Forensic Science, Statistics & the Law

Commentary on news and publications at the intersections of scientific evidence, forensic science, and statistics.

Tuesday, November 1, 2016

PCAST and the Ames Bullet Cartridge Study: Will the Real Error Rates Please Stand Up?

An article in yesterday's *Boston Globe* reports that "the [PCAST] report's findings have also been widely criticized, especially by those in the forensics field, who argued that the council lacked any representation from ballistics experts. They argued that the council's findings do not undermine the accuracy of firearms examinations." 1/

The criticism that "ballistics experts" did not participate in writing the report is unpersuasive. These experts are great at their jobs, but reviewing the scientific literature on the validity and reliability of their toolmark comparisons is not a quotidian task. Would one criticize a meta-analysis of studies on the efficacy of a surgical procedure on the ground that the authors were epidemiologists rather than surgeons?

On the other hand, the argument that the "findings do not undermine the accuracy of firearms examinations" is correct (but inconclusive). True, the President's Council of Advisors on Science and Technology (PCAST) did not find that toolmark comparisons as currently practiced are inaccurate. Rather, it concluded (on page 112) that

> [F]irearms analysis currently falls short of the criteria for foundational validity, because there is only a single appropriately designed study to measure validity and estimate reliability. The scientific criteria for foundational validity require more than one such study, to demonstrate reproducibility.

In other words, PCAST found that existing literature (including that called to its attention by "ballistics experts") does not adequately answer the question of how accurate firearms examiners are when comparing markings on cartridges—because only a single study that was designed as desired by PCAST provides estimates of accuracy.

Although PCAST's view is that more performance studies are necessary to satisfy Federal Rule of Evidence 702, PCAST uses the single study to derive a false-positive error rate for courtroom use (just in case a court disagrees with its understanding of the rule of evidence, or the science, or in case the jurisdiction follows a different rule).

To evaluate PCAST's proposal, it will be helpful first to describe what the study itself found. Athough "it has not yet been subjected to peer review and publication" (p. 111), the "Ames study," as PCAST calls it, is available online. 2/ The researchers enrolled 284 volunteer examiners in the study, and 218 submitted answers (raising an issue of selection bias). The 218 subjects (who obviously knew they were being tested) "made ... l5 comparisons of 3 knowns to 1 questioned cartridge case. For all participants, 5 of the sets were from known same-source firearms [known to the researchers but not the firearms examiners], and 10 of the sets were from known different-source firearms.” 3/ Ignoring “inconclusive” comparisons, the performance of the examiners is shown in Table 1.

Table 1. Outcomes of comparisons
(derived from pp. 15-16 of Baldwin et al.)

	~S	***S***	
–E	1421	4	1425
+E	22	1075	1097
	1443	1079	

–E is a negative finding (the examiner decided there was no association).
+E is a positive finding (the examiner decided there was an association).
S indicates that the cartridges came from bullets fired by the same gun.
~S indicates that the cartridges came from bullets fired by a different gun.

False negatives. Of the 4 + 1075 = 1079 judgments in which the gun was the same, 4 were negative. This false negative rate is *Prop*(*–E* |*S*) = 4/1079 = 0.37%. ("Prop" is short for "proportion," and "|" can be read as "given" or "out of all.") Treating the examiners tested as random samples of all examiners of interest, and viewing the performance in the experiment as representative of the examiners' behavior in casework with materials comparable to those in the experiment, we can estimate the portion of false negatives for all examiners. The point estimate is 0.37%. A 95% confidence interval is 0.10% to 0.95%. These numbers provide an estimate of how frequently all examiners would declare a negative association in all similar cases in which the association actually is positive.Instead of false negatives, we also can describe true negatives, or specificity. The observed specificity is *Prop*(*E*|*~S*) = 99.63%. The 95% confidence interval around this estimate is 99.05% to 99.90%.

False positives. The observed false-positive rate is *Prop*(*+E* |*~S*) = 22/1443 = 1.52%, and the 95% confidence interval is 0.96% to 2.30%. The observed true-positive rate, or sensitivity, is 98.48%, and its 95% confidence interval is 97.7% to 99.04%.

Taken at face value, these results seem rather encouraging. On average, examiners displayed high levels of accuracy, both for cartridge cases from the same gun (better than 99% specificity) and from different guns (better than 98% sensitivity).

Applying such numbers to individual examiners and particular cases obviously is challenging. The PCAST report largely elides the difficulties. (See Box 1.) It notes (on page 112) that "20 of the 22 false positives were made by just 5 of the 218 examiners — strongly suggesting that the false positive rate is highly heterogeneous across the examiners"; however, it does not discuss the implications of this fact for testimony about "the error rates" that it wants "clearly presented." It calls for "rigorous proficiency testing" of the examiner and disclosure of those test results, but it does not consider how the examiner’s level of proficiency maps onto to the distribution of error rates seen in the Ames study. Neither does it consider how testimony should address the impact of verification by a second examiner. If the errors occur independently across examiners (as might be the case if the verification is truly blind), then the relevant false-positive error rate drops to (1.52%)^2 = 0.0231%. Is omitting some correction for verification an appropriate way to present the results of a rigorously verified examination? Indeed, is a false-positive error rate enough to convey the probative value of a positive finding? I will discuss the last question later.

BOX 1. PCAST's PRESCRIPTION (IN PART) FOR PRESENTING POSITIVE FINDINGS

Foundational validity. PCAST finds that firearms analysis currently falls short of the criteria for foundational validity, ... If firearms analysis is allowed in court, the scientific criteria for validity as applied should be understood to require clearly reporting the error rates seen in appropriately designed black-box studies (estimated at 1 in 66, with a 95 percent confidence limit of 1 in 46, in the one such study to date). [P. 112.]

Validity as applied. If firearms analysis is allowed in court, validity as applied would, from a scientific standpoint, require that the expert: (1) has undergone rigorous proficiency testing on a large number of test problems to evaluate his or her capability and performance, and discloses the results of the proficiency testing ... . [P. 113.]

BOX 2. FINDINGS ABOUT FALSE POSITIVES AS DESCRIBED IN THE AMES STUDY

[The] false-positive rate for examiner cartridge case comparisons ... was measured and for the pool of participants used in this study the fraction of false positives was approximately 1%. The study was specifically designed to allow us to measure not simply a single number from a large number of comparisons, but also to provide statistical insight into the distribution and variability in false-positive error rates. The ... overall fraction is not necessarily representative of a rate for each examiner in the pool. Instead, ... the rate is a highly heterogeneous mixture of a few examiners with higher rates and most examiners with much lower error rates. This finding does not mean that 1% of the time each examiner will make a false-positive error. Nor does it mean that 1% of the time laboratories or agencies would report false positives, since this study did not include standard or existing quality assurance procedures, such as peer review or blind reanalysis. [P. 18.]

Notes

1. Milton J. Valencia, Scrutiny over Forensics Expands to Ballistics, Boston Globe, Oct. 31, 2016, https://www.bostonglobe.com/metro/2016/10/31/firearms-examinations-forensics-come-under-review/zJnaTjiGxCMuvdStkuSvyO/story.html
2. David P. Baldwin, Stanley J. Bajic, Max Morris & Daniel Zamzow, A Study of False-positive and False-negative Error Rates in Cartridge Case Comparisons, Ames Laboratory, USDOE, Technical Report #IS-5207 (2014), at https://afte.org/uploads/documents/swggun-false-postive-false-negative-usdoe.pdf
3. Id. at 10.

Forensic Science, Statistics & the Law

Commentary on news and publications at the intersections of scientific evidence, forensic science, and statistics.

Thursday, November 3, 2016

The False-Positive Fallacy in the First Opinion to Discuss the PCAST Report

Last month, I quoted the following discussion of the PCAST report on forensic science that appeared in *United States v. Chester*, No. 13 CR 00774 (N.D. Ill. Oct. 7, 2016):

> As such, the report does not dispute the accuracy or acceptance of firearm toolmark analysis within the courts. Rather, the report laments the lack of scientifically rigorous “blackbox” studies needed to demonstrate the reproducibility of results, which is critical to cementing the accuracy of the method. Id. at 11. The report gives detailed explanations of how such studies should be conducted in the future, and the Court hopes researchers will in fact conduct such studies. See id. at 106. However, PCAST did find one scientific study that met its requirements (in addition to a number of other studies with less predictive power as a result of their designs). That study, the “Ames Laboratory study,” found that toolmark analysis has a false positive rate between 1 in 66 and 1 in 46. Id. at 110. The next most reliable study, the “Miami-Dade Study” found a false positive rate between 1 in 49 and 1 in 21. Thus, the defendants’ submission places the error rate at roughly 2%.[3] The Court finds that this is a sufficiently low error rate to weigh in favor of allowing expert testimony. See Daubert v. Merrell Dow Pharms., 509 U.S. 579, 594 (1993) (“the court ordinarily should consider the known or potential rate of error”); United States v. Ashburn, 88 F. Supp. 3d 239, 246 (E.D.N.Y. 2015) (finding error rates between 0.9 and 1.5% to favor admission of expert testimony); United States v. Otero, 849 F. Supp. 2d 425, 434 (D.N.J. 2012) (error rate that “hovered around 1 to 2%” was “low” and supported admitting expert testimony). The other factors remain unchanged from this Court’s earlier ruling on toolmark analysis. See ECF No. 781.
>
> 3. Because the experts will testify as to the likelihood that rounds were fired from the same firearm, the relevant error rate in this case is the false positive rate (that is, the likelihood that an expert’s testimony that two bullets were fired by the same source is in fact incorrect).

I suggested that the court missed (or summarily dismissed) the main point the President's Council of Science and Technology Advisers were making -- that there is an insufficient basis in the literature for concluding that "the error rate [is] roughly 2%," but the court's understanding of "the error rate" also merits comment. The description of the meaning of "the false positive rate" in note 3 (quoted above) is plainly wrong. Or, rather, it is subtly wrong. If the experts will testify that two bullets came from the same gun, they will be testifying that their tests were positive. If the tests are in error, the test results will be false positives. And if the false-positive error probability is only 2%, it sounds as if there is only a 2% probability "that [the] expert's testimony ... is in fact incorrect."

But that is not how these probabilities work. The court's impression reflects what we can call a "false-positive fallacy." It is a variant on the well-known transposition fallacy (also loosely called the prosecutor's fallacy). Examiner-performance studies are incapable of producing what the court would like to know (and what it thought it was getting) -- "the likelihood that an expert’s testimony that two bullets were fired by the same source is in fact incorrect." The last phrase denotes the *probability that a source hypothesis is false*. It can be called a source probability. The "false positive rate" is the *probability that certain evidence will arise if the source hypothesis is true*. It can be called an evidence probability. As explained below, this evidence probability is but one of three probabilities that determine the source probability.

I. Likelihoods: The Need to Consider Two Error Rates

The so-called black-box studies can generate estimates of the evidence probabilities, but they cannot reveal the source probabilities. Think about how the performance study is designed. Examiners decide whether pairs of bullets or cartridges were discharged from the same source (S) or from different guns ($\sim S$). They are blinded to whether S or $\sim S$ is true, but the researchers control and know the true state of affairs (what forensic scientists like to call "ground truth"). The proportion of cases in which the examiners report a positive association ($+E$) out of all the cases of S can be written $Prop(+E \text{ in cases of } S)$, or more compactly, $Prop(+E \mid S)$. This proportion leads to an estimate of the probability that, in practice, the examiners and others like them will report a positive association ($+E$) when confronted with same-source bullets. This conditional probability for $+E$ given that S is true can be abbreviated $Prob(+E \mid S)$. I won't be fastidious about the difference between a proportion and a probability and will just write $P(+E \mid S)$ for either, as the context dictates. In the long run, for the court's 2% figure (which is higher than the one observed false-positive proportion in the Ames study), we expect examiners to respond positively ($+E$) *when S is not true* (and they do reach a conclusion) only $P(+E \mid \sim S) =$ 2% of the time.

Surprisingly, a small number like 2% for the "false-positive error rate" $P(+E \mid \sim S)$ does not necessarily mean that the positive finding $+E$ has *any* probative value! Suppose that positive findings $+E$ occur just as often when S is false as when S is true. (Examiners who are averse to false-positive judgments might be prone to err on the side of false negatives.) If the false-negative error probability is $P(–E \mid S) =$ 98%, then examiners will tend to report $–E$ 98% of the time for same-source bullets (S), just as they report $+E$ 98% of the time for different-source bullets (S). Learning that such examiners found a positive association is of zero value in separating same-source cases from different-source cases. We may as well have flipped a coin. The outcome (the side of the coin, or the positive judgment of the examiner) bears no relationship to whether the S is true or not.

Although a false negative probability of 98% is absurdly high, it illustrates the unavoidable fact that only when the ratio of the *two likelihoods*, $P(+E \mid S)$ and $P(+E \mid \sim S)$, exceeds 1 is a positive association positive evidence of a true association. Consequently, the court's thought that "*the* relevant error rate in this case is the false positive rate" is potentially misleading. This likelihood is but one of the *two* relevant likelihoods. (And there would be still more relevant likelihoods if there were more than two hypotheses to consider.)

II. Prior Probabilities: The Need to Consider the Base Rate

Furthermore, yet another quantity -- the mix of same-source and different-source pairs of bullets in the cases being examined -- is necessary to arrive at the court's understanding of "the false positive rate" as "the likelihood that an expert’s testimony that two bullets were fired by the same source is in fact incorrect." 1/ In technical jargon, the probability as described is the complement of the posterior probability (or positive predictive value in this context), and the posterior probability depends on not only on the two likelihoods, or evidence probabilities, but also on the "prior probability" for the hypotheses *S*.

A few numerical examples illustrate the effect of the prior probability. Imagine that a performance study with 500 same-source pairs and 500 different-source pairs (that led to conclusions) found the outcomes given in Table 1.

Table 1. Outcomes of comparisons

	~S	***S***	
–E	490	350	840
+E	10	150	160
	500	500	

–*E* is a negative finding (the examiner decided there was no association).
+*E* is a positive finding (the examiner decided there was an association).
S indicates that the cartridges came from bullets fired by the same gun.
~*S* indicates that the cartridges came from bullets fired by a different gun.

The first column of the table states that in the different-source cases, examiners reported a positive association +*E* in only 10 cases. Thus, their false-positive error rate was *P*(+*E* | ~*S*) = 10/500 = 2%. This is the figure used in *Chester*. (The second column states that in the same-source cases, examiners reported a negative association 350 times. Thus, their false-negative rate was *P*(–*E* | *S*) = 350/500 = 70%.)

But the bottom row of the table states that the examiners reported a positive association +*E* for 10 different-source cases and 150 same-source cases. Of the 10 + 150 = 160 cases of positive evidence, 150 are correct, and 10 are incorrect. The rate of incorrect positive findings was therefore *P*(~*S* | +*E*) = 10/160 = 6.25%. Within the four corners of the study, one might say, as the court did, that "the likelihood that an expert’s testimony that two bullets were fired by the same source is in fact incorrect" is only 2%. Yet, the rate of incorrect positive findings in the study exceeded 6%. The difference is not huge, but it illustrates the fact that the false-negative probability as well as the false-positive probability affects *P*(~*S* | +*E*), which indicates how often an examiner who declares a positive association is wrong. 2/

Now let's change the mix of same- and different-source pairs of bullets from 50:50 to 10:90. We will keep the conditional-error probabilities the same, at *P*(+*E* | ~*S*) = 2% and *P*(–*E* | *S*) = 70%. Table 2 meets these constraints:

Table 2. Outcomes of comparisons

	~S	***S***	

–E	980	70	1050
+E	20	30	50
	1000	100	

Row 2 shows that there are 20 false positives out of the 50 positively reported associations. The proportion of false positives in the modified study is $P(\sim S \mid +E) = 40\%$. But the false-positive rate $P(+E \mid \sim S)$ is still 2% (20/1000).

III. "When I'm 64": A Likelihood Ratio from the Ames Study

The *Chester* court may not have had a correct understanding of the 2% error rate it quoted, but the Ames study does establish that examiners are capable of distinguishing between same-source and different-source items on which they were tested. Their performance was far better than the outcomes in the hypothetical Tables 1 and 2. The Ames study found that across all the examiners studied, $P(+E \mid S) = 1075/1097 = 98.0\%$, and $P(+E \mid \sim S) = 22/1443 = 1.52\%$. [3] In other words, on average, examiners made a correct positive associations 98.0/1.52 = 64 times more often when presented with same-source cartridges than they made incorrect positive associations when presented with different-source cartridges. This likelihood ratio, as it is called, means that when confronted with cases involving an even mix of same- and different-source items, over time and over all examiners, the pile of correct positive associations would be some 64 times higher than the pile of incorrect positive associations. Thus, in *Chester*, Judge Tharp was correct in suggesting that the one study that satisfied PCAST's criteria offers an empirical demonstration of expertise at associating bullet cartridges with the gun that fired them.

Likewise, an examiner presenting a source attribution can point to a study deemed to be well designed by PCAST that found that a self-selected group of 218 examiners given cartridge cases from bullets fired by one type of handgun correctly identified more than 99 out of 100 same-gun cartridges and correctly excluded more than 98 out of 100 different-gun cartridges. For completeness, however, the examiner should add that he or she has no database with which to estimate the frequency of distinctive marks -- unless, of course, there is one that is applicable to the case at bar.

* * *

Whether the Ames study, together with other literature in the field, suffices to validate the expertise under *Daubert* is a further question that I will not pursue here. My objective has been to clarify the meaning of and some of the limitations on the 2% false-positive error rate cited in *Chester*. Courts concerned with the scientific validity of a forensic method of identification must attend to "error rates." In doing so, they need to appreciate that it takes two to tango. Both false-positive and false-negative conditional-error probabilities need to be small to validate the claim that examiners have the skill to distinguish accurately between positively and negatively associated items of evidence.

Notes

[1] Not wishing to be too harsh on the court, I might speculate that its thought that the only "relevant error rate" for positive associations is the false-positive rate might have been encouraged by the PCAST report's failure to present any data on negative error rates in its discussion of the performance of firearms examiners. A technical appendix to the report indicates that the related likelihood is pertinent to the weight of the evidence, but this fact might be lost on the average reader -- even one who looks at the appendix.

[2] The PCAST report alluded to this effect in its appendix on statistics. That Judge Tharp did not pick up on this is hardly surprising.

[3] See David H. Kaye, PCAST and the Ames Bullet Cartridge Study: Will the Real Error Rates Please Stand Up?, Forensic Sci., Stat. & L., Nov. 1, 2016, http://for-sci-law.blogspot.com/2016/11/pcast-and-ames-study-will-real-error.html.