THE BURDEN OF PROOF AND THE PRESENTATION OF FORENSIC RESULTS

Edward K. Cheng[*]

INTRODUCTION

To reach a conclusion of "match" or "non-match," a forensic analyst necessarily applies some threshold or burden of proof. To reach a conclusion of "guilty" or "not guilty" in a criminal trial, a jury uses the forensic conclusion along with other evidence, but also applies a burden of proof — the familiar beyond a reasonable doubt standard. What is the relationship between these two burdens of proof?[1] Should forensic "matches" be made according to the beyond a reasonable doubt standard? If not, then what standard should forensic examiners use?[2] Although seemingly technical at first, this question implicates some of the criminal justice system's deepest values. Applying the wrong forensic threshold could conceivably water down the reasonable doubt standard. Worse yet, by manipulating the threshold, forensic examiners could effectively usurp some of the jury's role in deciding guilt or innocence.

The recent Organization of Scientific Area Committees (OSAC) Letter[3] that is the subject of this *Harvard Law Review Forum* Commentary Series represents a laudable first step in addressing these questions. The Letter reaches two important conclusions: First, it states that the reasonable doubt standard does not preclude the use of different statistical procedures.[4] Second, it argues that "report of a match without more information about the probability of a match . . . would not fulfill the expert's role of impartially and adequately educating the trier of fact."[5] Both of these statements are facially correct,

---

[*] Professor of Law, Vanderbilt Law School. Thanks to David Kaye, Paul Edelman, and Alicia Solow-Niederman for helpful conversations and comments, and to Rachel Johnston for research assistance.

[1] *See* Memorandum from the Legal Res. Comm. to the Org. of Sci. Area Comms. for Forensic Sci., Nat'l Inst. of Standards & Tech., Question on Hypothesis Testing in ASTM 2926-13 and the Legal Principle that False Convictions Are Worse than False Acquittals (rev. ed. Oct. 7, 2016), *reprinted in* 130 HARV. L. REV. F. 137 (2017) [hereinafter Legal Resource Committee Memorandum].

[2] Some researchers have argued that forensic examiners should avoid conclusions like "match" or "non-match" altogether. *See* Geoffrey Stewart Morrison et al., Letter to the Editor, *A Comment on the PCAST Report: Skip the "Match"/"Non-Match" Stage*, 272 FORENSIC SCI. INT'L e7 (2017). Those arguments, however, are beyond the scope of this Commentary, which assumes the current practice of reaching conclusions.

[3] Legal Resource Committee Memorandum, *supra* note 1, at 2.

[4] *See id.*

[5] *Id.* at 7.

but ultimately the Letter fails to connect the dots completely. Indeed, as this Commentary will show, the two conclusions are fundamentally linked: if the jury receives proper contextual information — specifically, the likelihood ratio associated with the "match" or "non-match" — then the burden of proof does not require a specific statistical procedure.

The surprising answer to the original question about what is the "right" threshold is that there is actually no "right" threshold, nor do we need one. As long as the jury receives the likelihood ratio, it does not matter what threshold the forensic examiner uses.[6] The likelihood ratio, which measures the evidentiary worth or probative value of the forensic conclusion, incorporates the stringency or laxity of the threshold used by the examiner. Armed with that additional information, the jury can weigh the forensic conclusion along with everything else presented at trial using the proper burden of proof. The key lesson is that this contextual information matters, and it matters a lot. If forensic examiners present mere conclusions, then we do have to worry about the threshold. Without additional contextual information, the jury must weigh the conclusion in some generic way, and the forensic examiner's threshold can hijack the proof process. But given a measure of probative value like the likelihood ratio, the jury has enough information to maintain firm control.

The remainder of this Commentary develops these ideas with greater sophistication. Part I explores the fundamental tradeoff between false positives and false negatives in decisionmaking, and observes that every statistical test represents a tradeoff between these two types of error. Part II introduces likelihood ratios, and describes how they account for the false-positive–false-negative tradeoff in a forensic test, freeing the forensic examiner to use whatever threshold he likes. Part III briefly concludes by noting some practical challenges to this solution to the threshold problem.

THE ERROR TRADEOFF

In law, we often talk about error as if it were monolithic, but there are in fact two separate kinds: false positives and false negatives.[7]

---

[6] *Cf.* D.H. Kaye, *The Relevance of "Matching" DNA: Is the Window Half Open or Half Shut?*, 85 J. CRIM. L. & CRIMINOLOGY 676, 694–95 (1995) (making a somewhat analogous argument about likelihood ratios in debates about DNA profiling procedures).

[7] There are actually four potential error rates: Let $+$ and $-$ denote positive and negative test results, and C and $\overline{C}$ denote the presence and absence of the condition of interest. In this Commentary, we focus on the false positive rate, $P(+|\overline{C})$, and the false negative rate, $P(-|C)$. There are, however, two other error rates: the false discovery rate, $P(\overline{C}|+)$, and the false omission rate, $P(C|-)$. These latter two error rates require knowledge of the prevalence or base rate of the

Moreover, any diagnostic or decision test will necessarily trade these errors off against each other. For example, suppose we want to decide whether two samples of glass came from the same window pane based on their chemical compositions. The resulting analysis will have some random error, caused by, for example, imprecision in the instruments or variation in the manufacturing process, so a forensic examiner will inevitably observe some difference. How close then is close enough for a "match"? An exacting threshold, requiring a difference extremely close to zero, will minimize false positives, but will pay the price in false negatives. A lax threshold will do the opposite. Thus, given any chosen method of analysis, we actually have a whole range of possible tests, each corresponding to a different threshold, and accordingly, a different false-positive–false-negative (FPFN) tradeoff. This tradeoff is so fundamental that researchers have long assessed diagnostic methods using receiver-operator characteristic (ROC) curves, which essentially plot false positive rate against false negative rate.[8] It is the ROC curve that characterizes the method, not a specific false positive or false negative rate.

Classical hypothesis testing, like every other decision rule, makes a specific FPFN tradeoff. Long use and the strong conventions surrounding statistical significance may serve to hide its choices, but they operate in the background nevertheless. Classical hypothesis testing preferences the null hypothesis (H_0), which typically represents the status quo or what one is trying to disprove, over the alternative (H_1), which typically is what one is trying to prove. Specifically, it sets a maximum false positive rate, conventionally 0.05, and then seeks to minimize the false negative rate subject to this condition.[9] There are, however, no guarantees about the false negative rate: it may be small, large, or absurdly large.[10] The priority is that the false positive rate is

---

condition studied, and thus are not inherent properties of the test. Again, our focus here will only be on the false positive and false negative rates.

[8] Conventionally, ROC curves plot the false positive rate on the x-axis and the true positive rate (or sensitivity) on the y-axis. Since the false negative rate $P(-\mid C)$ is just one minus the true positive rate $P(+\mid C)$, the ROC curve contains the same information as a false positive rate vs. false negative rate plot.

[9] Following the earlier footnote, let $+$ and $-$ denote a rejection and acceptance, respectively, of the null hypothesis. Further denote the null hypothesis as $\overline{C}$ (that is, the absence of a condition) and the alternative hypothesis as C. Then classical hypothesis testing ensures that $P(+\mid \overline{C}) \leq \alpha$, where α is often set to 0.05.

[10] One way in which large false negative rates occur is when a method is too imprecise to be practically useful. Consider this extreme example: Suppose we have a useless instrument that always returns the same value no matter what we test. Classical hypothesis testing requires that we control the false positive rate — the probability of declaring a positive when a condition is actually absent, $P(+\mid \overline{C})$. But since we have a useless instrument, the only way to guarantee a low false positive rate is to never declare a positive at all. In other words, the rule becomes "always report negative." But if we always report negatives, then our test will always miss true positives, mean-

guaranteed to be small, and this setup is useful in certain scientific contexts.[11]

When a forensic examiner uses classical hypothesis testing to compare glass samples, he therefore implicitly makes a choice between false positives and false negatives. One natural setup parallels the "equivalence testing" methods used when generic drugs are tested for pharmaceutical equivalence: the null is that the two glass samples are different, whereas the alternative is that they are the same, where "sameness" is defined as being within some tolerance range.[12] Under these conditions, classical hypothesis testing ensures some maximum false positive rate, and then tries to minimize false negatives. The ASTM standard on glass comparisons also involves classical hypothesis testing, but curiously makes the opposite FPFN tradeoff.[13] The ASTM standard seems to define the null as "same," and the alternative as "different," and so it effectively caps the false negative rate (concluding that the glass is different when it is the same), while making no guarantees about the false positive rate (concluding that the glass is the same when it is different).[14] Given the criminal forensic context, this ASTM setup is arguably backwards, since most would agree that we should worry far more about false positives (that is, false matches) than false negatives.[15] Regardless, the key point remains that whatever test one uses, one is making a value judgment about the appropriate tradeoff between false positives and false negatives.

Even Bayesian approaches ultimately suffer this fate. A Bayesian approach would eschew rejecting or accepting a null hypothesis, and

---

ing that the false negative rate, $P(-\mid C)$, is 100%. Thus, we have a test in which the false positive rate is controlled, but which teaches us nothing.

[11] Notably, this asymmetric setup means that while rejection of the null hypothesis may be significant evidence in favor of the alternative, acceptance of the null may be only weak (if any) evidence in favor of the null. *E.g.*, Douglas G. Altman & J. Martin Bland, *Statistics Notes: Absence of Evidence Is Not Evidence of Absence*, 311 BRIT. MED. J. 485 (1995).

[12] *See generally* R. Clifford Blair & Stephen R. Cole, *Two-Sided Equivalence Testing of the Difference Between Two Means*, 1 J. MOD. APPLIED STAT. METHODS 139, 139 (2002) (describing equivalence testing); Giselle B. Limentani et al., *Beyond the* t*-Test: Statistical Equivalence Testing*, 77 ANALYTICAL CHEMISTRY 221A, 223A (2005) (same).

[13] *See* ASTM INT'L, ASTM E2926-13 STANDARD TEST METHOD FOR FORENSIC COMPARISON OF GLASS USING MICRO X-RAY FLUORESCENCE (μ-XRF) SPECTROMETRY §§ 10.7.3.1–.2 (2013).

[14] The ASTM standard seems to use a one-population *t*-test with a three standard deviation cutoff, corresponding to a 0.01, rather than 0.05, probability of error. *See id.* § 10.7.3.2. The one-population *t*-test here is technically wrong, because the population means for both specimens need to be estimated and thus there are two sources of uncertainty. *Cf.* David H. Kaye, *Reflections on Glass Standards: Statistical Tests and Legal Hypotheses*, 27 STATISTICA APPLICATA 173, 175 (2015) (making a similar point about ASTM E1967-11a, dealing with glass refractive index comparisons). A two-population *t*-test is therefore arguably preferable, but we will ignore this technicality going forward.

[15] *Cf.* Kaye, *supra* note 14, at 179 (making a similar point about ASTM E1967-11a).

instead focus on estimating the difference between the two samples and quantifying the associated uncertainty.[16] Yet, to render a "match" or "non-match" determination, an expert must ultimately apply a threshold to that estimate, and that choice of threshold trades off false positives and false negatives.[17]

THE LIKELIHOOD RATIO

If experts must invariably make a value judgment as to the importance of false positives versus false negatives, what FPFN tradeoff should they choose? Here, the initial answer is far from clear. First, just because the ultimate burden of proof is beyond a reasonable doubt does not mean that evidence must meet that standard to be admissible. To be relevant, the evidentiary rules merely require that evidence have "any tendency to make a fact more or less probable,"[18] because "a brick is not a wall."[19] Even the reliability requirements under *Daubert* only need to be established to a preponderance standard under Rule 104(a).[20] Second, even if we wanted to "match" the burden used for individual evidentiary elements to the ultimate burden, the relationship between the two is fraught with peril, as seen in other evidentiary contexts.[21] Finally, courts have long maintained that it is the jury's prerogative to define reasonable doubt.[22] So neither a top-down imposition of a particular FPFN tradeoff nor granting unbridled discretion to the expert seems to be the proper response.

Fortunately, there exists an elegant path around this morass. As it turns out, as long as juries receive a key piece of information about a forensic conclusion — the likelihood ratio — experts may choose any threshold they please. This likelihood ratio not only provides jurors

---

[16] *See* John K. Kruschke, *Bayesian Estimation Supersedes the* t-*Test*, 142 J. EXPERIMENTAL PSYCHOL. 573 (2013) (advocating for Bayesian estimation over classical *t*-tests).

[17] Again, we assume here that the experts continue to testify to forensic conclusions, rather than merely provide the jury with quantitative estimates.

[18] FED. R. EVID. 401(a).

[19] FED. R. EVID. 401 advisory committee's note (quoting CHARLES TILFORD MCCORMICK, MCCORMICK ON EVIDENCE § 152, at 317 (Edward W. Cleary ed., 2d ed. 1972)).

[20] Bourjaily v. United States, 483 U.S. 171, 175 (1987); *see also* FED. R. EVID. 104(a).

[21] For example, suppose that one could prove three elements to a probability of 0.6, so that each individually satisfied the preponderance standard. If all three elements are independent, however, the probability of their conjunction is $0.6 \times 0.6 \times 0.6 = 0.216$, which does not satisfy the preponderance standard. This problem is known as the Conjunction Paradox and has spawned an extensive literature. *See, e.g.*, Edward K. Cheng, *Reconceptualizing the Burden of Proof*, 122 YALE L.J. 1254, 1263 (2013); Mark Spottswood, *Unraveling the Conjunction Paradox*, 15 LAW, PROBABILITY & RISK 259 (2016).

[22] For example, courts have firmly eschewed quantification of the beyond a reasonable doubt standard. Peter Tillers & Jonathan Gottfried, *Case Comment* — United States v. Copeland*, 369 F. Supp. 2d 275 (E.D.N.Y. 2005): A Collateral Attack on the Legal Maxim that Proof Beyond a Reasonable Doubt Is Unquantifiable?*, 5 LAW, PROBABILITY & RISK 135, 135–37 (2006).

with a convenient measure of the probative value of the forensic conclusion, but also ensures that the jury — and not the expert — is the only actor applying the ultimate burden of proof.

A likelihood ratio has the following form:

$$\frac{P(E \mid H_1)}{P(E \mid H_0)}$$

where E represents the evidence observed, and H_1 and H_0 are the competing hypotheses. In the forensic testing context, E is the outcome of the forensic test — namely a declared "match" or "non-match." H_1 is the prosecution's story, which is that the defendant is guilty, or perhaps more modestly the fact that glass found on the defendant is the same glass from the crime scene. H_0 is the defense's position.

Scholars have long argued that the likelihood ratio is the mathematical representation of relevance.[23] Conceptually, evidence makes material facts more or less probable because the probability of seeing such evidence is more likely under one side's story than the other's. If the evidence is just as likely to surface under the prosecution's story as the defendant's story, it does not help us distinguish between the two. The likelihood ratio also features prominently in Bayes' Rule. For Bayesians, it is how one updates one's prior probability ratio, the persuasive balance between the two hypotheses before seeing the evidence, to obtain the posterior probability ratio, the persuasive balance after seeing the evidence. To wit,

$$\frac{P(H_1 \mid E)}{P(H_0 \mid E)} = \frac{P(E \mid H_1)}{P(E \mid H_0)} \times \frac{P(H_1)}{P(H_0)}$$

More importantly for our purposes, the likelihood ratio is not just related to the FPFN tradeoff, but in fact serves as its mathematical equivalent. Indeed, one scholar has suggested that ROC curves be transformed into plots of the positive likelihood ratio (the evidentiary worth of a "match") versus the negative likelihood ratio (the evidentiary worth of a "non-match").[24] This equivalence can be understood intuitively with a simple example: Stringent tests minimize false positives at the expense of false negatives. The probative value of a positive ("match") from a stringent test is therefore relatively high, because we can be confident that a declared "match" is the real thing. By contrast the probative value of a negative ("non-match") is relatively lower, since there are more false negatives muddling the waters. The rela-

---

[23] *See, e.g.*, David H. Kaye, Comment, *Quantifying Probative Value*, 66 B.U. L. REV. 761, 762 (1986); Richard O. Lempert, *Modeling Relevance*, 75 MICH. L. REV. 1021, 1022–32 (1977).

[24] Nils P. Johnson, *Advantages to Transforming the Receiver Operating Characteristic (ROC) Curve to Likelihood Ratio Co-Ordinates*, 23 STAT. MED. 2257, 2258 (2004).

tionship between false positive and false negative rate parallels the relationship between the positive likelihood ratio and the negative likelihood ratio.[25]

This relationship between the FPFN tradeoff and the likelihood ratio is extremely good news. No matter what tradeoff a forensic examiner makes between false positives and false negatives, that choice will be reflected in the probative value accorded the forensic conclusion. If the examiner chooses a stringent standard and gets a "match," then the jury can weigh that "match" heavily. If the examiner chooses a lax standard and gets a "match," then the jury can weigh it more skeptically. In either case, the jury can give the forensic evidence proper weight and then apply the burden of proof independently and without interference from the expert.[26] Thus, as long as forensic results are accompanied by their likelihood ratios, we avoid the need to match the forensic standard with the burden of proof.

CONCLUSION AND SOME PRACTICAL ISSUES

This Commentary has shown that any conclusion of "match" or "non-match" implicitly requires a judgment about the proper tradeoff between false positives and false negatives. Determining what tradeoff best coheres with the beyond a reasonable doubt standard (or any ultimate legal burden), however, is complicated and controversial. Fortunately, one can avoid this problem by presenting the likelihood ratio associated with the forensic conclusion to the factfinder. The likelihood ratio not only provides a mathematical representation of the result's probative value, but also naturally incorporates the choice of FPFN tradeoff made by the forensic examiner into the probative value. As long as the likelihood ratio is given, a forensic examiner can apply whatever FPFN tradeoff the examiner chooses.

Although in theory likelihood ratios solve the problem of "matching" a forensic examiner's FPFN tradeoff with the burden of proof, there remain several important practical issues. First, the solution pre-

---

[25] In mathematical terms, the two pairs are transformations of each other. Recall that $FPR = P(+ \mid \overline{C})$ and $FNR = P(- \mid C)$. Then,

$$LR_+ = \frac{P(+ \mid C)}{P(+ \mid \overline{C})} = \frac{1 - FNR}{FPR}$$

and

$$LR_- = \frac{P(- \mid C)}{P(- \mid \overline{C})} = \frac{FNR}{1 - FPR}$$

See also id. at 2258–59 (making a similar derivation).

[26] Under a Bayesian framework, each piece of information updates the prior probability ratio ($P(H_1)/P(H_0)$) until the jury has a final posterior probability ratio ($P(H_1 \mid E)/P(H_0 \mid E)$). The jury then decides how high a ratio is necessary in order for the prosecution to win under the reasonable doubt standard.

supposes that such likelihood ratios exist. Many forensic fields still lack the population data on their techniques that would be needed to calibrate the likelihood ratios. Some, such as toolmarks or handwriting, are subjective, making even the collection of such empirical data difficult. A likelihood ratio–based solution will require the use of more objective forensic techniques and the collection of such data. Second, whether courts would mandate the presentation of likelihood ratios remains an open question. Scholars have long argued in favor of presenting forensic results using likelihood ratios,[27] and indeed some forensic communities in Europe have embraced them,[28] but by and large in the United States, forensic results are still presented as bald conclusions. Finally, the solution requires that juries or other legal actors comprehend and properly use likelihood ratios. An increasingly complex literature has emerged on lay understanding of likelihood ratios and how such quantitative information is best presented.[29] Research thus far has yielded no easy answers, with Professor William C. Thompson and Eryn J. Newman recently concluding that the best presentation method may depend on context and the specific forensic discipline,[30] and even then, the definition of "best" is debatable.[31]

Conceding these practical issues, however, the more critical point is how likelihood ratios can provide jurors sufficient information to wrestle back control of the proof process from experts. The upshot is that a forensic examiner's choice of threshold does not necessarily distort the burden of proof. Does the legal system need continued research to determine the best ways to present likelihood ratios? Do we need to

---

[27] DAVID H. KAYE ET AL., THE NEW WIGMORE ON EVIDENCE: EXPERT EVIDENCE, at chs. 13–14 (2016); COMMUNICATING THE RESULTS OF FORENSIC SCIENCE EXAMINATIONS: FINAL TECHNICAL REPORT FOR NIST AWARD 70NANB12H014, at 2 (Cedric Neumann, Anjali Ranadive & David H. Kaye eds., 2015).

[28] Colin Aitken et al., Guest Editorial, *Expressing Evaluative Opinions: A Position Statement*, 51 SCI. & JUST. 1, 1–2 (2011).

[29] *See* William C. Thompson & Eryn J. Newman, *Lay Understanding of Forensic Statistics: Evaluation of Random Match Probabilities, Likelihood Ratios, and Verbal Equivalents*, 39 LAW & HUM. BEHAV. 332 (2015); *see also* K.A. Martire et al., *On the Interpretation of Likleihood Ratios in Forensic Science Evidence: Presentation Formats and the Weak Evidence Effect*, 240 FORENSIC SCI. INT'L 61 (2014); Kristy A. Martire et al., *The Expression and Interpretation of Uncertain Forensic Science Evidence: Verbal Equivalence, Evidence Strength, and the Weak Evidence Effect*, 37 LAW & HUM. BEHAV. 197 (2013). *See generally* Jonathan J. Koehler, *On Conveying the Probative Value of DNA Evidence: Frequencies, Likelihood Ratios, and Error Rates*, 67 U. COLO. L. REV. 859 (1996) (discussing the problems of presentation of DNA statistics).

[30] Thompson & Newman, *supra* note 29, at 339.

[31] Studies show that juries will give different weight to different presentations of the same mathematically equivalent material. Jury sensitivity to the strength of statistical evidence also varies by presentation method. *See* sources cited *supra* note 29. But we ultimately have an Archimedean point problem here, because we do not know which juror reaction is correct. Some researchers have used Bayesian reasoning as the goal, but even then, the presence of different priors and other factors makes assessment of mock juror behavior on an absolute scale difficult.

develop methods to teach jurors how to use statistical information? The answer to these questions is clearly yes. But teaching and helping jurors are projects with which the legal system is familiar. The key is that likelihood ratios present a clear path to improving the use of forensics testimony in court.