

Koehler, J. J. (1997). Why DNA Likelihood Ratios Should Account for Error (even when a National Research Council report says they should not). *Jurimetrics Journal*, 37, 425-437.

**WHY DNA LIKELIHOOD RATIOS SHOULD ACCOUNT FOR
ERROR (EVEN WHEN A NATIONAL RESEARCH
COUNCIL REPORT SAYS THEY SHOULD NOT)**

Jonathan J. Koehler¹

March 14, 1997

¹ Jonathan J. Koehler is Associate Professor of Behavioral Decision Making, Graduate School of Business and Law School, The University of Texas at Austin. This paper is based on a presentation given by the author at the International Conference on Forensic Statistics, Edinburgh, Scotland, July, 1996.

Address for correspondence: CBA 5.202, University of Texas at Austin, Austin, TX 78712-1175. 512-471-7856 (phone), 512-471-0587 (fax), koehler@mail.utexas.edu (e-mail).

ABSTRACT

A likelihood ratio (LR) analysis shows that the possibility of error limits the strength of DNA evidence in the same way that it limits the strength of other kinds of legal evidence, such as eyewitness testimony. However, a 1996 report by the National Research Council recommends against estimating an error rate to help identify the probative value of DNA evidence. The Committee's arguments are identified and critiqued. It is counter-argued that error rate data derived from broad reference classes (e.g., "all DNA laboratories") provide a relevant starting point for estimating the risk of error in individual cases. LRs that fail to incorporate this estimate may be misleading.

I. INTRODUCTION

The legal and scientific communities have become increasingly interested in the use of likelihood ratios (LRs) to describe the strength of scientific evidence presented at trial.²

Specifically, many have argued that the strength of DNA evidence should be reported as LRs.³

A LR reflects the likelihood of evidence under one hypothesis relative to its likelihood under another hypothesis.⁴ Thus, the LR for an item of evidence E relative to two mutually exclusive and exhaustive hypotheses H and \bar{H} is $LR = \frac{P(E | H)}{P(E | \bar{H})}$.² LRs support H to the extent that they are greater than 1, and support \bar{H} ³ to the extent that they are less than 1.

In sections II and III of this article, I argue that a LR analysis of the strength of DNA evidence shows that the possibility of various types of error limits the strength of DNA evidence in the same way that it limits the strength of other kinds of legal evidence, such as eyewitness testimony. Although this would seem to be an elementary point, most existing recommendations for the construction of DNA LRs fail to account for the possibility of error. Indeed, a 1996 report by the National Research Council (NRC-II) concludes that LRs that do

² Ward Edwards, Comment, 66 BOS. U. L. REV., 623 (1986); Richard O. Lempert, Modeling Relevance, 75 MICH. L. REV. 1021 (1977); David H. Kaye, Quantifying Probative Value, 66 BOS. U. L. REV., 761 (1986); Thomas D. Lyon and Jonathan J. Koehler, The Relevance Ratio: Evaluating the Probative Value of Expert Testimony in Child Sexual Abuse Cases, CORNELL L. REV. (In press); David Schum & Anne Martin, Formal and Empirical Research on Cascaded Inference in Jurisprudence, 17 LAW & SOC. REV. 105 (1982). For a discussion of related definitions of probative value, see Richard D. Friedman, A Close Look at Probative Value, 66 BOS. U. L. REV. 733 (1986), and David H. Kaye, Comment: Quantifying Probative Value, 66 BOS. UNIV. L. REV. 761 (1986).

³ C. G. G. AITKEN, STATISTICS AND THE EVALUATION OF EVIDENCE FOR FORENSIC SCIENTISTS (1995); Donald A. Berry, I. W. Evett, & R. Pinchin, Statistical Inference in Crime Investigations Using Deoxyribonucleic Acid Profiling, 41 APPL. STAT. 499 (1992); COMMITTEE ON DNA FORENSIC SCIENCE THE EVALUATION OF FORENSIC DNA EVIDENCE (Prepublication Copy) (1996) ("The likelihood ratio, then, is a way of summarizing the DNA evidence" p. 5-4); Ian W. Evett, J. Scrange, & R. Pinchin, An Illustration of the Advantages of Efficient Statistical Methods for RFLP Analysis in Forensic Science, 52 AM. J. HUM. GENET. 498 (1993); BERNARD ROBERTSON & G. A. VIGNAUX: INTERPRETING EVIDENCE: EVALUATING FORENSIC SCIENCE IN THE COURTROOM (1995).

⁴ David H. Kaye, The Relevance of "Matching" DNA: Is the Window Half Open or Half Shut? 85 J. CRIM. LAW & CRIMINOL. 676, 684 (1995); David Schum, Diverse Models of Evidence and Inference: Probability and the Processes of Discovery, Proof, and Choice, 66 BOS. U. L. REV. 825, 861 (1986).

not incorporate an error rate estimate "are appropriate for explaining the significance of [DNA] data."⁵ In Section IV, NRC-II's arguments related to the treatment of error in DNA testing are identified and critiqued. I argue that error rate data derived from broad reference classes (e.g., "all DNA laboratories") provide a relevant starting point for estimating the risk of error in individual cases. LRs that fail to incorporate such an estimate are likely to overstate the probative value of a reportedly matching DNA evidence, and may therefore be misleading.

II. LIKELIHOOD RATIOS: EYEWITNESS EVIDENCE

When an eyewitness identifies a suspect as the man he saw robbing a bank, we understand that this identification has probative value with respect to the question "Who robbed the bank?" At the same time, we understand that this report is imperfectly probative. This is largely because the witness may be lying or mistaken. Although reasons for lying are often hard to imagine, a more reasonable explanation for why an identified suspect may not, in fact, be the perpetrator is that an error occurred.

It is well known that there are two types of errors associated with inferential reasoning tasks such as attempts to determine the identity of the person who robbed a bank. These are false negative and false positive errors. False negative errors occur when a true hypothesis has been rejected; false positive errors occur when a false hypothesis has been accepted. Thus, when the person who actually did commit a robbery is cleared by an eyewitness in the context of a police lineup, a false negative error has occurred. Conversely, when the eyewitness identifies an innocent person from the lineup as the bank robber, a false positive error has occurred. In cases that include an eyewitness identification, the defense often suggests that a false positive error has occurred.

⁵ **COMMITTEE ON DNA FORENSIC SCIENCE THE EVALUATION OF FORENSIC DNA EVIDENCE (Prepublication Copy) at p. 6-30 (1996).**

For purposes of this discussion, it is helpful to divide false positive errors into two classes. One class of error occurs when the characteristics or features described by the witness belong to the suspect, but do not belong to the perpetrator. This may be called a "*feature-based error*." For example, the witness may incorrectly believe that the perpetrator had a tattoo of a heart on his left bicep when, in fact, the perpetrator actually had a bloody bruise on his right forearm. But when a suspect who has a tattoo of a heart on his left bicep is located, the witness may falsely identify the suspect because he matches up with this distinctive feature. Feature-based errors can occur for various reasons. The witness may have had a partial view of the perpetrator, been exposed to the perpetrator for a very brief time, or misremembered features of the perpetrator.

A second class of error occurs when the witness accurately identifies a shared characteristic of the suspect and perpetrator, but mistakenly concludes that the suspect and perpetrator are one and the same. This may be called an "*evaluative error*." For example, the witness may correctly recall that the perpetrator had a jagged facial scar, but may incorrectly infer that an innocent suspect who happens to share this feature is the perpetrator. Here, the shared trait is coincidental.

As indicated at the outset, the probative strength of eyewitness reports and other types of evidence can, in principle, be captured by a likelihood ratio, $LR = \frac{P(E | H)}{P(E | \bar{H})}$.⁴ When constructing a LR, E and H must be defined with care. In the case of eyewitness reports, E = "Eyewitness reports that certain features of the suspect 'match' features of the perpetrator." Eyewitnesses often go further and offer the evaluative conclusion that, to a reasonable degree of certainty, the suspect is the person they saw. H = "the suspect is the person who the eyewitness saw," and \bar{H} = "the suspect is not the person who the eyewitness saw."

The eyewitness's report has substantial probative value because the numerator of the

LR presumably is high (near 1) and because the denominator of the LR is low (near 0). Importantly, small absolute deviations from 0 in the denominator can have a large impact on the LR. When the numerator of the LR is fixed at, say, .99, denominators of .01 vs. .0001 produce strikingly different LRs (i.e., 99, vs. 9,900 respectively). Comparable deviations in the numerator have less impact.⁶ For example, when the denominator is fixed at .01, LR numerators of .99 and .9999 produce similarly large LRs (99 vs. 99.99 respectively). Thus, production of an accurate estimate of a LR's order of magnitude depends critically on whether all factors that may significantly affect the denominator of the LR have been identified, and included in the computation.

III. LIKELIHOOD RATIOS: DNA EVIDENCE

Now suppose that the evidence E is not an eyewitness's report that the suspect "matches" the perpetrator. Instead, let E be a report from a DNA analyst that the suspect's DNA "matches" the DNA found in genetic evidence that was recovered from a crime scene. Just as eyewitnesses frequently go beyond this type of feature-matching claim to offer evaluative conclusions in their testimony ("He's the one I saw"), DNA analysts frequently testify that, to a reasonable degree of scientific certainty, the suspect is the source of the recovered genetic evidence.⁷

Notice the parallels between the evidence E in the eyewitness and DNA contexts. In the eyewitness context, E is defined as a report of a match between features of the suspect and

⁶ This difference occurs because a small absolute deviations such that given by decreasing the LR denominator from .01 to .0001 are actually quite large in terms of magnitude, whereas the equivalent absolute deviation changes in the LR numerator given by increasing .99 to .9999 is quite small in terms of magnitude. Specifically, a change from .01 to .0001 describes a 100-fold decrease in magnitude, and a change from .99 to .9999 describes a 1.01-fold increase in magnitude.

⁷ For example, see State v. Bloom 516 N.W.2d 159 (Minn. 1994).

those of the perpetrator. When the evidence is defined this way, we implicitly recognize the possibility of both feature-based and evaluative errors. Likewise, in the criminal DNA context, when E is defined as a report of a match between the suspect's DNA and the DNA of evidence recovered from a crime scene, we recognize the possibilities that (a) the DNA samples actually do not match (feature-based error), and (b) the DNA samples match, but do not share a common source (evaluative error).

The important point that emerges from this comparison of eyewitness evidence and DNA evidence is that construction of LR's designed to reflect the strength of these types of evidence must account for the possibilities of both feature-based and evaluative errors. We would not think of providing LR's for eyewitness evidence premised on the assumption that the eyewitness has not committed one or more feature-based errors,⁸ and we should not do so in the DNA context either.⁹

IV. TREATMENT OF FEATURE-BASED ERROR IN THE 1996 NATIONAL RESEARCH COUNCIL (NRC-II) REPORT

In the spring of 1996, the National Research Council released its second report on the use of DNA evidence in the courtroom.¹⁰ This report (NRC-II) emphasizes the statistical

⁸ In one celebrated decision, the California Supreme Court expressly warned about the dangers of providing jurors with a numerical index related to the probative value of eyewitness evidence when that value assumes that the characteristics were "correctly observed and accurately described" by the witness. People v. Collins, 68 Cal 2d 319, 330, 66 Cal Rptr. 497, 438 P.2d 33, 36 A.L.R.3rd 1176 (1968).

⁹ Most commentaries on the use of LR's in forensic science do not incorporate the risk of feature-based error. See Bernard Devlin, Neil Risch, Katherine Roeder, *Statistical Evaluation of DNA Fingerprinting: A Critique of the NRC's Report*, 259 SCIENCE 748 (1993); Ian Evett, *DNA Statistics: Putting the Problems Into Perspective*, 33 JURIM. J. 139 (1992); D. Jarjoura, J. Jamison, and S. Androulakakis, *Likelihood Ratios for Deoxyribonucleic Acid (DNA) Typing in Criminal Cases*, 39 J. FOREN. SCI. 64 (1994); Katherine Roeder, *DNA Fingerprinting: A Review of the Controversy*, 9 STAT. SCI. 222 (1994). For analyses of the issue as it arose in People (CA) v. Simpson (1995), see Jonathan J. Koehler, *On Conveying the Probative Value of DNA Evidence: Frequencies, Likelihood Ratios and Error Rates*, 67 U. COLORADO L. REV. 859 (1996); William C. Thompson, *DNA Evidence in the O. J. Simpson Trial*, 67 U. COLO. L. REV. 827 (1996).

¹⁰ COMMITTEE ON DNA FORENSIC SCIENCE THE EVALUATION OF FORENSIC DNA EVIDENCE (Prepublication Copy) (1996).

aspects of DNA evidence, and includes a series of recommendations for the presentation of DNA evidence in court.

NRC-II expressly rejects construction of DNA LR's that allow for the possibility of feature-based error.¹¹ The arguments that NRC offers in defense of its position are identified and critiqued below.

1. Proficiency tests do not measure error rates.

"The objective of both proficiency-testing and auditing is to improve laboratory performance by identifying problems that need to be corrected. Neither is designed to measure error rates" (p. 3-5).

Some of those who favor incorporation of feature-based errors in numerical indexes that purport to describe the strength of DNA evidence have recommended estimating the frequency of such errors via proficiency testing.¹² In the typical proficiency test, DNA analysts are provided with sets of genetic samples (e.g., blood or semen stains) and asked to determine which match and which do not. On occasion, laboratories have reported that samples match when, in fact, the samples were from different sources.¹³

Even if the designers of proficiency tests are primarily interested in using the tests to improve laboratory performance, the frequency of false-match reports is a byproduct of the testing process. Consequently, NRC-II's point that proficiency tests are not designed to measure error rates is irrelevant. Moreover, because most proficiency tests are relatively easy

¹¹ NRC-II uses the term "error" to refer to "feature-based error."

¹² Jonathan J. Koehler, *Error and Exaggeration in the Presentation of DNA Evidence*, 34 JURIM. J. 21 (1993); Lawrence Mueller, *The Use of DNA Typing in Forensic Science*, 3 Accountability in Research 1 (1993); William C. Thompson, *Evaluating the Admissibility of New Genetic Identification Tests: Lessons From the "DNA War."* 84 J. CRIM. LAW & CRIMINOL. 22 (1993).

¹³ Jonathan J. Koehler, *Error and Exaggeration in the Presentation of DNA Evidence*, 34 JURIM. J. 21 (1993); COMMITTEE ON DNA TECHNOLOGY IN FORENSIC SCIENCE, DNA TECHNOLOGY IN FORENSIC SCIENCE (1992).

in the sense that they are nonblind, internal tests that use large, clean stains, performance on these tests would appear to provide a generously low estimate of (criminal) casework error rates. For these reasons, NRC-II's implicit suggestion that the error rates observed in proficiency tests are irrelevant to an estimate of the error rate that might be observed in actual case work is unconvincing.

2. Error rates are never known because they are always in a state of flux.

"Estimating rates at which nonmatching samples are declared to match from historical performance on proficiency tests is almost certain to yield wrong values. When errors are discovered, they are investigated thoroughly so that corrections can be made" (p. 3-10)

First, NRC-II does not offer citations or evidence for its claim that proficiency test errors are thoroughly investigated. Indeed, there is evidence that errors made on proficiency tests are often denied, minimized, or ignored.¹⁴ For example, a series of errors in the Collaborative Testing Services proficiency tests were discussed recently.¹⁵ However, these errors have received no attention among DNA scientists. Indeed, it is not uncommon for DNA scientists to argue that such errors are impossible or unheard of.¹⁶

Second, imagine if the U.S. Federal Aviation Administration made a parallel claim about airlines: "Estimating rates at which airlines crash from historical performance is almost certain to yield wrong values. When errors are discovered, they are investigated thoroughly so

¹⁴ William C. Thompson, *Worthwhile DNA Questions (Letter)*, 77 JUDICATURE 57 (1993). But see John W. Hicks, *John Hicks Responds (Letter)*, 77 JUDICATURE 57 (1993).

¹⁵ Jonathan J. Koehler, Audrey Chia, and J. Sam Lindsey, *The Random Match Probability (RMP) in DNA Evidence: Irrelevant and Prejudicial?* 35 JURIM. J. 201 (1995).

¹⁶ Bruce Budowle and J. Stafford, *Response to Expert Report by D. L. Hartl*, 18 CRIME LAB. DIGEST 101, 104 (1991) ("To date, there has never been any evidence in these situations [i.e., those that include reference sample profiles] to support that operator error (i.e., inadvertent mixing of samples) has occurred.") See Jonathan J. Koehler, *Error and Exaggeration in the Presentation of DNA Evidence*, 34 JURIM. J. 21 (1993) for other courtroom statements made by DNA experts about the impossibility of error in DNA profiling.

that corrections can be made."

For any phenomenon that operates in a non-static environment, historical data always yield "wrong values" about future performance. Nevertheless, such data are useful for estimating future performance. Suppose that 1% or .1% of all airlines erred--i.e., crashed--last year. Assuming that all of these crashes were "investigated thoroughly so that corrections can be made," few among us would argue that flying is safe on grounds that "estimating crash rates from historical performance is almost certain to yield wrong values." In short, NRC-II's attempt to dismiss existing error rate data on grounds that error rates are in flux is unconvincing.

My final comment in connection with this point concerns the false-positive errors committed by Cellmark in 1988 and 1989 as discussed in NRC-II. NRC-II notes that Cellmark made 2 errors in its first set of 125 proficiency test samples, but zero errors in 450 additional tests since that time. The report concludes: "Clearly an estimate of .35% (2/575) is inappropriate as a measure of the chance of error at Cellmark today."¹⁷

Holding aside comments about what should count as a "test sample,"¹⁸ this conclusion is not statistically sound. Even if one only attended to Cellmark's error-free performance on its most recent 450 tests, one could not reject the null hypothesis that Cellmark's error rate is 0.35% or more at conventional confidence levels. Using the formula NRC-II provides on the very page where the Cellmark case is discussed, the 95% upper bound error rate associated with a performance of 0 incorrect match reports out of 450 match reports is 0.66%. Because 0.35% is well within the 95% confidence interval that ranges from 0.00-0.66%, it cannot be

¹⁷ COMMITTEE ON DNA FORENSIC SCIENCE THE EVALUATION OF FORENSIC DNA EVIDENCE (Prepublication Copy) at 3-11 (1996).

¹⁸ William C. Thompson, Subjective Interpretation, Laboratory Error and the Value of Forensic DNA Evidence: Three Case Studies, 96 GENETICA 153 (1995).

rejected as a measure of the chance of error at Cellmark.¹⁹

3. An estimate of the chance of error is wanted for this case, not for cases in general.

"The question to be decided is not the general error rate for a laboratory or laboratories over time but rather whether the laboratory doing DNA testing in this particular case made a critical error. This risk of error in any particular case depends on many variables ... and there is no simple equation to translate these variables into the probability that a reported match is spurious. . . . The risk of error is properly considered case by case" (p. 3-10, 3-11).

The claim that industry-wide error rate data cannot be used to estimate case-specific error rates is a version of the "base rate fallacy" that has received much attention in the behavioral decision making literature.²⁰ The fallacious argument runs roughly as follows: In a probabilistic decision task in which both background frequency and individuating information are available, the frequency information (i.e., the base rate information) should be disregarded because it is insufficiently case-specific.²¹

From a Bayesian perspective, frequency information should inform decision makers'

¹⁹ The Cellmark case may not fairly represent the current industry-wide error rate in DNA testing. A broader range of tests are needed that examine recent performance in a large number of forensic DNA laboratories. An unsolved philosophical question is "How long should we attend to test data from the past for purposes of extrapolation to the present and future?" On the one hand, sticking too closely to past performance data may mean that insufficient attention is given to recent improvements. On the other hand, historical data do not become irrelevant simply because some improvements have been made. Perhaps the appropriate error rate model is one in which historical data signal current error rates, but this signal weakens over time as a function of the number and types of changes that are made. Such a model should also allow for the appearance of a new set of errors that may arise as procedures are modified.

²⁰ Maya Bar-Hillel, The Base-rate Fallacy in Probability Judgments, 44 *Acta Psychologica* 211 (1980); RICHARD NISBETT AND LEE ROSS, *HUMAN INFERENCE: STRATEGIES AND SHORTCOMINGS OF SOCIAL JUDGMENT* (1980); Amos Tversky and Daniel Kahneman, Evidential Impact of Base Rates, in *JUDGMENT UNDER UNCERTAINTY: HEURISTICS AND BIASES* (EDS. D. KAHNEMAN, P. SLOVIC, & A. TVERSKY) (1982). For a recent review, see Jonathan J. Koehler, The Base Rate Fallacy Reconsidered: Normative, Descriptive and Methodological Challenges, 19 *BEHAV. & BRAIN SCI.* 1 (1996).

²¹ There is also a descriptive element to the base rate fallacy, namely, that people actually do disregard base rates in probabilistic decision tasks.

prior odds ratios, and individuating information should inform their likelihood ratios.²² Failure to consider the frequency of a phenomenon when predicting its chance of occurrence in a specific instance can lead to inverse fallacies in which people mistakenly conclude that the denominator of the LR is identical to the denominator of the posterior odds ratio.

Consider the following example. Suppose decision makers are told that the probability that a person who does not have the HIV virus will nonetheless test positive for the HIV virus is 1%,²³ some may invert this conditional probability statement and conclude that there is a 1% chance that a person who tests positive for HIV does not have the virus. However, this conclusion is unlikely to be true. The former probability is $P(\text{Test Positive} | \text{No HIV})$ and the latter probability is its inverse $P(\text{No HIV} | \text{Test Positive})$.⁷ As the odds form of Bayes' Theorem indicates, an estimate of $P(\text{No HIV} | \text{Test Positive})$ ⁸ requires an estimate of the prior probability that a member of the population would have HIV. If the prior probability were, say, 2%, and if we assume that the false negative rate of the HIV test is identical to the false positive error rate provided (i.e., 1%), then $P(\text{No HIV} | \text{Test Positive})$ ⁹ is not 1% as given by

²² There are cases in which it is unclear whether information informs prior probabilities or likelihoods. Terry Connolly, *Are Base Rates a Natural Category of Information?* 19 BEHAV. & BRAIN SCI. 19 (1996). But the important point is that normative theory requires that the information be used.

²³ This probability, $P(\text{Test Positive} | \text{No HIV})$, ~~Error! Main Document Only~~ is the false positive error rate of the test.

inverse fallacy reasoning.²⁴ Unfortunately, inverse fallacies appear to be common,²⁵ and have been observed in statements made by DNA experts at trial.²⁶

The point here is that a failure to consider the general frequency of a phenomenon (such as HIV infection or laboratory error) when predicting its chance of occurrence in a specific instance can lead to poor probabilistic estimates. In the case of laboratory error, there is a real danger that the possibility of error will not be considered (i.e., will be assigned an implicit probability of zero) when case-specific reasons for believing that an error did occur are absent. But a failure to consider the non-zero laboratory error base rate could lead to exaggerated estimates of the strength of the DNA evidence.

Individuating features associated with a particular case should not be ignored when those features are demonstrably related to a case-specific DNA error rate estimate. Where data

²⁴ According to the odds form of Bayes' Theorem,

$$\frac{P(\text{HIV} | \text{Test Positive})}{P(\text{No HIV} | \text{Test Positive})} = \frac{P(\text{HIV})}{P(\text{No HIV})} \times \frac{P(\text{Test Positive} | \text{HIV})}{P(\text{Test Positive} | \text{No HIV})} \quad \text{Error! Main Document Only.}$$

$$\frac{P(\text{HIV} | \text{Test Positive})}{P(\text{No HIV} | \text{Test Positive})} = \frac{.02}{.98} \times \frac{.99}{.01} = \frac{.0198}{.0098} \quad \text{Error! Main Document Only.}$$

∴ P(HIV | Test Positive) = 66.9%, P(No HIV | Test Positive) = 33.1%. **Error! Main Document Only. In cases such as this where the false positive and false negative error rates are equal, the denominator of the LR equals the denominator of the posterior odds ratio only when the prior odds ratio is one (i.e., when the prior probability of the event in question is 50%).**

²⁵ Maya Bar-Hillel, *The Base-rate Fallacy in Probability Judgments*, 44 *ACTA PSYCHOLOGICA* 211 (1980); Ward Casscells, Arno Schoenberger, & Thomas B. Graboys, *Interpretation by Physicians of Clinical Laboratory Results*, 299 *NEW ENG. J. MED.* 999 (1978); Leda Cosmides and John Tooby, *Are Humans Good Intuitive Statisticians After All? Rethinking Some Conclusions From the Literature on Judgment Under Uncertainty*, 58 *COGNITION* 1 (1996); Gerd Gigerenzer and Ulrich Hoffrage, *How to Improve Bayesian Reasoning Without Instruction: Frequency Formats*, 102 *PSYCHOLOGICAL REVIEW* 684 (1995); Robert M. Hamm, *Explanations for Common Responses to the Blue/Green Cab Probabilistic Inference Word Problem*, 72 *PSYCHOLOGICAL REPORTS* 219 (1993); David H. Kaye and Jonathan J. Koehler, *Can Jurors Understand Probabilistic Evidence?* 154 *J. ROYAL STAT. SOC. (SERIES A)* 75, 77 (1991) (Introduced the term "inversion fallacy"); William C. Thompson and Edward L. Schumann, *Interpretation of Statistical Evidence in Criminal Trials: The Prosecutors' Fallacy and the Defense Attorney's Fallacy*, 11 *LAW & HUMAN BEHAV.* 167 (1987); Christopher R. Wolfe, *Information Seeking on Bayesian Conditional Probability Problems: A Fuzzy-trace Theory Account*, 8 *J. BEHAV. DEC. MAKING* 85 (1995).

²⁶ Jonathan J. Koehler, *Error and Exaggeration in the Presentation of DNA Evidence*, 34 *JURIM. J.*, 21 (1993).

exist, they should be used to modify the general laboratory error base rate. But care must be taken to ensure that the individuating feature data are not used selectively, and that the individuating features identified are distinctive and diagnostic.

Even when favorable individuating features exist (e.g., the analyst is experienced, the analyst was observed, the samples were clean, the results were double-checked, etc.), these may not provide a sufficient basis for concluding that the chance of error in the instant case differs from the chance of error across the industry. First, the favorable features may not be distinctive in the sense that they may be standard within the industry. When this is the case, even highly favorable characteristics should not be used to modify the industry-wide base rate because their influence has already been incorporated into the observed values. Second, there is a danger of selective use of feature data. Even when features exist that do differentiate the focal laboratory (or analyst) from the industry, care must be taken to ensure that a fair cross selection of features have been considered. Because both favorable and unfavorable differentiating features can be identified in almost every case, it would not be appropriate to consider only those differentiating features that are favorable (or unfavorable) to bolster a claim that a case-specific error rate is better (or worse) than the industry average.

4. The use of an industry-wide error rate is unfair to the good laboratories.

"The pooling of proficiency-test results across laboratories ... as a means of estimating an "industry-wide" error rate ... could penalize the better laboratories; multiple errors on a single test by one laboratory could substantially affect the overall estimated false-match error rate" (p. 3-10).

This point is true but unpersuasive as a justification for ignoring industry-wide error rates. Why should laboratory-favorable assumptions be made at trial? DNA analysts commonly contend that their individual error rate is less than that of the average analyst as

measured by proficiency tests.²⁷ But it can't be that all analysts are above average and, depending on how average is defined, it may not be possible for even a majority of analysts to be above average.²⁸ If a DNA analyst offers that he/she is substantially more accurate than the average analyst, then he/she should be encouraged (or perhaps required) to present appropriately supporting data. If it is too burdensome to produce such data, then the appropriate thing to do is to use data from the average analyst to estimate the individual analyst's error rate.

5. A relevant error rate estimate cannot be provided because it would require too many proficiency tests.

"To estimate accurately, from proficiency test results, the overall rate at which a laboratory declares nonmatching samples to match ... would require a laboratory to undergo an unrealistically large number of proficiency trials" (p. 3-10).

Implicit in this point is the assumption that the correct level of reference class specificity for error rate estimation is "the laboratory." Presumably, then, error rates derived from "the DNA industry" reference class are too broad, and error rates derived from "the individual analyst within the focal laboratory" are unnecessarily narrow. But on what grounds can we justify a binary division of the notion of relevance? On what grounds can we say that "the laboratory" is the right level of specificity for estimating error rates rather than, say, "the individual analyst within the focal laboratory"?

²⁷ **The I'm-better-than-average phenomenon is hardly unique to DNA analysts. Numerous studies and polls reveal that most people believe they are more intelligent, more fair-minded, less prejudiced, better drivers, better at their job, and less wrinkled than their peers. See THOMAS GILOVICH, HOW WE KNOW WHAT ISN'T SO, at 77 (1991); Ola Svenson, Are We All Less Risky and More Skillful Than Our Fellow Drivers? 47 ACTA PSYCHOLOGICA 143 (1981); Diane White, Wrinkled Boomers Won't Face Truth, BOSTON GLOBE E1 (July 18, 1996).**

²⁸ **If average is defined as the median, then it is not possible for a majority of analysts to have better-than-average error rates. If average is defined as the mean, and if performance is highly skewed, then it would be possible for a large majority of analysts to have better-than-average error rates.**

Regardless of which reference class is used, the data may not account for all variables that influence the risk of error in the instant case. But absent reason to believe that the missing variables shift the probabilities in a particular direction, the base rates identified through scientific study of broad reference classes are not only relevant, but they yield a best-guess probability for the instant case.²⁹ Thus, although it may be impractical to conduct a sufficiently large number of proficiency tests to produce error rate estimates based on a highly specific reference classes, proficiency testing at a more aggregated level can still provide error rate estimates that are useful in individual cases.

6. Aggregation of an error rate estimate with a coincidental match estimate into a single statistic deprives a jury of valuable information.

"But withholding the components of the summary statistic from the judge or jury would deprive the trier of fact of the opportunity to evaluate separately the possibility that the profiles match by coincidence as opposed to the possibility that they are reported to match by reason of laboratory or handling error" (p. 3-9 - 3-10).

Even if a broadly acceptable error rate estimate could be produced, NRC-II is reluctant to allow this estimate to be aggregated with an estimate of the coincidental match frequency (i.e., the random match probability) to yield a numerical index of the strength of the DNA evidence.

NRC-II correctly points out that aggregation denies jurors the opportunity to evaluate separately the individual components of the aggregated index. But why is separate evaluation desirable? As discussed earlier, if Evidence E refers to "a reported DNA match," then the denominator of the LR that reflects the probative value of E should take into account all of the important ways in which a reported match might occur when, in fact, the reported matchee is

²⁹ Paul E. Meehl and A. Rosen, *Antecedent Probability and the Efficiency of Psychometric Signs, Patterns, or Cutting Scores*, 52 *PSYCH. BULL.* 94 (1955).

not the source. That is, the LR denominator is the disjunctive probability associated with the various ways in which a suspect may reportedly match a DNA sample recovered from a crime scene given that he is not, in fact, its source. Once the disjunctive estimate has been produced, there is little additional value in providing the individual values that were used to create the disjunctive estimate.

In response, one could argue that production of the disjunctive probability requires knowledge of the individual probabilities that comprise it. In principle, this is true. However, when those individual probabilities are several orders of magnitude smaller than at least one other individual probability in the group, they contribute so little to the disjunctive estimate that they may be ignored. Where DNA evidence is involved, it will often be true that the risk of laboratory error is obviously several orders of magnitude larger than the risk of a coincidental match (which is often on the order of 1 in many millions or billions).³⁰ In these cases, it is hard to see what is gained by giving triers of fact the opportunity to evaluate separately the possibilities of laboratory error and coincidental.

A simple analogy may be useful. When an engineer must decide whether a double linked chain will break when pulled, and he knows that one link on the chain is several orders of magnitude weaker than the other, it is not important for the engineer to have data on the strong link. Because the ability of the chain to withstand the pull is similar to the ability of the weakest link to withstand the pull, the engineer knows all that he needs to know when he knows the strength of the weakest link. This alone enables him to generate an estimate of the risk that the chain will break, and there is little to be gained by a separate examination of the breakage risks associated with the stronger link. Similarly, a trier of fact gains little from the opportunity to conduct separate examinations of the risks associated with laboratory error and

³⁰ To my knowledge, no one has suggested that any DNA laboratory or analyst has an error rate on the order of one in millions or billions.

the risk associated with a coincidental match.

This view was recently rejected by a judge in the first evidentiary hearing on the admissibility of PCR DNA technology in U.S. Federal Court. The judge opined that the probabilities associated with error and coincidental match should not be combined because a jury must first decide whether there is error and--if it concludes that error has not occurred--must then assess the strength of the DNA evidence based on the chance that the match is or is not coincidental.³¹ This sequential approach to decision making violates some of the most elementary rules of probability. Specifically, it can lead to conjunctive and disjunctive fallacies,³² and to the selection of dominated alternatives.³³

For example, suppose a jury is instructed to return a verdict for the defendant if it believes that either circumstance A or circumstance B is more likely than not to have occurred. Suppose further that the jury considers each circumstance sequentially, and concludes that $P(A) = .40$ and $P(B) = .40$. Because the jury believes that neither of the circumstances is more likely than not, it returns a verdict for the plaintiff.

Although this jury may have been true to the instructions it received, it is not clear that a verdict for the plaintiff is the appropriate one from a probabilistic accuracy standpoint. Depending on the nature of the dependence between circumstances A and B, the disjunctive probability of A or B is $.4 \leq P(A \vee B) \leq .8$.¹⁰ If the jury believes that A and B are mutually exclusive (i.e., $P(A \cap B) = 0$)¹¹ it should conclude that the disjunctive probability is $.80$.³⁴ Such

³¹ **U.S. v. Shea, Day 4, Afternoon session, Transcript of hearing before the Honorable Paul J. Barbadoro (August 22, 1996), p. 188.**

³² **Amos Tversky and Daniel Kahneman, Extensional vs. Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment, 91 PSYCH. REV. 293 (1983); Amos Tversky and Eldar Shafir, The Disjunction Effect in Choice Under Uncertainty, 3 PSYCHOLOGICAL. SCI. 305 (1992).**

³³ **Amos Tversky and Daniel Kahneman, Judgment Under Uncertainty: Heuristics and Biases, 185 SCIENCE 1124 (1974).**

³⁴ **$P(A \vee B) = P(A) + P(B) - P(A \cap B)$. Error! Main Document Only. If A & B are mutually exclusive, $P(A \vee B) = .4 + .4 - 0 = .8$. Error! Main Document Only.**

a conclusion would be more consistent with a verdict for the defendant than for the plaintiff. Similarly, jurors who follow the U.S. Federal Judge's guidelines and assess the strength of a DNA match in accordance via a sequential processing rule rather than the rules of probability may end up with an inflated sense of the value of the evidence against the defendant relative.

7. Error rates are irrelevant because questions of accuracy can be resolved through retesting.

"[T]here is no need to debate differing estimates of false-match error rates when the question of a possible false match can be put to direct test. . . [R]etesting provides an opportunity to identify and correct errors that might have been made during the course of analysis" (p. 3-11).

Retesting is an excellent suggestion and should reduce the chance that a false match report will arise in court. But retesting does not protect against errors that may occur during the earliest stages of handling (e.g., mislabeling). Even when early stage errors have not occurred, retesting may not provide as much protection against the risk of duplicate errors as some think. The risk of duplicate error lies somewhere between the risk of a single error E and the risk of duplicate error in the special case where errors are independent across laboratories, E^2 . Although no studies have examined error dependence across laboratories, the fact that some laboratories have committed some of the same errors in proficiency testing suggests that errors may not be independent.³⁵

V. CONCLUSION

By carefully defining "the evidence," a LR analysis shows that the possibility of feature-based and evaluative errors limits the strength of DNA evidence in much the same way as it limits the strength of eyewitness testimony. It is no more scientifically acceptable to

³⁵ See Jonathan J. Koehler, Audrey Chia, and J. Sam Lindsey, *The Random Match Probability (RMP) in DNA Evidence: Irrelevant and Prejudicial?* 35 *JURIM. J.* 201 (1995).

exclude feature-based errors from LR computations that purportedly reflect the strength of the evidence than it would be to compute the strength of eyewitness testimony based on the assumption that the witness correctly observed and reported the characteristics at issue.

With this in mind, NRC-II's position on error rates is puzzling. It recommends against incorporating error rate estimates into LR computations. Most of the reasons it offers in support of this recommendation have more to do with identifying inevitable imperfections in the estimates than with supporting the logic of excluding from consideration information about ceilings on the diagnostic strength of the evidence.

As LRs find their way into the courtroom with increasing frequency, it is important to know that the values forensic scientists offer fairly represent the strength of the evidence.³⁶ But because DNA LRs that assume away feature-based errors will often be substantially larger than LRs that do not, there is a real risk that our legal decision makers may be misled.

³⁶ Professor David Kaye, a member of NRC-II, recently wrote: "the forensic expert faces the unique challenge of presenting the scientific evidence bearing on the hypothesis so that the jury will appreciate its actual probative value," David H. Kaye, *Criminology: The Relevance of "Matching" DNA: Is the Window Half Open or Half Shut?* 85 J. CRIM. LAW & CRIMINOL. 676, 688 (1995).