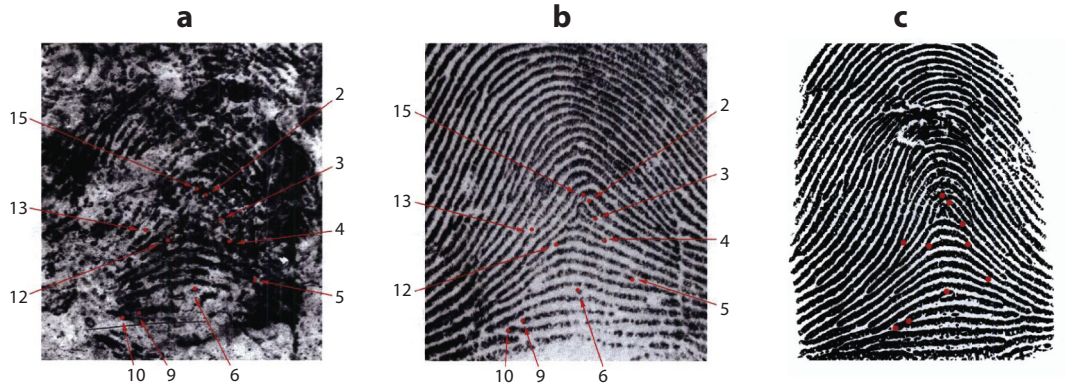

Figure 2

Figures from the US Department of Justice Office of the Inspector General's review of the Brandon Mayfield case (Fine 2006). (*a*) A subset of the minutiae (numbered *red dots*) identified on the latent print. (*b*) Corresponding features on the Mayfield known print. (*c*) Corresponding features on the known print of the believed true source, Ouhnane Daoud.

probability model. Despite community interest in developing likelihood ratios for pattern evidence, these disciplines have proven to be the most difficult in which to apply the likelihood ratio approach. There are several reasons for this. First, the evidence in this case is typically a high-dimensional and complex representation of an impression (often an image). Second, there is a wide variety of features that one might look at. Given the high dimension of the data and the large number of features, a third challenge is finding probability models that can describe the variation in identified features in repeated impressions from a single source or across multiple sources. A final challenge, similar to that mentioned above for trace evidence, is the difficulty in identifying a relevant reference population that can be used to assess the possibility of a coincidental match, which is needed for the denominator of the likelihood ratio.

There is currently considerable research aimed at developing likelihood ratios for pattern evidence. The most progress has been made to date in latent print analysis (Neumann et al. 2012, 2015). Features are identified—this can be done by computer or by an examiner—and each is represented by a vector of its characteristics (feature type, orientation, location/shape relative to nearby features). Though these representations seem promising, it remains challenging to assess the amount of variation in configurations of minutiae that can occur naturally (i.e., due to distortion) and to characterize the amount of variation that might be expected in a relevant population of nonmatching sources. It seems clear that there is considerable work required before the likelihood ratio estimates for pattern evidence will be ready for use in court.

3.4. Score-Based Likelihood Ratios

Given the challenges associated with developing formal probability model–based likelihood ratios for many evidence types, especially among the pattern evidence disciplines, some researchers have begun to consider an alternative approach known as score-based likelihood ratios (see, e.g., Hepler et al. 2012). The basic idea is to avoid developing a formal probability for the evidence E_x, E_y and instead focus on a measure of the similarity of E_x and E_y, often quantified as a score $S(E_x, E_y)$. Typically, low scores indicate very similar impressions, and high scores indicate considerable differences. Then a score-based likelihood ratio, which we denote by SLR below in order to avoid

any confusion, can be defined as

$$\mathrm{SLR}_1 = \frac{\Pr(S(E_x, E_y) = s \,|\, H_p)}{\Pr(S(E_x, E_y) = s \,|\, H_d)}. \tag{5}$$

The numerator requires knowing the distribution of scores among repeated observations obtained from a single source. The denominator requires assessing the distribution of scores for pairs of observations that come from different sources. A particular challenge here is that it is difficult to imagine that a single different-source distribution could be obtained that would be equally relevant to any crime scene evidence. The relative likelihood of obtaining a low score from a different source may be very different depending on the features of the crime scene evidence E_x. This could be expressed explicitly by redefining the score-based likelihood ratio as

$$\mathrm{SLR}_2 = \frac{\Pr(S(E_x, E_y) = s \,|\, E_x, H_p)}{\Pr(S(E_x, E_y) = s \,|\, E_x, H_d)}, \tag{6}$$

where the conditioning makes explicit the need to account for whether the crime scene sample is particularly complex or unique. Note that this is implicitly accounted for in the more formal likelihood ratio approach described in Sections 3.1–3.3.

4. ADDITIONAL ISSUES

There are compelling reasons to believe the likelihood ratio is a very useful way to summarize the strength of forensic evidence of the types considered in Section 3. The likelihood ratio approach provides a natural, intuitive framework for evaluating the evidence under the two hypotheses that a forensic examiner must address. Some of the technical challenges in implementing the likelihood ratio approach for different evidence types are described above. There are also conceptual issues that must be addressed in thinking about the statistical analysis and presentation of evidence.

4.1. Multiple Types of Evidence

This article focuses on a particular piece of evidence in an investigation, e.g., a fingerprint or a set of glass fragments. Of course, it is quite common for multiple types or pieces of evidence to be considered in a case. To begin to think about this, consider a scenario in which there are two different types of evidence, which we denote E_1 and E_2. Both E_1 and E_2 involve observations from the crime scene and from the suspect, but we do not make the x, y notation explicit here. The likelihood ratio framework would assess the strength of the combined evidence through $LR = \Pr(E_1, E_2 \,|\, H_p) / \Pr(E_1, E_2 \,|\, H_d)$. Formal evaluation of this likelihood ratio now requires a joint probability model for the two evidence types. Given the challenges (identified in the previous section) associated with developing probability models for a single evidence type, it seems unlikely that such joint distributions will be available in the near future. Clearly, the task of evaluating this likelihood ratio and assessing the evidence simplifies considerably if the two evidence types can be viewed as independent under both H_p and H_d. If so, then one can compute the likelihood ratio separately for each, and the product of those two likelihood ratios would yield the joint likelihood ratio. It is more challenging to consider how to proceed if the evidence types are not independent and joint probability models are not available. In that case one can write the likelihood ratio in one of two ways, either

$$LR = \frac{\Pr(E_1 \,|\, H_p)}{\Pr(E_1 \,|\, H_d)} \frac{\Pr(E_2 \,|\, E_1, H_p)}{\Pr(E_2 \,|\, E_1, H_d)} \tag{7}$$

or the reverse, in which the likelihood ratio for E_2 is evaluated first, and then E_1 is evaluated conditional on E_2. Conditional probability models for one type of evidence given another would be complex to develop, and indeed, one might wonder whether an E_2 analyst would have the expertise to incorporate the information from evidence of the type E_1.

4.2. Context and Other Information

In introducing the likelihood ratio, we noted that it is straightforward in concept to allow the likelihood ratio to depend on other information I beyond the hypothesis under consideration, $LR = \Pr(E|I, H_p)/\Pr(E|I, H_d)$. This raises the critical question of what information ought to be included in I. For many evidence types, there is relevant information that may impact the ability of the examiner to carry out an analysis. Examples include the substrate in which a fingerprint was left (blood, ink) and the location from which a tire track or shoeprint image was extracted. At the same time, there is much information that is not directly relevant to the forensic comparison and therefore should probably not be included in I. For example, knowledge that the shoe being examined came from an individual running away from the crime scene is not directly relevant to the question of whether the shoe is a possible source of the shoeprint found at the crime scene. Note that such information may be relevant for other aspects of the case, e.g., for jurors trying to assess the innocence or guilt of the suspect, but may detract from an objective assessment of the shoeprint evidence. The question of defining task-relevant information for different forensic domains is a topic of current consideration in the forensic community. The reader is directed to Dror et al. (2015) and NCFS (2015) for more information.

4.3. Uncertainty in the Likelihood Ratio

The likelihood ratio in a given situation depends strongly on assumptions and choices made by its developer. For a start, the likelihood ratio will depend on the features or samples selected for evaluation. The number of glass fragments included in an analysis can impact the likelihood ratio, as can the number of chemical elements considered. The likelihood ratio will also be sensitive to any assumptions made regarding the probability distribution of the evidence under H_p or H_d. It will also be sensitive to assumptions made about the relevant reference population under H_d. It is thus valuable for the developers of likelihood ratio methodologies to carefully assess the sensitivity of results to the choices that are made. Such sensitivity analyses may even be suggested as a natural part of a statistical summary of a forensic examination.

We have seen that, as well as being sensitive to modeling choices and assumptions, the likelihood ratio typically depends on a number of parameters. These parameters can be estimated, in which case the likelihood ratio may be sensitive to the choice of estimators, or they can be averaged over in a Bayesian analysis, in which case the result may be sensitive to the prior distribution used for the parameters. Some have begun to advocate examining the uncertainty in the likelihood ratio by considering a range of parameter values or prior distributions. This topic is controversial, however, with Taroni et al. (2016) arguing that only a single likelihood ratio that averages over all sources of uncertainty is appropriate.

4.4. Understanding and Interpreting Likelihood Ratios

This article focuses on the logical advantages of assessing forensic evidence through the use of likelihood ratios. There are numerous studies, however, that have demonstrated the difficulty lay audiences have in understanding the likelihood ratio. One common error is for a lay audience to

reverse the conditioning of the evidence and the hypothesis. Thus, where the proper interpretation of the likelihood ratio is that it describes the relative likelihood of obtaining the evidence under the two hypotheses (same source and different source), some listeners will interpret the resulting ratio as informing about the relative likelihood of the two hypotheses given the evidence. Of course, the latter can only be obtained if one starts by initially specifying the a priori relative likelihood of the two hypotheses, which is not an easy thing to do. Even if a juror understands the likelihood ratio, there is evidence that jurors do not update their opinions in the way that would be prescribed by Bayes' Theorem (Martire et al. 2013, Thompson & Newman 2015).

5. DISCUSSION

The NRC (2009) report identified a number of concerns with the state of forensic evidence analysis in the United States. These include a lack of standardization in the examination procedures that are used, techniques and procedures being used in some disciplines that are not based on peer-reviewed research, the need for studies on the reliability and accuracy of forensic techniques, the need for studies on human observer bias and other sources of human error, inconsistent practices in reporting and testifying about forensic science, and the need for enhanced probabilistic and statistical methods for improved forensic inferences. The call for more statistical research featured prominently in the report, which recognized the critical role of statisticians for well-designed studies of reliability and accuracy, for well-designed studies of observer bias, and for developing probability and statistical tools that can be used to assess forensic evidence and provide reliable information about the certainty of the conclusions. This review article has focused primarily on the likelihood ratio as a potentially useful tool for the interpretation of forensic evidence. Likelihood ratios are currently used in the analysis of DNA evidence, where they are well accepted, especially for samples of sufficient size known to contain genetic material from a single source. More complex samples (those with limited material to work with and/or those that may contain material from more than one source) are a topic of current research (Steele & Balding 2014). The use of the likelihood ratio for other evidence types has lagged but is currently under development. The European Network of Forensic Science Institutes has recently issued guidelines for the evaluation of evidence that strongly endorse the use of the likelihood ratio framework (ENFSI 2015). There is a need for statisticians to play a greater role in forensic science; there are numerous challenges associated with studying the accuracy and reliability of current forensic practice and in developing novel methods and approaches that can enhance accuracy and reliability.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I am grateful for conversations with JoAnn Buscaglia, Tom Busey, Alicia Carriquiry, Steve Fienberg, Austin Hicklin, Karen Kafadar, Cedric Neumann, and Chris Saunders that have influenced this work. They are of course not responsible for any errors. The article also benefited a great deal from comments received during the review process. This research was partially funded through Cooperative Agreement #70NANB15H176 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University; the University of California, Irvine; and the University of Virginia.

LITERATURE CITED

Aitken CGG, Lucy D. 2004. Evaluation of trace evidence in the form of multivariate data. *Appl. Stat.* 53:109–22

Almirall J, Trejos T. 2006. Advances in the forensic analysis of glass fragments: a review with focus on refractive index and elemental analysis. *Forensic Sci. Rev.* 18(2):74–95

Butler JM. 2015. *Advanced Topics in Forensic DNA Typing: Interpretation*. San Diego: Academic Press

Carriquiry AL, Daniels M, Stern HS. 2000. *Statistical treatment of class evidence: Trace element concentrations in bullet lead*. Technical Report, Iowa State University

Daubert v. Merrell Dow Pharmaceuticals, Inc. 509 U.S. 579 (1993)

Dettman JR, Cassabaum AA, Saunders CP, Snyder DL, Buscaglia J. 2014. Forensic discrimination of copper wire using trace element concentrations. *Anal. Chem.* 86:8176–82

Dror IE, Charlton D, Peron A. 2006. Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Sci. Int.* 156:74–78

Dror IE, Thompson WC, Meissner CA, Kornfield I, Krane D, et al. 2015. Content management toolbox: a linear sequential unmasking (LSU) approach for minimizing cognitive bias in forensic decision making. *J. Forensic Sci.* 60:1111–12

Eckholm E. 2016. Texas panel calls for an end to criminal IDs via bite mark. *New York Times*, Feb. 12, p. A10

ENFSI (Eur. Netw. Forensic Sci. Inst.). 2015. *ENFSI Guideline for Evaluative Reporting in Forensic Science.* Wiesbaden, Ger.: ENFSI. **http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf**

Expert Work. Group Hum. Factors Latent Print Anal. 2012. *Latent Print Examination and Human Factors: Improving the Practice through a Systems Approach.* Gaithersburg, MD: Natl. Inst. Stand. Technol. and Washington, DC: Natl. Inst. Justice. **http://www.crime-scene-investigator.net/LatentPrintExaminationHumanFactors.pdf**

FBI (Fed. Bur. Investig.). 2015. *FBI testimony on microscopic hair analysis contained errors in at least 90 percent of cases in ongoing review.* Press Release, April 20. **https://www.fbi.gov/news/pressrel/press-releases/fbi-testimony-on-microscopic-hair-analysis-contained-errors-in-at-least-90-percent-of-cases-in-ongoing-review**

Fenton N, Neil M, Berger D. 2016. Bayes and the law. *Annu. Rev. Stat. Appl.* 3:51–77

Fienberg SE, ed. 1989. *The Evolving Role of Statistical Assessments as Evidence in the Courts*. New York: Springer

Fine GA. 2006. *A Review of the FBI's Handling of the Brandon Mayfield Case (Unclassified and Redacted)*. Washington, DC: US Dep. Justice Off. Insp. Gen. **https://oig.justice.gov/special/s0601/final.pdf**

Frye v. United States, 293 F. 1013 (1923)

Garris MD, McCabe RM. 2000. *NIST Special Database 27: fingerprint minutiae from latent and matching tenprint images.* Tech. Rep. NISTIR 6534. Natl. Inst. Stand. Technol., Gaithersburg, MD

Hepler AB, Saunders CP, Davis LJ, Buscaglia J. 2012. Score-based likelihood ratios for handwriting evidence. *Forensic Sci. Int.* 219:129–40

Kelly H, Bright J, Buckleton JS, Curran JM. 2014. A comparison of statistical models for the analysis of complex forensic DNA profiles. *Sci. Justice* 54(1):66–70

Lehmann EL, Romano JP. 2008. *Testing Statistical Hypotheses*. New York: Springer. 3rd ed.

Lindley DV. 1977. A problem in forensic science. *Biometrika* 64:207–13

Martire KA, Kemp RI, Watkins I, Sayle MA, Newell BR. 2013. The expression and interpretation of uncertain forensic science evidence: verbal equivalence, evidence strength, and the weak evidence effect. *Law Hum. Behav.* 37:197–207

NCFS (Natl. Comm. Forensic Sci.). 2015. *Ensuring that forensic analysis is based upon task-relevant information.* Views doc., Natl. Inst. Stand. Technol., Gaithersburg, MD

NCFS (Natl. Comm. Forensic Sci.). 2016. *Views of the Commission regarding use of the term reasonable scientific certainty*. Views doc., Natl. Inst. Stand. Technol., Gaithersburg, MD

NRC (Natl. Res. Counc.). 1996. *The Evaluation of Forensic DNA Evidence*. Washington, DC: Natl. Acad. Press

NRC (Natl. Res. Counc.). 2003. *The Polygraph and Lie Detection*. Washington, DC: Natl. Acad. Press

NRC (Natl. Res. Counc.). 2004. *Forensic Analysis: Weighing Bullet Lead Evidence*. Washington, DC: Natl. Acad. Press

NRC (Natl. Res. Counc.). 2008. *Ballistic Imaging*. Washington, DC: Natl. Acad. Press

NRC (Natl. Res. Counc.). 2009. *Strengthening Forensic Science in the United States: A Path Forward*. Washington, DC: Natl. Acad. Press

Neumann C, Champod C, Yoo M, Genessay T, Langenburg G. 2015. Quantifying the weight of fingerprint evidence through the spatial relationship, directions and types of minutiae observed on fingermarks. *Forensic Sci. Int.* 248:154–71

Neumann C, Evett IW, Skerrett J. 2012. Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm. *J. R. Stat. Soc. A* 175:371–415

Neumann C, Stern H. 2016. Forensic examination of fingerprints: past, present and future. *Chance* 29(1):9–16

PCAST (Pres. Counc. Advis. Sci. Technol.). 2016. *Report to the President: Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. Washington, DC: Executive Off. Pres. **http://www.crime-scene-investigator.net/PDF/forensic-science-in-criminal-courts-ensuring-scientific-validity-of-feature-comparison-methods.pdf**

Pepe MS. 2003. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York: Oxford Univ. Press

Saks MJ, Albright T, Bohan TL, Bierer BE, Bowers CM, et al. 2016. Forensic bitemark identification: weak foundations, exaggerated claims. *J. Law Biosci.* **https://doi.org/10.1093/jlb/lsw045**

Steele CD, Balding DJ. 2014. Statistical evaluation of forensic DNA profile evidence. *Annu. Rev. Stat. Appl.* 1:361–84

Taroni F, Bozza S, Biedermann A, Aitken C. 2016. Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. *Law Prob. Risk* 15(1):1–16

Thompson WC, Newman EJ. 2015. Lay understanding of forensic statistics: evaluation of random match probabilities, likelihood ratios, and verbal equivalents. *Law Hum. Behav.* 39:332–49

Ulery BT, Hicklin RA, Buscaglia J, Roberts MA. 2011. Accuracy and reliability of forensic latent fingerprint decisions. *PNAS* 108:7733–38

Ulery BT, Hicklin RA, Buscaglia J, Roberts MA. 2012. Repeatability and reproducibility of decisions by latent fingerprint examiners. *PLOS ONE* 7:e32800

Zweig MH, Campbell G. 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.* 39(4):561–77