ASA Board Policy Statement on Forensic Science Reform

April 17, 2010

The 2009 National Academies' report, *Strengthening Forensic Science in the United States: A Path Forward,*[1] identified many serious deficiencies in the nation's forensic science system and called for major reforms and new research. The report came after years of critiques of specific forensic science practices as well as calls for reform but especially broke new ground by offering a comprehensive review and adding the authority of the National Academies.

Statisticians have played an important role in this constructive criticism and can play an important role in the reform urged by the National Academies' report. Indeed, the *Strengthening Forensic Science* report cites examples of the lack of sufficient recognition for sources of variability and their effects on uncertainties in forensic science analyses. Statisticians are vital to establishing measurement protocols, quantifying uncertainty, designing experiments for testing new protocols or methodologies and analyzing data from such experiments.

The American Statistical Association Board of Directors recognizes the urgent need to improve forensic science because of its pivotal role in our judicial system and therefore endorses *Strengthening Forensic Science in the United States: A Path Forward*[1] and the recommendations therein. To better achieve many of the report recommendations, the report urges the establishment of a separate institute for forensic science (Recommendation 1 of the report). The board notes that sound statistical practices are essential for the proposed institute to achieve its mission. Specific examples include:
1. Current and newly developed forensic practices should be assessed using properly designed experiments and data analytic methods.
2. Statistical methods based on established principles and procedures should be used for the analysis of data, including estimated error rates.

3. Novel methods (beyond variants of established methods) developed for the analysis of data should be reviewed in mainstream scientific journals that include statistically qualified experts as reviewers.
4. Modern statistical quality control and quality assurance procedures should be used to assure that measurements, procedures, and testimony are of high quality.
5. Proficiency tests should use accepted statistical designs that are, whenever possible, double blind to avoid testing-response-grading biases.
6. All expert reports should be available to interested parties and sufficient supporting data and information provided to permit independent review (including replication and verification of findings).

Background

The 2009 National Academies' Report *Strengthening Forensic Science in the United States: A Path Forward* provided 13 recommendations including the establishment of an independent body, the National Institute for Forensic Science, to facilitate the development of scientific research and standard practices in forensic science. The report describes the following requirements for the institute:

- It must be an independent federal agency established to address the needs of the forensic science community
- It must have a culture that is strongly rooted in science, with strong ties to the national research and teaching communities, including federal laboratories
- It must have strong ties to state and local forensic entities, as well as to the professional organizations within the forensic science community
- It must not be in any way committed to the existing system, but should be informed by its experiences
- It must not be part of a law-enforcement agency
- It must have the funding, independence, and sufficient prominence to raise the profile of the forensic science disciplines and push effectively for improvements
- It must be led by persons who are skilled and experienced in developing and executing national strategies and plans for standards setting; managing accreditation and testing processes; and developing and implementing rulemaking, oversight, and sanctioning processes

The *Strengthening Forensic Science* noted that no federal agency exists that meets these well-considered and important criteria and therefore recommended the development of a new and separate body. We support the *Strengthening Forensic Science* recommendation for developing the institute. We also second their emphasis on the institute having the independence necessary to produce the needed scientific outcomes. Any perception of outside influence on the institute's products will undermine its credibility. Such independence is a key principle for statistical agencies as made clear in the National Academies' *Principles and Practices for a Federal Statistical Agency* (Fourth Edition, 2009). Indeed, although the proposed Institute is not a statistical agency, much of the content of *Principles and Practices* is relevant to an institute of forensic science.

For the statistical community it is especially critical that the new agency use appropriate statistical practices to raise the level of forensic science in the United States. Here we elaborate on six sound statistical practices listed above as essential for the proposed institute to achieve its mission.

1. The need for well-designed experiments – Current forensic practices have not always been supported by valid assessments that yield defendable and transparent error rates. We view this as a critical need. As Donald Kennedy, then Editor-in-Chief of *Science*, noted in an Editorial,[2] "It's not that fingerprint analysis is unreliable. The problem, rather, is that its reliability is unverified either by statistical models of fingerprint variation or by consistent data on error rates. Nor does the problem with forensic methods end there. The use of hair samples in identification and the analysis of bullet markings exemplify kinds of 'scientific' evidence whose reliability may be exaggerated when presented to a jury." The following examples demonstrate how assessments that have been done are too often flawed:

- The so-called FBI 50K fingerprint comparison study was particularly weak. In an attempt to establish the uniqueness of fingerprints, the FBI contracted with a company to examine 50,000 fingerprints against each other and quantitatively assess the degree of similarity. David H. Kaye exposed this test as unsound.[3] Quoting from the abstract of his article: "Forensic scientists or analysts concerned with 'individualization' often presume that features such as fingerprint minutia are unique to each individual. In the United States, defendants in criminal cases have been demanding proof of such assumptions. In at least two cases, the government of the United States has successfully relied on an unpublished statistical study prepared specifically for litigation to demonstrate the uniqueness of fingerprints. This article suggests that the study is neither designed nor executed in a way that can show whether an individual's fingerprint impressions are unique." Issues with the 50K study include comparing a digitized image of a fingerprint to itself rather than a second fingerprint of the same finger (even though the latter is the relevant comparison), using unrealistic estimates for standard error, and poor modeling of the underlying distribution used too make inferences.
- In a recent review of fingerprint validation, Haber and Haber[4] conclude: "We analyze evidence for the validity of the standards underlying the conclusions made by fingerprint examiners. We conclude that the kinds of experiments that would establish the validity of ACE-V [Analysis-Comparison-Evaluation-Verification – the current standard fingerprint methodology] and the standards on which conclusions are based have not been performed. These experiments require a number of prerequisites, which also have yet to be met, so that the ACE-V method currently is both untested and untestable."

2. Use of well-accepted statistical methods for analysis of data – It is critical that appropriate statistical methods be used to analyze data obtained in support of forensic methods. The validity of these methods should be demonstrated, preferably in peer-reviewed statistical or mainstream scientific journals before being used in litigation. This has not always been the case.

- Compositional analysis of bullet lead (CABL): The FBI practice of comparing crime scene bullets with bullets found in the possession of a potential suspect illustrates the consequences of a poorly designed analysis. The "working hypothesis" justifying CABL is that the chemical concentration of the lead used to make a 'batch' of bullets provide a unique signature, so bullets that come from the same batch of lead should have the same

concentrations of certain trace elements. To show a low error rate for matching bullets the FBI said that it selected one bullet from each of 1837 cases and experimental bullets randomly and matched them to each other. The FBI claimed the bullets were chosen to be representative of the population of manufactured bullets, but also acknowledges that the bullets in this set were "selected". Spiegelman and Kafadar provided indications that the "selection" was neither random nor representative.[5] Consequently, the way that these bullets were chosen led to an indefensibly low error rate (see Ch 3 of Reference 6). Finally, the "statistical test" used to compare bullets was an unjustified modification of Student's t test. The reaction from the scientific community and the media ultimately led the FBI to both abandon the procedure and issue a letter to many convicts that the testimony used against them did not have scientific support.

3. Rigorous review of new data analysis methods – Novel methods for analysis of data in cases do not always have support that would pass scientific muster if subject to peer review. Two illustrations are:

- The FBI had used an ad-hoc data clustering method ("chaining") in CABL that led to clustering together bullets of very different compositions that were claimed to have come from the same batch.[5,6] The 2004 NRC report[6] showed a high rate of false matches; as a result, chaining is no longer used by the FBI.
- DNA profiling is a powerful tool for identification when a single source of DNA is present in an evidence sample (or a resolvable mixture of multiple sources). But no consensus yet exists on the analysis of more complex mixtures of DNA (using the current, 15-year old STR methodology) where "allelic dropout" is present due to poor quality or limited quantity of sample. In 2006, the DNA commission of the International Society of Forensic Genetics issued a report on the situation.[7] Its abstract states: "The purpose of the group was to agree on guidelines to encourage best practice that can be universally applied to assist with mixture interpretation. … Our discussions have highlighted a significant need for continuing education and research into this area. We have attempted to present a consensus from experts but to be practical we do not claim to have conveyed a clear vision in every respect in this difficult subject. For this reason, we propose to allow a period of time for feedback and reflection by the scientific community." Despite the continuing lack of consensus regarding the analysis of complex DNA mixtures, crime laboratory technicians often make strong and unqualified statistical statements in court about the strength of such evidence using ad hoc and unsupported statistical methods.

4. Modern statistical quality control and quality assurance procedures – Forensic laboratories should have in place appropriate quality control procedures to ensure high-quality measurements, standardized procedures, and valid testimony.

- The Clinical Laboratory Improvement Amendments (CLIA), passed by Congress in 1988, established "quality standards for all laboratory testing to ensure the accuracy, reliability and timeliness of patient test results regardless of where the test was performed."[8] Forensic laboratories are explicitly exempt from the CLIA standards (as are some other categories such as research laboratories that "do not report patient-specific results"). The College of American Pathologists do regulate some forensic practices such as Forensic Pathology, but such regulation external to the profession is the exception

rather than the rule in forensic science.

5. Double-blind proficiency testing – Existing forensic associations recognize the need for proficiency testing. Unfortunately existing proficiency tests do not always mirror the level of complexity found in actual practice and are rarely (if ever) double blind. As is well known in medical research the latter can lead to biased evaluations. Examples establishing the need for more challenging tests and the potential value of blind tests are described below:

- Historically, even well established areas of forensic science did not implement appropriate proficiency testing until relatively recently. For example, in 1995 the Collaborative Testing Service (CTS) administered a fingerprint proficiency test. According to David Grieve, then editor *Journal of Forensic Identification*,[9] "the CTS latent print proficiency test was designed, assembled, and reviewed by those representing the IAI [International Association for Identification], thus making it the first such examination authorized by the association." Its results were unanticipated and illustrate how important such tests are: "Of the 156 respondents, only 68, or 44%, had correctly identified the five latent impressions as well as correctly noted the two eliminations."[9] Grieve went on to described the reaction of the forensic community to the results of the CTS test as ranging from "shock to disbelief." [9]

- A 2008 Champion paper[10] by Adina Schwartz includes a quotation that addresses the importance of appropriate testing, "One examiner who took the 2006 CTS cartridge case test commented, 'This test was straightforward and very easy. It took only a few minutes to make correct associations using toolmarks devoid of subclass influence. … I suggest that you consider making the test more of a challenge in order to determine an error rate really reflective of actual casework where borderline cases are not uncommon.'" This is an example of a test that is too easy and not blind in any manner.

- A study published by Dror, Charlton, and Péron in 2006[11] demonstrates how strong contextual biases can be and thus how important blinding is. In the study, they told five experienced fingerprint experts from around the world (including the USA) that they were to look at a reference fingerprint from Brandon Mayfield (the American attorney wrongfully identified as matching a latent fingerprint found in the 2004 Madrid terrorist train bombing) to see if they thought there was a match between his print and the 2004 Madrid latent. Three experts said there was no match and one was "not sure." The participants were in fact shown prints (reference and latent) from their own cases (not the Spanish train suspect) where they had previously declared a match. Four of the five participants changed their opinion, suggesting the existence of contextual bias.

- Double blind proficiency studies have long been used to assess the accuracy of many types of diagnostic and screening procedures. A survey by Gastwirth (1987)[12] provides a number of examples, including the 1984 paper by Morgan[13] both of which demonstrate the long recognized need for double blind testing.

6. Public Availability of Expert Reports: Any statistician who has tried to obtain supporting data for a published paper but met resistance from an uncooperative author, knows the difficulty of verifying or testing the conclusions in that paper. Unlike civil cases, discovery in criminal cases is often much more limited, and similar problems arise. It is also the case that some law enforcement organizations will conduct studies to support a methodology but not make the

supporting data available to scientists interesting in reviewing their findings, as the following example illustrates:

- In the early 1990s, when the FBI RFLP population database (consisting in part of samples from FBI agents) was the primary basis for published theoretical analyses used justifying case work calculations, the FBI refused to make its database available to independent scholars who wished to subject those published analyses to critical scrutiny (unless ordered in some cases by a court and even then protective orders were sought to prevent further dissemination of the database. (Note: this was not an offender or forensic casework database, but a database collected solely for statistical analyses.) One statistician who encountered this problem was Seymour Geisser: "After submitting his article to the American Journal of Human Genetics, Professor Geisser was asked to obtain permission from the FBI to use their original data rather than the data submitted by the FBI to defense attorneys in court cases. Geisser then requested this data from Dr. Budowle, the top FBI DNA scientist. The FBI informed Geisser that (1) the FBI had made commitments earlier to other scientists (Chakraborty, Devlin, Risch, and Weir) and therefore his study must not conflict with their studies, (2) the FBI data may be used only in a joint collaboration with Dr. Budowle, (3) the use of the data was restricted to this one paper, and (4) all authors must agree to the entire contents of a final manuscript prior to submission to a journal."[14] (See also Reference15, and especially footnote 23 therein.)

---

[1] http://www.nap.edu/catalog.php?record_id=12589.

[2] Kennedy, D., "Forensic Science: Oxymoron?," *Science,* (2003), **302**, p. 1625.

[3] Kaye, D.H., "Questioning a Courtroom Proof of the Uniqueness of Fingerprints," *International Statistical Review* (2003), **71**.3, p 521-533.

[4] Haber, L., and Haber, R.N., "Scientific validation of fingerprint evidence under Daubert," *Law, Probability and Risk* (2008), **7**, p. 87.)

[5] Spiegelman, C.H., and Kafadar, K, "Data Integrity and the Scientific Method: the Case of Bullet Lead Data as Forensic Evidence," *Chance* (2006), **19**.2, p. 17-25.

[6] Forensic Analysis: Weighing Bullet Lead Evidence, National Research Council, 2004; http://www.nap.edu/catalog.php?record_id=10924.

[7] Gill, P., et al. "DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures," *Forensic Science International* (2006), **160** p. 90–101.

[8] See, for example, http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/IVDRegulatoryAssistance/ucm124105.htm.

[9] Grieve, D., "Possession of Truth," *Journal of Forensic Identification* (1996), **46,** p. 521.

[10] Schwartz, A, "Challenging Firearms and Toolmark Identification-Part One," *The Champion* (2008), **XXXII**.8, p. 10-19; and "Challenging Firearms and Toolmark Identification-Part Two," *The Champion* (2008)*,* **XXXII**.9, p. 44-52.

[11] Dror, I.E., Charlton, D., Péron, A.E., "Contextual information renders experts vulnerable to making erroneous identifications," *Forensic Science International* (2006), **156**.1, p. 74-78

[12] Gastwirth, J.L., "The Statistical Precision of Medical Screening Procedures: Application to Polygraph and AIDS Antibodies Test Data," *Statistical Science* (1987), **2**, p. 213-238.

[13] Morgan, J.P. "Problems of mass urine screening for misused drugs." *Journal of Psychoactive Drugs*, (1984), **16**, p. 305-317.

[14] Giannelli, P.C., "Book Review: The DNA Story: An Alternative View," [book review of "And the Blood Cried out by Harlan Levy,"] *The Journal of Criminal Law and Criminology,* (1997) **88**.1, p. 380-422.

[15] Thompson, W. C., "Evaluating the Admissibility of New Genetic Identification Tests: Lessons from the DNA War," *The Journal of Criminal Law and Criminology* (1993), **84**.1, p. 22-104.