

ESTHER SUAREZ
Prosecutor of Hudson County
Attorney ID 023161997
595 Newark Avenue
Jersey City, New Jersey 07306
(201) 795-6400
Attorney for Plaintiff-Respondent
BY: Asst. Prosecutor Stephanie Davis Elson
Attorney ID 05182000

STATE OF NEW JERSEY,	:	SUPERIOR COURT OF NEW JERSEY
Plaintiff-Respondent,	:	APPELLATE DIVISION
	:	
	:	DOCKET NO. A-004207-19T2
	:	
v.	:	
	:	
COREY PICKETT,	:	
Defendant-Movant.	:	
	:	

STATE'S BRIEF AND APPENDIX IN SUPPORT OF MOTION FOR
RECONSIDERATION

ESTHER SUAREZ
Prosecutor of Hudson County
Attorney ID 023161997
Attorney for Plaintiff-Respondent
595 Newark Avenue
Jersey City, New Jersey 07306

STEPHANIE DAVIS ELSON
Assistant Prosecutor
Attorney ID 005182000
selson@hcpo.org
On the Brief

TABLE OF CONTENTS

STATEMENT OF PROCEDURAL HISTORY AND FACTS.....1

LEGAL ARGUMENT.....2

AS THE OPINION IN THIS MATTER RELIED
ON INCORRECT INFORMATION AND FAILED TO
APPRECIATE THE SIGNIFICANCE OF PROBATIVE,
COMPETENT EVIDENCE, THE MOTION FOR
RECONSIDERATION SHOULD BE GRANTED.....2

A. PROBLEMS DOCUMENTED IN FST AND STRMIX
WERE DISCOVERED BY TESTING, NOT SOURCE
CODE REVIEW, AND BOTH PROGRAMS ARE VALID
AND CONTINUE TO BE USED.....3

B. THE COURT’S RELIANCE ON PCAST IS
IMPROPER INASMUCH AS PCAST HAS BEEN
DENOUNCED BY THE RELEVANT SCIENTIFIC COMMUNITY.....7

C. THE COURT IMPROPERLY DISREGARDED RELEVANT
RULINGS FROM OTHER JURISDICTIONS IN WHICH
SIMILAR CLAIMS REGARDING TRUEALLELE
SOURCE CODE WERE REJECTED BASED ON
HEARINGS WHICH WERE TESTIMONIAL AND
ADVERSARIAL IN NATURE.....12

CONCLUSION.....17

TABLE TO APPENDICES

New York State DNA Subcommittee letter
dated December 4, 2017..... Ma1-2

ANSI-ASQ National Accreditation Board letter
dated January 16, 2018..... Ma3-4

New York State Inspector General letter
dated February 20, 2018..... Ma5

United States Department of Justice Statement
on the PCAST Report: Forensic Science in
Criminal Courts: Ensuring Scientific
Validity of Feature-Comparison Methods..... Ma6-31

National District Attorneys Association letter dated November 16, 2016.....	Ma32-40
FBI Comments on: President's Council of Advisors on Science and Technology REPORT TO THE PRESIDENT Forensic Science in Federal Criminal Courts: Ensuring Scientific Validity of Pattern Comparison Methods.....	Ma41
Department of Forensic Sciences Science Advisory Board's Statement with regard to the PCAST Report.....	Ma42-47
American Society of Crime Laboratory Directors, Inc. Statement on September 20, 2016 PCAST Report on Forensic Science.....	Ma48-49
An Addendum to the PCAST Report on Forensic Science in Criminal Courts.....	Ma50-58
<u>Finding the Way Forward for Forensic Science in the US - A Commentary on the PCAST Report, I.W. Evett, C.E.H. Berger, J.S. Buckleton, C. Champod, G. Jackson, Forensic Science International 278 (2017) 16-23.....</u>	Ma59-66
<u>DNA Commission of the International Society for Forensic Genetics: Recommendations on the validation of software programs performing biostatistical calculations for forensic genetics applications, M.D. Coble, J. Buckleton, J.M. Butler, T. Egeland, R. Fimmers, P. Gill, L. Gusmão, B. Guttman, M. Krawczak, N. Morling, W. Parson, N. Pinto, P.M. Schneider, S.T. Sherry, S. Willuweit, M. Prinz, Forensic Science International: Genetics 25 (2016) 191-197.....</u>	Ma67-73
<u>Commonwealth of Pennsylvania v. Michael Robinson, CC 201307777, December 7, 2015.....</u>	Ma74-78
<u>State of Georgia v. Monte Baugh and Thaddeus Howell, CASE NO: 2017-CR-618.....</u>	Ma79-93

STATEMENT OF PROCEDURAL HISTORY AND FACTS

The State relies on the Statements of Procedural History and of Facts previously detailed in the prior submissions, with the addition of the following:

On February 3, 2021, this Court issued a per curiam opinion reversing the trial court's denial of the motion to compel source code and related materials and remanding the matter to the trial court. This Court directed the trial court to compel discovery of TrueAllele's source code and related materials pursuant to a protective order, then complete its gate keeping function at the continued Frye hearing.

The State asks this Court to reconsider its opinion requiring the production of source code and related materials.

LEGAL ARGUMENT

AS THE OPINION IN THIS MATTER RELIED ON INCORRECT INFORMATION AND FAILED TO APPRECIATE THE SIGNIFICANCE OF PROBATIVE, COMPETENT EVIDENCE, THE MOTION FOR RECONSIDERATION SHOULD BE GRANTED.

Pursuant to R. 2:11-6, within 10 days of the entry of a judgment or order, a party may apply for reconsideration. A motion for reconsideration is appropriate where the court based its decision "upon a palpably incorrect or irrational basis," the court "did not consider, or failed to appreciate the significance of probative, competent evidence," or the party "wishes to bring new or additional information to the [c]ourt's attention which it could not have provided on the first application[.]'" Cummings v. Bahr, 295 N.J. Super. 374, 384 (App.Div. 1996)(quoting D'Atria v. D'Atria, 242 N.J. Super. 392, 401-02 (Ch.Div. 1990)). The State respectfully submits that this Court's opinion was based on palpably incorrect information and this Court failed to appreciate the significance of probative, competent evidence. As such, the State is seeking reconsideration of this Court's opinion requiring the production of source code.

The hallmark of the admissibility of scientific evidence in a criminal prosecution is general acceptance within a relevant scientific community. Experience shows how this is achieved. Once a new technique is introduced to the scientific community,

it is tested again and again. Its output is studied and scrutinized by experts in the field. It is the subject of peer-reviewed studies. It is challenged, and if need be, refined. Once it has achieved general acceptance in the scientific community, it may be admitted in a court of law.

There is little disagreement among scientists in any field as to how this works: science means testing. This is true in science and in everyday life. A person who owns Microsoft Word validates the program by testing it: keystrokes become words and sentences. The process of validating this comparatively simple set of functions need only go beyond entering keystrokes when someone familiar with the process is able to observe output that is unexpected or does not make sense. In the absence of this, reverse engineering is not required: review of the program's source code is divorced from the process of validating it.

A. PROBLEMS DOCUMENTED IN FST AND STRMIX WERE DISCOVERED BY TESTING, NOT SOURCE CODE REVIEW, AND BOTH PROGRAMS ARE VALID AND CONTINUE TO BE USED.

This Court's opinion failed to appreciate the significance of probative, competent evidence as to the manner in which errors have been discovered in FST and STRmix. In Buckleton JS, Curran J, Taylor D, Bright J-A, What can forensic probabilistic genotyping software developers learn from significant software failures?, *WIREs Forensic Sci.* 2020; e1398, (Ra587-594) the authors noted that they are not aware of any documented example

of discovery of a miscode by way of code review. As further detailed:

The closest we can find is the rediscovery of an undocumented minor routine in the Forensic Statistical Tool software created by the Office of Chief Medical Examiner of the City of New York (Adams et al., 2018; Buckleton & Curran, 2020). This minor and largely innocuous routine was rediscovered by testing and subsequently confirmed in the code. In our experience this is the normal sequence. The testing identifies an unusual behavior in the software, the cause of which is subsequently found in the code once both a suitable test example is available, and a portion of the code comes under scrutiny.

This is the method described by Grgicak et al. (2020) in their development of NOCit and has been the process for all the miscodes we have found in STRmix, the two we found in Lab Retriever, and the one in EuroForMix. The latter two software are open source. (emphasis added)(Ra591)¹

It cannot be ignored that it was testing that led to the discovery of errors, not source code review.

In addressing the arguments of the defense and amici that demand comparison on the reliability of TrueAllele to that of FST and STRmix, this Court pointedly noted that "LAS highlight[ed] the discontinued FST program as a cautionary tale" and referenced the "since-discontinued FST program." The clear implication of these statements is that FST is no longer in use due to discovery of source code errors or because it was discredited. Nether assumption is true. The New York Forensic

¹ The State hereby incorporates by reference the appendix previously submitted to this Court.

"Ma" refers to the motion appendix.

Science Commission and its DNA Subcommittee is the governing body that regulates all of the forensic laboratories in New York State. Before the New York City Office of the Chief Medical Examiner (hereinafter "OCME") could utilize FST, they first had to obtain approval from the Subcommittee and the Commission.

In 2017, the New York Legal Aid Society filed a complaint with the New York Inspector General making allegations against OCME with regard to its use of FST. OCME, as it has always done and continues to do, vigorously refuted the claims and stood by the validity of FST. The DNA Subcommittee, in a detailed letter to the Inspector General, "found no merit in the allegations regarding the OCME's scientific processes." (Ma1-2) The ANSI-ASQ National Accreditation Board also submitted a letter to the Inspector General indicating that it found the allegations of the Legal Aid Society and Federal Defenders of New York to be unfounded. (Ma3-4) As a result of the information received, the Inspector General denied the request to open an investigation. (Ma5)

To this day, OCME stands by the validity of FST and courts continue to rule it is admissible. Indeed, the recent decision by the Second Circuit Court appeals in United States v. Jones, 965 F.3d 149 (2020) clearly shows that FST has not been discredited. Defendant Jones moved before the United States District Court, S.D. New York to exclude any evidence at trial

produced by FST and requested a Daubert hearing on the issue. The District Court gave the testimony of Mr. Adams short shrift. It is clear that, when compared to the substantive testimony of Dr. Craig O'Connor and Dr. Adele Mitchell, the court found Mr. Adams' findings unpersuasive. Ultimately, the District Court found, and the Court of Appeals found no error in the determination, "that FST is sufficiently accepted—both in its admission in scores of New York State cases and in 'the fact that the FST has been approved for use in casework by members of the relevant scientific community and subjected to peer review'... to warrant its admission." Id. at 162.

Any claim that FST was discontinued due to flaws in the source code is equally incorrect. As quoted in a 2017 New York Times article, Director Timothy Kupferschmid explained that FST was well-tested and valid, and "that the lab was adopting newer methods to align with changing FBI standards."² The director analogized the switch to a vehicle upgrade indicating that while a new vehicle may work better, the old vehicle still worked great. Any claim that FST is no longer in use due to discovery of source code errors or because it was discredited is wholly inaccurate. In fact, the FST continues to be used for new comparisons to evidence cases at the OCME that were originally tested prior to 2017.

² [nytimes.com/2017/09/04/nyregion/dna-analysis-evidence-new-york-disputed-techniques.html](https://www.nytimes.com/2017/09/04/nyregion/dna-analysis-evidence-new-york-disputed-techniques.html)

B. THE COURT'S RELIANCE ON PCAST IS IMPROPER INASMUCH AS PCAST HAS BEEN DENOUNCED BY THE RELEVANT SCIENTIFIC COMMUNITY.

Throughout this Court's opinion, there are multiple references to and reliance upon the policies set forth in the 2016 report generated by the President's Council of Advisors on Science and Technology (hereinafter "PCAST"). The report focuses on six "forensic feature-comparison methods" that attempt to determine whether evidentiary samples can be associated with source samples based on the presence of similar patterns, characteristics, features, or impressions. (PCAST report pg. 23) Among the methods addressed are DNA analysis of single-source and simple mixture samples, and DNA analysis of complex mixture samples. The report primarily addresses the reliability of these disciplines for purposes of admissibility under Federal Rule 702, and by implication, its state equivalents. However, the PCAST report has been denounced by the forensic science and associated law enforcement community. The Department of Justice announced that it would not follow PCAST's recommendations. (Ma6)

Among the expressed concerns is the pervasive bias and lack of independence apparent throughout the report. The PCAST report repeatedly demands that studies used to determine and/or establish the scientific validity of feature comparison disciplines must be conducted by those independent of

individuals or entities who may have some stake in the outcome. However, the very composition of the PCAST violates this principle. As noted in the letter from the National District Attorneys Association, "the PCAST membership included several who are far from 'independent' and who have a direct 'stake in the outcome.' A significant example is Eric Lander, Co-Chair of PCAST, and Chair of the working group, who is also a Member of the Board of Directors of the Innocence Project, an organization that has argued for years that the forensic feature comparison disciplines have failed to demonstrate their scientific validity and are, in part, responsible for numerous wrongful convictions." (Ma32)

While PCAST's membership consisted of individuals who are distinguished in their fields, the working group (and PCAST at large) included no forensic scientists. It was made up of six PCAST members (none with forensic laboratory experience), ten judges, two law school professors, and two college professors. (Ma32) Additionally, the report does not include a bibliography/appendix of the literature upon which it relied on in support of its findings and conclusions as required in a properly conducted scientific literature review, instead it made citations to the literature in footnotes. The report only offers, in Appendix B, a list of "Additional Experts Providing Input." While PCAST did solicit literature references from

various forensic organizations, the report does not indicate which of these the PCAST relied upon, considered or even read. (Ma32)

As noted by the FBI in its response, "the report makes broad, unsupported assertions regarding science and forensic science practice." The report states that "the *only way*" to establish "validity as applied" is through proficiency testing, requiring a measurement of how often the examiner gets the correct answer. This is fundamentally at odds with a report of the National Academy of Sciences. (Ma41) The FBI response also expresses concern that the report also creates its own criteria for scientific validity and then proceeds to apply these tests to the listed forensic science disciplines, but fails to provide scientific support that these criteria are well accepted within the scientific community. Notably, "PCAST defines their internally developed criteria as 'scientific criteria' by which forensic feature-comparison methods must be supported by. However, PCAST does not apply its own criteria consistently or transparently. The PCAST criteria define 'black box' studies as the benchmark to demonstrate foundational validity, but provide no clarification on how many studies are needed or why some studies that have been conducted do not meet their criteria. These criteria seem to be subjectively derived and are therefore inconsistent and unreliable." (Ma41) The scientific findings of

the report are questionable given that the report itself was not peer-reviewed prior to its release. Ironically, one of the report's criteria for any study to be acceptable in determining validity was that it be peer-reviewed.

Members of the relevant scientific community published an article, DNA Commission of the International Society for Forensic Genetics: Recommendations on the validation of software programs performing biostatistical calculations for forensic genetics applications, M.D. Coble, J. Buckleton, J.M. Butler, T. Egeland, R. Fimmers, P. Gill, L. Gusmão, B. Guttman, M. Krawczak, N. Morling, W. Parson, N. Pinto, P.M. Schneider, S.T. Sherry, S. Willuweit, M. Prinz, *Forensic Science International: Genetics* 25 (2016) 191-197, to "present recommendations for the minimum requirements to validate bio-statistical software to be used in forensic genetics." (Ma67-73) The International Society for Forensic Genetics convened a DNA Commission to establish validation guidelines for such software. As stated in Recommendation 7, "[t]he DNA Commission does not consider examination of source code to be a useful fact-finding measure in a legal setting. A rigorous validation study (both developmental and internal) should be sufficient to reveal shortcomings or errors in coding." (Ma70)

After input "from stakeholders", PCAST released an addendum to the report in January 2017. (Ma50-58) Despite this addendum,

there was still concern from the forensic community. As noted in Finding the Way Forward for Forensic Science in the US - A Commentary on the PCAST Report, I.W. Evett, C.E.H. Berger, J.S. Buckleton, C. Champod, G. Jackson, *Forensic Science International* 278 (2017) 16-23, (Ma59-66) the report and the addendum fail to recognize advances in the logic of forensic inference that have taken place over the last 50 years or so. This serious omission has led PCAST to a "narrowly-focused and unhelpful view of the future of forensic science." (Ma59-60)

Given the lack of support and confidence in the PCAST report from the forensic science community, any reliance on its mandates and guidelines by this Court is problematic. In addressing peer-reviewed publications as to TrueAllele's reliability, this Court noted that Dr. Perlin had authored six himself and provided guidance on the seventh. This Court then cited to the PCAST report's determination that a software developer's participation in such studies is an impediment to reliable validation. The PCAST determination is rather hypocritical given the PCAST membership included members who are far from independent and who have a direct stake in the outcome.

Additionally, Dr. Perlin's involvement in the peer-reviewed publications should not be considered damning given the very nature of such publications. The Supreme Court of Nebraska addressed this very same argument in State v. Simmer, 304 Neb.

369, 386 (2019), a case involving the use of TrueAllele wherein the court heard testimony from both Dr. Perlin and Mr. Adams. As the Nebraska Supreme Court noted, peer-reviewed publication is valuable because it places research in the public domain, permits evaluation, and permits criticism. In that way, a developer's involvement is of no moment when there is peer-review because research validity is open to public scrutiny and subject to challenge by those independent from the research. The "peer-review" process entails the scientist describing their research methods, results and conclusions in a scientific paper which is submitted to a journal for publication. An editor of the journal has at least two independent and anonymous scientists in the relevant field read the paper, assess its merits and advise on the suitability of the paper for publication. The paper is then either accepted, rejected or sent back to the author for edits and another round of review.

Notably, none of the publications at issue in this matter were authored solely by Dr. Perlin. Other independent members of the scientific community attached their names to and staked their reputations on the reliability of TrueAllele.

C. THE COURT IMPROPERLY DISREGARDED RELEVANT RULINGS FROM OTHER JURISDICTIONS IN WHICH SIMILAR CLAIMS REGARDING TRUEALLELE SOURCE CODE WERE REJECTED BASED ON HEARINGS WHICH WERE TESTIMONIAL AND ADVERSARIAL IN NATURE.

Lastly, this Court failed to fully consider pertinent

information with regard to opinions from other jurisdictions rulings on the production of source code. Claiming the creation of an authority "house of cards", this Court determined that after the decision in Foley, subsequent courts have placed great emphasis on that court's ruling. This court stated that other jurisdictions failed to scrutinize computer science or source code in their determinations. This is factually inaccurate.

A number of courts in other jurisdictions held testimonial hearings wherein the issue of the admissibility of TrueAllele and the production of source code were addressed. Unlike this matter, the courts were provided the opportunity to hear the testimony of witnesses from both sides whose ideas were challenged through vigorous cross-examination. Although many courts cited to the Foley opinion, they did so only after having heard testimony from witnesses on both sides of the source code issue. In State v. Shaw, No. CR-13-575691 (Ohio C.P. Ct. Cuyahoga Cnty. Oct. 10, 2014)(Ra148-173), the court held a Daubert hearing. During the hearing, Dr. Perlin and Jay Caponera testified on behalf of the State while Dr. Ranajit Chakraborty and Dr. Dan Krane testified on behalf of the defense. The issue of source code production was part of the hearing. After hearing testimony from all of the experts, the court ruled that the State had "established that the TrueAllele methodology and the State's witness are reliable without the use of source code."

(Ra167-168)

In Commonwealth of Virginia v. Matthew Brady, (Ra34-40) the court held a testimonial hearing that addressed the source code issue. Dr. Perlin and Dr. Susan Greenspoon testified on behalf of the State while Dr. Kirk Lohmuller testified on behalf of the defense. The defense also called a crime lab director, Brad Jenkins, to testify. After hearing testimony from all of the experts, TrueAllele was determined to be reliable. The court noted that much had been made during the hearing about the inability to thoroughly test the TrueAllele protocol because its source code is unknown. However, the court found that validation studies have been performed with positive results showing that TrueAllele is not junk science. The studies have shown that it is reliable. (Ra36-37)

In Commonwealth of Pennsylvania v. Michael Robinson, (Ma74-78) the court held a testimonial hearing that addressed the source code issue. Dr. Perlin testified on behalf of the State while Dr. Ranajit Chakraborty testified on behalf of the defense. Defendant Robinson alleged that TrueAllele's reliability could not be evaluated without the source code. After a two day hearing, the court determined that the source code was not material to the defendant's ability to pursue a defense. (Ma76) The discovery motion as to the production of source code was denied. (Ma74)

Not only have courts in other jurisdictions held testimonial motions wherein the issue of source code was addressed, but several of those matters involved the experts consulted in this matter. In Washington v. Emmanuel Fair, No. 10-1-09274-5 SEA, January 2017 (Ra203-221), a number of experts testified at a pretrial hearing wherein the production of source code was at issue. Dr. Perlin and Jay Caponera testified on behalf of the State while Mr. Adams, Dr. Dan Krane, Dr. Kirk Lohmuller, and Mr. Brian Ferguson testified on behalf of the defense. After hearing testimony from all of the experts, TrueAllele was found to be reliable without necessitating the production of source code.

In State of Tennessee v. Demontez Watkins, Case No. 2017-C-1811, December 17, 2018 (Ra174-202), the court held a Daubert hearing wherein source code production was at issue. Dr. Perlin testified on behalf of the State while Mr. Adams testified on behalf of the defense. The court stated that one of the main challenges raised by the defense and addressed by Mr. Adams at the hearing was a lack of transparency as to TrueAllele's proprietary source code. (Ra192) The court also noted that the defense raised concern about "bugs" in the TrueAllele system. However, Mr. Adams admitted that, as in this case, he had not conducted any testing of the TrueAllele system. (Ra198) The court ruled that TrueAllele's analysis was reliable and noted

that the criticisms raised by the defense go towards the weight of the evidence, not admissibility. (Ra201)

In State of Georgia v. Monte Baugh and Thaddeus Howell, (Ma79-93), the court held a pretrial hearing wherein source code production was at issue. Dr. Perlin testified on behalf of the State while Mr. Adams testified on behalf of the defense. After hearing testimony from the experts, TrueAllele was determined to be reliable without the necessity of producing the source code.

In State of Nebraska v. Charles Simmer, Case ID CR16-1634, February 2, 2018 (Ra100-112), the court held a Daubert hearing. The issue of source code was part of the hearing. Dr. Perlin testified on behalf of the State while Mr. Adams testified on behalf of the defense. After hearing testimony from the witnesses, the request for source code was denied. The court ruled that the DNA analysis conducted by using TrueAllele probabilistic genotyping software was admissible at trial. Defendant Simmer appealed his conviction. The sole issue on appeal was whether the court erred in admitting the DNA analysis. State of Nebraska v. Charles M. Simmer, 304 Neb. 369 (2019). The Supreme Court of Nebraska found no abuse of discretion and affirmed the decision of the lower court. In its opinion, the Nebraska Supreme Court discussed the testimony of the witnesses at length, detailing the arguments of both sides as to the source code issue. Id. at 372-381. The Court

specifically stated that it was not persuaded that the validation studies were inadequate because the likelihood ratios generated by TrueAllele cannot be confirmed as accurate, a position advocated by the assertions of Mr. Adams. Id. at 387-388.

Given all of the foregoing, the state respectfully submits that the motion for reconsideration should be granted.

CONCLUSION

For the foregoing reasons, the State respectfully urges this Court to reconsider its opinion of February 3, 2021 and affirm the denial of the motion to compel production of source code.

Respectfully submitted,
ESTHER SUAREZ
Prosecutor of Hudson County

By: /s/ Stephanie Davis Elson
STEPHANIE DAVIS ELSON
(005182000)
Assistant Prosecutor
selson@hcpc.org

CERTIFICATION

I hereby certify that this motion is submitted in good faith and not made for purposes of delay.

/s/ Stephanie Davis Elson
STEPHANIE DAVIS ELSON
(005182000)
Assistant Prosecutor

DATED: February 16, 2021



DNA Subcommittee



December 4, 2017

DWIGHT ADAMS, PH.D.
CHAIR
University of Central Oklahoma

MARK BATZER, PH.D.
Louisiana State University

FREDERICK BIEBER, PH.D.
Harvard Medical School

ERIC BUEL, PH.D.

ALLISON EASTMAN, PH.D.
Forensic DNA Consulting, LLC

KENNETH KIDD, PH.D.
Yale University School of Medicine

AMANDA C. SOZER, PH.D.
SNA International

Michael C. Green, Esq.
Chair, New York State Commission on Forensic Science
c/o NYS Division of Criminal Justice Services
80 South Swan Street
Albany, NY 12210

Dear Chairman Green:

In your letter dated September 29, 2017, the Commission requested that the DNA Subcommittee review correspondence dated September 1, 2017 that was sent to the New York State Inspector General (IG) by the Legal Aid Society and Federal Defenders of New York. That correspondence made serious allegations against the New York City Medical Examiner's Office (OCME). The DNA Subcommittee reviewed that correspondence, and the October 18, 2017 response from the OCME. In addition, the Subcommittee collectively reviewed approximately 1,700 pages of supporting documentation provided by the Office of Forensic Services.

The DNA Subcommittee met on November 3, 2017 and commenced discussion of this matter in executive session. Due to the copious amount of materials provided, and the serious nature of the allegations contained within the September 1, 2017 letter, the DNA Subcommittee needed additional time to evaluate and discuss the allegations before it could provide a critical assessment to the Commission. Working individually, and in small groups of two or three members, the DNA Subcommittee reviewed and evaluated the allegations regarding OCME's scientific practices contained within the documentation provided to it. After careful analysis of the claims and associated documentation, the DNA Subcommittee met again on December 1, 2017. During executive session, the DNA Subcommittee finalized its assessment and offers the following comments.

On April 5, 2011, the OCME brought online a version of the Forensic Statistical Tool (FST) software. The use of FST software had been approved by the DNA Subcommittee and the Commission on Forensic Science. On April 6, 2011, the OCME identified an issue that resulted in an unexpected statistical outcome,¹ and the OCME immediately took FST

¹ The unexpected statistical outcome, which would only occur in instances in which a minimum allele frequency was used for rare alleles, was discovered while some enhancements were being made to the developmental version of the software.

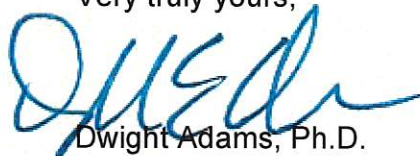
offline to make the necessary changes. The OCME indicated that no casework results were reported using FST Version 1.0. The DNA Subcommittee concluded that there was no "significant malfunction" as asserted in the letter to the IG. Version 2.0 of FST was brought online on June 29, 2011. This enhanced version capped the likelihood ratio (LR) as 1.0 at any locus with a summed allele frequency ≥ 0.97 . Version 2.5 of FST, put in place in July 2013, included minor enhancements that reportedly modified the "landing page" of the software and made minor changes to certain terminology. The OCME performed adequate performance checks prior to the use of Versions 2.0 and 2.5. Accordingly, the DNA Subcommittee does not believe that any re-validation was required as the performance checks were sufficient.

In its October 18, 2017 response to the IG, the OCME correctly stated that some uncertainty will always be present in estimation of various parameters used in forensic DNA analysis, including estimates of allele frequencies in various populations. This is also true for estimates of allelic drop-out rates in low level mixtures. As OCME noted in its October 18, 2017 letter, drop-out rates were estimated empirically, as described in several public presentations in 2010 to the DNA Subcommittee. The DNA Subcommittee again concludes that the OCME used reasonable scientific methods to estimate the role of allelic drop-out in the FST software, and that the resulting FST calculations are not inherently biased as alleged in the September 1, 2017 letter to the IG.

In addition to raising issues with FST, allegations were made regarding the OCME's Low Copy Number (LCN) methodology. Based on the validations performed by the OCME, the DNA Subcommittee believes that the OCME could, using their LCN methodology, potentially identify a major contributor to a DNA mixture regardless of the number of minor contributors. The OCME validated its use of 31 PCR cycles in its LCN methodology. The DNA Subcommittee concludes it was appropriate for the OCME to use 31 PCR cycles in accordance with the OCME's validated casework protocols.

In sum, the DNA Subcommittee finds no merit in the allegations regarding the OCME's scientific processes contained in the September 1, 2017 letter sent to the IG.

Very truly yours,

A handwritten signature in blue ink, appearing to read "D. Adams", is written over the typed name.

Dwight Adams, Ph.D.
Chair

cc: Members, DNA Subcommittee
Members, NYS Commission on Forensic Science
Gina L. Bianchi, Esq., Special Counsel to the Commissioner



January 16, 2018

Timothy Kupferschmid
New York City Office of Chief Medical Examiner
Department of Forensic Biology
421 East 26th Street
New York, NY 10016

Delivered via email: TKupferschmid@ocme.nyc.gov

Dear Director Kupferschmid,

This letter provides the results of our investigation in response to the complaint lodged against the New York City Office of the Chief Medical Examiner Department of Forensic Biology by *The Legal Aid Society and Federal Defenders of New York*.

I have reviewed available information related to the complaint and determined the following:

Allegation 1: The laboratory did not properly validate the revised Forensic Statistical Tool (FST) software.

ANAB Response: The laboratory evaluated the modifications to the FST software to confirm that the changes did not have an adverse effect on the performance of the previously validated method.

The allegation is unfounded.

Allegation 2: The laboratory selectively flattened data during the FST validation “to generate their desired result” was not reported to the Commission on Forensic Science.

ANAB Response: The New York State DNA Subcommittee review the original FST software validation and approved the FST software for use. In response to this complaint, the DNA Subcommittee also reviewed the laboratory’s subsequent evaluation of the modifications to the software. The DNA Subcommittee concluded that “the resulting FST calculations are not inherently biased as alleged.”

The allegation is unfounded.

Allegation 3: A laboratory employee made false statements to members of the Commission on Forensic Science about the validation of Low Copy Number (LCN) testing methodology.

ANAB is Now the Home of



LABORATORY
ACCREDITATION
BUREAU



www.anab.org | Milwaukee, WI | Alexandria, VA | Fort Wayne, IN | Cary, NC


Ma3

ANAB Response: There is insufficient evidence to support the allegation that a false statement was made by the laboratory employee in response to the Commission's inquiry. There appears to be a difference in the interpretation of the question asked and the response provided.

ANAB considers the complaint against the New York City Office of the Chief Medical Examiner Department of Forensic Biology closed.

Should you have further questions regarding this matter, please contact me at (414) 501-5361 or psale@anab.org.

Sincerely,



Pamela L. Sale

Vice President, Forensics

cc: The Legal Aid Society
Federal Defenders of New York
Brian Gestring, NY DCJS, Office of Forensic Services
ANAB office



STATE OF NEW YORK
OFFICE OF THE INSPECTOR GENERAL
OFFICE OF THE WELFARE INSPECTOR GENERAL
OFFICE OF THE WORKERS' COMPENSATION FRAUD INSPECTOR GENERAL

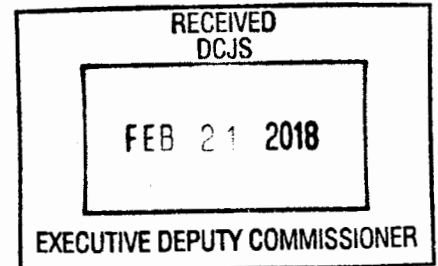
EMPIRE STATE PLAZA
AGENCY BLDG. 2, 16TH FLOOR
ALBANY, NEW YORK 12223
(518) 474-1010

65 COURT STREET, 5TH FLOOR
BUFFALO, NEW YORK 14202
(716) 847-7118

61 BROADWAY, SUITE 2100
NEW YORK, NEW YORK 10006
(212) 635-3150

CATHERINE LEAHY SCOTT
INSPECTOR GENERAL

February 20, 2018



Michael C. Green, Esq.
Executive Deputy Commissioner
New York State Division of Criminal Justice Services
Alfred E. Smith State Office Building
80 South Swan Street
Albany, New York 12210

Re: NYS IG 3401-197-2017

Dear Mr. Green:

Please be advised that my office will not be opening an investigation in the above-referenced matter.

Should you have any questions or concerns, please do not hesitate to contact me.

Respectfully submitted,

Catherine Leahy Scott
Inspector General

Cc. John Czajka, Esq.
General Counsel

United States Department of Justice Statement on the PCAST Report: *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*

In September 2016, the President’s Council of Advisors on Science and Technology (“PCAST”) released its report, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*.¹ The stated purpose of the Report was to determine what additional scientific steps could be taken after publication of the 2009 National Research Council Report² to ensure the validity of forensic evidence used in the legal system.³ PCAST identified what it saw as two important gaps: 1) the need for clarity about scientific standards for the validity and reliability of forensic methods; and 2) the need to evaluate specific methods to determine whether they had been scientifically established as valid and reliable.⁴ The Report “aimed to close these gaps” for a number of what it described as “feature comparison methods.”⁵ These are methods for comparing DNA samples, latent fingerprints, firearm marks, footwear patterns, hair, and bitemarks.⁶

Unfortunately, the PCAST Report contained several fundamentally incorrect claims. Among these are: 1) that traditional forensic pattern comparison disciplines, as currently practiced, are part of the scientific field of metrology; 2) that the validation of pattern comparison methods can *only* be accomplished by strict adherence to a non-severable set of experimental design criteria; and 3) that error rates for forensic pattern comparison methods can *only* be established through “appropriately designed” black box studies.

The purpose of this statement is to address these claims and to explain why each is incorrect. After the PCAST Report was released, the Department of Justice (“Department”) announced that it would not follow PCAST’s recommendations.⁷ The Report was criticized by a number of commentators and organizations outside of the Department for its analysis, conclusions, factual inaccuracies, and other mistakes.⁸ Formally addressing PCAST’s incorrect claims has become

¹ PRESIDENT’S COUNCIL OF ADVISORS ON SCI. & TECH., EXEC. OFFICE OF THE PRESIDENT, *FORENSIC SCI. IN CRIM. COURTS: ENSURING SCI. VALIDITY OF FEATURE COMPARISON METHODS* (2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final [<https://perma.cc/VJB4-5JVQ>] [hereinafter PCAST REPORT].

² NAT’L RES. COUNCIL, NAT’L ACAD’ S., *STRENGTHENING FORENSIC SCI. IN THE UNITED STATES: A PATH FORWARD* 122 (Nat’l Acad. Press 2009).

³ PCAST REPORT, *supra* note 1, at 1.

⁴ *Id.*

⁵ *Id.* In this statement, we use the term “pattern comparison,” rather than PCAST’s chosen term, “feature comparison” to describe the general nature of the methods discussed.

⁶ *Id.* Department of Justice laboratories do not practice what PCAST described as “bitemark analysis.”

⁷ Gary Fields, *White House Advisory Council Is Critical of Forensics Used in Criminal Trials*, WALL ST. J. (Sept. 20, 2016, 4:25 PM), <https://www.wsj.com/articles/whitehouse-advisory-council-releases-report-critical-of-forensics-used-in-criminal-trials-a1474394743> [<https://perma.cc/N9KM-NHJL>].

⁸ *See, i.e.*, I.W. Evett et al., *Finding a Way Forward for Forensic Science in the US—A Commentary on the PCAST Report*, 278 *FORENSIC SCI. INT’L* 16, 22–23 (2017); Letter from Michael A. Ramos, President, Nat’l Dist. Attorneys Ass’n, to President Barack Obama (Nov. 16, 2016), <http://tinyurl.com/hczkt3k>; Ass’n of Firearms and Toolmark Examiners (AFTE) Response to PCAST Report on Forensic Sci. (October 31, 2016), <https://afte.org/uploads/documents/AFTE-PCAST-Response.pdf>; Org. of Sci. Area Committees (OSAC) Firearms and Toolmarks Subcommittee Response to the President’s Council of Advisors on Sci. and Tech. (PCAST) Call for Additional References Regarding its Rep. “Forensic Sci. in Crim. Courts: Ensuring Sci. Validity of Feature-

increasingly important as a number of recent federal and state court opinions have cited the Report as support for limiting the admissibility of firearms/toolmarks evidence in criminal cases.⁹ Accordingly, the Department offers its view on these claims.

I. PCAST’s Claim that “Feature Comparison Methods” are Metrology

Several times throughout its Report, PCAST claimed that forensic “feature comparison methods belong to the scientific discipline of metrology.”¹⁰ (Metrology is the science of measurement and its application.) The accuracy of this assertion is critically important because if forensic pattern comparison methods are *not* metrology, then the fundamental premise PCAST used to justify its “guidance concerning the scientific standards for [the] scientific validity”¹¹ of forensic pattern comparison methods is erroneous. And if that premise is flawed, then key elements of the Report have limited relevance to the methods that PCAST addressed.

PCAST cited a single source in support of its linchpin claim that pattern comparison methods are metrology. That authority, the *International Vocabulary of Metrology*¹² (“VIM”), refutes the claim.

On this point, PCAST states:

Within the broad span of forensic disciplines, we chose to narrow our focus to techniques that we refer to here as forensic “feature-comparison” methods . . . because . . . they all belong to the same broad scientific discipline, *metrology*, which is “the science of measurement and its application,” in this case to measuring and comparing features.¹³

Later in its Report, PCAST claimed:

Comparison Methods (December 14, 2016), https://theiai.org/docs/20161214_FATM_Response_to_PCAST.pdf; Org. of Sci. Area Committees (OSAC) Friction Ridge Subcommittee Response to Call for Additional References Regarding: President’s Council of Advisors on Sci. and Tech. Rep. to the President (December 14, 2016), https://www.nist.gov/system/files/documents/2016/12/16/osac_friction_ridge_subcommittees_response_to_the_presidents_council_of_advisors_on_science_and_technologys_pcast_request_for_additional_references_-_submitted_december_14_2016.pdf; International Ass’n for Identification (IAI) Comments on the PCAST Report from the IAI FW/TT Sci. and Prac. Subcommittee (undated), https://theiai.org/docs/8.IAI_PCAST_Response.pdf; American Soc’y of Crime Laboratory Directors (ASCLD) Statement on September 20, 2016 PCAST Report on Forensic Sci. (September 30, 2016), <https://pceinc.org/wp-content/uploads/2016/10/20160930-Statement-on-PCAST-Report-ASCLD.pdf>.

⁹ *U.S. v. Odell Tony Adams*, 2020 U.S. Dist. LEXIS 45125 (D. Oregon); *U.S. v. Shipp*, 2019 U.S. Dist. LEXIS 205397 (E.D.N.Y.); *U.S. v. Davis*, 2019 U.S. Dist. LEXIS 155037 (W.D. Va.); *U.S. v. Tibbs*, 2019 D.C. Super LEXIS 9 (D.C. 2019); *Williams v. U.S.*, 210 A.3d 734 (D.C. Ct. App. 2019); *U.S. v. Jovon Medley*, PWG 17-242 (D. Md., April 24, 2018); *People v. Azcona*, 2020 Cal. App. LEXIS 1173 (Cal. Ct. App.); *State v. Barquet*, DA No. 2392544-1D (Multnomah County, Oregon November 12, 2020); *People v. A.M.*, 2020 N.Y. Misc. LEXIS 2961 (Sup. Ct. Bronx, N.Y. 2020); *State v. Goodwin-Bey*, Case No. 1531-CR00555-01 (Greene County, Mo., Dec. 16, 2016).

¹⁰ PCAST REPORT, *supra* note 1, at 23, 44 n.93, 143.

¹¹ *Id.* at x, 2, 4, 7, 21, 43.

¹² INT’L VOCABULARY OF METROLOGY – BASIC AND GENERAL CONCEPTS AND ASSOCIATED TERMS (VIM 3rd edition) JCGM 200 (2012), <https://www.ceinorme.it/en/normazione-en/vim-en/vim-content-en.html>.

¹³ PCAST REPORT, *supra* note 1, at 23 (citing the VIM) (emphasis original).

[F]eature-comparison methods belong squarely to the discipline of metrology—the science of measurement and its application.¹⁴

Again, the Report provided only a general citation to the VIM in support.¹⁵

The VIM makes no reference to forensic science or what PCAST described as “feature comparison methods.” Further, the document provides no examples of the types of scientific disciplines, technologies, or applied knowledge that constitute metrology. Most fundamentally, however, the VIM’s terms and definitions affirmatively *refute* PCAST’s claim that “feature comparison methods” are metrology. The VIM defines “measurement” as follows:

Measurement

process of experimentally obtaining *one or more quantity values* that can reasonably be attributed to a quantity

NOTE 1 Measurement *does not apply to nominal properties*.

NOTE 2 Measurement implies *comparison of quantities or counting of entities*

NOTE 3 Measurement presupposes a description of the quantity commensurate with the intended use of a measurement result, a measurement procedure, and a calibrated measuring system operating according to the specified measurement procedure, including the measurement conditions.¹⁶

The term “quantity” is defined in the VIM as follows:

Quantity

property of a phenomenon, body, or substance, where *the property has a magnitude* that can be *expressed as a number* and a reference.¹⁷

Finally, a “nominal property” is defined as:

Nominal Property

property of a phenomenon, body, or substance, where the property has no magnitude

EXAMPLE 1 Sex of a human being

EXAMPLE 2 Colour of a paint sample

EXAMPLE 3 Colour of a spot test in chemistry

EXAMPLE 4 ISO two-letter country code

EXAMPLE 5 Sequence of amino acids in a polypeptide

NOTE 1 *A nominal property has a value, which can be expressed in words, by alphanumerical codes, or by other means.*¹⁸

¹⁴ *Id.* at 44.

¹⁵ *Id.*

¹⁶ VIM, *supra* note 12, at (2.1) (emphasis added).

¹⁷ *Id.* at (1.1) (emphasis added).

¹⁸ *Id.* at (1.30) (emphasis added).

According to the VIM, “measurement” is a process for obtaining a “quantity value.” A “quantity” is the property of a phenomenon, body, or substance that has a magnitude expressed as a number. Measurement, however, does not apply to “nominal” properties—features of a phenomenon, body, or substance that have no magnitude. “Nominal” properties have a value expressed in words, codes, or by other means.

As their reflexive description makes clear, forensic pattern comparison methods *compare* the features/characteristics and overall patterns of a questioned sample to a known source; they do not *measure* them.¹⁹ Any measurements made during the comparison process involve general class characteristics. However, the distinctive features or characteristics that examiners observe in a pattern form the primary basis for a source identification conclusion. These features or characteristics are not “measured.”

During the examination process, forensic examiners initially focus on the general patterns observed in a trace sample. Next, they look for successively more detailed and distinctive features or characteristics. Once those properties are observed and documented, a visual comparison is made between one or more trace samples and/or one or more known sources. The method of comparison is observational, not based on measurement. Correspondence or discordance between class and sub-class features or characteristics of a trace sample and a known source are documented in “nominal” terms—not by numeric values. Finally, examination conclusions are provided in reports and testimony in words (nominal terms), not as measurements (magnitudes).

The conclusion categories described in the Department’s Uniform Language for Testimony and Reports (ULTRs) illustrate this point.²⁰ Pattern comparison ULTR conclusions are reported and expressed in nominal terms such as “source identification,” “source exclusion,” “inclusion,” “exclusion,” and “inconclusive.” Conclusions offered by examiners in the traditional forensic pattern disciplines are not expressed or reported as a measurement or a magnitude. To the contrary, the ULTRs specifically describe the nominal nature of the conclusions offered, along with restrictions on the use of certain terms that might otherwise imply reliance on measurement or statistics. For example, the following language is taken from the Department’s Latent Print Discipline ULTR:

A conclusion provided during testimony or in a report is ultimately an examiner’s decision and is *not based on a statistically-derived or verified measurement or comparison* to all other friction ridge skin impression features. Therefore, an examiner shall not:

- assert that a ‘source identification’ or a ‘source exclusion’ conclusion is based on the ‘uniqueness’ of an item of evidence.

¹⁹ See, e.g., BRADFORD T. ULERY, ET AL., ACCURACY AND RELIABILITY OF FORENSIC LATENT PRINT DECISIONS, 108 PROC. OF THE NAT’L ACAD. OF SCI. 7733, 7733 (May 10, 2011) (“Latent print examiners compare latents to exemplars, using their expertise rather than a quantitative standard to determine if the information content is sufficient to make a decision.”).

²⁰ See U.S. DEP’T OF JUST., UNIFORM LANGUAGE FOR TESTIMONY AND REPORTS (ULTRs), www.justice.gov/forensics.

- use the terms ‘individualize’ or ‘individualization’ when describing a source conclusion.
- assert that two friction ridge skin impressions originated from the same source to the exclusion of all other sources.²¹

A separate limitation in all Department pattern ULTRs directs that “[a]n examiner shall not provide a conclusion that includes a statistic or numerical degree of probability except when based on relevant and appropriate data.”²²

Aside from PCAST’s reference to the VIM, it offers a *single argument*—confined to a footnote—that pattern comparison methods are metrology:

That forensic feature-comparison methods belong to the field of metrology is clear from the fact that NIST—whose mission is to assist the Nation by “advancing measurement science, standards and technology,” and which is the world’s leading metrological laboratory—is the home within the Federal government for research efforts on forensic science. NIST’s programs include internal research, extramural research funding, conferences, and preparation of reference materials and standards . . . Forensic feature-comparison methods involve determining whether two sets of features agree within a given measurement tolerance.²³

This statement is both a non-sequitur and factually inaccurate. PCAST’s claim that NIST is the “world’s leading metrological laboratory” and “is the home within the Federal government for research efforts on forensic science” has no logical nexus to its further claim that forensic pattern comparison methods—as currently practiced—are metrology. Obviously, a laboratory’s status as a leader in the field of metrology and the fact that it conducts forensic research does not somehow transform the subject matter studied into metrology. In addition, PCAST’s claim that “feature-comparison methods involve determining whether two sets of features agree within a given measurement tolerance” is simply not accurate.

As noted, the features or characteristics in a pattern are not “measured” and determined to be “within a given measurement tolerance.” Rather, the combination of class characteristics and distinctive sub-class features within patterns are visually analyzed, compared, and evaluated for correspondence or discordance with a known source. An examiner *does* form an opinion whether “two sets of features agree”;²⁴ however, that opinion is *not* based on whether those features agree “within a given *measurement* tolerance.” Instead, examiners analyze, compare, evaluate, and

²¹ *See Id.* (Emphasis added).

²² *Id.*

²³ PCAST REPORT, *supra* note 1, at 44 n.93.

²⁴ To “agree,” the features observed in the compared samples need not be identical. For example, in latent print examination, due to the pliability of skin, two prints from the same source will not appear to be identical. Surface type, transfer medium, and development method—among other factors—will affect the appearance of the friction ridge features. Because of these factors, examiners must determine whether the observed differences are within the range of variation that may be seen in different recorded impressions from the same source. This also applies to facial comparison—the same face will appear different when the subject’s expression changes.

express their conclusions in nominal terms—not magnitudes. Therefore, contrary to PCAST’s claim, forensic pattern comparison disciplines—as currently practiced—are *not* metrology.

From a legal perspective, however, that fact has no bearing on their admissibility. The Supreme Court made clear in *Daubert v. Merrell Dow Pharms., Inc.* and *Kumho Tire Co. v. Carmichael*²⁵ that judges “cannot administer evidentiary rules under which a gatekeeping obligation depend[s] upon a distinction between ‘scientific’ knowledge and ‘technical’ or ‘other specialized’ knowledge”²⁶” The Court emphasized that trial judges, as part of their gatekeeping function, should not attempt to compartmentalize and shoehorn expert testimony into separate and mutually exclusive bins or boxes of knowledge that is then rigidly analyzed as “scientific,” “technical,” or “specialized.”²⁷ As the Court noted, such efforts would range from difficult to impossible and would inevitably produce no clear lines of distinction capable of case-specific application.

To emphasize this point, the *Kumho Tire* Court cautioned, “We do not believe that Rule 702 creates a schematism that segregates expertise *by type* while mapping certain kinds of questions to certain kinds of experts. Life and the legal cases that it generates are too complex to warrant so definitive a match.”²⁸ Rather than promoting impractical efforts at binning separate categories of knowledge, the Court stressed that the touchstone for the admissibility of expert knowledge under FRE 702—whatever its epistemic underpinning—is relevance and reliability.²⁹

Reliable evidence must be grounded in *knowledge*, whether scientific, technical, or specialized in nature.³⁰ The term knowledge “ ‘applies to any body of known facts or to any body of ideas inferred from such facts or accepted as truths on good grounds.’ ”³¹ The Court hastened to add that no body of knowledge—including scientific knowledge—can or must be “known” to a certainty.³² In addition, the *Kumho Tire* Court stressed that the assessment of reliability may appropriately focus on the personal knowledge, skill, or experience of the expert witness.³³

²⁵ 509 U.S. 579 (1993); 526 U.S. 137 (1999).

²⁶ See *Kumho Tire*, 526 U.S. 137 at 148.

²⁷ See Thomas S. Kuhn, *Reflections on my Critics*, in CRITICISM AND THE GROWTH OF KNOWLEDGE 231, 263 (Imre Lakatos & Alan Musgrave eds., Cambridge Univ. Press, 1965) (“Most of the puzzles of normal science are directly presented by nature, and all involve nature indirectly. Though different solutions have been received as valid at different times, *nature cannot be forced into an arbitrary set of conceptual boxes.*”) (Emphasis added).

²⁸ *Kumho Tire*, at 151 (emphasis added).

²⁹ *Daubert*, 509 U.S. at 589; see also *U.S. v. Mitchell*, 365 F.3d 215, 244 (3rd Cir. 2004) (“That a particular discipline is or is not ‘scientific’ tells a court little about whether conclusions from that discipline are admissible under Rule 702 . . . Reliability remains the polestar.”); *U.S. v. Herrera*, 704 F.3d 480, 486 (7th Cir. 2013) (“[E]xpert evidence is not limited to ‘scientific’ evidence, however such evidence might be defined. It includes any evidence created or validated by expert methods and presented by an expert witness that is shown to be reliable.”).

³⁰ *Id.* at 590 (emphasis added). See also *Restivo v. Hessemann*, 846 F.3d 547, 576 (2d Cir. 2017) (“Rule 702 ‘makes no relevant distinction between ‘scientific’ knowledge and ‘technical’ or ‘other specialized’ knowledge, and ‘makes clear that any such knowledge might become the subject of expert testimony.’ *Kumho Tire Co.*, 526 U.S. at 147.”).

³¹ *Daubert*, *supra* note 25, at 590 (citing WEBSTER’S THIRD NEW INT’L DICTIONARY 1252 (Merriam-Webster Inc.1986).

³² *Id.*

³³ *Kumho Tire*, 526 U.S. at 150 (“[T]he relevant reliability concerns may focus upon personal knowledge or experience.”).

As the *Daubert* and *Kumho Tire* decisions made clear, an expert’s opinion may—but need not—be derived from or verified by measurement or statistics. Experience, either alone or in conjunction with knowledge, skill, training, or education, provides an equally legitimate legal foundation for expert testimony. This fact is reflected in the Comment to FRE 702, which states:

Nothing in this amendment is intended to suggest that experience alone—or experience in conjunction with other knowledge, skill, training, or education—may not provide a sufficient foundation for expert testimony. To the contrary, the text of Rule 702 expressly contemplates that an expert may be qualified on the basis of experience. In certain fields, experience is the predominant, if not sole, basis for a great deal of reliable expert testimony.³⁴

Finally, a forensic expert’s reasoning process is typically inductive,³⁵ (and thereby potentially fallible) and her opinion may be offered in categorical form.³⁶ In the domain of forensic science, a “source identification”³⁷ conclusion is the result of an inductive reasoning process³⁸ that makes

³⁴ FED. RULE OF EVIDENCE 702 advisory committee’s note to 2000 amendment.

³⁵ See NEWTON C.A. DA COSTA & STEVEN FRENCH, *SCI. AND PARTIAL TRUTH: A UNITARY APPROACH TO MODELS AND SCI. REASONING* 130-159 (Oxford Univ. Press 2003) for a formal treatment of pragmatic inductive inference.

³⁶ See FED. RULE OF EVIDENCE 704 (the “Ultimate Issue Rule”); see also *U.S. v. Sherwood*, 98 F.3d 402, 408 (9th Cir. 1996) (fingerprint source identification); *U.S. v. Williams*, 2013 U.S. Dist. LEXIS 120884 (D. Hawaii); *U.S. v. McClusky*, 954 F. Supp. 2d 1224 (D. N. M. 2013); *U.S. v. Davis*, 602 F. Supp.2d 658 (D. Md. 2009) (forensic DNA source attribution); *Revis v. State*, 101 So.3d 247 (Ala. Ct. App. 2011) (firearms/toolmarks source identification).

³⁷ Eoghan Casey & David-Olivier Jaquet-Chiffelle, *Do Identities Matter?* 13 *POLICING: A JOURNAL OF POL’Y & PRAC.* 21, 21 (March 2019) (“Identification is the decision process of establishing with sufficient confidence (not absolute certainty), that some identity-related information describes a specific entity in a given context, at a certain time.”).

³⁸ See COLIN AITKEN ET AL., *COMMUNICATING AND INTERPRETING STAT. EVIDENCE IN THE ADMIN. OF CRIM. JUST., I. FUNDAMENTALS OF PROBABILITY AND STAT. EVIDENCE IN CRIM. PROC., GUIDANCE FOR JUDGES, LAWYERS, FORENSIC SCIENTISTS AND EXPERT WITNESSES*, ROYAL STAT. SOC’Y 14 (November 2010),

<http://www.rss.org.uk/Images/PDF/influencing-change/rss-fundamentals-probability-statistical-evidence.Pdf>

Most inferential reasoning in forensic contexts is inductive. It relies on evidential propositions in the form of empirical generalisations . . . and it gives rise to inferential conclusions that are ampliative, probabilistic and inherently defeasible. This is, roughly, what legal tests referring to “logic and common sense” presuppose to be the lay fact-finder’s characteristic mode of reasoning. Defeasible, ampliative induction typifies the eternal human epistemic predicament, of reasoning under uncertainty to conclusions that are never entirely free from rational doubt.

PAUL ROBERTS & COLIN AITKEN, *COMMUNICATING AND INTERPRETING STAT. EVIDENCE IN THE ADMIN. OF CRIM. JUST., 3. THE LOGIC OF FORENSIC PROOF — INFERENTIAL REASONING IN CRIM. EVIDENCE AND FORENSIC SCI., GUIDANCE FOR JUDGES, LAWYERS, FORENSIC SCIENTISTS AND EXPERT WITNESSES*, ROYAL STAT. SOC’Y 43 (March 2014), <https://www.maths.ed.ac.uk/~cgga/Guide-3-WEB.pdf>.

Events or parameters of interest, in a wide range of academic fields (such as history, theology, law, forensic science), are usually not the result of repetitive or replicable processes. These events are singular, unique, or one of a kind. It is not possible to repeat the events under identical conditions and tabulate the number of occasions on which some past event actually occurred. The use of subjective probabilities allows us to consider probability for events in situations such as these.

COLIN AITKEN & FRANCO TARONI, *STAT. AND THE EVALUATION OF EVIDENCE FOR FORENSIC SCIENTISTS* 22-23 (Wiley 2nd Ed. 2004); See also DA COSTA, *supra* note 35, at 8-20 for a formal treatment of pragmatic probability;

no claim of certainty.³⁹ During an examination, two items are examined for a sufficient combination of corresponding features. If correspondence is observed,⁴⁰ an examiner must determine whether that correspondence provides extremely strong support for the proposition that the items came from the same source and extremely weak or no support for the proposition that the items came from a different source.⁴¹

If an examiner determines that there *is* sufficient correspondence such that she (based on her knowledge, training, experience, and skill) would not expect to find the same combination of features repeated in another source and there is insufficient disagreement to conclude that the combination of features came from a different source, then the examiner inductively infers (from the observed data) that the items originated from the same source.⁴²

Importantly, however, an examiner makes no claim that the observed combination of corresponding features (class and individual characteristics) are “unique”⁴³ in the natural world, or that the examiner can universally “individualize”⁴⁴ the item or person from which the pattern originated. In addition, given the limitations of inductive reasoning, an examiner cannot logically “exclude all other” potential sources of the item.⁴⁵ Accordingly, ULTR documents that authorize

“Probability can be ‘objective’ (a logical measure of chance, where everyone would be expected to agree to the value of the relevant probability) or ‘subjective,’ in the sense that it measures the strength of a person’s belief in a particular proposition.”

³⁹ See N. Malcolm, *Certainty and Empirical Statements*, 51 MIND, 18-46, 41 (1942) (“If any statement is capable of demonstrative proof, then it is not an empirical statement, but an a priori statement.”)

⁴⁰ Christophe Champod & Ian Evett, *A Probabilistic Approach to Fingerprint Evidence*, J. OF FORENSIC IDENTIFICATION, 101-22, 103 (2001) (“The question for the scientist is not ‘are this mark and print identical’ but, ‘given the detail that has been revealed and the comparison that has been made, what inference might be drawn in relation to the propositions that I have set out to consider.’”).

⁴¹ See WILLIAM THOMPSON ET AL., FORENSIC SCI. ASSESSMENTS: A QUALITY AND GAP ANALYSIS (2017), at 66 (2017), https://mcprodnaas.s3.amazonaws.com/s3fs_public/reports/Latent%20Fingerprint%20Report%20FINAL%209_14.pdf?i9xGS_EyMHnIPLG6INIUyZb66L5cLdlb.

Because ridge features have been demonstrated to be highly variable, an examiner may well be justified in asserting that a particular feature set is rare, even though there is no basis for determining exactly how rare. And an examiner may well be justified in saying that a comparison provides “strong evidence” that the prints have a common source, even though there is no basis for determining exactly how strong.

⁴² See David Kaye, *Probability, Individualization, and Uniqueness in Forensic Sci. Evidence: Listening to the Academies*, 75 BROOK. L. REV. 1163, 1176 (2010) (“In appropriate cases . . . it is ethical and scientifically sound for an expert witness to offer an opinion as to the source of the trace evidence. Of course, it would be more precise to present the random-match probability instead of the qualitative statement, but scientists speak of many propositions that are merely highly likely as if they have been proved. They are practicing rather than evading science when they round off in this fashion.”).

⁴³ Champod, *supra* note 40, at 103 (“Every entity is unique; no two entities can be ‘Identical’ to each other because an entity may only be identical to itself. Thus, to say ‘this mark and this print are identical to each other’ invokes a profound misconception: they might be indistinguishable but they cannot be identical.”).

⁴⁴ Kaye, *supra* note 42, at 1166 (“[I]ndividualization—the conclusion that ‘this trace came from this individual or this object’—is not the same as, and need not depend on, the belief in universal uniqueness. Consequently, there are circumstances in which an analyst reasonably can testify to having determined the source of an object, whether or not uniqueness is demonstrable.” The Department uses the term “identification” rather than “individualization.”).

⁴⁵ Champod, *supra* note 40, at 104-105.

a “source identification”⁴⁶ conclusion also prohibit claims that two patterns originated from the same source “to the exclusion of all other sources.” They also preclude assertions of absolute/100% certainty, infallibility, or an error rate of zero.⁴⁷ Federal courts have found these limitations to be reasonable and appropriate constraints on expert testimony.⁴⁸

The empirically-informed inductive process through which a qualified forensic pattern examiner forms and offers an opinion is the product of technical and specialized knowledge under Rule 702,⁴⁹ grounded in science, but ultimately based on an examiner’s training, skill, and experience—not statistical methods or measurements. Moreover, the classification of a “source identification,” “source exclusion,” “inconclusive,” or other conclusion is ultimately an examiner’s *decision*. Thus, PCAST’s claim that the traditional forensic pattern comparison disciplines—as currently practiced—are metrology is plainly incorrect.

II. PCAST’s Claim that Forensic “Feature Comparison” Methods Can Only be Validated Using Multiple “Appropriately Designed” Independent Black Box Studies

In its Report, PCAST claimed that it compiled and reviewed more than 2,000 forensic research papers.⁵⁰ From that number—based on its newly-minted criteria—PCAST determined that only

We cannot consider the entire population of suspects - the best we can do is to take a *sample*... We use our observations on the sample, whether formal or in formal, to draw inferences about the *population*. No matter how large our sample, it is not possible for us to say that we have eliminated every person in the population with certainty. . . . This is the classic scientific problem of *induction* that has been considered in the greatest depth by philosophers.

⁴⁶ See also *Kaye*, *supra* note 42, at 1185 (“Radical skepticism of all possible assertions of uniqueness is not justified. Absolute certainty (in the sense of zero probability of a future contradicting observation) is unattainable in any science. But this fact does not make otherwise well-founded opinions unscientific or inadmissible. Furthermore, whether or not global uniqueness is demonstrable, there are circumstances in which an analyst can testify to scientific knowledge of the likely source of an object or impression.”).

⁴⁷ <https://www.justice.gov/olp/uniform-language-testimony-and-reports>.

⁴⁸ *U.S. v. Hunt*, 2020 U.S. Dist. LEXIS 95471 (W.D. Okla. 2020) (“The Court finds that the limitations . . . prescribed by the Department of Justice are reasonable, and that the government’s experts should abide by those limitations.”); *U.S. v. Harris*, 2020 U.S. Dist. LEXIS 205810 (D.C. 2020) (“This Court believes . . . that the testimony limitations as codified in the DOJ ULTR are reasonable and should govern the testimony at issue here. Accordingly, the Court instructs [the witness] to abide by the expert testimony limitations detailed in the DOJ ULTR.”).

⁴⁹ See e.g., *U.S. v. Harris*, 2020 U.S. Dist. LEXIS 205810 (D.C. 2020) (firearms/toolmarks); *U.S. v. Hunt*, 2020 U.S. Dist. LEXIS 95471 (W.D. Okla. 2020); latent prints); *U.S. v. Johnson*, 2019 U.S. Dist. LEXIS 39590 (S.D.N.Y. 2019) (firearms/toolmarks); *U.S. v. Simmons*, 2018 U.S. Dist. LEXIS 18606 (E.D. Va. 2018) (firearms/toolmarks); *U.S. v. Otero*, 849 F. Supp. 2d 425 (D. N.J. 2012) (firearms/toolmarks); *U.S. v. Mouzone*, 696 F. Supp. 2d 536 (D. Md. 2009) (firearms/toolmarks); *U.S. v. Glynn*, 578 F. Supp. 2d 567 (S.D.N.Y. 2008) (firearms/toolmarks); *U.S. v. Monteiro*, 407 F. Supp. 2d 351 (D. Mass. 2006) (firearms/toolmarks); *U.S. v. Herrera*, 704 F.3d 480 (7th Cir. 2013) (latent prints); *U.S. v. Baines*, 573 F.3d 979 (10th Cir. 2009) (latent prints); *U.S. v. Mosley*, 339 Fed. Appx. 568 (6th Cir. 2009) (latent prints); *U.S. v. Mitchell*, 365 F.3d 215 (3rd Cir. 2004) (latent prints); *U.S. v. Jones*, 2003 U.S. App. LEXIS 3396 (4th Cir. 2003) (latent prints); *U.S. v. Navarro-Fletes*, 49 Fed. Appx. 732 (9th Cir. 2002) (latent prints); *U.S. v. Mercado-Gracia*, 2018 U.S. Dist. LEXIS 192973 (D. N.M. 2018) (latent prints); *U.S. v. Bonds*, 2017 U.S. Dist. LEXIS 166975 (N.D. Ill. 2017) (latent prints); *U.S. v. Kreider*, 2006 U.S. Dist. LEXIS 63442 (W.D.N.Y. 2006) (latent prints); *U.S. v. Plaza*, 188 F. Supp. 2d 549 (E.D. Pa. 2002) (latent prints).

⁵⁰ PCAST REPORT, *supra* note 1, at 2.

three of those 2,000+ studies were “appropriately designed”—two for latent prints and one for firearms/toolmarks.⁵¹ According to PCAST, “the foundational validity of a subjective method can *only* be established through multiple, appropriately designed black box studies.”⁵² To be “appropriately designed,” a study must adhere to a strict set of six, non-severable criteria.⁵³ PCAST claimed that absent conformity with each of these requirements a “feature-comparison” method cannot be considered scientifically valid.⁵⁴

PCAST’s six criteria for an “appropriately designed” black box study are as follows:

Scientific validation studies — intended to assess the validity and reliability of a metrological method for a particular forensic feature comparison application — must satisfy a number of criteria.

(1) The studies must involve a sufficiently large number of examiners and must be based on sufficiently *large* collections of *known* and *representative* samples from *relevant* populations to reflect the range of features or combinations of features that will occur in the application. In particular, the sample collections should be:

(a) representative of the quality of evidentiary samples seen in real cases. (For example, if a method is to be used on distorted, partial, latent fingerprints, one must determine the *random match probability* — that is, the probability that the match occurred by chance—for distorted, partial, latent fingerprints; the random match probability for full scanned fingerprints, or even very high quality latent prints would not be relevant.)

(b) chosen from populations relevant to real cases. For example, for features in biological samples, the false positive rate should be determined for the overall US population and for major ethnic groups, as is done with DNA analysis.

(c) large enough to provide appropriate estimates of the error rates.

(2) The empirical studies should be conducted so that neither the examiner nor those with whom the examiner interacts have any information about the correct answer.

(3) The study design and analysis framework should be specified in advance. In validation studies, it is inappropriate to modify the protocol afterwards based on the results.

(4) The empirical studies should be conducted or overseen by individuals or organizations that have no stake in the outcome of the studies.

(5) Data, software and results from validation studies should be available to allow other scientists to review the conclusions.

(6) To ensure that conclusions are reproducible and robust, there should be multiple studies by separate groups reaching similar conclusions.⁵⁵

⁵¹ *Id.* at 91 (latent prints) (firearms/toolmarks) at 111.

⁵² *Id.* at 9 (emphasis original).

⁵³ *Id.* at 52-53.

⁵⁴ *Id.* at 68.

⁵⁵ *Id.* at 52-53 (emphasis original).

To be clear, none of these criteria standing alone are novel or controversial. However, PCAST failed to cite a single authority that supports its sweeping claim that the collective and *non-severable* application of *all* of these experimental design requirements in multiple black box studies is the *sine qua non* for establishing the scientific validity of forensic “feature comparison” methods. Indeed, the sources that PCAST did cite only serve to undermine its position. In footnote 118 of its Report, PCAST claimed: “The analogous situation in medicine is a clinical trial to test the safety and efficacy of a drug for a particular application.”⁵⁶ This is a reference to *post hoc* changes in the analysis of a study that may compromise its validity. PCAST offered a handful of Food and Drug Administration (FDA) validation guidance documents to support its analogy.⁵⁷ However, the first two cited sources refute PCAST’s claim that method validation studies must adhere to a strict set of mandatory criteria. On that point, the documents offer the following disclaimer in bold and prominent display: “**Contains Non-Binding Recommendations.**” Additionally, the first two cited sources include a call-out box that states, in relevant part:

This guidance represents the Food and Drug Administration's (FDA's) current thinking on this topic. It does not create or confer any rights for or on any person and does not operate to bind FDA or the public. *You can use an alternative approach if the approach satisfies the requirements of the applicable statutes and regulations.* If you want to discuss an *alternative approach*, contact the FDA staff responsible for implementing this guidance.⁵⁸

Similarly, the first page of *Statistical Principles for Clinical Trials*, contains nearly identical language:

This guidance represents the Food and Drug Administration's (FDA's) current thinking on this topic. It does not create or confer any rights for or on any person and does not operate to bind FDA or the public. *An alternative approach may be used if such approach satisfies the requirements of the applicable statutes and regulations.*⁵⁹

Moreover, the *Adaptive Designs* document states, “The use of the word *should* in Agency guidance means that something is suggested or recommended, but not required.”⁶⁰ In addition, the *Design Considerations for Pivotal Clinical Investigations for Medical Devices* document states:

⁵⁶ *Id.* at 52.

⁵⁷ U.S. DEP’T OF HEALTH AND HUM. SERV. FOOD AND DRUG ADMIN. CTR. FOR DEVICES AND RADIOLOGICAL HEALTH, and THE CTR. FOR BIOLOGIC EVALUATION AND RES., DESIGN CONSIDERATIONS FOR PIVOTAL CLINICAL INVESTIGATIONS FOR MED. DEVICES: GUIDANCE FOR INDUSTRY, CLINICAL INVESTIGATORS, INST. REV. BOARDS AND FOOD AND DRUG ADMIN. STAFF (November 7, 2013); U.S. DEP’T OF HEALTH AND HUM. SERV., CTR. FOR DEVICES AND RADIOLOGICAL HEALTH, and CTR. FOR BIOLOGICS EVALUATION AND RES.: ADAPTIVE DESIGNS FOR MED. DEVICE CLINICAL STUD. (July 27, 2016); and U.S. DEP’T OF HEALTH AND HUM. SERV., CTR. FOR DRUG EVALUATION AND RES., and CTR. FOR BIOLOGICS EVALUATION AND RES.: GUIDANCE FOR INDUSTRY E9 STAT. PRINCIPLES FOR CLINICAL TRIALS (September 1998).

⁵⁸ DESIGN CONSIDERATIONS, *supra* note 57, at 4; ADAPTIVE DESIGNS, *supra*, note 57, at 2 (emphasis added).

⁵⁹ GUIDANCE FOR INDUSTRY, *supra* note 57, at 1 (emphasis added).

⁶⁰ ADAPTIVE DESIGNS, *supra* note 57, at 2 (emphasis added).

Although the Agency has articulated policies related to design of studies intended to support specific device types, and a general policy of tailoring the evidentiary burden to the regulatory requirement, *the Agency has not attempted to describe the different clinical study designs that may be appropriate to support a device pre-market submission, or to define how a sponsor should decide which pivotal clinical study design should be used to support a submission for a particular device.* This guidance document describes different study design principles relevant to the development of medical device clinical studies that can be used to fulfill pre-market clinical data requirements. *This guidance is not intended to provide a comprehensive tutorial on the best clinical and statistical practices for investigational medical device studies.*⁶¹

Finally, PCAST's purportedly mandatory criteria for pattern comparison method validation is inconsonant with the regulatory definition of "Valid Scientific Evidence" in the FDA's *Design Considerations* document:

Valid scientific evidence is evidence from well-controlled investigations, partially controlled studies, studies and objective trials without matched controls, well-documented case histories conducted by qualified experts, and reports of significant human experience with a marketed device, from which it can fairly and responsibly be concluded by qualified experts that there is reasonable assurance of the safety and effectiveness of a device under its conditions of use. *The evidence required may vary according to the characteristics of the device, its conditions of use, the existence and adequacy of warnings and other restrictions, and the extent of experience with its use.* Isolated case reports, random experience, reports lacking sufficient details to permit scientific evaluation, and unsubstantiated opinions are not regarded as valid scientific evidence to show safety or effectiveness. Such information may be considered, however, in identifying a device the safety and effectiveness of which is questionable.⁶²

The FDA's validation guidance clearly states that no single experimental design is either essential or required. To the contrary, the documents take pains to stress that it may be appropriate to utilize various study designs when validating medical devices or clinical drugs. The FDA also emphasized the non-binding nature of its guidance, which contains no prescriptive requirements or mandatory criteria. Finally, the applicable federal regulation instructs that "valid scientific evidence" may be generated by a variety of study designs and that the evidence required for validation may vary by the nature of the device, the conditions of use, and experience.

a. Forensic Laboratory Standards

Laboratory accreditation standards in the field of forensic science address the issue of method validation. The international standard applicable to all testing and calibration laboratories—

⁶¹ DESIGN CONSIDERATIONS, *supra* note 57, at 4 (emphasis added).

⁶² *Id.* at 9 (quoting 21 CFR 860.7(c)(1)) (emphasis added).

including crime labs—is ISO 17025.⁶³ This document guides the core activities and management operations of laboratories engaged in a diverse range of scientific inquiry. This includes clinical testing and diagnostics, research and development, and forensic science, among many other fields. Identical requirements apply to all testing and calibration laboratories, regardless of whether they analyze clinical samples, groundwater, or forensic evidence.

ISO generally defines validation as “confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled.”⁶⁴ A method has been validated per ISO/IEC 17025 when “the specified requirements are adequate for an intended use.”⁶⁵ Section 7.2.2 of ISO 17025 is the applicable requirement for validating test methods. It provides that “validation shall be as extensive as is necessary to meet the needs of the given application or field of application.”⁶⁶

In contrast to PCAST’s prescriptive stance, ISO does not dictate *how* labs must validate their methods, *which* criteria must be employed, or *what* experimental design must be followed. Instead, ISO simply requires that “[t]he performance characteristics of validated methods, as assessed for the intended use, shall be relevant to the customer’s needs and consistent with specified requirements.” The selection of those requirements, the chosen experimental design, and the extent of the validation performed, is the responsibility of each laboratory. The pragmatic and flexible nature of method validation is also emphasized by other international scientific organizations.⁶⁷

⁶³ ISO/IEC 17025:2017, ISO, <https://www.iso.org/obp/ui/#iso:std:iso-iec:17025:ed-3:v1:en> [<https://perma.cc/C4V5-2RU4>].

⁶⁴ See *id.* § 3.9; ISO/IEC 9000:2015 § 3.8.13, ISO, <https://www.iso.org/obp/ui/#iso:std:iso:9000:ed-4:v1:en> [<https://perma.cc/7E5R-MMDH>].

⁶⁵ ISO/IEC 17025:2017, *supra* note 63, § 3.9.

⁶⁶ *Id.* § 7.2.2.1.

⁶⁷ For example, in the United Kingdom, the Forensic Science Regulator publishes the FORENSIC CODE OF PRACTICE AND CONDUCT (“Code”), which states:

The functional and performance requirements for interpretive methods are less prescriptive than for measurement-based methods. They concentrate on the competence requirements for the staff involved and how the staff shall demonstrate that they can provide consistent, reproducible, valid and reliable results that are compatible with the results of other competent staff.

FORENSIC SCI. REGULATOR, CODES OF PRAC. AND CONDUCT FOR FORENSIC SCI. PROVIDERS AND PRAC. IN THE CRIM. JUST. SYS. § 20.9.1 (2016).

Like ISO, the Code sets forth a non-prescriptive, non-exclusive combination of measures that may be used to validate interpretive methods. These include blind confirmation by a second examiner, inter-laboratory comparisons and proficiency tests, and the in-house use of competency tests. The Code also states that an interpretive method “shall require only the relevant subset of . . . parameters and characteristics for measurement-based methods.” § 20.9.1 & .2. Finally, an equally flexible view of validating interpretive methods is shared by Australia’s National Association of Testing Authorities (NATA). NATA recognizes that the validation of interpretive methods “is more challenging and less proscriptive than it is for analytical methods.” However, validity can be established “if the analyst or examiner repeatedly obtains correct results for positive and negative known tests.” In addition, NATA correctly concedes that certain validation parameters “are not relevant in subjective tests.” NAT’L ASS’N OF TESTING AUTHS., TECHNICAL NOTE 17: GUIDELINES FOR THE VALIDATION AND VERIFICATION OF QUANTITATIVE AND QUALITATIVE TEST METHODS § 5 (2013) at § 5-5.1.

b. Authorities in Experimental Design

The need for pragmatic flexibility in validating test methods is also stressed by authoritative sources in the field of experimental design. These experts advise that there are no rigid rules and that the most suitable approach depends on a variety of factors and circumstances. For example, Westgard, in his classic text, *Basic Method Validation*, states, “Method validation should be a standard laboratory process, but the process *need not be exactly the same for every laboratory or for every method validated* by a laboratory.”⁶⁸ He also emphasizes the individual nature of validation: “Develop a validation plan on the basis of the characteristics of the test and method that will be critical for its successful application *in your laboratory*.”⁶⁹ Finally, Westgard notes that “[e]ach laboratory situation may be different, therefore different adaptations are possible in different laboratories. The approach we advocate is to maintain the principles of the method validation process, while making the experimental work as efficient and practical as possible.”⁷⁰

Creswell, another leading expert on research design, emphasizes the contingent nature of various approaches and decisions:

In planning a research project, researchers need to identify whether they will employ a qualitative, quantitative, or mixed methods approach. This approach is based on bringing together a worldview or assumptions about research, a specific design, and research methods. Decisions about choice of an approach are further influenced by the research problem or issue being studied, the personal experiences of the researcher, and the audience for whom the researcher writes.⁷¹

A group of legal academics has also observed, “There is no one best way to study a phenomenon of interest. Each methodological choice involves trade-offs.”⁷² Trade-offs, in turn, require flexibility, which is necessitated by the pull of competing interests, existing resources, and countless operational considerations.⁷³

Perhaps most notably, a leading treatise in the field of metrology states, “The situation regarding the frequency of validation is comparable for the situation for the appropriate amount of validation; there are *no firm and generally applicable rules, and only recommendations can be offered* that help the person responsible for validation with a competent assessment of the particular situation.”⁷⁴

⁶⁸ WESTGARD, *BASIC METHOD VALIDATION* 198 (Westgard QC Inc., 3rd ed. 2008) (emphasis added).

⁶⁹ *Id.* at 203 (emphasis added).

⁷⁰ *Id.* at 205.

⁷¹ JOHN W. CRESWELL, *RESEARCH DESIGN: QUALITATIVE, QUANTITATIVE, AND MIXED METHOD APPROACHES* 21 (4th ed. 2014).

⁷² 1 DAVID L. FAIGMAN ET AL., *MODERN SCI. EVIDENCE: THE LAW AND SCI. OF EXPERT TESTIMONY, STAT. & RES. METHODS* § 1:22 (2010).

⁷³ GEOFFREY MARCZYK ET AL., *ESSENTIALS OF RES. DESIGN AND METHODOLOGY* 137 (2005) (“The most obvious limitation of studies that employ a randomized experimental design is their logistical difficulty. Randomly assigning participants in certain settings (e.g., criminal justice, education) may often be unrealistic, either for logistical reasons or simply because it may be considered inappropriate in a particular setting. Although efforts have been made to extend randomized designs to more real-world settings, it is often not feasible. In such cases, the researcher often turns to quasi-experimental designs.”).

⁷⁴ CZICHOS ET AL., *SPRINGER HANDBOOK OF METROLOGY AND TESTING* 86 (Springer 2011) (emphasis added).

On this point, the American Association for the Advancement of Science (AAAS) recently published a study on latent fingerprint examination.⁷⁵ The authors disagreed with PCAST's premise that only those research projects "intentionally and appropriately designed" should be considered when assessing evidential support for method validation.⁷⁶ Instead, the AAAS discussed the concept of "convergent validity," an approach that draws conclusions about method validity from the body of relevant literature as a whole. This approach acknowledges that various study designs have different strengths and weaknesses.⁷⁷ It also recognizes that some studies can reinforce others and collectively support conclusions not otherwise warranted.⁷⁸

In sum, the sources cited by PCAST, the relevant international standard, and noted authorities in the fields of experimental design all refute its claim that only multiple black studies that strictly adhere to its six non-severable criteria may be used to validate forensic pattern comparison methods. Instead, they emphasize the absence of strict rules, the need for pragmatic flexibility, and an adaptive, context-based approach for testing a method's fitness for purpose.

III. PCAST's Claim that Error Rates for Forensic Pattern Comparison Methods Must be Established Using *Only* Black Box Studies

The Department fully agrees with PCAST's statement that "all laboratory test and feature comparison analyses have non-zero error rates."⁷⁹ That said, the more difficult questions are: *Can* such rates be accurately determined? *How* can that be accomplished? And to *whom*, *where*, and to *what* activities may such rates be validly applied?

PCAST claimed that error rates for subjective forensic pattern comparison methods must be *solely* determined through black box studies.⁸⁰ It also asserted that forensic examiners who took no part in those studies should testify that those study-derived rates apply to their work in the case at hand.⁸¹ There are significant practical and scientific problems with these specious claims. Most fundamentally, no single error rate is generally applicable to all laboratories, all examiners, and all cases in which a particular method is used. Error rates derived from any given study are the output of numerous different inputs. Rates will vary depending on a multitude of factors immanent in a study's design, participants, rules, execution, and the model chosen for data generation and statistical summation.

⁷⁵ See THOMPSON ET AL., *supra* note 41.

⁷⁶ *Id.* at 44. ("[W]e consider all studies that examine the accuracy of latent print examiners, rather than focusing just on those that are 'intentionally and appropriately designed' for a particular purpose. Our goal is to draw conclusions from the literature as a whole, recognizing (consistent with the concept of convergent validity) that studies will have different strengths and limitations, and that the literature as a whole will have strengths and limitations.").

⁷⁷ *Id.* ("Our goal is to draw conclusions from the literature as a whole, recognizing (consistent with the concept of convergent validity) that studies will have different strengths and limitations, and that the literature as a whole will have strengths and limitations.").

⁷⁸ *Id.* at 94. ("[We] determined that the evaluation of individual publications, one at a time, was not an effective approach to reviewing this literature. This atomistic approach ignores the concept of convergent validity- i.e., the possibility that various publications, each with distinct limitations when considered by itself, can reinforce each other and collectively support conclusions that would not be warranted on the basis of a single article.").

⁷⁹ PCAST REPORT, *supra* note 1, at 3, 29.

⁸⁰ *Id.* at 46, 51, 111, 112, 116, 143, 147, 150.

⁸¹ *Id.* at 56, 66, 112, 147, 150. *But see* ULERY ET AL., *supra* note 19, at 7734 ("Combining [experimental study] results among multiple agencies with heterogeneous procedures and types of casework would be problematic.").

In the experimental context, inputs are the assumptions and choices that researchers make and the actions they take to answer the questions of interest. These include: the study’s internal design—its structure and scope; its experimental conditions; its participants—including their number, experience, and skill; how they are selected; their risk tolerance or aversion; whether they know they are being tested; the requirements of the laboratory quality systems in which they work; how closely test conditions mimic those requirements/systems; instructions researchers provide to participants; the number and type of comparisons conducted; the nature of the test samples used; how representative those samples are to evidence encountered during actual casework; how different answers are classified; and the statistical model(s) selected and employed to describe the results—to name a few.

Similar points were recently made by a well-known academic psychologist and commentator. Although noting the desirability of valid error rates, he also conceded that practical and scientific problems with generating such rates abound:

Providing “an error rate” for a forensic domain may be misleading because it is a function of numerous parameters and depends on a variety of factors. An error rate varies by difficulty of the decision. . . . Error rates are going to be higher for difficult cases, but lower for easier cases . . . An error rate will also vary across individuals. Some experts have higher error rates, and others, lower error rates. This can be a function of training background . . . as well as cognitive aptitude, motivation, ideology, experience, etc. Therefore, error rates may give insights into forensic domains in general, but may say very little about a specific examiner’s decision in a particular case. Hence, an average error rate for an average expert, in an average case, may not be informative (may even be misleading) for evaluating a specific expert examiner, doing a specific case.⁸²

The American Association for the Advancement of Science’s (AAAS) recent report, *Forensic Science Assessments: A Quality and Gap Analysis – Latent Fingerprint Examination*,⁸³ also cautioned against generalizing study-derived error rates to unrelated case scenarios. The report stated, “[I]t is unreasonable to think that the ‘error rate’ of latent fingerprint examination can meaningfully be reduced to a single number or even a single set of numbers.”⁸⁴ The AAAS found that “[t]he probability of error in a particular case may vary considerably depending on the difficulty of the comparison. Factors such as the quality of the prints, the amount of detail present, and whether the known print was selected based on its similarity to the latent will all be important.”⁸⁵

The AAAS also noted that black box studies “can in principle determine the relative strength of different analysts and the relative difficulty of different comparisons, however the relationship of

⁸² Itiel Dror, *The Error in “Error Rates”: Why Error Rates Are So Needed Yet So Elusive*, 65 JOURNAL OF FORENSIC SCIENCES 5, 15-16 (2020).

⁸³ THOMPSON ET AL., *supra* note 41, at 46. With relevance to the points raised in Section I, the AAAS report stated, “Because the characteristics of fingerprints are unlikely to be statistically independent, it will be difficult to determine the frequency of any particular combinations of features. While research of this type is important, it is unlikely to yield quick answers.” At 22.

⁸⁴ *Id.* at 45.

⁸⁵ *Id.* at 58.

such findings to the error rate in a specific case is problematic.”⁸⁶ One concern was that study participants know they are being tested, which could affect their performance.⁸⁷ Another was that decision thresholds used by participants in controlled studies may differ from those used during actual casework. In sum, the report concluded that “the existing studies generally do not fully replicate the conditions that examiners face when performing casework.”⁸⁸ Consequently, “the error rates observed in these studies *do not necessarily reflect the rate of error in actual practice.*”⁸⁹

PCAST also claimed that forensic examiners should testify that error rates from black box studies apply to their individual casework. This raises additional concerns about the relevance of rates generated by a discrete reference class of study participants to *all* forensic examiners who practice that method. This, in turn, raises larger questions about the overall external validity of black box studies. PCAST failed to squarely address these fundamental concerns about scientific relevance and general applicability.

As alluded to in the AAAS Report, the reference class of examiner-participants in a given black box study cannot be used as a valid proxy for the class of *all* such examiners.⁹⁰ Allen and Pardo have separately noted, “The reference class problem demonstrates that objective probabilities based on a particular class of which an item . . . [in our context, an examiner] is a member cannot typically (and maybe never) capture the probative value of that evidence for establishing facts relating to a specific event.”⁹¹ They continue, adding, “There is only one empirically objective reference class—the event itself. Among the various other reference classes, there is no other unique class that will capture the probative value of the evidence.”⁹² In short, error rates will vary based on the chosen reference class of examiners. As such, rates generated by examiners who participate in a given study cannot be generalized to and adopted by different examiners as *their* local error rate for unrelated casework scenarios.⁹³

⁸⁶ *Id.*

⁸⁷ *Id.* See also ULERY ET AL., *supra* note 19, at 7734 (“Ideally, a study would be conducted in which participants were not aware that they were being tested.”).

⁸⁸ THOMPSON ET AL., *supra* note 41, at 46.

⁸⁹ *Id.* (Citing Haber and Haber, 2014; Koehler, 2017; Thompson et al., 2014) (emphasis added); see also Ulery et al., *supra* note 19, at 7734 (“There is substantial variability in the attributes of latent prints, in the capabilities of latent print examiners, in the types of casework received by agencies, and the procedures used among agencies. *Average measures of performance across this heterogeneous population are of limited value*—but do provide insight necessary to understand the problem a scope future work.”) (Emphasis added); BALDWIN ET AL., A STUDY OF FALSE POSITIVE AND FALSE NEGATIVE ERROR RATES IN CARTRIDGE CASE COMPARISON 18 (2014), <https://www.ncjrs.gov/pdffiles1/nij/249874.pdf>: (“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.”).

⁹⁰ See generally, Allen, Ronald; Pardo, Michael, *The Problematic Value of Mathematical Models of Evidence*, 36 J. OF LEGAL STUD. 107-140, 122 (January 2007) (“[G]enerally if not always there is a practically unbounded set of reference classes with probabilities within those reference classes ranging from zero to one, and nothing privileges any particular class.”).

⁹¹ *Id.* at 114.

⁹² *Id.* at 123.

⁹³ See also ULERY ET AL., *supra* note 19, at 7738 (“The rates measured in this [latent print black box] study provide useful reference estimates that can inform decision making and guide future research: the results are *not representative of all situations, and do not account for operational context and safeguards.*”) (Emphasis added).

A concern closely related to the reference class problem is the external or ecological validity of error rates generated through black box studies. External validity refers to whether an experiment accurately and adequately represents the subject matter, activities, and types of individuals studied. “If a study is externally valid, its findings can be generalized to other populations (of people, objects, organizations, times, places, etc.).”⁹⁴ Conversely, if a study lacks external validity, its findings cannot be generalized and applied to different people, places, and circumstances.

It is beyond dispute that black box studies do not reflect the numerous factors at play in actual casework. The reasons are many: They are performed outside of a laboratory’s quality assurance system; there is no verification and review by a second examiner;⁹⁵ study directives may deviate from participants’ work-related procedures and protocols;⁹⁶ sample quantity, quality, and analytical difficulty may differ from that typically encountered in actual casework; classification decisions may be dictated by study directives; and participants know they are being tested. In addition, black box studies may include a wide range of participants with differing levels of knowledge, skill, experience, training, and risk tolerance/aversion.⁹⁷ On this point, it is important to note that in pattern comparison black box studies performed to date, false positive errors have clustered among a small number of participants.⁹⁸ Moreover, in one latent print black box study

⁹⁴ FAIGMAN ET AL., *supra* note 72, at § 5:39.

⁹⁵ See BALDWIN ET AL., *supra* note 89, at 18:

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 result is that we can tell that the overall fraction is not necessarily representative of a rate for each examiner in the pool. Instead, examination of the data shows that 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. What this result does suggest is that quality assurance is extremely important in firearms analysis and that an effective QA system must include the means to identify and correct issues with sufficient monitoring, proficiency testing, and checking in order to find false-positive errors that may be occurring at or below the rates observed in this study.*

(Emphasis added).

See also ULERY ET AL., *supra* note 19, at 7735 (“In no case did two examiners make the same false positive error [out of six total in the study]. Five errors occurred on image pairs where a large majority of examiners correctly excluded; one occurred on a pair where the majority of examiners made inconclusive decisions. *This suggests that these erroneous individualizations would have been detected if blind verification were routinely performed.*”) (Emphasis added).

⁹⁶ See ULERY ET AL., *supra* note 19, at 7734 (“Combining results among multiple agencies with heterogeneous procedures and types of casework would be problematic.”).

⁹⁷ *Id.* at 7737 (“Examiner skill varied substantially.”); BALDWIN, ET AL., *supra* note 89, at 18 (“[E]xamination of the data shows that the rate is a highly heterogeneous mixture of a few examiners with higher rates and most examiners with much lower error rates.”).

⁹⁸ BALDWIN ET AL., *supra* note 89, at 3, 18 (“[E]xamination of the data shows that the [false positive] rate is a highly heterogeneous mixture of a few examiners with higher rates and more examiners with much lower rates”); ULERY ET AL., *supra* note 19, at 7735, 7738 (the 6 false positive errors were committed by 5 examiners from a total of 169 study participants). In addition, (“Most of the false positive errors involved latents on the most complex combination

discussed by PCAST, when a second examiner performed the verification of the first examiner's results under non-biased conditions, all false positive results reported by the first examiners were detected.⁹⁹

A different study examined the repeatability and reproducibility of decisions made by latent print examiners.¹⁰⁰ Participants examined approximately one-hundred image pairs of latent prints.¹⁰¹ Six false positive errors were committed by five (out of one-hundred sixty-nine) examiners in the initial test.¹⁰² Seventy-two examiners participated in the retest.¹⁰³ None of the six false positive errors were reproduced by a different examiner in the initial test and none of the four false positive errors was repeated by the same examiner in the retest.¹⁰⁴ The study concluded that “blind verification [by a second examiner] should be highly effective at detecting such errors.”¹⁰⁵

PCAST's claim that forensic pattern comparison error rates can *only* be derived from black box studies and that examiners must testify that those rates apply to the case at hand is scientifically erroneous. Black box error rates cannot travel from place to place and equally apply from case to case. In sum, these rates cannot be generalized to different laboratories, examiners, and casework situations.¹⁰⁶

a. Alternative Experimental Designs

The PCAST Report also criticized forensic studies that employed what it described as a “closed-set” experimental design. In closed-set studies, a small number of samples generate many comparisons in which the source of the questioned items is always present.¹⁰⁷ PCAST noted that this creates internal dependencies among comparisons. It expressed concern that this type of experimental design may underestimate false-positive error rates. PCAST focused its critique on

of processing and substrate included in the study.”); Ulery et al., *Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners*. PLoS ONE 7(3): e32800. Doi: 10.1371/journal.pone.0032800 (2012), available at: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0032800>.

⁹⁹ PACHECO ET AL., MIAMI-DADE RES. STUDY FOR THE RELIABILITY OF THE ACE-V PROCESS: ACCURACY & PRECISION IN LATENT FINGERPRINT EXAMINATIONS 2, 7, 66 (2014), <https://www.ncjrs.gov/pdffiles1/nij/grants/248534.pdf>. (“Of the 42 erroneous identifications reported in both Phase 1 and Phase 2, seventeen of these errors occurred during Phase 2 ACE trials. The seventeen erroneous identifications were sent to fourteen of the 63 participants for verification in Phase 3, and fifteen responses for the seventeen erroneous identifications were returned. None of the fourteen participants agreed with the initial erroneous identification; twelve participants disagreed a total of thirteen times and two participants reported an inconclusive decision.”).

¹⁰⁰ Ulery, et al., *Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners*. PLoS ONE 7(3): e32800. Doi: 10.1371/journal.pone.0032800 (2012), <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0032800>.

¹⁰¹ *Id.* at 3.

¹⁰² *Id.* at 3, 6.

¹⁰³ *Id.* at 3.

¹⁰⁴ *Id.* at 6, 9.

¹⁰⁵ *Id.* at 9.

¹⁰⁶ See ULERY ET AL., *supra* note 19, at 7734 (“There is substantial variability in the attributes of latent prints, in the capabilities of latent print examiners, in the types of casework received by agencies, and the procedures used among agencies. Average measures of performance across this heterogeneous population are of limited value—but do provide insight necessary to understand the problem and scope future work.”).

¹⁰⁷ PCAST REPORT, *supra* note 1, at 86.

several studies conducted in the firearms/toolmarks discipline. While PCAST criticized the closed-set design of these studies, it failed to consider their purpose, substance, and utility.

The studies PCAST reviewed used consecutively manufactured firearms to produce the test samples provided to participants. Consecutively manufactured firearms are known to bear subclass characteristics. These are machined marks that carry over from one manufactured part of a firearm (i.e. breech face) to the next with little variation. It should be noted that subclass characteristics are unlikely to appear in real casework. Nevertheless, using test samples made from consecutively manufactured parts creates a challenging “worst-case-scenario” of best non-matching patterns. This can create comparison scenarios for examiners that are more difficult than those typically encountered during actual casework. In addition, a number of these studies used more “questioned” than “known” samples. As a result, participants were unable to determine a few correct answers and simply deduce the rest. Finally, because these studies used samples produced by consecutively manufactured parts, it was equally important to know whether participants could correctly associate questioned samples with known sources as it was to know whether those samples would be falsely identified. As a result, the studies included at least one known source with each questioned sample.

An additional benefit of a closed-set design is that it simulates real casework. In black box studies, the questioned samples are examined independently of each other—not as a set. During actual casework, however, examiners are not faced with completely independent comparison scenarios. Questioned samples and known items are typically collected and examined as a group, a circumstance that is mimicked by closed-set study designs. If one goal of method validation is to partially replicate casework conditions, then it is important to supplement black box studies with closed-set or partially open experimental designs.

There is no question that black box studies generate valuable information about examiner performance and decision thresholds under specified experimental conditions. Nevertheless, forensic method validation cannot be performed in a singular and one-dimensional manner. Studies of various design, scope, and substance all add value in the quest to better understand the circumstances under which error occurs and how it can be minimized. These efforts have been enhanced by a variety of experimental designs that have posed different questions to seek different types of answers.

To date, there have been approximately twenty firearms/toolmarks studies primarily focused on sample classification decisions and resulting error rates.¹⁰⁸ These studies used various experimental designs (black box, closed-set, partially-open, set-to-set), but have all resulted in a false positive error rate ranging from 0% to just over 1.0%.¹⁰⁹ The overall consistency of these findings when considered as a whole is a good indicator of what the *Daubert* Court described as a method’s “potential rate of error.”¹¹⁰ Importantly, this aggregate rate is very low, giving the

¹⁰⁸ See Appendices “A” and “B” to this statement.

¹⁰⁹ It is important to note that the composite upper range of approximately one percent false positive error in these studies does not mean that one percent of the time each examiner will make a false positive error, or that one percent of the time labs would report false positives, since these studies did not use standard quality assurance procedures, such as peer review and blind reexamination. See BALDWIN ET AL., *supra* note 89, at 18.

¹¹⁰ *Daubert*, 509 U.S. at 594.

overall indication that examiners are very accurate and make few source identification errors. Finally, it is worth noting that PCAST opined that an acceptable error rate should be less than 5%.¹¹¹ The aggregate false positive error rate in firearms/toolmarks studies to date falls well below that figure.

b. The Rate of Error vs. the Risk of Error

Despite the focus on the general frequencies at which various errors occur, the overall *rate* of error has little relevance to the critical question posed in most criminal litigation: What is the *risk* that error occurred in the case at hand? A 1996 report by the National Research Council, *The Evaluation of Forensic DNA Evidence* (“NRC II”),¹¹² recognized this important distinction. The NRC II observed, “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.”¹¹³

The NRC II committee specifically rejected a recommendation that laboratories use proficiency tests as the exclusive means for error rate determination—a proposal offered in a prior NRC committee report on forensic DNA evidence (NRC I, 1992), co-chaired by PCAST Co-Chair, Dr. Eric Lander. On this point, the NRC II committee stated:

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. A laboratory is not likely to make the same error again, so the error probability is correspondingly reduced.¹¹⁴

The committee also noted, “The risk of error is properly considered case by case, taking into account the record of the laboratory performing the tests, the extent of redundancy, and the overall quality of the results.”¹¹⁵ Moreover, the NRC II found it unnecessary to debate differing estimates of error when concerns about a false inclusion can be easily resolved by retesting the evidence.¹¹⁶ The NRC II’s view on error rates is shared by many leading scientists, statisticians, and forensic practitioners.¹¹⁷

¹¹¹ PCAST REPORT, *supra* note 1, at 151-52.

¹¹² NAT’L RES. COUNCIL, NAT’L ACADS., *THE EVALUATION OF FORENSIC DNA EVIDENCE* 85–88 (1996).

¹¹³ *Id.* at 85.

¹¹⁴ *Id.* at 86 (emphasis added).

¹¹⁵ *Id.* at 87.

¹¹⁶ *Id.*

¹¹⁷ *See, e.g.*, JOHN S. BUCKLETON ET AL., *FORENSIC DNA EVIDENCE INTERPRETATION* 76–77 (2d ed. 2016) (noting that error and error rates should be examined on a case-by-case basis) (“Our view is that the possibility of error should be examined on a per-case basis and is always a legitimate defence explanation for the DNA result. . . . The answer lies, in our mind, in a rational examination of errors and the constant search to eliminate them.”); BERNARD ROBERTSON ET AL., *INTERPRETING EVIDENCE: EVALUATING FORENSIC SCI. IN THE COURTROOM* 138 (2d ed. 2016) (“It is correct . . . to say that the possibility of error by a laboratory is a relevant consideration. It is wrong, however, to assume that the probability of error in a given case is measured by the past error rate. The question is what the chance of error was on this occasion.”); I.W. Evett et al., *Finding a Way Forward for Forensic Science in the US—A Commentary on the PCAST Report*, 278 *FORENSIC SCI. INT’L* 16, 22–23 (2017) (suggesting that proficiency tests should be used to determine error rates and rejecting the use of “black box” studies in their calculation and courtroom presentation).

IV. Conclusion

In their response to the PCAST Report, Dr. Ian Evett and colleagues wrote, “The notion of an error rate to be presented to courts is misconceived because it fails to recognise that the science moves on as a result of proficiency tests. . . . [O]ur vision is not of the black-box/error rate but of continuous development through calibration and feedback of opinions.”¹¹⁸

This sentiment reflects the current lack of scientific consensus on how—and indeed whether—error rates can or should be determined for forensic pattern comparison methods. Black box error rates, although adding to the body of knowledge, are a mere snapshot in time, place, and circumstance that capture a unique set of experimental conditions. Moreover, PCAST’s notion of a single, generally applicable error rate wrongly assumes that such a figure can be generally applied to different evidence, examiners, and case circumstances.¹¹⁹

In conclusion, error rates derived from scientific studies of various size, scope, and experimental design *can and do* provide important information about the decision-making abilities and proclivities of examiner-participants. For most pattern comparison disciplines, extant studies show that examiners, on average, perform extremely well under a variety of experimental conditions. Competency and proficiency tests add to the body of knowledge by measuring how often examiners make correct decisions using known, ground truth samples. Verification by a second examiner, technical review, case controls, and other quality assurance measures used by accredited laboratories are critical components of risk management and mitigation. Lastly, as noted by the NRC, a wrongfully accused person’s best insurance against false incrimination is the opportunity to have the evidence retested. In most cases, the typically non-consumptive nature of forensic pattern examination easily facilitates this final safeguard.

¹¹⁸ Evett et al., *supra* note 8, at 22.

¹¹⁹ MARCZYK ET AL., *supra* note 73, at 180 (“Every study operates under a unique set of conditions and circumstances related to the experimental arrangement. The most commonly cited examples include the research setting and the researchers involved in the study. The major concern with this threat to external validity is that the findings from one study are influenced by a set of unique conditions, and thus may not necessarily generalize to another study, even if the other study uses a similar sample.”).

APPENDIX A

Lead Author	Source	Year	Number of Participants	False Positive Rate (%)	Comparison Type Cases/Bullets
*Brundage	AFTE Journal	1998	30 (Plus 37 Informal Participants)	0	Bullets
Bunch	AFTE Journal	2003	8	0	Cartridge Cases
DeFrance	AFTE Journal	2003	9	0	Bullets
Smith	AFTE Journal	2004	8	0	Both
*Hamby	AFTE Journal	2009	507 (Includes *Brundage (1998) Participants)	0	Bullets
Lyons	AFTE Journal	2009	22	1.2 ^a	Cartridge Cases
Mayland	AFTE Journal	2010	64	1.7 ^b	Cartridge Cases
Cazes	AFTE Journal	2013	68 (or 69)	0	Cartridge Cases
Fadul	AFTE Journal	2013	Phase 1: 217 Phase 2: 114	Phase 1: .064 ^c Phase 2: 0.18 ^c	Cartridge Cases
Fadul	NIJ (NCJRS)	2013	183	0.40 ^d	Bullets
Stroman	AFTE Journal	2014	25	0	Cartridge Cases
Baldwin	NIJ (NCJRS)	2014	218	1.0	Cartridge Cases
Kerkhoff	Science & Justice	2015	11	0	Both
Smith	JFS	2016	31	0.14 Cases 0 Bullets	Cartridge Cases Bullets
Duez	JFS	2018	46 Examiners 10 trainees	0 ^e	Cartridge Cases
Keisler	AFTE Journal	2018	126	0	Cartridge Cases
*Hamby	JFS	2019	619 (Includes *Brundage (1998) + Hamby (2009) Participants)	0.053% ^f	Bullets
Smith	JFS	2020	72	0.08	Bullets

*Brundage study was continued by Hamby who added additional participants and reported the combined data in Fall 2009 and 2019.

^a The error rate reported by the author appears to be (1-True Positive Rate). There were three false positive identifications made but the number of true negative comparisons is not reported. 259 correct positive identifications were made. The False Discovery Rate (FDR) for the study is $3/(3+259)= 1.1\%$.

^b The false positive error rate is not reported by the authors. There were three false positive identifications and 178 correct positive identifications made. The False Discovery Rate (FDR) for the study is $3/(3+178)= 1.7\%$ and is reported in the table above.

^c The error rates reported by the authors are roughly equivalent to the False Discovery Rates (FDR) for each of the study phases (FDR = .062% and 0.18% respectively).

^d Eleven false positives occurred. The false positive error rate is not reported by the authors. The error rate quoted is equivalent to the False Discovery Rate $=11/(11+2734)= 0.40\%$.

^e Two false positives were made by one trainee. None were made by the qualified examiners. The false positive rate does not include the trainee errors. If trainee data is included with that submitted by examiners, the False Positive Rate is $(2/112) = 1.8\%$.

^f The empirically observed false positive rate is 0%. Using Bayesian estimation methods, the authors' most conservative (worst case) estimate of the average examiner false positive error rate for the study is .053% with a 95% credible interval of $(1.1 \times 10^{-5}\%, 0.16\%)$.

APPENDIX B

Firearms/Toolmarks – Error Rate Studies (Bullets & Cartridge Cases)

1. Brundage, D. (Summer 1998). The Identification of Consecutively Rifled Gun Barrels, *AFTE Journal*, 30(3), 438-44 (Bullets).
2. Bunch, S.G., & Murphy, D.P. (Spring 2003). A Comprehensive Validity Study for the Forensic Examination of Cartridge Cases, *AFTE Journal*, 35(2), 201-03 (Cartridge Cases).
3. DeFrance, C.S. & Van Arsdale, M.D. (Winter 2003). Validation Study of Electrochemical Rifling, *AFTE Journal*, 35(1), 35-37 (Bullets).
4. Smith, E.D. (Fall 2004). Cartridge Case and Bullet Comparison Validation Study with Firearms Submitted in Casework, *AFTE Journal*, 36(4), 130-35 (Bullets and Cartridge Cases).
5. Hamby, J.E., Brundage, D.J., & Thorpe, J.W. (Spring 2009). The Identification of Bullets Fired from 10 Consecutively Rifled 9mm Ruger Pistol Barrels: A Research Project Involving 507 Participants from 20 Countries, *AFTE Journal*, 41(2), 99-110 (Bullets).
6. Lyons, D.J. (Summer 2009). The Identification of Consecutively Manufactured Extractors, *AFTE Journal*, 41(3), 246-56 (Cartridge Cases).
7. Mayland, B. & Tucker, C. (Spring 2012). Validation of Obturation Marks in Consecutively Reamed Chambers, *AFTE Journal*, 44(2), 167-69 (Cartridge Cases).
8. Cazes, M. & Goudeau, J. (Spring 2013). Validation Study Results from Hi-Point Consecutively Manufactured Slides, *AFTE Journal*, 45(2), 175-77 (Cartridge Cases).
9. Fadul Jr., T.G., Hernandez, G.A., Wilson, E., Stoiloff, S., & Gulati, S. (Fall 2013). An Empirical Study to Improve the Scientific Foundation of Forensic Firearm and Tool Mark Identification Utilizing 10 Consecutively Manufactured Slides, *AFTE Journal*, 45(4), 376-93 (Cartridge Cases).
10. Fadul Jr., T.G., Hernandez, G.A., Wilson, E., Stoiloff, S., & Gulati, S. (December 2013). An Empirical Study to Improve the Foundation of Firearm and Tool Mark Identification Utilizing Consecutively Manufactured Glock EBIS Barrels with the Same EBIS Pattern. <https://www.ncjrs.gov/pdffiles1/nij/grants/244232.pdf> (Bullets)
11. Stroman, A. (Spring 2014), Empirically Determined Frequency of Error in Cartridge Case Examinations Using a Declared Double Blind Format, *AFTE Journal*, 46(2), 157-75 (Cartridge Cases).
12. Baldwin, D.P., Bajic, S.J., Morris, M., & Zamzow, D. (April 7, 2014). A Study of False-Positive and False-Negative Error Rates in Cartridge Case Comparisons. <https://apps.dtic.mil/dtic/tr/fulltext/u2/a611807.pdf> (Cartridge Cases).

13. Kerkhoff, W. et al. (2015). Design and Results of an Exploratory Double Blind Testing Program in Firearms Examination, *Science & Justice*, 55, 514-19 (Bullets and Cartridge Cases).
14. Smith, T.P., Smith, A.G., & Snipes, J.B. (July 2016). A Validation Study of Bullet and Cartridge Case Comparisons Using Samples Representative of Actual Casework, *Journal of Forensic Sciences*, 61(4), 939-45 (Cartridge Cases).
15. Duez, P. et al. (July 2018). Development and Validation of a Virtual Examination Tool for Firearm Forensics, *Journal of Forensic Sciences*, Vol. 63(4), 1069-1084 (Cartridge Cases).
16. Keisler, M. et al. (Winter 2018). Isolated Pairs Research Study, *AFTE Journal*, 50(1), 56-58 (Cartridge Cases).
17. Hamby, J. et al. (March 2019). A Worldwide Study of Bullets Fired From 10 Consecutively Rifled 9MM Ruger Pistol Barrels—Analysis of Examiner Error Rates, *Journal of Forensic Sciences*, 64(2), 551-57 (Bullets).
18. Smith, J. (October 2020). Beretta Barrel Fired Bullet Validation Study, *Journal of Forensic Sciences*, 2020;00:1-10 <https://onlinelibrary.wiley.com/doi/full/10.1111/1556-4029.14604> (Bullets).



National District Attorneys Association
1400 Crystal Drive, Suite 330, Arlington, VA 22202
703.549.9222/703.836.3195 Fax
www.ndaa.org

November 16, 2016

The President of the United States
The White House
1600 Pennsylvania Avenue, NW
Washington, DC 20500

Reference: Report Entitled “Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods”

Dear President Obama:

On behalf of the National District Attorneys Association (NDAA), the nation’s largest prosecutor organization, representing 2,500 elected and appointed District Attorneys across the United States, as well as 40,000 assistant district attorneys, I write to you today regarding the Report to the President-Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods (“the Report”). The NDAA takes issue with, and has substantial concern about, the logic of the report and the manner in which it portrays several forensic disciplines.

First and foremost amongst NDAA’s concerns is the pervasive bias and lack of independence apparent throughout the report. The report repeatedly contends that studies used to determine and/or establish the scientific validity of feature comparison disciplines must be conducted by entities independent of those who may have some stake in the outcome. The composition of the PCAST, however, violates this very principle; the PCAST membership included several who are far from “independent” and who have a direct “stake in the outcome.” A significant example is Eric Lander, Co-Chair of PCAST, and Chair of the working group, who is also a Member of the Board of Directors of the Innocence Project, an organization that has argued for years that the forensic feature comparison disciplines have failed to demonstrate their scientific validity and are, in part, responsible for numerous wrongful convictions. There is no evidence the scientific basis for forensic feature comparisons are responsible for wrongful convictions.

Second, the working group (and PCAST at large) included no forensic scientists. Rather, it consisted of six PCAST members (none of whom have forensic laboratory experience), ten judges, two law school professors, and two college professors. In addition, the report does not include a bibliography/appendix of the literature upon which it relied on in support of its findings and conclusions. Instead, the report simply offers, in Appendix B, a list of (apparently hand-picked) “Additional Experts Providing Input.” It is true that PCAST solicited literature references from various forensic organizations. The Report, however, does not indicate which of these the PCAST relied upon, considered or even read.

Third, without a single citation to scientific authority, the PCAST Report simply declares that forensic feature comparison methods belong to the scientific field of “metrology (including statistics).” Metrology is the study of scientific measurement. Crime labs use forensic metrology for determining the measurement of blood alcohol content, quantitation of drugs in a toxicology sample, weight of a controlled substance and the barrel length of a firearm. In light of this contention, it is inexplicable that the PCAST’s working group included **no** metrologists.

In their current form, the feature comparison methods considered in the Report clearly do *not* fall under the field of metrology. Labeling them as such was a transparently strategic attempt to bring these methods under the ambit of *Daubert v. Merrell Dow Pharmaceuticals, Inc.* 509 U.S. 579, (1993), a requirement that “in a case involving *scientific* evidence, evidentiary reliability will be based on *scientific validity*.” The Report’s self-professed primary purpose was to define what *scientific validity* means.

According to PCAST, (again without citation to any literature or authority), *scientific validity* for subjective feature comparison methods can be established *only* through numerous, properly constructed, independent black box studies with a variety of samples from a diverse population of features. The report then posits that there is an insufficient number of these properly designed black box studies that comply with PCAST’s unilaterally imposed criteria to establish the scientific validity of several of the disciplines discussed. Based on that claim, the report then not-so-subtly urged that courts consider excluding results from these disciplines, while giving mere lip service to the notion that admissibility remains a question for courts, not PCAST, to determine.

By wrongly labeling the forensic feature comparison disciplines as belonging to the field of metrology, the report conveniently overlooks the ancient debate over precisely what constitutes “science.” The answer to this question depends fundamentally upon which philosopher one finds most compelling and which definition one finds most persuasive. (Indeed, the debate over exactly what constitutes “science” has been ongoing since the time of Aristotle and is far from settled.) Under many definitions, the feature comparison methods that are the subject of the report certainly incorporate *aspects* of science. These methods however, also *independently* constitute “*technical*” and “*specialized knowledge*” under Federal Rule of Evidence 702. Significantly, “*technical*” and “*specialized knowledge*” are *not* fields of knowledge for which *Daubert* requires *scientific validity*. See *Daubert*, fn. 8 (“Our discussion is limited here to the scientific context because that is the nature of the expertise offered here”; and fn. 9, “In a case involving scientific evidence, *evidentiary reliability* will be based upon *scientific validity*.”) (Emphasis original). In *Kumho Tire v. Carmichael*, 526 U.S. 137, 149, (1999), the Supreme Court recognized that distinction, holding that where the “factual basis, data, principles, methods (of technical or specialized knowledge) or their application are called sufficiently into question...the trial judge must determine whether the testimony has “a reliable basis in the knowledge and experience of [the relevant] discipline.”

Further illustrating the internal contradiction is the inconvenient truth that the same working group critics who have long argued that the feature comparison methods are *not* science *now* insist that they *are* in fact science. This change of heart, however, appears to have been driven solely by the strategic need to shoehorn these disciplines into *Daubert's* holding that, in the case of scientific evidence, legal reliability is synonymous with *scientific validity*. Having completed this maneuver, the Report then imposes its own outcome-determinative definition of *scientific validity* on each canvassed method. Finally, the Report declares each one invalid due to an insufficient number of properly qualified black box studies that meet PCAST's newly-minted set of criteria. This is a transparent effort to persuade courts that they should exclude this technical or specialized evidence because it is not scientifically valid as required by *Daubert*. As elucidated by *Kumho Tire*, however, *Daubert* does not require *scientific validity* in the case of technical or specialized evidence, even if it incorporates scientific aspects.

Complex Mixture DNA

In assessing the scientific validity of DNA analysis of single-source and simple mixture samples, the Report determines that as an objective method, each of the steps has been found to be “repeatable, reproducible and accurate.” Thus, the authors correctly conclude that analyses of single –source and simple mixture samples of two individuals are an objective scientific method whose foundational validity has been properly and irrefutably established.

Moving onto the analysis of “complex mixture samples,” the Report contrasts the analysis of such samples with the analyses of single-source and simple mixtures by suggesting that complex mixture analysis is not based on “precisely defined laboratory protocols” as single-source and simple mixture analyses are. Although it is certainly true that DNA interpretation rests solidly on a laboratory’s protocols developed after conducting internal validation studies, such “precisely defined protocols” are by no means limited to single-source and simple mixture samples. Furthermore, non-probabilistic genotyping methods of DNA interpretation – whether of single source, simple mixture, or complex mixtures – requires some level of interpretation by a trained, well-qualified DNA analyst.

The Report challenges the DNA analysis of complex mixture samples and erroneously concludes that the Combined Probability of Inclusion (CPI) approach to complex mixture analysis is an inadequately specified, subjective method that is not foundationally valid.

From the outset, the Report paints with an overly broad brush in defining a “complex mixture sample.” The Report defines a complex mixture as one with more than two contributors and states in entirely conclusory fashion that this type of mixture is inherently difficult to interpret. In defining complex mixtures so broadly, the Report fails to make a critical distinction between complex mixtures that have a discernable ratio of the various contributors – and therefore can be validly interpreted based on laboratory validation studies and standard operating protocols using a random match probability statistic, a likelihood ratio, or a CPI approach -- and those that do not have such discernable ratios.

DNA interpretations of complex mixtures with discernable contributor ratios are carried out daily by laboratories across the United States reporting accurate and reliable results. The Report ignores the fundamental difference between this type of complex mixture and those in which a greater-than-two-person mixture contains undiscernible ratios of contributors. Complex mixtures in which contributor ratios are not distinct demonstrate phenomena such as allele stacking or allelic dropout. Laboratories can overcome such interpretation challenges with rigorous internal laboratory validation studies, well-defined standard operating procedures, and rigorous training of the DNA analysts. The critical issue is not (or should not be) whether a particular method such as CPI is not scientifically valid (as it has been demonstrated to be valid when applied correctly) but whether that scientifically valid method has been applied correctly to the particular sample being analyzed.

As evidence of the putative unreliability of the CPI approach, the Report devotes significant discussion to what it describes as “systemic” problems with the subjective analysis of complex DNA mixtures. The Report cites purported failings of analyses conducted in Texas in 2015. The Report unfairly attributes the failings of the Texas laboratories -- in which dramatic shifts in statistics resulted from the laboratories changing the way in which they calculated the CPI statistics – on the CPI method itself. The Report broadly asserts that it was not until 2015 that attorneys learned for the first time “the extent to which DNA mixture analysis involved subjective interpretation” and that problems arose with CPI because existing guidelines did not clearly, adequately, or correctly specify the proper use or limitation of the approach. To cast doubt on the method itself based on an individual laboratory’s misapplication of the method is misguided at best or disingenuous at worst. Rather than spending pages detailing the occurrences in the Texas laboratories and concluding that the problem was “systemic” while dismissing those who reliably interpret complex DNA mixtures, the Report should have relied upon articles published in peer-reviewed journals by experts in the field describing the proper use and limitations of the CPI method to interpret complex DNA mixture profiles.

Four publications describe the proper, scientifically valid use of CPI.¹ Dr. John Butler devotes parts of several chapters in his 2015 publication on advanced topics in DNA interpretation specifically to the use and limitations of CPI in complex DNA mixture interpretation.² The Report gives but a passing nod to the comprehensive methodology paper published in BMC

¹ Budowle, B, Onorato AJ, Callaghan TF, Della Manna A, Gross AM, Guerreri RA, et al. Mixture Interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework. *J. Forensic Sci.* (2009); 54:810-21; Butler JM. *Advanced Topics in Forensic DNA Typing: Interpretation*. Oxford: Elsevier; 2015; Scientific Working Group on DNA Analysis Methods (SWGDM). *SWGDM Interpretation Guidelines for Autosomal STR* Scientific Working Group on DNA Analysis Methods (SWGDM) *SWGDM Interpretation Guidelines for Autosomal Typing by Forensic DNA Laboratories*. 2010.; Bieber F, Buckleton J, Budowle B, Butler, J and Coble M, *Evaluation of Forensic DNA Mixture Evidence: Protocol for Evaluation, Interpretation, and Statistical Calculations using the Combined Probability of Inclusion*, *BMC Genetics*, 2016

² Butler JM. *Advanced Topics in Forensic DNA Typing: Interpretation*. Oxford: Elsevier;

Genetics in 2016³ that provides a detailed, specific set of rules for the use of CPI which the PCAST Report erroneously claims is lacking. The purpose of the article was to assist forensic laboratories that use CPI by providing a formal protocol for the proper use of CPI. The forensic DNA community has met the criteria set out by the PCAST Report by providing standardized protocols and methodology for the proper use of the CPI in complex mixture analysis. Nonetheless, ignoring published scientific literature, the Report, inexplicably concludes that the interpretation of complex DNA mixtures with the CPI statistic is inappropriately subjective and “clearly not foundationally valid.” To make such a sweeping claim in the face of publications authored by experts in the field seriously undermines confidence in the Report’s objectivity and reliability.

Latent Print Discipline

The report concludes that the use of latent fingerprint analysis satisfies the requirements of scientific reliability. The Report goes on to suggest that judges insist that jurors be apprised of error rates, which are the subject of significant scientific/technical disagreement. This is an example of the Report's confusion of the roles of experts, counsel, the judge, and the jury. Error rate issue is an issue of fact -- for experts to testify about and juries to resolve -- not one of law.

In addition, although NDAA concurs with the Report’s conclusion that latent prints are a scientifically reliable discipline, that concurrence is based on a great deal of scientific and technical validation that goes well beyond the two black box studies cited in the Report. Also indicative of the internal incoherence of the Report’s methodology is its failure to apply its own criteria for evaluation of black box studies to the studies cited on latent fingerprint analysis. That is, having set out criteria for the assessment of black box studies (and having artificially and unnecessarily limited the scope of potential validation for latent fingerprint analysis to black box studies), the Report inexplicably fails to apply those criteria to the black box studies it cites in support of the scientific reliability of latent fingerprint analysis.

Firearms Analysis

The science of tool mark identification, specifically firearms, is based on the premise that a tool mark can be individualized to the specific tool that produced it. Firearms identification involves the microscopic examination and comparison of cartridge casings and expended bullets to each other, and to test fires produced from known firearms. The unique features of each firearm, as designed by the firearm manufacturer, are transferred to the cartridge case and bullet whenever a weapon is fired. The cartridge case or shell is impressed with marks from contact with the metal surfaces of the gun’s firing and loading mechanisms, including the firing pin, breech face, ejector, extractor and magazine. In addition to marks left on the cartridge casing,

³ Bieber F, Buckleton J, Budowle B, Butler, J and Coble M, Evaluation of Forensic DNA Mixture Evidence: Protocol for Evaluation, Interpretation, and Statistical Calculations using the Combined Probability of Inclusion, BMC Genetics, (2016) 17:125.

as a fired bullet travels down the barrel of a gun, it will pick up impressed and striated tool marks (lands and grooves) that are generated by the working surface of the rifled bore of the barrel.

The PCAST findings with respect to firearms are especially puzzling as the Association of Firearm and Tool mark Examiners (AFTE) provided to the PCAST a comprehensive list of over 40 peer-reviewed published studies supporting the foundational aspects of the discipline and answering questions relating to other aspects of the discipline. This information is available at <https://afte.org/resources/wggun-ark>. This research includes a significant number of comprehensive experimental models involving close to a thousand examiners from the US and across the globe. The varied experimental models included numerous “consecutively manufactured barrel” tests, in which manufacturers provided a series of consecutively manufactured firearm barrels, which would be expected to be virtually identical. Trained examiners were asked to examine unknown fired bullets to determine whether they could correctly identify those bullets as having been fired from the barrel of a particular firearm. Other tests involved the effect of consecutive firing of firearms to determine how the wear on barrels and breech faces would affect the identification of fired bullets and cartridge casings. Still other tests involved microscopic studies of the reproducibility of tool marks on high velocity bullets fired through a single machine gun barrel. Various tests used double-blind procedures and studied false-positive and false-negative error rates and compared automated analyses systems to those of trained human examiners. The studies demonstrated that unique features of each firearm are transferred to cartridge casings and bullets and that trained examiners are able to correctly link related tool marks to the tool, i.e., the firearm that produced it with a high degree of accuracy.

PCAST, however, is critical of these studies. PCAST arbitrarily defined the acceptable parameters of validation studies and determined that the types cited by AFTE failed to meet those parameters. In comments regarding several cited studies, PCAST implies that these particular types of firearm validation studies are not challenging and the participants can determine the correct response by a process of elimination. Yet the PCAST members are neither forensic firearm scientists performing casework nor did they participate as examiners in these validation studies. PCAST unilaterally dismisses all work that does not comport with its arbitrary, singular experimental design. Years of research conducted prior to the PCAST report have established the scientific foundational validity of firearm/tool mark analysis.

Forensic Odontology

Forensic dentists are highly-trained medical professionals and their methods employ well-documented and well-understood medical and forensic techniques. Forensic dentists undergo standard medical dental training during which they take the same courses as medical students in pharmacology, physiology, histology, and anatomy of the oral and facial structure.

By virtue of their experience reading x-rays and performing surgeries, forensic dentists are experts in comparing dentitions, pattern, and are well-versed in the injury and healing properties of human skin.

Forensic dentists perform bite mark evidence collection through the use of highly specialized photography and harvest injured skin from deceased victims. They analyze bite marks using very specific criteria and highly specialized computer programs and tools.

Best practices for comparisons include blinded suspect sample collection and a “lineup” of potential suspects. Board certified forensic odontologists undergo a rigorous training and examination process by the American Board of Forensic Odontology.

Studies cited by the PCAST Report in support of its rejection of forensic odontology have been thoroughly discredited in court. For example, both the cadaver studies and 2-D and 3-D studies by Mary and Peter Bush were poorly designed and executed and as a result, did not reliably demonstrate anything. The AAFS study was similarly flawed. The authors admit that the small number of participants and mid-study rule changes, among other problems, meant the study proved only the obvious fact that the best possible evidence should be used when conducting bite mark analysis and comparison.

Forensic odontology is an important tool, for both prosecution and defense, especially in child abuse cases. These cases commonly involve a limited number of people who have access to the child and comparisons between this “closed population” of suspects can often reliably exclude all but one suspect who may be included as a possible perpetrator based on specific similarities between the suspect’s dentition and the bite mark injury. Judges, juries, potential defendants and victims all need this valuable tool in the pursuit of justice. PCAST’s study of historic cases in which convictions were vacated do not address vast improvements in forensic odontology and are not relevant to forensic practices today.

Closing

Finally, it should be noted that the Report applies only selectively its assertion that numerous peer reviewed and published studies are required. In several instances (for example, cognitive bias) the Report relies upon a single study on an isolated topic that has not been replicated by other researchers and generalizes the single study’s findings to all analogous forensic disciplines. The Report does this despite its requirement that proponents of a particular discipline support their claims with numerous peer-reviewed studies. Cherry-picking studies that report findings that support the report’s positions, but that fail to satisfy the report’s own criteria for feature comparison methods, further exposes the Report’s biases and, in doing so undermines its credibility.

Throughout its report, PCAST announces, by fiat, certain broad and sweeping definitions and sets of criteria without a single attribution to extant scientific authority in support of these assertions. Among these are its definitions of scientific validity (for both objective and

subjective methods); validity as applied; and the assertion that the only means by which these scientific concepts can be established is via multi-part tests, apparently created adhoc by the PCAST working group.

In its report, PCAST provides three types of evidence that it argues undermines, “from a scientific standpoint,” “the continuing validity of conclusions that were not based on appropriate empirical evidence.” These are Innocence Project exonerations; the 2009 NRC Report; and “the scientific review in this report by PCAST, the leading scientific advisory body established by the Executive Branch, finding that some forensic feature-comparison methods lack foundational validity.”

PCAST’s attempt to bootstrap its own qualifications as justification for the exclusion of feature comparison evidence, and its attempt to appeal to the reader’s deference to its own political authority, is the height of irony (and hypocrisy) for a group that criticizes feature comparison methods because of their reliance on *skill and experience* rather than upon foundational authorities.

In addition, while criticizing the feature comparison disciplines for failing to rely on adequate empirical evidence to establish their foundational validity, PCAST, ironically, feels no need to rely upon any foundational scientific material to support its *own* numerous scientific edicts. Instead, PCAST bases its assertions on “the *ipse dixit*” of its own alleged expertise in this field. Setting aside that PCAST has no forensic expertise *per se*, the *ipse dixit* of the expert is not a sufficient basis upon which to admit scientific testimony in a courtroom. Likewise, it offers no reason to credit the assertions made in its Report.

In the end, the report offers an appeal to its own authority as a justification for courts to rely on its recommendations to exclude feature comparison evidence. Not only is this dangerous but it is well beyond the Report’s purview. Assertions by the Attorney General and the FBI Director that they will not heed the report’s recommendations constitute a powerful repudiation of the methods and conclusions of the PCAST process. Experience shows these disciplines offer reliable and powerful evidence in a court of law. It is therefore entirely inappropriate for the report to suggest otherwise to this country’s courts.

To address legitimate questions surrounding forensic science, NDAA supports establishment of an Office of Forensic Science within the Department of Justice as recommended by Senators Cornyn and Leahy in 2014 in the Criminal Justice and Forensic Science Reform Act of 2014. One of the Act’s recommendations is a Comprehensive Research Strategy and Agenda for fostering and improving peer-reviewed scientific research relating to the forensic science disciplines, including research addressing validity, reliability, and accuracy in the forensic science disciplines. It is our understanding that PCAST has been tasked with generating a research strategy within the Office of Science and Technology Policy (OSTP) under your Office. An Office of Forensic Science, in our opinion, should be charged with these tasks in order to help facilitate all the partners collaboratively within the forensic community and the Department of Justice. In our view, the Department of Justice is better suited for this task than the OSTP, due to the

broad range of subjects it is asked to study such as climate change, antibiotic resistance and education. We support peer-reviewed scientific research relating to the forensic science disciplines to continue to improve validity, reliability, and accuracy.

Sincerely,

A handwritten signature in cursive script that reads "Michael A. Ramos". The signature is written in a dark ink and is positioned above the printed name.

Michael A. Ramos

President

National District Attorneys Association



Comments on:
President's Council of Advisors on Science and Technology
REPORT TO THE PRESIDENT
Forensic Science in Federal Criminal Courts: Ensuring Scientific
Validity of Pattern Comparison Methods

The FBI agrees with the authors of the President's Council of Advisors on Science and Technology (PCAST) report that forensic science plays a critical role in the criminal justice system, and therefore needs to be held to high standards. Further, the FBI agrees with the PCAST report as well as the 2009 National Research Council report (2009 NAS report) that significant funding is needed to develop stronger ties between the academic research community and the forensic science community. It is inherent within science that over time, our knowledge of a subject evolves. It is critical that continued research be pursued in order to ensure that forensic science meets the high standards necessary to be used in a court of law.

However, the FBI disagrees with many of the scientific assertions and conclusions of the report. The report makes broad, unsupported assertions regarding science and forensic science practice. For example, the report states that "the *only* way" to establish "validity as applied" is through proficiency testing, and requires a measurement of how often the examiner gets the correct answer, which is fundamentally at odds with a report of the National Academy of Sciences.¹

The report also creates its own criteria for scientific validity and then proceeds to apply these tests to seven forensic science disciplines, failing to provide scientific support that these criteria are well accepted within the scientific community. In fact, PCAST defines their internally developed criteria as "scientific criteria" by which forensic feature-comparison methods must be supported by. However, PCAST does not apply its own criteria consistently or transparently. The PCAST criteria define "black box" studies as the benchmark to demonstrate foundational validity, but provide no clarification on how many studies are needed or why some studies that have been conducted do not meet their criteria. These criteria seem to be subjectively derived and are therefore inconsistent and unreliable.

The report does not mention numerous published research studies which seem to meet PCAST's criteria for appropriately designed studies providing support for foundational validity. That omission discredits the PCAST report as a thorough evaluation of scientific validity.

The report proposes federal government criminal justice related-databases should be made available to researchers for independent studies while consistently overlooking the legal authorization and limitations set out in statutes and regulations regarding the use of such databases.

Finally, the report ignores important differences between forensic science disciplines, conflating fundamental differences between class-level and identification-level evidence, leading to troubling generalized conclusions about all forensic science disciplines.

¹ National Academy of Sciences, *STRENGTHENING FORENSIC SCIENCE IN THE UNITED STATES*, 9 (2009); *see also id.* At 207 ("[b]lind proficiency testing is recommended . . . not as a way to determine error rates, but as a more precise test of a worker's accuracy.")

September 20, 2016

Department of Forensic Sciences Science Advisory Board's Statement with regard to the PCAST Report

Introduction

On September 20, 2016, the US President's Council of Advisors on Science and Technology (PCAST) published a report on *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods* in response to the President's question as to whether there are additional steps that could help ensure the validity of forensic evidence in the Nation's legal system.

As appropriate to the disciplines offered by the Department of Forensic Sciences, the Advisory Board will address the disciplines of Forensic Biology (DNA), Latent Fingerprint Analysis, and Firearms Analysis. The Board has decided to address these disciplines separately, beginning with Forensic Biology. The other disciplines will be addressed in the next few meetings.

DNA

According to published reviews of this report (e.g., [1-4]), the PCAST report presents a flawed paradigm for forensic evaluation, misapplies statistics and the notion of probability, ignores existing data and literature in forensic science, and, as a result, state that the PCAST report is scientifically unsound.

The PCAST report concludes that the DNA analysis of single-source specimen and simple mixtures of two contributors is a foundationally valid and reliable method, yet raises several concerns about the interpretation of complex DNA mixtures (pp. 75-83). Regarding the latter, the report concludes (page 82):¹

Objective analysis of complex DNA mixtures with probabilistic genotyping software is relatively new and promising approach. Empirical evidence is required to establish the foundational validity of each such method within specified ranges. At present, published evidence supports the foundational validity of analysis, with some programs, of DNA mixtures of 3 individuals in which the minor contributor constitutes at least 20 percent of the intact DNA in the mixture and in which the DNA amount exceeds the minimum required level for the method. The range in which foundational validity has been established is likely to grow as adequate evidence for more complex mixtures is obtained and published.

We, the Science Advisory Board, state that at the time of this writing, the range in which foundational validity has been established for the interpretation of complex mixtures at DFS using

¹ Note that an addendum to the report that appeared in January 2017 reached a slightly different conclusion (page 8):

PCAST found that empirical testing of PG [probabilistic genotyping] had largely been limited to a narrow range of parameters (number and ratios of contributors). We judged that the available literature supported the validity and reliability of PG for samples with three contributors where the person of interest comprises at least 20% of the sample. Beyond this approximate range (i.e. with a larger number of contributors or where the person of interest makes a lower than 20% contribution to the sample), however, there has been little empirical validation.

probabilistic genotyping² extends from DNA mixtures of 2 individuals up to DNA mixtures of 5 individuals. The PCAST notion of a lower limit percentage of the minor contributor as a criterion for deciding whether a DNA profile is interpretable or uninterpretable is scientifically unsound. The scientific criterion for making this decision is the quantity of information in the electropherogram(s) for a particular contributor.³ DFS has a valid pre-evaluation phase in place for making this decision.

More specifically, an internal validation study conducted by the DNA analysts at DFS⁴ consisting of over 10,000 comparisons to 100 DNA mixtures ranging from 2 contributors to 5 contributors has addressed each of the PCAST concerns listed below (PCAST, pp. 79-80).

These probabilistic genotyping software programs clearly represent a major improvement over purely subjective interpretation. However, they still require careful scrutiny to determine (1) whether the methods are scientifically valid, including defining the limitations on their reliability (that is, the circumstances in which they may yield unreliable results) and (2) whether the software correctly implements the methods. This is particularly important because the programs employ different mathematical algorithms and can yield different results for the same mixture profile. (PCAST, page 79)

The internal validation study conducted at DFS demonstrates that the interpretation of complex mixtures using STRmix™ version 2.4 in conjunction with GlobalFiler™ PCR Amplification Kit and 3500/3500xL Genetic Analyzer is scientifically valid for mixtures of 2 to 5 individuals.

To test the correctness of the software's implementation of the method, the DFS internal validation study reproduced the likelihood ratio values for each locus of a single-source profile in quadruple, once for each of four allele frequency databases. These results confirm that the software correctly implements the method.

Appropriate evaluation of the proposed methods should consist of studies by multiple groups, not associated with the software developers, that investigate the performance and define the limitations of programs by testing them on a wide range of mixtures with different properties. In particular, it is important to address the following issues:

- (1) *How well does the method perform as a function of the number of contributors to the mixture? How well does it perform when the number of contributors to the mixture is unknown? (PCAST, page 79)*

² Note that probabilistic genotyping does not identify contributors with 100% certainty. Instead it applies mathematical models and probability theory to assign probabilities to the observed peak heights given different sets of potential contributors. The conclusion is therefore probabilistic, taking the form of a likelihood ratio.

³ The quantity of information in the electropherogram(s) for a particular contributor depends on the quantity of data and the information known about the mixture.

⁴ The DFS internal validation study strictly follows the FBI approved SWGDAM Guidelines for the Validation of Probabilistic Genotyping Systems available at https://docs.wixstatic.com/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf (accessed January 2, 2018). It was approved by the Technical Leader on 1/7/2016 for the Identifiler Plus PCR Amplification kit and on 2/24/2017 for the GlobalFiler PCR Amplification kit. A summary of the results is available at <https://dfs.dc.gov/page/fbu-validation-studiesperformance-checks> (accessed January 5, 2018), and these results have been published in a peer-reviewed journal as part of a larger compilation of results from STRmix™ internal validation studies [5].

The DFS internal validation study tested the performance of the method for 40 mixtures with 2 contributors, and 20 mixtures each for 3, 4 and 5 contributors. These mixtures varied in DNA quantity and mixture proportions to represent the typical profiles⁵ encountered by the laboratory. The method correctly and reliably produced the expected results for each of the different number of contributors tested.

In addition, the results of the FBI internal validation study on the performance of STRmix™ version 2.3.06 contains a total of 290 mixtures with 2, 3, 4, and 5 contributors, for each of which the software proved to be appropriately sensitive and specific [6].

In casework, the number of contributors is always unknown (e.g., [7]). The DNA analyst assigns the number of contributors based on the number of peaks and the peak height information in the electropherogram.

To test the effect of an incorrect assignment of the number of contributors, the DFS internal validation study included the following tests:

- 10 mixtures each with 1, 2, 3 and 4 contributors were incorrectly interpreted as having 2, 3, 4 and 5 contributors, respectively; and
- 3 mixtures each with 2 and 3 contributors, and 4 mixtures each with 4 and 5 contributors were incorrectly interpreted as having 1, 2, 3 and 4 contributors, respectively

Each mixture was then evaluated against each of the known contributors and against 134 known non-contributors.

Overestimation of the number of contributors correctly produced likelihood ratios greater than 1 for the known contributors. It produced a few likelihood ratios greater than 1 for known non-contributors, but their order of magnitude is much lower than the likelihood ratios produced for the known contributors.⁶

Underestimation of the number of contributors did not have any influence on the likelihood ratios for the known major and minor contributors. It correctly produced lower likelihood ratios for the known trace contributors.

The FBI internal validation study included similar tests on an additional 30 mixtures which produced the same expected trends as the DFS internal validation results [6].

(2) How does the method perform as a function of the number of alleles shared among individuals in the mixture? Relatedly, how does it perform when the mixtures include related individuals? (PCAST, page 79)

The DFS internal validation study performed sensitivity and specificity studies on mixtures with different amounts of alleles shared among the contributors across the loci. These tests correctly and reliably produced the expected results. Given that continuous probabilistic genotyping models take allele sharing into account in their peak height models, this method can handle the entire range of possible allele sharing among the DNA's contributors.

⁵ This includes partial profiles.

⁶ Note that DFS has defined likelihood ratios between 1 and 100 as being “uninformative” based on the results of their internal validation study.

With regard to related individuals, the FBI internal validation study tested the method on mixtures with 3 contributors that consisted of 2 parents and 1 child. This type of mixture entails a risk of an underestimation of the number of contributors if only the number of peaks is counted and peak height information is disregarded. An underestimation of the number of contributors has no impact on the likelihood ratios of the known major and minor contributors, yet lowers the likelihood ratio for the known trace contributor.

(3) How well does the method perform—and how does accuracy degrade—as a function of the absolute and relative amounts of DNA from the various contributors? For example, it can be difficult to determine whether a small peak in the mixture profile represents a true allele from a minor contributor or a stutter peak from a nearby allele from a different contributor. (Notably, this issue underlies a current case that has received considerable attention.) (PCAST, page 79)

The DFS internal validation study included sensitivity and specificity studies on DNA mixtures of varying amounts of DNA. These ranged from an average peak height of about 20 rfu to >25,000 rfu (saturation). The mixture ratios ranged from 25:1 to 1:1 for two person mixtures, with the full range in between for three, four and five person mixtures. As expected for all methods, this method correctly and reliably produced uninformative results for contributors with very low template. For contributors with higher template, this method correctly and reliably produced high likelihood ratios greater than 1 for known contributors, and low likelihood ratios less than 1 for known non-contributors, which clearly separated the results of the known contributors from the results of the known non-contributors. On the high-template end, the method correctly interprets the profile qualitatively for saturated profiles.

Probabilistic genotyping does not determine whether a small peak in the mixture profile represents a true allele from a minor contributor or a stutter peak from a nearby allele from a different contributor. It takes all reasonable possibilities into account, and assigns probabilities to the observations given each of the possibilities. In other words, it assigns weights to the different possibilities, and must therefore not choose between the category of a true allele and the category of a stutter peak.

(4) Under what circumstances—and why—does the method produce results (random inclusion probabilities) that differ substantially from those produced by other methods? (PCAST, page 80)

The method used by DFS uses a fully continuous probabilistic genotyping model to produce likelihood ratios which express the relative support the DNA typing results provide for one proposition with regard to an alternative proposition. A likelihood ratio is a different statistical quantity from a random match probability or a combined probability of inclusion, and will therefore produce different numerical results than either of the latter quantities. In addition, a fully continuous model can produce likelihood ratios that are different from likelihood ratios obtained from a binary model or a semi-continuous model: the reason for these differences is that a fully continuous model takes into account all of the available peak height information above the analytical threshold in the electropherogram, whereas binary and semi-continuous models only take a very limited amount of this information into account (e.g., comparing observed peak heights to a stochastic threshold), if at all. Hence a fully continuous model will produce results different from those produced by binary and semi-continuous models in circumstances where the electropherogram contains peak height information that is taken into account by the fully

continuous model and not taken into account by the binary and semi-continuous models. Taking into account more information makes this method produce higher likelihood ratios in support of the DNA contribution of known contributors and lower likelihood ratios (or exclusions) in support of no DNA contribution of known non-contributors (e.g., [8-11]). This is the expected performance for all likelihood ratio methods.

Most importantly, current studies have adequately explored only a limited range of mixture types (with respect to number of contributors, ratio of minor contributors, and total amount of DNA). The two most widely used methods (STRMix and TrueAllele) appear to be reliable within a certain range, based on the available evidence and the inherent difficulty of the problem. Specifically, these methods appear to be reliable for three-person mixtures in which the minor contributor constitutes at least 20 percent of the intact DNA in the mixture and in which the DNA amount exceeds the minimum level required for the method. (PCAST, page 80)

The DFS internal validation study has shown that STRmix™ version 2.4 is reliable for DNA mixtures with 2, 3, 4 and 5 contributors. Independently, the FBI internal validation study has shown that STRmix™ version 2.3.06 is reliable for DNA mixtures with 2, 3, 4, and 5 contributors [6]. The results of additional internal validation studies of STRmix™ conducted by other laboratories can be found at <https://johnbuckleton.wordpress.com/strmix/strmix-validations/> (accessed October 24, 2017).

Again, we note that the PCAST notion of a lower limit percentage of the minor contributor as a criterion for deciding whether a DNA profile is interpretable or uninterpretable is scientifically unsound. The scientific criterion for making this decision is the quantity of information in the electropherogram(s) for a particular contributor (e.g. [12]).

References:

- [1] I.W. Evett, C.E.H. Berger, J. Buckleton, C. Champod, G. Jackson, Finding the way forward for forensic science in the US - A commentary on the PCAST report, Forensic Science International 278 (2017) 16-23.
- [2] G.S. Morrison, D.H. Kaye, D.J. Balding, D. Taylor, A.P. Dawid, C.G.G. Aitken, S. Gittelsohn, G. Zadora, B. Robertson, S. Willis, S. Pope, M. Neil, K.A. Martire, A. Hepler, R.D. Gill, A. Jamieson, J.d. Zoete, R.B. Ostrum, A. Caliebe, A comment on the PCAST report: Skip the 'match'/'non-match' stage, Forensic Science International 272 (2017) e7-e9.
- [3] B. Budowle, Response to the PCAST report, 2017.
- [4] J. Buckleton, U.S. v. Benito Valdez, 2017.
- [5] J.-A. Bright, R. Richards, M. Kruijver, H. Kelly, C. McGovern, A. Magee, A. McWhorter, A. Cieccko, B. Peck, C. Baumgartner, C. Buettner, S. McWilliams, C. McKenna, C. Gallacher, B. Mallinder, D. Wright, D. Johnson, D. Catella, E. Lien, C. O'Connor, G. Duncan, J. Bundy, J. Echard, J. Lowe, J. Stewart, K. Corrado, S. Gentile, M. Kaplan, M. Hassler, N. McDonald, P. Hulme, R.H. Oefelein, S. Montpetit, M. Strong, S. Noël, S. Malsom, S. Myers, S. Welti, T. Moretti, T. McMahon, T. Grill, T. Kalafut, M.M.

- Greer-Ritzheimer, V. Beamer, D. Taylor, J. Buckleton, Internal validation of STRmix™ - A multi laboratory response to PCAST, *Forensic Science International: Genetics*, in press (2018).
- [6] T.R. Moretti, R.S. Just, S.C. Kehl, L.E. Willis, J. Buckleton, J.-A. Bright, D. Taylor, A.J. Onorato, Internal validation of STRmix™ for the interpretation of single source and mixed DNA profiles, *Forensic Science International: Genetics* 29 (2017) 126-144.
- [7] J.-A. Bright, D. Taylor, C. McGovern, S. Cooper, L. Russell, D. Abarno, J. Buckleton, Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles, *Forensic Science International: Genetics* 23 (2016) 226-239.
- [8] T.W. Bille, S.M. Weitz, M.D. Coble, J. Buckleton, J.-A. Bright, Comparison of the performance of different models for the interpretation of low level mixed DNA profiles, *Electrophoresis* 35 (2014) 3125-3133.
- [9] H. Kelly, J.-A. Bright, J. Buckleton, J.M. Curran, A comparison of statistical models for the analysis of complex forensic DNA profiles, *Science & Justice* 54 (2014) 66-70.
- [10] J.-A. Bright, I.W. Evett, D. Taylor, J.M. Curran, J. Buckleton, A series of recommended tests when validating probabilistic DNA profile interpretation software, *Forensic Science International: Genetics* 14 (2015) 125-131.
- [11] D. Taylor, Using continuous DNA interpretation methods to revisit likelihood ratio behaviour, *Forensic Science International: Genetics* 11 (2014) 144-153.
- [12] J.-A. Bright, D. Taylor, S. Gittelsohn, J. Buckleton, The paradigm shift in DNA profile interpretation, *Forensic Science International: Genetics* 31 (2017) e24-e32.



AMERICAN SOCIETY OF CRIME LABORATORY DIRECTORS, INC.

139 A Technology Drive Garner, NC 27529



September 30, 2016

Statement on September 20, 2016 PCAST Report on Forensic Science

On September 20, 2016, the President's Council of Advisors on Science and Technology (PCAST) issued the report to the President, "**Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods,**" which contained seven (7) Scientific Findings and eight (8) Recommendations on the scientific validity of forensic sciences involving feature-comparisons.

The ASCLD Board of Directors has reviewed the official report from PCAST, and finds that while we do agree with some aspects of the PCAST report, we respectfully disagree with many of the Findings and Recommendations including the overarching methodology with which the analysis was performed.

ASCLD strongly agrees that additional financial investment from the Federal government into forensic science is sorely needed. From foundational and applied research funding to investment into operational capacity building and technological advancement, a strong financial investment from the Federal government is critical.

ASCLD also agrees that additional research can always be performed to further demonstrate the appropriate weight that should be afforded to the feature comparison disciplines, both in the capability of the science itself and in the capability of those that conduct examinations. This is how science evolves. PCAST's dismissal, however, of a wealth of existing research because it does not meet an arbitrary criteria of black box studies with an ideal sample size is unhelpful. ASCLD is aware that more than 2,000 post-2009 articles were submitted to PCAST for review during this year-long effort. Additionally, the former OSTP Subcommittee on Forensic Science Interagency Working Groups, AAAS, and several industry working groups either have or are currently developing extensive bibliographies, many of which do not appear to have been reviewed or given credibility.

ASCLD disagrees with discarding these studies as not credible simply for lack of black box studies or ideal sample size. ASCLD concurs that black box and white box studies are significantly important and helpful. Indeed, we sincerely appreciate that the Council highlighted a firearms study in which ASCLD participated. ASCLD does not agree, however, that black box studies are the singular method through which to judge an entire forensic discipline's reliability. ASCLD does not dispute that the proposed methodologies incorporated in the report are highly aspirational and rigorous; however, ASCLD is concerned that a one-size-fits-all approach is not always appropriate due to the specific research needs and unique evidence sample traits of each discipline. These disciplines have previously withstood both scientific and judicial scrutiny, aiding investigators, prosecutors, and defense attorneys throughout the criminal justice system.

In addition to the methodology of PCAST's review, ASCLD wishes to express concern over the following:

- **Practitioner involvement.** The report seems to favor that all scientific evaluation activities be performed completely separate from scientists with direct forensic science experience. ASCLD strongly disagrees with the removal of forensic scientists from the evaluation of scientific integrity or technical merit of analyses. ASCLD supports the involvement of academic scientists in the process, but strongly disagrees that these evaluations should be performed in a vacuum devoid of participation by the forensic scientists who can impart an applied knowledge and understanding to the research. Science is not specific or unique to academia or industry. It is the intersection of both that ensures true advancement and the collaboration of both paradigms is paramount to the continued improvement of forensic science.
- **OSAC “independence.”** ASCLD disagrees with the assertion that the NIST OSAC must be staffed with more “independent” scientists. ASCLD believes independence has already been demonstrated by the current OSAC composition, as several existing industry standards have already been referred to standards development organizations for revision in order to incorporate suggested improvements by OSAC units. ASCLD acknowledges there is an important need for input in OSAC from statisticians, metrologists, academic scientists, cognitive behavioral scientists, and legal experts; however, there is no evidence that the current process is broken or needs revision. In fact, ASCLD believes that great success has been shown in OSAC when these resources are engaged early in the process when standards and guidelines are in the development stage at the subcommittee level rather than later in the approval process only.
- **DNA mixture interpretation.** The report determines that, “...the interpretation of complex DNA mixtures with the CPI statistic has been an inadequately specified—and thus inappropriately subjective—method. As such, the method is clearly not foundationally valid.” ASCLD concurs with PCAST to the extent that the principle issue is the subjectivity and variability in the application of mixture interpretation guidelines within the community. ASCLD, however, urges PCAST to consider that the CPI statistic itself: (1) does not interpret complex DNA mixtures and; (2) is a valid statistical tool when properly applied to some DNA mixtures. The use of the CPI statistic is valid and fundamentally sound for use with DNA mixtures where all allelic peaks - after accounting for potential allele stacking and peak height variability - remain above the stochastic threshold. In summary, it is the inappropriate use of the CPI statistic by some practitioners rather than the CPI statistic itself that is not foundationally valid. As the PCAST report correctly acknowledges, new probabilistic software tools have been developed and are being made available to practitioners in an effort to achieve greater consistency in mixture interpretation. The use of new software tools, however, does not necessarily increase the objectivity of the analysis.
- **Simple proficiency tests.** The report indicates that the forensic community prefers proficiency tests not to be too challenging. ASCLD does not agree with this characterization of the entire community, regardless of who made the statement. ASCLD believes the majority of the forensic science community has, and continues, to implement rigorous quality assurance systems that include proficiency testing schemes that resemble the level of difficulty of casework.

While ASCLD has expressed disagreement with a number of aspects of the PCAST report on forensic science, we also wish to convey our desire to work collaboratively with PCAST and other federal agencies on continuing to improve forensic science. ASCLD remains committed to providing excellence in forensic science through leadership and innovation and encouraging the highest standards of practice in the field. The Board of Directors looks forward to continuing to partner with all members of the criminal justice community and any other group with the same interests.

AN ADDENDUM TO THE PCAST REPORT ON FORENSIC SCIENCE IN CRIMINAL COURTS

On September 20, 2016, PCAST released its unanimous report to the President entitled “*Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods.*” This new document, approved by PCAST on January 6, 2017, is an addendum to the earlier report developed to address input received from stakeholders in the intervening period.

Background

PCAST’s 2016 report addressed the question of when expert testimony based on a forensic feature-comparison method should be deemed admissible in criminal courts.¹ We briefly summarize key aspects of the previous report.

Forensic feature-comparison methods

PCAST chose to focus solely on forensic feature-comparison methods. These methods seek to determine whether a questioned sample is likely to have come from a known source based on shared features in certain types of evidence. Specific methods are defined by such elements as:

- (i) the type of evidence examined (e.g., DNA, fingerprints, striations on bullets, bitemarks, footwear impressions, head-hair);
- (ii) the complexity of the sample examined (e.g., a DNA sample from a single person vs. a three-person mixture in which a person of interest may have contributed only 1%); and
- (iii) whether the conclusion concerns only “class characteristics” or “individual characteristics” (e.g., whether a shoeprint was made by a pair of size 12 Adidas Supernova Classic running shoes vs. whether it was made by a *specific* pair of such running shoes).

The U.S. legal system recognizes that scientific methods can assist the quest for justice, by revealing information and allowing inferences that lie beyond the experience of ordinary observers. But, precisely because the conclusions are potentially so powerful and persuasive, the law requires scientific testimony be based on methods that are scientifically valid and reliable.²

Requirement for empirical testing of subjective methods

In its report, PCAST noted that the *only* way to establish the scientific validity and degree of reliability of a *subjective* forensic feature-comparison method—that is, one involving significant human judgment—is to test it *empirically* by seeing how often examiners actually get the right answer. Such an empirical test of a subjective forensic feature-comparison method is referred to as a “black-box test.” The point reflects a central tenet underlying all science: *an empirical claim cannot be considered scientifically valid until it has been empirically tested.*

If practitioners of a subjective forensic feature-comparison method claim that, through a procedure involving substantial human judgment, they can determine with reasonable accuracy whether a particular type of evidence came from a particular source (e.g., a specific type of pistol or a specific pistol), the claim cannot be considered scientifically valid and reliable until one has tested it by (i) providing an adequate number of examiners with an adequate number of test problems that resemble those found in forensic practice and (ii) determining whether they get the right answer with acceptable

¹ As noted in the report, PCAST did not address the use of forensic methods in criminal *investigations*, as opposed to in criminal prosecution in courts.

² See discussion of the Federal Rules of Evidence in Chapter 3 of PCAST’s report.

frequency for the intended application.³ While scientists may debate the precise design of a study, there is no room for debate about the absolute requirement for empirical testing.

Importantly, the test problems used in the empirical study define the specific bounds within which the validity and reliability of the method has been established (e.g., is a DNA analysis method reliable for identifying a sample that comprises only 1% of a complex mixture?).

Evaluation of empirical testing for various methods

To evaluate the empirical evidence supporting various feature-comparison methods, PCAST invited broad input from the forensic community and conducted its own extensive review. Based on this review, PCAST evaluated seven forensic feature-comparison methods to determine whether there was appropriate empirical evidence that the method met the threshold requirements of “scientific validity” and “reliability” under the Federal Rules of Evidence.

- In two cases (DNA analysis of single-source samples and simple mixtures; latent fingerprint analysis), PCAST found that there was clear empirical evidence.
- In three cases (bitemark analysis; footwear analysis; and microscopic hair comparison), PCAST found *no empirical studies whatsoever* that supported the scientific validity and reliability of the methods.
- In one case (firearms analysis), PCAST found only one empirical study that had been appropriately designed to evaluate the validity and estimate the reliability of the ability of firearms analysts to associate a piece of ammunition with a specific gun. Because scientific conclusions should be shown to be reproducible, we judged that firearms analysis currently falls short of the scientific criteria for scientific validity.
- In the remaining case (DNA analysis of complex mixtures), PCAST found that empirical studies had evaluated validity within a limited range of sample types.

Responses to the PCAST Report

Following the report’s release, PCAST received input from stakeholders, expressing a wide range of opinions. Some of the commentators raised the question of whether empirical evidence is truly needed to establish the validity and degree of reliability of a forensic feature-comparison method.

The Federal Bureau of Investigation (FBI), which clearly recognizes the need for empirical evidence and has been a leader in performing empirical studies in latent-print examination, raised a different issue. Specifically, although PCAST had received detailed input on forensic methods from forensic scientists at the FBI Laboratory, the agency suggested that PCAST may have failed to take account of some relevant empirical studies. A statement issued by the Department of Justice (DOJ) on September 20, 2016 (the same day as the report’s release) opined that:

The report does not mention numerous published research studies which seem to meet PCAST’s criteria for appropriately designed studies providing support for foundational validity. That omission discredits the PCAST report as a thorough evaluation of scientific validity.

Given its respect for the FBI, PCAST undertook a further review of the scientific literature and invited a variety of stakeholders—including the DOJ—to identify any “published . . . appropriately designed

³ The size of the study (e.g., number of examiners and problems) affects the strength of conclusions that can be drawn (e.g., the upper bound on the error rate). The acceptable level of error rate depends on context.

studies” that had not been considered by PCAST and that established the validity and reliability of any of the forensic feature-comparison methods that the PCAST report found to lack such support. As noted below, DOJ ultimately concluded that it had no additional studies for PCAST to consider.

PCAST received written responses from 26 parties, including from Federal agencies, forensic-science and law-enforcement organizations, individual forensic-science practitioners, a testing service provider, and others in the US and abroad.⁴ Many of the responses are extensive, detailed and thoughtful, and they cover a wide range of topics; they provide valuable contributions for advancing the field. PCAST also held several in-person and telephonic meetings with individuals involved in forensic science and law enforcement. In addition, PCAST reviewed published statements from more than a dozen forensic-science, law-enforcement and other entities.⁵ PCAST is deeply grateful to all who took the time and effort to opine on this important topic.

In what follows, we focus on three key issues raised.

[Issue: Are empirical studies truly necessary?](#)

While forensic-science organizations agreed with the value of empirical tests of subjective forensic feature-comparison methods (that is, black-box tests), many suggested that the validity and reliability of such a method could be established *without* actually empirically testing the method in an appropriate setting. Notably, however, none of these respondents identified any *alternative* approach that could establish the validity and reliability of a subjective forensic feature-comparison method.

PCAST is grateful to these organizations because their thoughtful replies highlight the fundamental issue facing the forensic sciences: *the role of empirical evidence*. As noted in PCAST’s report, forensic scientists rightly point to several elements that provide critical foundations for their disciplines. However, there remains confusion as to whether these elements can suffice to establish the validity and degree of reliability of particular methods.

- (i) The forensic-science literature contains many papers describing variation among features. In some cases, the papers argue that patterns are “unique” (e.g., that no two fingerprints, shoes or DNA patterns are identical if one looks carefully enough). Such studies can provide a valuable *starting point* for a discipline, because they suggest that it may be worthwhile to attempt to develop reliable methods to identify the source of a sample based on feature comparison. However, such studies—no matter how extensive—can *never* establish the validity or degree of reliability of any particular method. Only empirical testing can do so.
- (ii) Forensic scientists rightly cite examiners’ experience and judgment as important elements in their disciplines. PCAST has great respect for the value of examiners’ experience and judgment: they are critical factors in ensuring that a scientifically valid and reliable method is practiced correctly. However, experience and judgment alone—no matter how great—can *never* establish the validity or degree of reliability of any particular method. Only empirical testing of the method can do so.⁶

⁴ www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_2016_additional_responses.pdf.

⁵ www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_2016_public_comments.pdf.

⁶ Some respondents, such as the Organization of Scientific Area Committees’ Friction Ridge Subcommittee, suggested that forensic science should be considered as analogous to medicine, in which physicians often treat patients on the basis of experience and judgment even in the absence of established empirical evidence. However, the analogy is inapt. Physicians act with a patient’s consent for the patient’s benefit. There is no legal requirement, analogous to the requirement imposed upon expert testimony in court by the Federal Rules of Evidence, that physician’s actions be based on “reliable principles and methods.” Physicians may rely on hunches; experts testifying in court about forensic feature-comparison methods may not.

- (iii) Forensic scientists cite the role of professional organizations, certification, accreditation, best-practices manuals, and training within their disciplines. PCAST recognizes that such practices play a critical role in any professional discipline. However, the existence of good professional practices alone—no matter how well crafted—can *never* establish the validity or degree of reliability of any particular method. Only empirical testing of the method can do so.

PCAST does not diminish in any way the important roles of prior research and other types of activities within forensic science and practice. Moreover, PCAST expresses great respect for the efforts of forensic practitioners, most of whom are devoted public servants. It is important to emphasize, however, contrary to views expressed by some respondents, that there is no “hierarchy” in which empirical evidence is simply the best way to establish validity and degree of reliability of a subjective feature-comparison method. In science, empirical testing is the only way to establish the validity and degree of reliability of such an empirical method.

Fortunately, empirical testing of empirical methods is feasible. There is no justification for accepting that a method is valid and reliable in the absence of appropriate empirical evidence.

[Issue: Importance of other kinds of studies](#)

In its response to PCAST’s call for further input, the Organization of Scientific Area Committees’ Friction Ridge Subcommittee (OSAC FRS), whose purview includes latent-print analysis, raised a very important issue:

While the OSAC FRS agrees with the need for black box studies to evaluate the overall validity of a particular method, the OSAC FRS is concerned this view could unintentionally stifle future research agendas aimed at dissecting the components of the black box in order to transition it from a subjective method to an objective method. If the PCAST maintains such an emphasis on black box studies as the *only* means of establishing validity, the forensic science community could be inundated with predominantly black box testing and potentially detract from progress in refining other foundational aspects of the method, such as those previously outlined by the OSAC FRS, in an effort to identify ways to emphasize objective methods over subjective methods (see www.nist.gov/topics/forensic-science/osac-research-development-needs.) Given the existing funding limitations, this will be especially problematic and the OSAC FRS is concerned other foundational research will thus be left incomplete.

PCAST applauds the work of the friction-ridge discipline, which has set an excellent example by undertaking both (i) path-breaking black-box studies to establish the validity and degree of reliability of latent-fingerprint analysis, and (ii) insightful “white-box” studies that shed light on how latent-print analysts carry out their examinations, including forthrightly identifying problems and needs for improvement. PCAST also applauds ongoing efforts to transform latent-print analysis from a subjective method to a fully objective method. In the long run, the development of objective methods is likely to increase the power, efficiency and accuracy of methods—and thus better serve the public.

In the case of subjective methods whose validity and degree of reliability have already been established by appropriate empirical studies (such as latent-print analysis), PCAST agrees that continued investment in black-box studies is likely to be less valuable than investments to develop fully objective methods. Indeed, PCAST’s report calls for substantial investment in such efforts.

The situation is different, however, for subjective methods whose validity and degree of reliability has not been established by appropriate empirical studies. If a discipline wishes to offer testimony based on a subjective method, it must first establish the method's validity and degree of reliability—which can only be done through empirical studies. However, as the OSAC FRS rightly notes, a discipline could follow an alternative path by abandoning testimony based on the subjective method and instead developing an objective method. Establishing the validity and degree of reliability of an objective method is often more straightforward. PCAST agrees that, in many cases, the latter path will make more sense.

[Issue: Completeness of PCAST's evaluation](#)

Finally, we considered the important question, raised by the DOJ in September, of whether PCAST had failed to consider “numerous published research studies which seem to meet PCAST's criteria for appropriately designed studies providing support for foundational validity.”

PCAST re-examined the five methods evaluated in its report for which the validity and degree of reliability had not been fully established. We considered the more than 400 papers cited by the 26 respondents; the vast majority had already been reviewed by PCAST in the course of the previous study. At the suggestion of John Butler of the National Institute of Standards and Technology (NIST), we also consulted INTERPOL's extensive summary of the forensic literature to identify additional potentially relevant papers.⁷ Although our inquiry was undertaken in response to the DOJ's concern, DOJ informed PCAST in late December that it had no additional studies for PCAST to consider.

Bitemark analysis

In its report, PCAST stated that it found no empirical studies whatsoever that establish the scientific validity or degree of reliability of bitemark analysis as currently practiced. To the contrary, it found considerable literature pointing to the unreliability of the method. None of the respondents identified any empirical studies that establish the validity or reliability of bitemark analysis. (One respondent noted a paper, which had already been reviewed by PCAST, that studied whether examiners agree when measuring features in dental casts but did not study bitemarks.) One respondent shared a recent paper by a distinguished group of biomedical scientists, forensic scientists, statisticians, pathologists, medical examiners, lawyers, and others, published in November 2016, that is highly critical of bitemark analysis and is consistent with PCAST's analysis.

Footwear analysis

In its report, PCAST considered feature-comparison methods for associating a shoeprint with a specific shoe based on randomly acquired characteristics (as opposed to with a class of shoes based on class characteristics). PCAST found no empirical studies whatsoever that establish the scientific validity or reliability of the method.

The President of the International Association for Identification (IAI), Harold Ruslander, responded to PCAST's request for further input. He kindly organized a very helpful telephonic meeting with IAI member Lesley Hammer. (Hammer has conducted some of the leading research in the field—including a 2013 paper, cited by PCAST, that studied whether footwear examiners reach similar conclusions when they are presented with evidence in which the identifying features have already been identified.)

⁷ The INTERPOL summaries list 4232 papers from 2010-2013 and 4891 papers from 2013-2016, sorted by discipline, see www.interpol.int/INTERPOL-expertise/Forensics/Forensic-Symposium.

Hammer confirmed that no empirical studies have been published to date that test the ability of examiners to reach correct conclusions about the source of shoeprints based on randomly acquired characteristics. Encouragingly, however, she noted that the first such empirical study is currently being undertaken at the West Virginia University. When completed and published, this study should provide the first actual empirical evidence concerning the validity of footwear examination. The types of samples and comparisons used in the study will define the bounds within which the method can be considered reliable.

Microscopic hair comparison

In its report, PCAST considered only those studies on microscopic hair comparison cited in a recent DOJ document as establishing the scientific validity and reliability of the method. PCAST found that none of these studies provided any meaningful evidence to establish the validity and degree of reliability of hair comparison as a forensic feature-comparison method. Moreover, a 2002 FBI study, by Houck and Budowle, showed that hair analysis had a stunningly high error rate in practice: Of hair samples that FBI examiners had found in the course of actual casework to be microscopically indistinguishable, 11% were found by subsequent DNA analysis to have come from different individuals.

PCAST received detailed responses from the Organization of Scientific Area Committees' Materials Subcommittee (OSAC MS) and from Sandra Koch, Fellow of the American Board of Criminalistics (Hairs and Fibers). These respondents urged PCAST not to underestimate the rich tradition of microscopic hair analysis. They emphasized that anthropologists have published many papers over the past century noting differences in average characteristics of hair among different ancestry groups, as well as variation among individuals. The studies also note intra-individual differences among hair from different sites on the head and across age.

While PCAST agrees that these empirical studies describing hair differences provide an encouraging starting point, we note that the studies do not address the validity and degree of reliability of hair comparison as a forensic feature-comparison method. What is needed are empirical studies to assess how often examiners incorrectly associate similar but distinct-source hairs (i.e., false-positive rate). Relevant to this issue, OSAC MS states: "Although we readily acknowledge that an error rate for microscopic hair comparison is not currently known, this should not be interpreted to suggest that the discipline is any less scientific." In fact, this is the central issue: the acknowledged lack of any empirical evidence about false-positive rates indeed means that, as a *forensic feature-comparison method*, hair comparison lacks a scientific foundation.

Based on these responses and its own further review of the literature beyond the studies mentioned in the DOJ document, PCAST concludes that there are no empirical studies that establish the scientific validity and estimate the reliability of hair comparison as a forensic feature-comparison method.

Firearms analysis

In its report, PCAST reviewed a substantial set of empirical studies that have been published over the past 15 years and discussed a representative subset in detail. We focused on the ability to associate ammunition not with a class of guns, but with a specific gun within the class.

The firearms discipline clearly recognizes the importance of empirical studies. However, most of these studies used flawed designs. As described in the PCAST report, "set-based" approaches can inflate examiners' performance by allowing them to take advantage of internal dependencies in the data. The

most extreme example is the “closed-set design”, in which the correct source of each questioned sample is always present; studies using the closed-set design have underestimated the false-positive and inconclusive rates by more than 100-fold. This striking discrepancy seriously undermines the validity of the results and underscores the need to test methods under appropriate conditions. Other set-based designs also involve internal dependencies that provide hints to examiners, although not to the same extent as closed-set designs.

To date, there has been only one appropriately designed black-box study: a 2014 study commissioned by the Defense Forensic Science Center (DFSC) and conducted by the Ames Laboratory, which reported an upper 95% confidence bound on the false-positive rate of 2.2%.⁸

Several respondents wrote to PCAST concerning firearms analysis. None cited additional appropriately designed black-box studies similar to the recent Ames Laboratory study. Stephen Bunch, a pioneer in empirical studies of firearms analysis, provided a thoughtful and detailed response. He agreed that set-based designs are problematic due to internal dependencies, yet suggested that certain set-based studies could still shed light on the method if properly analyzed. He focused on a 2003 study that he had co-authored, which used a set-based design and tested a small number of examiners (n=8) from the FBI Laboratory’s Firearms and Toolmarks Unit.⁹ Although the underlying data are not readily available, Bunch offered an estimate of the number of truly independent comparisons in the study and concluded that the 95% upper confidence bound on the false-positive rate in his study was 4.3% (vs. 2.2% for the Ames Laboratory black-box study).

The Organization of Scientific Area Committee’s Firearms and Toolmarks Subcommittee (OSAC FTS) took the more extreme position that all set-based designs are appropriate and that they reflect actual casework, because examiners often start their examinations by sorting sets of ammunition from a crime-scene. OSAC FTS’s argument is unconvincing because (i) it fails to recognize that the results from certain set-based designs are wildly inconsistent with those from appropriately designed black-box studies, and (ii) the key conclusions presented in court do not concern the ability to sort collections of ammunition (as tested by set-based designs) but rather the ability to accurately associate ammunition with a specific gun (as tested by appropriately designed black-box studies).

Courts deciding on the admissibility of firearms analysis should consider the following scientific issues:

- (i) There is only a single appropriate black-box study, employing a design that cannot provide hints to examiners. The upper confidence bound on the false-positive rate is equivalent to an error rate of 1 in 46.
- (ii) A number of older studies involve the seriously flawed closed-set design, which has dramatically underestimated the error rates. These studies do not provide useful information about the actual reliability of firearms analysis.
- (iii) There are several studies involving other kinds of set-based designs. These designs also involve internal dependencies that can provide hints to examiners, although not to the same extent that closed-set designs do. The large Miami-Dade study cited in the PCAST report and the small studies cited by Bunch fall into this category; these two studies have upper confidence bounds corresponding to error rates in the range of 1 in 20.

From a scientific standpoint, scientific validity should require at least two properly designed studies to ensure reproducibility. The issue for judges is whether one properly designed study, together with

⁸ PCAST also noted that some studies combine tests of both class characteristics and individual characteristics, but fail to distinguish between the results for these two very different questions.

⁹ PCAST did not select the paper for discussion in the report owing to its small size and set-based design, although it lists it.

ancillary evidence from imperfect studies, adequately satisfies the legal criteria for scientific validity. Whatever courts decide, it is essential that information about error rates is properly reported.

DNA analysis of complex mixtures

In its report, PCAST reviewed recent efforts to extend DNA analysis to samples containing complex mixtures. The challenge is that the DNA profiles resulting from such samples contain many alleles (depending on the number of contributors) that vary in height (depending on the ratios of the contributions), often overlap fully or partially (due to their “stutter patterns”), and may sometimes be missing (due to PCR dropout). Early efforts to interpret these profiles involved purely subjective and poorly defined methods, which were not subjected to empirical validation. Efforts then shifted to a quantitative method called combined probability of inclusion (CPI); however, this approach also proved seriously problematic.¹⁰

Recently, efforts have focused on an approach called probabilistic genotyping (PG), which uses mathematical models (involving a likelihood-ratio approach) and simulations to attempt to infer the likelihood that a given individual’s DNA is present in the sample. PCAST found that empirical testing of PG had largely been limited to a narrow range of parameters (number and ratios of contributors). We judged that the available literature supported the validity and reliability of PG for samples with three contributors where the person of interest comprises at least 20% of the sample. Beyond this approximate range (i.e. with a larger number of contributors or where the person of interest makes a lower than 20% contribution to the sample), however, there has been little empirical validation.¹¹

A recent controversy has highlighted issues with PG. In a prominent murder case in upstate New York, a judge ruled in late August (a few days before the approval of PCAST’s report) that testimony based on PG was inadmissible owing to insufficient validity testing.¹² Two PG software packages (STRMix and TrueAllele), from two competing firms, reached differing¹³ conclusions about whether a DNA sample in the case contained a tiny contribution (~1%) from the defendant. Disagreements between the firms have grown following the conclusion of the case.

PCAST convened a meeting with the developers of the two programs (John Buckleton and Mark Perlin), as well as John Butler from NIST, to discuss how best to establish the range in which a PG software program can be considered to be valid and reliable. Buckleton agreed that empirical testing of PG software with different kinds of mixtures was necessary and appropriate, whereas Perlin contended that empirical testing was unnecessary because it was mathematically impossible for the likelihood-ratio approach in his software to incorrectly implicate an individual. PCAST was unpersuaded by the latter argument. While likelihood ratios are a mathematically sound concept, their application requires

¹⁰ Just as the PCAST report was completed, a paper was published that proposed various rules for the use of CPI. See Bieber, F.R., Buckleton, J.S., Budowle, B., Butler, J.M., and M.D. Coble. “Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion.” *BMC Genetics*. bmcbgenet.biomedcentral.com/articles/10.1186/s12863-016-0429-7. While PCAST agreed that these rules are *necessary*, PCAST did not review whether these rules were sufficient to ensure reliability and took no position on this question.

¹¹ The few studies that have explored 4- or 5-person mixtures often involve mixtures that are derived from only a few sets of people (in some cases, only one). Because the nature of overlap among alleles is a key issue, it is critical to examine mixtures from various different sets of people. In addition, the studies involve few mixtures in which a sample is present at an extremely low ratio. By expanding these empirical studies, it should be possible to test validity and reliability across a broader range.

¹² See McKinley, J. “Judge Rejects DNA Test in Trial Over Garrett Phillips’s Murder.” *New York Times*, August 26, 2016, www.nytimes.com/2016/08/27/nyregion/judge-rejects-dna-test-in-trial-over-garrett-phillips-murder.html. The defendant was subsequently acquitted.

¹³ Document updated on January 17, 2017.

making a set of assumptions about DNA profiles that require empirical testing.¹⁴ Errors in the assumptions can lead to errors in the results. To establish validity with a range of parameters, it is thus important to undertake empirical testing with a *variety* of samples in the relevant range.¹⁵

PCAST received thoughtful input from several respondents. Notably, one response¹⁶ suggested that the relevant category for consideration should be expanded from “complex mixtures” (defined based on the number of contributors) to “complex samples” (defined to include also samples with low amounts of template, substantial degradation, or significant PCR inhibition, all of which will also complicate interpretation). We agree that this expansion could be useful.

The path forward is straightforward. The validity of specific PG software should be validated by testing a diverse collection of samples within well-defined ranges. The DNA analysis field contains excellent scientists who are capable of defining, executing, and analyzing such empirical studies.

When considering the admissibility of testimony about complex mixtures (or complex samples), judges should ascertain whether the published validation studies adequately address the nature of the sample being analyzed (e.g., DNA quantity and quality, number of contributors, and mixture proportion for the person of interest).

Conclusion

Forensic science is at a crossroads. There is growing recognition that the law requires that a forensic feature-comparison method be established as scientifically valid and reliable before it may be used in court and that this requirement can only be satisfied by actual empirical testing. Several forensic disciplines, such as latent-print analysis, have clearly demonstrated that actual empirical testing is feasible and can help drive improvement. A generation of forensic scientists appears ready and eager to embrace a new, empirical approach—including black-box studies, white-box studies, and technology development efforts to transform subjective methods into objective methods.

PCAST urges the forensic science community to build on its current forward momentum. PCAST is encouraged that NIST has already developed an approach, subject to availability of budget, for carrying out the functions proposed for that agency in our September report.

In addition, progress would be advanced by the creation of a cross-cutting Forensic Science Study Group—involving leading forensic and non-forensic scientists in equal measure and spanning a range of feature-comparison disciplines—to serve as a scientific forum to discuss, formulate and invite broad input on (i) empirical studies of validity and reliability and (ii) approaches for new technology development, including transforming subjective methods into objective methods. Such a forum would complement existing efforts focused on developing best practices and informing standards and might strengthen connections between forensic disciplines and other areas of science and technology. It might be organized by scientists in cooperation with one or more forensic and non-forensic science organizations—such as DFSC, NIST, IAI, and the American Association for the Advancement of Science.

¹⁴ Butler noted that one must make assumptions, for each locus, about the precise nature of reverse and forward stutter and about the probability of allelic dropout.

¹⁵ Butler noted that it is important to consider samples with different extents of allelic overlap among the contributors.

¹⁶ This response was provided by Keith Inman, Norah Rudin and Kirk Lohmueller.



Review Article

Finding the way forward for forensic science in the US—A commentary on the PCAST report

I.W. Evett^{*a}, C.E.H. Berger^b, J.S. Buckleton^{c,d}, C. Champod^e, G. Jackson^f^a Principal Forensic Services Ltd., 34 Southborough Road, Bickley, Bromley, Kent, BR1 2EB, United Kingdom^b Institute for Criminal Law and Criminology, Faculty of Law, Leiden University, PO Box 9520, 2300 RA Leiden, The Netherlands^c Environmental Science & Research Ltd, Private Bag 92021, Auckland 1142, New Zealand^d Department of Statistical Genetics, University of Washington, Box 357232 Seattle, WA 98195-7232, United States^e Ecole des Sciences Criminelles, Faculty of Law, Criminal Justice and Public Administration, Université de Lausanne, Batochime – quartier Sorge, CH-1015

Lausanne-Dorigny, Switzerland

^f Abertay University, Dundee, DD1 1HG, United Kingdom

ARTICLE INFO

Article history:

Received 16 March 2017

Received in revised form 30 April 2017

Accepted 18 June 2017

Available online 26 June 2017

Keywords:

Forensic inference

Evidence

Comparison methods

Probability

Likelihood ratio

ABSTRACT

A recent report by the US President's Council of Advisors on Science and Technology (PCAST), (2016) has made a number of recommendations for the future development of forensic science. Whereas we all agree that there is much need for change, we find that the PCAST report recommendations are founded on serious misunderstandings. We explain the traditional forensic paradigms of *match* and *identification* and the more recent foundation of the logical approach to evidence evaluation. This forms the groundwork for exposing many sources of confusion in the PCAST report. We explain how the notion of treating the scientist as a black box and the assignment of evidential weight through error rates is overly restrictive and misconceived. Our own view sees inferential logic, the development of calibrated knowledge and understanding of scientists as the core of the advance of the profession.

© 2017 Elsevier B.V. All rights reserved.

Contents

In Memoriam	17
1. Introduction	17
2. The logical approach	17
2.1. Framework of circumstances	17
2.2. Propositions	17
2.3. Probability of the observations	17
3. The match paradigm	18
4. The identification paradigm	18
5. Misconceptions, fallacies and confusions in the PCAST report	18
5.1. Confusion between the match and identification paradigms	18
5.2. Judgement	19
5.3. Subjective versus Objective	19
5.4. Transposed conditional	19
5.5. "Probable match"	20
5.6. Foundational validity and accuracy	20
5.7. The PCAST paradigm	21
5.8. The scientist as a "black box"	21
5.9. Black box studies	21
5.10. Governance	22

* Corresponding author.

E-mail address: ianevett@btinternet.com (I.W. Evett).

6.	Our view of the future	22
6.1.	Logical inference	22
6.2.	Calibration	22
6.3.	Knowledge and data	22
7.	Conclusion	23
	References	23

In Memoriam

This paper is dedicated to the memory of Bryan Found who did so much to advance the profession of forensic scientist through his work on calibrating and enhancing the performance of experts under controlled conditions. He will be sorely missed.

1. Introduction

This paper is written in response to a recent report on forensic science of the US President's Council of Advisors on Science and Technology (PCAST) [1]. There have already been several responses to the report from the forensic community [2–7] which have resulted in an addendum to the report [8]. Our main concern is that the report (and its addendum) fails to recognise the advances in the logic of forensic inference that have taken place over the last 50 years or so. This is a serious omission which has led PCAST to a narrowly-focused and unhelpful view of the future of forensic science.

The structure of our paper is as follows. In Section 2 we briefly outline our view of the requirements imposed by logic on the assessment of the probative value of evidence. This allows us to set up a framework against which we can contrast some of the suggestions of the report. In Sections 3 and 4 we briefly explain the notions of “match” and “identification” paradigms that have underpinned much of forensic inference over the last century or so. Section 5 will point out misconceptions, fallacies, sources of confusion and improper terminology in the PCAST report. Our contrasting view of the future path for forensic science follows in Section 6.

2. The logical approach

Much has been written over the past 40 years on inference in forensic science. The frequency of appearance of articles, papers and books on the topic has increased markedly in recent years. Practically all of this material is founded on a logical, probabilistic approach to the assessment of the probative value of scientific observations [9,10]. The PCAST report mentions this body of work only briefly and pays scant attention to its principles [11], which we list and explain briefly as follows.

2.1. Framework of circumstances

It is necessary to consider the evidence within a framework of circumstances.

A simple example will illustrate this. Imagine that a sample¹ has been obtained from a crime scene which yielded a DNA profile from which the genotype of the originator of the sample has been inferred. A suspect for the crime is known to have the same genotype. Because the alleles revealed by a DNA profile will be found in different proportions in different ethnic groups, it is relevant to the assessment of the probative value of this

correspondence of genotypes that a credible eyewitness of the crime said that the offender was of a particular ethnic appearance.

It follows that, when presenting an evaluation, the scientist should clearly state the framework of circumstances that are relevant to their assessment of the probative value of the observations, with a caveat that, if details of the circumstances change, the evaluation must be revisited.

2.2. Propositions

The probative value of the observations cannot be assessed unless two propositions are addressed.

In a criminal trial, these will represent what the scientist believes the prosecution may allege and a sensible alternative that represents the defence position.² In taking account of both sides of the argument, the scientist is able to assess the evidence in a balanced, justifiable way and display to the court an unbiased approach, irrespective of which side calls the witness.

Propositions may be formed at any of at least four levels in a hierarchy of propositions [12–14]. These levels are termed offence, activity, source and sub-source. We do not discuss these in any depth here. Most of the PCAST report appears to address questions at the source or sub-source level. Examples of these would be:

1. Sub-source: The DNA came from the person of interest (POI),³ or
2. Source: This fingerprint was made by the POI.

2.3. Probability of the observations

It is necessary for the scientist to consider the probability⁴ of the observations given the truth of each of the two propositions in turn.

The ratio of these two probabilities is widely known as the *likelihood ratio* (LR) and this is a measure of the weight of evidence that the observations provide in addressing the issue of which of the propositions is true. A likelihood ratio greater than one provides support for the truth of the prosecution proposition. A likelihood ratio less than one provides support for the truth of the defence proposition.

It cannot be sufficiently emphasized that it is the scientist's role to provide expert opinion on the probability of the *observations* given the proposition. The role of assigning a value to the probability of the *proposition* given the observations is that of the jury in a criminal trial. This probability will take account, not just of the scientific observations, but also of all of the other evidence presented at court.

² We recognise that the scientist, particularly at an early stage of proceedings, may not know the position that defence will take. It is common practice for the scientist to adopt what appears to be a reasonable proposition, given what is known of the circumstances—making it clear that this is provisional and subject to change at any time.

³ A source level DNA proposition would specify the nature of the recovered material, e.g. “the semen came from the POI”.

⁴ This could be a probability density, depending on the nature of the observations. But the principle remains unchanged.

¹ The term “sample” is used generically to describe what is available for forensic examination. The term is not used here to suggest any statistical sampling process.

3. The match paradigm

In most forensic comparisons, one of the items will be from a known origin (such as: a reference sample for DNA profiling from a particular individual; a pair of shoes from a suspect; a set of control fragments of glass from a broken window). The other will be from an unknown, or disputed origin (such as: DNA recovered from a crime scene; a footwear mark from the point of entry at a burglary; or a few small fragments of glass recovered from the clothing of a suspect). It is convenient to refer to these as the *reference* and *questioned* samples, respectively. The matter of interest to the court relates to the origin of the questioned sample. This question will be addressed scientifically by carrying out observations on both samples. These observations may be purely qualitative: such as, for example, the shapes of the loops of letters such as “y” and “g” in a passage of handwriting. They may be quantitative and discrete, such as the alleles in a DNA STR profile. Or they may be quantitative and continuous, such as the refractive index of glass fragments. The match paradigm calls for a judgement, by the scientist, as to whether or not the two sets of observations agree within the range of what would be expected if the questioned sample had come from the same origin as the reference sample. The basis for that judgement may, in the case of quantitative observations, be based on a set of pre-determined criteria; but where the observations are qualitative such criteria may be vague or purely judgemental.

If the two sets of observations are considered to be outside the range of what may have been expected if the two samples had come from the same source then the result may be reported as a “non-match”. Depending on the nature of the observations, this provides the basis for a strong implication that the questioned and reference samples came from different sources. In many instances this conclusion will be non-controversial in the sense that prosecution and defence will be content to accept it.

However, when the result of the comparison is a “match” it does not logically follow that the two samples do share the same source or even that they are likely to be from the same source. It is possible that the two samples came from two different sources that, by coincidence, have similar properties. Throughout the history of forensic science there has been the notion – often imperfectly expressed – that the smaller the probability of such a coincidence, the greater the evidential value to be associated with the observed match. In DNA profiling, for example, we encounter the notion of a “match probability”. The implication of this approach is that the jury should assign an evidential weight that is related to the inverse of the match probability.

The logical approach has done much to clarify the rather woolly inference that historically has been associated with the match paradigm but it has also demonstrated the considerable advantages of the single stage approach implied by the assignment of weight through the calculation of the likelihood ratio, over the rather clumsy and inefficient two-stage approach implied by the match paradigm. This has already been pointed out by Morrison et al. [4].

4. The identification paradigm

Historically, fingerprint comparison was seen to be the gold standard by which the power of any other forensic technique could be judged. The paradigm here was the notion of “identification”⁵ or

“individualization” (the terms are used synonymously here). Provided that sufficient corresponding detail was observed, the outcome of a comparison between a fingerprint of questioned origin and a print taken from a known person would be reported as a categorical opinion: the two were definitely made by the same person.

So, the match and identification paradigms are related with the difference that in the latter the scientist is allowed to state that the match probability is so infinitesimally small that it is reasonable to conclude that the two items came from the same source. Historically, many examiners would have claimed that the source was established with certainty to the exclusion of all others.

The identification paradigm went largely unchallenged for many years until later in the 20th century when its logical basis was questioned (see, for example, [16] or more recently [17,18]) and also when, in a number of high profile cases, misidentifications with serious consequences were exposed.

An example of the paradigm is given in box 6, p. 137 of the PCAST report (DOJ proposed uniform language) (emphasis added).

The examiner may state that it is his/her opinion that the shoe/tire *is the source of the impression* because there is sufficient quality and quantity of corresponding features such that the examiner would not expect to find that same combination of features repeated in another source. This is the highest degree of association between a questioned impression and a known source.

The PCAST report rightly indicates that the conclusions conveying “100 percent certainty” or “zero or negligible error rates” are not scientifically defensible. Such conclusions tend to overestimate the weight to be assigned to the forensic observations.

5. Misconceptions, fallacies and confusions in the PCAST report

The most serious weakness in the PCAST report is their flawed paradigm for forensic evaluation. Unfortunately, the report contains more misconceptions, fallacies, confusions and improper wording. In this section we will discuss the main problems with the report.

5.1. Confusion between the match and identification paradigms

This is the first source of confusion in the report. For example, from p. 90 of the report (emphasis added):

An FBI examiner concluded with “100 percent certainty” that the fingerprint *matched* Brandon Mayfield . . . even though Spanish authorities were unable to confirm the *identification*.

On p. 48 we find (emphasis added):

To meet the scientific criteria of foundational validity, two key elements are required:

(1) a reproducible and consistent procedure for (a) identifying features within evidence samples; (b) comparing the features in two samples; and (c) determining based on the similarity between the features in two samples, whether the samples should be declared to be a proposed *identification* (“*matching rule*”).

We have seen that declaring a match and declaring an identification are not the same thing. Declaring a match implies nothing about evidential weight whereas declaring an identification implies evidential weight amounting to complete certainty.

The PCAST report proposes an approach that is fusion of the match and identification paradigms. See, from p. 45/46:

⁵ Kirk [15] defined the term identification as only placing an object in a restricted class. The criminalist would, for example, identify a particular mark as a fingerprint. Individualization was defined by Kirk as establishing which finger left the mark. An opinion of the kind “this latent mark was made by the finger which made this reference print” is an individualization.

Because the term “match” is likely to imply an inappropriately high probative value, a more neutral term should be used for an examiner’s belief that two samples came from the same source. We suggest the term “proposed identification” to appropriately convey the examiner’s conclusion, along with the possibility that it might be wrong. We will use this term throughout the report.

If a scientist says that the questioned and reference samples match, the immediate inference to be drawn from this (as we have explained) is that they might have come from the same source but it is also true that they might not have come from the same source. These two statements make no implication with regard to evidential weight. Weight only comes from the second stage of the paradigm which entails coming up with some impression of rarity. The identification paradigm, on the other hand, is different in that implies a statement of certainty: the two samples certainly came from the same source.

The PCAST paradigm requires that the scientist should make a categorical statement (an identification) that cannot be justified on logical grounds as we have already explained. Most scientists would be comfortable with the notion of observing that two samples *matched* but would, rightly, refuse to take the logically unsupportable step of inferring that this observation amounts to an *identification*.

5.2. Judgement

The report emphasises the value of empirical data (emphasis added):

The frequency with which a particular pattern or set of features will be observed in different samples, which is an essential element in drawing conclusions, *is not a matter of ‘judgment’*. It is an empirical matter *for which only empirical evidence is relevant*. ([1], p. 6)

This denial of the importance of judgement betrays a poor understanding of the nature of forensic science. We offer a simple example.

Mr POI is the suspect for a crime who was arrested at time T in location Z . Some questioned material has been found on the clothing of Mr POI which is to be compared with reference material taken from the crime scene. Denote the observations on the two samples by y and x respectively. Whichever paradigm we follow, we are interested in the probability of finding material with observations y on the clothing of Mr POI if he had nothing to do with the crime. Ideally, of course, we would like a survey carried out near to time T and in the general region of Z and of people of a socio-economic group Q that would include Mr POI. But this is, of course unrealistic. What we do have is a survey of materials on clothing carried out at some earlier time T' and at another location Z' and of a slightly different socio-economic group Q' . Who is to make a judgement on the relevance of this survey data to the case at hand? We would argue that this is where the knowledge and understanding of the forensic scientist is of crucial importance.

The reality is, of course, that the perfect database never exists. The council is wrong: it is most certainly *not* the case that “only empirical evidence” is relevant. Without downplaying the importance of data collections, they can only inform judgement—it is judgement that is paramount and informed judgement is founded in reliable knowledge.

5.3. Subjective versus Objective

PCAST give their definition of the distinction between “objectivity” and “subjectivity” p. 5—footnote 3.

Feature-comparison methods may be classified as either objective or subjective. By objective feature-comparison methods, we mean methods consisting of procedures that are each defined with enough standardized and quantifiable detail that they can be performed by either an automated system or human examiners exercising little or no judgment. By subjective methods, we mean methods including key procedures that involve significant human judgment . . .

What is suggested is that many of the decisions be moved from the examiner to the procedure and/or software. The procedure or software will have been written by one or more people and the decisions about what models are used or how decisions are made are now enshrined in paper or code. Hence all the subjective judgements are now made by this person or group of people via the paper or code. Whereas this approach could be viewed as repeatable and reproducible, the objectivity is illusory.

In the US environment, subjectivity has been associated with bias and sloppy thinking, and objectivity with an absence of bias and rigorous thinking. It is worthwhile examining whence the fear of subjectivity arises. There is considerable proof that humans are susceptible to quite a number of cognitive effects many of which can affect judgement. We suspect that the fear is that these effects bias the decisions in ways that are detrimental to justice. Hence, it is bias arising from cognitive effects that is the enemy, not subjectivity.

If we return to the concept of enforced precision, we could assume that trials could be conducted on such a system and that the outputs could be calibrated. Such a system could be of low susceptibility to bias arising from cognitive effects. We suspect that these are the goals sought by PCAST. We certainly could support calibrating subjective judgements but we see little value in pretending that writing them down or coding them makes them objective.

5.4. Transposed conditional

We are concerned by the report’s poor use of the notion of probability. In particular we note in the report many instances where the fallacy of the transposed conditional either occurs explicitly or is implied. We have seen that the logic of forensic inference directs us to assign a value to the probability of the observations given the truth of a proposition. The probability of the truth of a proposition is for the jury *not* the scientist. Confusion between these two different probabilities has been called the “prosecutor’s fallacy” [19]. We prefer the term *transposed conditional* because, in our experience, the fallacy is regularly committed by prosecutors, defence attorneys, the judiciary and the media alike.

The fallacy is widespread, even though it can be grounds for a retrial if given in testimony by an expert witness. The document [20] that attempts to explain DNA statistics to defence attorneys in the US describes – incorrectly – a likelihood ratio for a mixture profile as:

4.73 quadrillion times more likely⁶ to have originated from [suspect] and [victim/complainant] than from an unknown individual in the U.S. Caucasian population and [victim/complainant].” ([20], p. 52)

⁶ We are fully aware of the distinction made in statistical theory between “likelihood” and “probability”. We believe that attempting to explain that distinction in this paper would cause more confusion than the worth of it. It is our experience that in courts of law the two terms are taken to be synonymous.

This is a classic example of the transposed conditional. It is a transposition of the likelihood ratio, which would be more correctly presented as follows:

The DNA profile is 4.73 quadrillion times more likely to be obtained if the DNA had originated from the suspect and the victim/complainant rather than if it had originated from an unknown individual in the U.S. Caucasian population and the victim/complainant.

The contrast between these two statements, though apparently subtle, is profound. The first is an expression of the probability (or odds) that a particular proposition is true—this, we have seen, is the probability that the jury must address, not the scientist.⁷ The second considers the probability of the *observations*, given the truth of one proposition then the other, which is the appropriate domain for the expertise of the scientist. It is important to realise that the first statement is not a simple rephrasing of the second statement. Whereas the second may be a valid representation of the scientist's evaluation in a given case, the first most definitely cannot be.

Consider the following quote from the first paragraph on footwear methodology in the PCAST report ([1], p. 114):

Footwear analysis is a process that typically involves comparing a known object, such as a shoe, to a complete or partial impression found at a crime scene, to assess whether the object is likely to be the source of the impression.

This is wrong. We state again that it is not for the scientist to present a probability for the truth of the proposition that the object was the source of the impression. The scientist addresses the probability of the outcome of the comparison *if* the object were the source of the impression: this probability forms the numerator of the likelihood ratio. Just as important, of course, is the probability of the outcome of the comparison *if* some other object were the source of the impression. The latter forms the denominator of the likelihood ratio. It is the two probabilities, taken together, that determine the evidential weight in relation to the two propositions of interest to the court.

The PCAST report sentence clearly states that the objective of the footwear analysis is to present a probability for the proposition given the observations, and not for the observations given the proposition. This is clearly a transposition of the conditional.

Similarly, the scientist is not in a position to consider the probability addressed in the following ([1], p. 65 and repeated on p. 146):

... determining, based on the similarity between the features in two sets of features, whether the samples should be declared to be likely to come from the same source . . .

We have seen that it is not for the scientist to consider the probability that the samples came from the same source given the observation of a “match”. It is another example of the fallacy of the transposed conditional.

This confusion is systematic in the original report and we note that it continues into the addendum ([8], p. 1) (emphasis added): These methods seek to determine whether a questioned sample *is likely to come* from a known source based on shared features in certain types of evidence.

We have seen that this is most certainly *not* what a feature-comparison should aspire to. It is not the role of the forensic

scientist to offer a probability for the proposition that a questioned sample came from a given source since this would require the scientist to take account of all of the non-scientific information which properly lies within the domain of the jury.

The need for precision of language when presenting probabilities is exemplified by two quotations from the report. First, from p. 8 when talking about the interpretation of a DNA profile:

Could a suspect's DNA profile be present within the mixture profile? And, what is the probability that such an observation might occur by chance?

As we read it, this second sentence can be taken to mean:

What is the probability that such an observation would be made if the suspect's DNA were not present in the mixture?

Within the logical paradigm, this is a legitimate question to ask—it is the probability of the observations given that one of the propositions were true.

However, later in the report we find (p. 52):

the random match probability—that is, the probability that the match occurred by chance”.

There is an economy of phrasing here that obscures meaning and the reader could be forgiven for believing that the question implied by the second phrase is:

What is the probability that the two samples had come from different sources and matched by chance?

This is a probability of a proposition (the two samples came from different sources) given the observation (a match) and would imply a transposed conditional. We are aware that the council may respond that this is not at all what they meant—to which we would respond that the council should have been far more careful in its phraseology.

5.5. “Probable match”

In giving their definition of the distinction between “objectivity” and “subjectivity” p. 5—see footnote 3 the report states:

how to determine whether the features are sufficiently similar to be called a probable match.

The council do not say what they mean by a “probable match” but it seems to us that it is another example of confusion between the match and identification paradigms. Following the match paradigm there is no such thing as a probable match—the two samples either match or they do not.

5.6. Foundational validity and accuracy

The report distinguishes two types of scientific validity: “foundational validity” and “validity as applied”. We confine ourselves to the first of these (p. 4):

Foundational validity for a forensic-science method requires that it be shown based on empirical studies to be *repeatable, reproducible, and accurate*, at levels that have been measured and are appropriate to the intended application. Foundational validity, then, means that a method can, *in principle*, be reliable.

Repeatability refers to the ability of the same operator with the same equipment to obtain the same (or closely similar) results when repeating analysis of the same material. Reproducibility refers to the ability of the equipment to obtain the same (or closely similar) results with different operators. As such, both are

⁷ In Bayesian terms, the first statement is one of posterior odds. This can be derived from the second statement either by assigning prior odds of one (which would be highly prejudicial in most criminal trials) or by making the mistake of transposing the conditional. Neither is acceptable behaviour for a scientist.

expressions of precision, which is how close each measurement or result is to the others.

Accuracy is a measure of how close one or a set of measurements is to the true answer. This has an obvious meaning when we know or could know the true answer. We could imagine some measurement such as the weight of an object where that object has been weighed by some very advanced technique and we can accept that as the “true” weight. We wish then to consider the accuracy of some other, perhaps cheaper, technique. We could assess the accuracy of this second technique by using it to weigh the object multiple times and observing the deviation of the results from the “true” weight of the object.

For some questions in forensic science, such as “How much heroin is in this seized sample?” or “How much ethanol is in this blood sample?”, the notion of the accuracy of an applied analytical technique is relevant because it is possible to assess a technique’s accuracy using trials with known quantities of heroin or ethanol. However, when it comes to answering a question such as “What is the probability that there would have been a match with a suspect’s shoe if it did not make the mark at the scene of crime?”, then there is no sense in which there is a “true answer”. The values that experts assign for such probabilities will vary depending on the specific knowledge of the experts and the nature of any databases that experts may use to inform their probabilities.

We could use a weather forecaster as an illustration. If she says that there is a 0.8 probability of a sunny day tomorrow, there can be no sense in which this is a “true” statement. Equally, if tomorrow brings rain, she is not “wrong” in any sense. Nor is she “inaccurate”. A probabilistic statement of this nature may be unhelpful or misleading, in the sense that it may lead us to make a poor decision, but it cannot be either true or false.

Once we abandon the idea of a true answer for probabilities, we are left with the difficult question of what we mean by accuracy. We suggest that the report does a disservice to the important task of calibrating probabilities by a simplistic allusion to accuracy.

The PCAST report says (p. 46):

Without appropriate estimates of accuracy, an examiner’s statement that two samples are similar – or even indistinguishable – is scientifically meaningless; it has no probative value, and considerable potential for prejudicial impact. Nothing – not training, personal experience nor professional practices – can substitute for adequate empirical demonstration of accuracy.

We have seen that the report is wrong here—it is not a matter of “accuracy” but of evidential weight.

5.7. The PCAST paradigm

The PCAST report proposes an approach that is fusion of the match and identification paradigms. See, from p. 45/46:

Because the term “match” is likely to imply an inappropriately high probative value, a more neutral term should be used for an examiner’s belief that two samples came from the same source. We suggest the term “proposed identification” to appropriately convey the examiner’s conclusion, along with the possibility that it might be wrong. We will use this term throughout the report.

First, we have seen that the term “match”, if used properly, makes no implication of probative value: it implies that the two samples might have come from the same source but also might have come from different sources. This is evidentially neutral. Second, we have seen that there is no place for the “examiner’s

belief that two samples came from the same source”: it is not for the scientist to assign a probability to the proposition that the two samples came from the same source.

Next we must consider what the council understand the phrase “proposed identification” to mean. Do they mean that, because it is an identification, it is a categorical opinion? Note that the qualifier “proposed” does not make the identification less than categorical – if it were probabilistic it could not be “wrong”.⁸ If it is not probabilistic then the scientist is to provide a categorical opinion while telling the court that he/she might be wrong! It is difficult to believe that any professional forensic scientist would be happy to be put in this position.

5.8. The scientist as a “black box”

On page 49 we find:

For subjective methods, procedures must still be carefully defined—but they involve substantial human judgment. For example, different examiners may recognize or focus on different features, may attach different importance to the same features, and may have different criteria for declaring proposed identifications. Because the procedures for feature identification, the matching rule, and frequency determinations about features are not objectively specified, the overall procedure must be treated as a kind of “black box” inside the examiner’s head.

The report justifiably emphasises weaknesses of qualitative opinions. The intuitive “black box” view of the scientist will certainly have been true in many instances in the past and, indeed, in certain quarters in the present day. But for us the solution is emphatically not to continue to treat this as an acceptable state of affairs for the future. The PCAST view appears to be “it’s a black box, so let’s treat it like a black box”. Our approach has been, and will continue, to break down intuitive mental barriers by expanding transparency, knowledge and understanding. We do not see the future forensic scientist as an *ipse dixit* machine—whatever the opinion, we expect the scientist to be able to explain it in whatever detail is necessary for the jury to comprehend the mental processes that led to it.

5.9. Black box studies

That the council intend the proposed identification to be categorical is clarified in the following from page 49 (emphasis added):

In black-box studies, many examiners are presented with many independent comparison problems – typically, involving “questioned” samples and one or more “known” samples – and asked to declare whether the questioned samples came from the same source as one of the known samples.⁹ The researchers then determine how often examiners reach erroneous conclusions.

PCAST proposes that the error rates from such experiments would be used to assign evidential value at court.

We are strongly against the notion that the scientist should be forced into the position of giving categorical opinions in this way. Whereas, we are strongly in favour of the notion of calibrating the

⁸ Though, of course, it would be logically incorrect because it would imply a transposed conditional.

⁹ In footnote 111 the report says: “Answers may be expressed in such terms as “match/no match/inconclusive” or “identification/exclusion/inconclusive”. This strengthens our belief that the council see match and identification as interchangeable”.

opinions of forensic scientists under controlled conditions we see those opinions expressed in terms of statements of evidential weight. We return to the subject of calibration later.

5.10. Governance

PCAST suggests that forensic science should be governed by those, such as metrologists, from outside the profession. This speaks to the view, reinforced by a very selective reference list, that the forensic science discipline is not to be trusted with developing procedures, testing them, and self-governance. We do not reject input from outside the profession: we welcome it. But our own observations are that those outside may be engaged to different extents, varying from a passing interest to years of study. They may be unduly influenced by headlines in newspapers highlighting or exaggerating deficiencies. On occasion, these same commentators from outside the profession may not recognise the limitations in their own knowledge base where it concerns specifically forensic aspects, may be reticent to consult subject matter experts from amongst practising scientists and may give well-intentioned, but erroneous, advice [1,21].

6. Our view of the future

6.1. Logical inference

The recommendations of the PCAST report are founded on a conflation of two classical forensic paradigms: match and identification. These paradigms are as old as forensic science but their inadequacies and illogicalities have been comprehensively exposed over the last 50 years or so. All of us maintain, and have done so in our writings, that the future of forensic science should be founded first on the notion of logical inference and second on the notion of calibrated knowledge. The former leads to a framework of principles (which have been adopted by ENFSI) and we are disappointed that PCAST has apparently chosen to ignore, or at most pay lip service to, this fundamental change. The second is a deeper and far richer concept than the profoundly limited notion of false-positive and false-negative error rates: this is the notion of *calibration*.

6.2. Calibration

We are most definitely in favour of the studying of expert opinion under controlled circumstances, see for example Evett [22] but proficiency testing is far more than the counting of errors. The PCAST black-box approach calls for a categorical opinion that is recorded as right or wrong but we have seen that forensic interpretation is far richer and more informative than simple yes/no answers. In a source level proficiency test we expect the participants to respond with a statement of evidential weight in relation to one of two clearly stated propositions. Support thus expressed for a proposition that is, in fact, false is undesirable because it is misleading—not “wrong”. Obviously, the desirable outcome of the proficiency test is a small value for the expected weight of evidence in relation to a false proposition. But whatever the outcome, the study must be seen as a learning exercise for all participants: the pool of knowledge has grown. The notion of an error rate to be presented to courts is misconceived because it fails to recognise that the science moves on as a result of proficiency tests. The work led by Found and Rogers [23] has shown how the profession of handwriting comparison in Australia and New Zealand has grown in stature because of the culture of advancing knowledge through repeated study under controlled conditions. To repeat then, our vision is not of the black-box/error rate but of continuous development through calibration and feedback of opinions.

A striking example of forensic calibration is the evolution of fingerprints evidence from the identification paradigm to the logical paradigm via mathematical modelling [24,25]. Instead of the categorical identification, we have a mathematical approach that leads to a likelihood ratio. The validation of such approaches is founded on two desiderata: we require large likelihood ratios in cases in which the prosecution proposition is true; and small likelihood ratios in cases in which the defence proposition is true. Investigation of performance in relation to these two desiderata is undertaken by considering two sets of comparisons: one set in which it is known that the two samples came from the same source; and one set in which it is known that the two samples came from different sources. There have been major advances over recent years in how the likelihood ratio distributions from such experiments may be compared and evaluated (Ramos [26], Brümmer [27] see also Robertson et al. [28] for a layman’s introduction to calibration). The elegance and performance of such methods far transcends the crude PCAST notion of “false-positive” and “false-negative” error rates.

6.3. Knowledge and data

The PCAST report focuses on “feature-comparison” methods and, as we have explained, this has meant that it is concerned with inference relating to source-level propositions. At this level, the report sees data as the sole means for assigning probabilities. An important part of the role of the forensic scientist is concerned with inference with regard to activity-level propositions. Consider, for example, a question of the form “what is the probability of finding this number of fragments of glass on Mr POI’s jacket if he is the person who smashed the window at the crime scene?” The answer is heavily dependent on circumstantial information (how large is the window? where was the person who smashed the window standing? was any implement used? how much time elapsed between the breaking of the window and the seizure of the jacket from Mr POI? etc.) and the variation in this between cases is vast. There is no single database to inform such probabilities. The scientist will, it is hoped, be thoroughly familiar with all of the published literature on glass transfer in crime cases [29] and may, if resources permit, carry out experiments that reproduce the current case circumstances. The knowledge and judgement of other scientists who have encountered similar questions is also relevant. We agree with PCAST that length of experience is not a measure of reliability of scientific opinion: the foundation is reliable knowledge. Too little effort has been devoted within the forensic sphere thus far to the harnessing of knowledge through knowledge based systems but see [29] for examples of how such a system was created for glass evidence interpretation.

We do not deny the importance of data collections but the view that data may replace judgement is misconceived. A data collection should be used to inform reliable knowledge—not replace it.

We have explained that our view of the scientist is the antithesis of the PCAST “black box” automaton. Although there is a need for data, PCAST are mistaken in seeing it as the be-all and end-all: qualitative judgement will always be at the centre of forensic science evidence evaluation. We reject the PCAST vision of the scientist who gives a categorical opinion and a statement about the probability that the opinion is wrong. We see the model scientist as deeply knowledgeable about her domain of expertise and able to rationalise the opinion in terms that the jury will understand. The principles have been expressed elsewhere [11] as balance, logic, robustness and transparency. There is no place for the black box. We agree that the scientist should be able to provide the court with evidence of performance under controlled conditions. Found and Rogers [23] have provided a model for handwriting comparison

and we see such approaches as extending into other areas: the emphasis is on calibration of probabilistic assessments.

7. Conclusion

The 44th US president's request was "to consider whether there are additional steps that could usefully be taken on the scientific side to strengthen the forensic-science disciplines and ensure the validity of forensic evidence used in the Nation's legal system" ([1], p. 1). We suggest that the report has very little emphasis on positive steps and does much to reinforce poor thinking and terminology.

Our own view of the future of forensic science is based on the principle that forensic inference should be founded on a logical framework for reasoning in the face of uncertainty. That framework is provided by probability theory coupled with the recognition that probability is necessarily subjective and conditioned by knowledge and judgement. It follows that our view of the forensic scientist is a knowledgeable, logical and reasonable person. Whereas data collections are valuable they should be viewed within the context of reliable knowledge. The overarching paradigm of reliable knowledge should be founded on the notion of knowledge management, including comprehensive systems for the calibration of expert opinion.

References

- [1] President's Council of Advisors on Science and Technology, Report to the president Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, Washington DC, 2016. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.
- [2] Federal Bureau of Investigation—FBI, Comments on: President's Council of Advisors on Science and Technology Report to the President on Forensic Science in Federal Criminal Courts: Ensuring Scientific Validity of Pattern Comparison Methods. September 20, 2016. www.fbi.gov/file-repository/fbi-pcast-response.pdf/view.
- [3] National District Attorneys Association—NDAA, Report Entitled Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. November 16, 2016. <http://www.ciclt.net/ul/ndaajustice/PCAST/NDAAPCASTResponseFINAL.pdf>.
- [4] G.S. Morrison, D.H. Kaye, D.J. Balding, D. Taylor, P. Dawid, C.G.G. Aitken, S. Gittelson, G. Zadora, B. Robertson, S. Willis, S. Pope, M. Neil, K.A. Martire, A. Hepler, R.D. Gill, A. Jamieson, J. de Zoete, R.B. Ostrum, A. Caliebe, A comment on the PCAST report: skip the match/non-match stage, *Forensic Sci. Int.* 272 (2017) e7–e9, doi:<http://dx.doi.org/10.1016/j.forsciint.2016.10.018>.
- [5] Association of Firearm and Tool Mark Examiners AFTE, Response to PCAST Report on Forensic Science. October 31, 2016. https://afte.org/uploads/documents/AFTE_PCAST_Response.pdf.
- [6] Bureau of Alcohol Tobacco Firearms and Explosives—ATF, ATF Response to the President's Council of Advisors on Science and Technology Report. September 21, 2016. https://www.theiai.org/president/20160921_ATF_PCAST_Response.pdf.
- [7] The International Association for Identification (IAI), IAI Response to the President's Council of Advisors on Science and Technology Report, 2016. https://www.theiai.org/president/IAI_PCAST_Response.pdf.
- [8] President's Council of Advisors on Science and Technology, An addendum to the PCAST report on forensic science in criminal courts, Washington DC, 2017. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_addendum_finalv2.pdf.
- [9] C.G.G. Aitken, F. Taroni, *Statistics and the Evaluation of Evidence for Forensic Scientists*, 2nd ed, John Wiley & Sons Ltd., Chichester, 2004.
- [10] C. Aitken, P. Roberts, G. Jackson, *Fundamentals of Probability and Statistical Evidence in Criminal Proceedings*, London, 2011. <http://www.rss.org.uk/uploadedfiles/userfiles/files/Aitken-Roberts-Jackson-Practitioner-Guide-1-WEB.pdf>.
- [11] Expressing evaluative opinions: a position statement, *Sci. Justice* 51 (1) (2011) 1–2, doi:<http://dx.doi.org/10.1016/j.scijus.2011.01.002>.
- [12] R. Cook, I.W. Evett, G. Jackson, P.J. Jones, J.A. Lambert, A model for case assessment and interpretation, *Sci. Justice* 38 (3) (1998) 151–156, doi:[http://dx.doi.org/10.1016/S1355-0306\(98\)72099-4](http://dx.doi.org/10.1016/S1355-0306(98)72099-4).
- [13] R. Cook, I.W. Evett, G. Jackson, P.J. Jones, J.A. Lambert, A hierarchy of propositions: deciding which level to address in casework, *Sci. Justice* 38 (4) (1998) 231–240, doi:[http://dx.doi.org/10.1016/S1355-0306\(98\)72117-3](http://dx.doi.org/10.1016/S1355-0306(98)72117-3).
- [14] R. Cook, I.W. Evett, G. Jackson, P.J. Jones, J.A. Lambert, Case pre-assessment and review in a two-way transfer case, *Sci. Justice* 39 (2) (1999) 103–111, doi:[http://dx.doi.org/10.1016/S1355-0306\(99\)72028-9](http://dx.doi.org/10.1016/S1355-0306(99)72028-9).
- [15] P.L. Kirk, The ontogeny of criminalistics, *J. Crim. Law Criminol. Police Sci.* 54 (1963) 235–238.
- [16] D.A. Stoney, What made us ever think we could individualize using statistics, *J. Forensic Sci. Soc.* 31 (2) (1991) 197–199, doi:[http://dx.doi.org/10.1016/S0015-7368\(91\)73138-1](http://dx.doi.org/10.1016/S0015-7368(91)73138-1).
- [17] A. Biedermann, S. Bozza, F. Taroni, Decision theoretic properties of forensic identification: underlying logic and argumentative implications, *Forensic Sci. Int.* 177 (2–3) (2008) 120–132, doi:<http://dx.doi.org/10.1016/j.forsciint.2007.11.008>.
- [18] A. Biedermann, S. Bozza, F. Taroni, The decisionalization of individualization, *Forensic Sci. Int.* 266 (2016) 29–38, doi:<http://dx.doi.org/10.1016/j.forsciint.2016.04.029>.
- [19] W.C. Thompson, E.L. Schumann, Interpretation of statistical evidence in criminal trials: the prosecutor's fallacy and the defence attorney's fallacy, *Law Hum. Behav.* 11 (3) (1987) 167–187, doi:<http://dx.doi.org/10.1007/BF01044641>.
- [20] E.H. Holder, M.L. Leary, J.H. Laub, DNA for the Defense Bar, U.S. Department of Justice Office of Justice Programs, Washington, DC, 2012.
- [21] National Research Council - Committee on DNA Technology in Forensic Science, *DNA Technology in Forensic Science*, National Academy Press, Washington, D.C, 1992.
- [22] I. Evett, The logical foundations of forensic science: towards reliable knowledge, *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370 (1674) (2015), doi:<http://dx.doi.org/10.1098/rstb.2014.0263>.
- [23] B. Found, D. Rogers, The initial profiling trial of a program to characterize forensic handwriting examiners' skill, *J. Am. Society of Questioned Document Examiners* 6 (2) (2003) 72–81.
- [24] C. Champod, C.J. Lennard, P.A. Margot, M. Stoilovic, *Fingerprints and other Ridge Skin Impressions*, CRC Press, Boca Raton, 2016.
- [25] C. Neumann, I.W. Evett, J. Skerrett, Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm, *J. Roy. Stat. Soc. Ser. A. (Stat. Soc.)* 175 (Part 2) (2012).
- [26] D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, C. Aitken, Information-theoretical assessment of the performance of likelihood ratio computation methods, *J. Forensic Sci.* 58 (6) (2013) 1503–1518, doi:<http://dx.doi.org/10.1111/1556-4029.12233>.
- [27] N. Brümmer, J. du Preez, Application-independent evaluation of speaker detection, *Comput. Speech Language* 20 (2006) 230–275, doi:<http://dx.doi.org/10.1016/j.csl.2005.08.001>.
- [28] G.A. Robertson, C.E.H. Vignaux, *Interpreting Evidence—Evaluating Forensic Science in the Courtroom*, 2nd ed., John Wiley & Sons, Ltd., Chichester, 2016.
- [29] J.M. Curran, T.N. Hicks, J.S. Buckleton, *Forensic Interpretation of Glass Evidence*, CRC Press LLC, Boca Raton, 2000.



Research paper

DNA Commission of the International Society for Forensic Genetics: Recommendations on the validation of software programs performing biostatistical calculations for forensic genetics applications



M.D. Coble^{a,*}, J. Buckleton^{b,c}, J.M. Butler^d, T. Egeland^e, R. Fimmers^f, P. Gill^{g,h},
L. Gusmão^{i,j,k}, B. Guttman^l, M. Krawczak^m, N. Morlingⁿ, W. Parson^{o,p}, N. Pinto^{j,k,q,r},
P.M. Schneider^s, S.T. Sherry^t, S. Willuweit^u, M. Prinz^v

^a National Institute of Standards and Technology, Applied Genetics Group, Gaithersburg, MD, USA

^b ESR, Private Bag 92021, Auckland 1142, New Zealand

^c National Institute of Standards and Technology, Statistical Engineering Division (Guest Researcher), Gaithersburg, MD, USA

^d National Institute of Standards and Technology, Special Programs Office, Gaithersburg, MD, USA

^e Norwegian University of Life Sciences, Oslo, Norway

^f Institute for Medical Statistics, Informatics, and Epidemiology, University Bonn, Germany

^g Norwegian Institute of Public Health, Oslo, Norway

^h University of Oslo, Oslo, Norway

ⁱ State University of Rio de Janeiro (UERJ), Rio de Janeiro, Brazil

^j IPATIMUP, Institute of Molecular Pathology and Immunology of the University of Porto, Portugal

^k Instituto de Investigação e Inovação em Saúde, University of Porto, Portugal

^l National Institute of Standards and Technology, Software and Systems Division, Gaithersburg, MD, USA

^m Institute of Medical Informatics and Statistics, Christian-Albrechts University of Kiel, Germany

ⁿ Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

^o Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria

^p Forensic Science Program, The Pennsylvania State University, PA, USA

^q Institute for Research and Innovation in Health (I3S), University of Porto, Porto, Portugal

^r Centre of Mathematics of the University of Porto, Porto, Portugal

^s Institute of Legal Medicine, Faculty of Medicine, University of Cologne, Germany

^t National Center for Biotechnology Information, National Library of Medicine, NIH, Bethesda, MD, USA

^u Department of Forensic Genetics, Institute of Legal Medicine and Forensic Sciences, Charité—Universitätsmedizin, Berlin, Germany

^v John Jay College of Criminal Justice, New York, USA

ARTICLE INFO

Article history:

Received 29 August 2016

Accepted 2 September 2016

Available online 4 September 2016

Keywords:

Biostatistical software

Software validation

Validation and verification

Forensic genetics

Software training and testing

ABSTRACT

The use of biostatistical software programs to assist in data interpretation and calculate likelihood ratios is essential to forensic geneticists and part of the daily case work flow for both kinship and DNA identification laboratories. Previous recommendations issued by the DNA Commission of the International Society for Forensic Genetics (ISFG) covered the application of bio-statistical evaluations for STR typing results in identification and kinship cases, and this is now being expanded to provide best practices regarding validation and verification of the software required for these calculations. With larger multiplexes, more complex mixtures, and increasing requests for extended family testing, laboratories are relying more than ever on specific software solutions and sufficient validation, training and extensive documentation are of utmost importance.

Here, we present recommendations for the minimum requirements to validate bio-statistical software to be used in forensic genetics. We distinguish between developmental validation and the responsibilities of the software developer or provider, and the internal validation studies to be performed by the end user. Recommendations for the software provider address, for example, the documentation of the underlying models used by the software, validation data expectations, version control, implementation and training support, as well as continuity and user notifications. For the internal validations the recommendations include: creating a validation plan, requirements for the range

* Corresponding author at: National Institute of Standards and Technology, Applied Genetics Group, 100 Bureau Drive MS 8314, Gaithersburg, MD 20899-8314, USA.
E-mail address: mcoble@nist.gov (M.D. Coble).

of samples to be tested, Standard Operating Procedure development, and internal laboratory training and education. To ensure that all laboratories have access to a wide range of samples for validation and training purposes the ISFG DNA commission encourages collaborative studies and public repositories of STR typing results.

Published by Elsevier Ireland Ltd.

1. Introduction

Forensic genetics is experiencing an increase in data volume and complexity, and the interpretation of these data is becoming more and more dependent upon the use of appropriate bio-statistical computer programs. Software for calculating likelihood ratios to evaluate trace evidence or competing kinship scenarios has been in use for many years now, and several groups have described validation exercises of either in-house, open source, or commercial software packages [1–15].

These publications vary notably in terms of the validation approach taken, and standardized reporting of which quality measures were invoked, which tests have been successfully completed, and which software documentation was available. This information is not only of interest to the forensic scientist but also to the legal community. For quality measures, a distinction must be drawn between the responsibility of the software developer or provider, e.g. for code review, version control, documentation of the underlying theory and validation against known data sets, and the responsibility of the end user, e.g. internal validation under local laboratory conditions, formulation of standard operating procedures (SOPs), and training and competency testing.

International industry standards apply to software validation, verification [16] and test documentation [17]. These standards can be simplified and extrapolated [18] to forensic genetics. For internal validation, the goal is similar to other analysis methods: to test the proper function and assess accuracy and limitations of the methods. Previous recommendations on forensic method validation and application of genetic analyses are useful to be read in conjunction with these guidelines [19–25].

The International Society for Forensic Genetics (ISFG) has convened a DNA Commission to establish validation guidelines for bio-statistical software to be used in forensic genetics. Examples include software to calculate statistics for: single-source samples, autosomal DNA mixtures of two or more individuals with no drop-out, or where drop-out and drop-in are possible, paternity and kinship testing, and haploid marker interpretation. The goal of the DNA Commission was to carve out a consensus view on the minimum requirements for the validation (is it doing the right thing?) and verification (is it doing the thing right?) of a software program (V&V) [16] and to describe the software test documentation (STD) [17] to be generated by the software provider. The DNA Commission differentiated developmental from internal (laboratory) validation and emphasizes that the software used is an integral part of the evidential process and should not be treated as a separate and isolated component.

2. Provider responsibilities and developmental validation

The software developer has the burden to specify and document the assumptions and genetic/statistical models underlying the software program and refer to mathematical/statistical proofs or provide these with the software. Prior to promoting their software for practical use, the provider or developer must conduct a developmental validation demonstrating that the intended calculations are being performed correctly and that they provide the expected results. The data sets used for validation should be made

publicly available alongside the validation results, as is outlined below.

2.1. Underlying models and developer's validation

Recommendation 1

Bio-statistical software for forensic genetic applications should be accompanied by scientific papers or information or guidance materials, such as a user manual, describing the underlying method. The population genetic and data model(s) used should be explicitly described and disclosed to allow the reproducibility of all the computations by other means (algebraic formulae, other software programs or statistical approaches) as publication in peer-reviewed journals

The DNA Commission encourages software providers or developers to report the theoretical assumptions underlying their product or refer to already published models. We also encourage the publication of the design and outcome of their developmental validation in peer-reviewed journals. We discourage insufficiently documented or described software where the end user cannot adequately explain to the trier of fact (e.g. judge or jury) the theoretical basis of the software used.

Recommendation 2

Bio-statistical software for forensic genetic applications should be validated according to particular requirements and specific intended use. The software developer's validation should use publicly-available data sets or disclose the used data set otherwise. The result of the software developer's validation and its environment (hardware and software dependencies) should be documented and disclosed

One of the principles of scientific research is that any new finding should be amenable to independent replication. The DNA Commission therefore encourages software providers or developers to verify and validate their software (e.g. by generating or using validation data sets with known outcomes) along with the parameters necessary for the software to work (e.g. population allele counts for frequency calculation). Verification may be assessed using code review. This information could then be publicized so as to support interested laboratories with their own internal training and explorative testing of the software.

The test cases of the validation data should be designed so as to cover all of the software functionality, to be complex enough to detect installation errors, and to be generic enough to also serve as a basis for testing the consistency of future versions of the validated software. Although the goal of internal validation is not to repeat developmental validation, making the data and parameters used for the latter publicly available may add extra benefit in that it would allow laboratories to investigate the local performance of the software under the conditions of the developmental validation, if they so wish. Validation test results should be documented (and disclosed) following a test plan [17] as well as system requirements and platform (hardware and software) specification.

The validity of the results obtained from a given validation data set should also be assessed by way of comparison to the results obtained through hand calculations of algebraic formulae (if possible), using alternative statistical approaches where applicable

e.g. paternity index or the Random Match Probability (RMP), qualitative conclusions drawn by trained analysts [7], or through the use of similar software [8]. Validation exercises should include simulated or real samples with a known underlying scenario. Simulations should cover all relevant aspects of the behavior of genotypes, e.g. mutations, silent alleles, marker linkage, linkage disequilibrium, or population substructure. All input and output data file formats should be documented and/or validated as well. Where applicable external references defining the file format should be included. The DNA Commission encourages the use of open and license free file formats.

Mixture analysis software should be validated on test data involving both known donors (contributors in the mixture which explain the hypothesis of the prosecution (Hp true)) and known non-donors (contributors not in the mixture which explains the hypothesis of the defense (Hd true)), with scenarios underlying the data that cover the range likely to be encountered in casework. The representativeness of the data should cover, as a minimum, the number of contributors, mixture ratios of contributors and DNA template amounts. False donors may be created by simulation or may be real. For the Hp true samples the LR should be largely above 1. The proportions of samples producing a LR less than 1 for Hp true and greater than 1 for Hd true should be noted. The results of these experiments should be disclosed. Circumstances where the LR is above 1 for Hd true or less than 1 for Hp true should be discussed.

For kinship testing software, computations should be performed comparing the likelihoods of the (available) individuals related through the pedigree A (Hp) or through the pedigree B (Hd), under the established assumptions of the program. Samples for Hp and Hd true can be obtained from casework or (preferably) from simulated data. Tests for different levels (and types) of kinship defining Hp and Hd should be computed. The results of these experiments should be disclosed, namely through the plotting of true and false positive rates (in the sense of adopting Hp) for various thresholds of the LR.

Examples of using ground-truth data to test the performance of software can be found in [8,14,26].

2.2. Version control

Recommendation 3

Each version and build of a software should be distinguishable by a version and build number. Each version and build of a software should be validated independently. Exceptions or exclusion of specific tests should be documented.

Software development is often incremental. Amendments to a program may involve alteration of the core algorithms or may be merely cosmetic (such as improving the user interface). If software has been developed in separate parts, any change to one part may bear a risk of consequential changes in the other parts. This has to be taken into account when validating revised software components separately, even though such partial testing may greatly lower the efforts for developmental validation and for internal revalidation by the laboratories.

Providers or developers must label their software by version numbers and a build number to completely identify the software. Every significant change to the code in a released version should be given a unique version number. Whereas additions to the code that, for example, only affect the display of results may not require a change in version number, systems should be in place ensuring that substantial changes cannot be made to the software without changing the version number. All material made available with regard to the developmental validation must be linked to the applicable version number. All software documentation also needs to be clearly tied to a specific version of the software.

Retired versions and documentation should be archived by the providers or developers so as to ensure the possibility of reusing these versions if required, e.g. for review of old cases. Many laboratories are moving towards the use of probabilistic software for mixture interpretation, and consequently often face requests from both prosecution and defense to re-interpret historical cases, especially where “inconclusive” results were obtained by other means of interpretation. We anticipate that future probabilistic bio-statistical software programs will necessitate the review of today’s interpretational methods. It is important to retain retired software versions and the associated documentation of these programs.

2.3. Education and training to the end user from the provider

Recommendation 4

The software provider or developer should create instructions on how to validate and configure the software prior to use in a laboratory. These instructions should form the basis of any internal validation plan to be designed by users.

Recommendation 5

Any bio-statistical software should be accompanied by a user manual enabling a trained user to understand and explain the principles of the software functions and to use the software correctly.

Recommendation 6

Any potential user should have access to sufficient knowledge to use the software in a reasonable way. It is the responsibility of the laboratory to make sure that it has sufficient training resources and provides sufficient support to users to demonstrate that a proposed implementation is ‘fit-for-purpose’.

Laboratories validating software should also create their own examples to test the limits of the software of interest. Guidance from the developer or provider could be valuable to allow the laboratory to develop the most sensible and efficient strategy for validation.

Implementation instructions of stand-alone software should include hardware specifications and troubleshooting information. It is anticipated that the known data sets (either generated by the provider or the testing laboratory) with previously established outcomes will be used to verify proper on-site performance as discussed in Recommendation 2.

User manuals should also have version control for the former to match the software actually in use. Every released version of the software should be accompanied by a comprehensive user manual, or a comprehensive description of the introduced modifications (in case of minor changes). The user manual should be linked to the software version (e.g. use of the software version number on every page). The manual should include a description of the theoretical basis of the software or references to publications or other work describing the basis of the implemented methods. Changes from previous versions should be detailed within the documentation. A separate version history listing the changes introduced for each version release should also be available.

The manual should be standalone or provide detailed references to the available literature. If training is a prerequisite for obtaining the software in the first place, then the manual should provide all instructions in conjunction with that training. In any case, trained users should understand the principles and limitations of the software sufficiently well to represent and explain the results in court. If training is not a prerequisite for obtaining the software, then the user manual must be sufficient that an untrained user can also competently use the software.

As far as training is concerned, the DNA Commission endorses practical in-house training sessions, or remote training (either live or recorded); or at a minimum adequate written material required to meet this recommendation.

Not only the laboratory and the prosecution, but also the defense must have access to suitable information, and the defense may need to investigate significant aspects of the performance of the software for a specific case. Scientists working for the defense should be allowed to attend training and should be permitted to obtain or purchase the software after meeting any training requirements.

2.4. Software updates and continuity

Recommendation 7

To ensure continued availability of software in the future, it is recommended that software source code is placed in a secure repository (e.g. GitHub or an escrow account) and that the algorithms are described in sufficient detail to allow for reimplementations. It is the responsibility of the customer of software to ensure that they have a legal basis to access the code in the event of a supplier ceasing to trade or withdrawing support

The DNA Commission does not consider examination of the source code to be a useful fact-finding measure in a legal setting. A rigorous validation study (both developmental and internal) should be sufficient to reveal shortcomings or errors in coding. There should be sufficient public information available to allow for independent reimplementations as described in recommendation 1. However, if requested by the legal system, the code should be made available subject to the software provider's legitimate copyright or commercial interests being safeguarded. Supervised access to the code under a "no copy" policy is acceptable.

If the software follows the open source principle, the DNA Commission encourages open-source developers to publish their source code using systems such as SoftwareX (<http://www.journals.elsevier.com/softwarex/>) as Supplementary data. Language specific repository systems such as CRAN (<https://cran.r-project.org/>) or general ones like GitHub (<https://github.com>) should be utilized where publishing is unsuitable or impossible.

Sharing of the source code can be useful for collaborative efforts or further development, improvements, or modifications. The sharing of source code does not release the developer from their obligation to rigorously document, verify and validate their software.

Recommendation 8

Custodians of software used for forensic genetics purposes should establish a system allowing them to notify users about quality assurance issues and updates. Software bugs (and their fixes) together with a list of changes should be disclosed.

During the time a given piece of software is in use, new limitations or programming faults almost inevitably will be discovered. The impact of such faults should be investigated by the providers or developers and disclosed together with the fix. However, it is important that knowledge of any newly arisen problems is shared transparently with end users and other stakeholders in the judicial process. Corrective actions must be triggered as needed and end users prevented from continued use of outdated or flawed versions. This requires, as a minimum, a link between the providers and developers on the one hand, and end users and interested third parties on the other that may even be unknown to the providers or developers themselves. This link could be drawn, for example, by a website where critical information is made available, or a registration system whereby the provider or developer can contact users directly.

2.5. Randomness

Recommendation 9

Software using algorithms with components of randomness, such as Monte Carlo methods or random permutations, should have a feature to set this function to a stable state/mode that allows for repeated testing or recalculation (e.g. the user should be able to set the seed for initiating a Monte Carlo process to allow for repeated analyses of the same data set).

Some software programs utilize randomness (e.g. model the drop-out probability as part of a Markov Chain Monte Carlo, determine a p-value by random permutation or random selection as part of a bootstrap process). These use a random number generator which starts from an initial number, known as the seed, and apply an algorithm that produces a sequence of numbers that have little relationship to each other. The series will eventually repeat itself, although usually only after a very long time.

It may be necessary to reproduce results after the fact and reanalyze one specific run in exactly the same fashion, for example as part of verifying the software after a change, or due to a retrospective investigation. Since this can only be achieved by using the same seed in the second run that was used in the first run, it is desirable that the seed is reported as part of the output of each run, and that the end user can set a particular seed for a run themselves, if they so wish.

3. Internal validation

Internal validation refers to empirical studies performed either within a laboratory or outsourced to a third party entity to ensure that the software runs properly within the relevant laboratory. It should cover a wide range of the functionality of the software and all relevant parameter settings of the software. Unless the software will only be used on pristine samples with complete genotypes, the validation needs to address variations in multiplexes, cycle numbers, clean-up chemistries, injection strategies, or equipment that may be used in casework. Internal validation should be planned carefully. The plan should include (at a minimum) the objectives outlined in recommendations 10 through 13. Developmental validation information should be gathered from the provider or developer and laboratories should be familiar with the content of this material before starting their internal validation.

The goal of an internal validation study is to explore the limitations of the software and test the reliability, robustness, and reproducibility of the system. Samples that mimic the types of cases encountered should be tested. These will primarily include "mock" samples. Real casework samples can also be used. The challenge with using real casework samples is that the "ground truth" composition of the mixture components may be difficult to determine, especially with very low level minor contributors.

Some laboratories may be restricted with their use of casework data for validation activities. Where previous interpretation methods resulted in an inclusion of a person of interest, broadly one should expect an inclusionary likelihood ratio for the interpretation of the same profile using probabilistic genotyping software.

3.1. Developing a plan and sample testing

Recommendation 10

Before initiating the validation of a software program, the laboratory should develop a documented validation plan. The software should have a completed and up to date developmental validation along with other supporting materials such as

publications describing the models, propositions and parameters used by the software and a user's manual.

Recommendation 11

The laboratory should test the software on representative data generated in-house with the reagents, detection instrumentation, and analysis software, used for casework. If a laboratory employs variable DNA typing conditions (e.g. within variation in the amplification and/or electrophoresis conditions to increase or decrease the sensitivity of detection of alleles and/or artifacts), then these types of profiles should also be tested as part of the internal validation plan.

Recommendation 12

The laboratory should consider the range of samples expected to be analyzed in casework to define the scope of application of the software. Internal validation should address (1) true donors and non-donors and/or (2) related and unrelated individuals across a range of situations that span or exceed the complexity of the cases likely to be encountered in casework.

Planning is crucial for any validation exercise to be successful. In addition to identifying suitable staff to conduct the necessary experiments, the information technology resources required for running the software should be scrutinized as well. Moreover, some of the experiments called for in Recommendation 11 may be redundant under certain circumstances. For example, if a laboratory is validating software for kinship analysis, then varying the amplification or electrophoresis conditions is usually unnecessary because only the specific alleles (and not the variation in peak heights) are required for software validation.

The consideration of both known contributors and known non-contributors is important to determine the limits of any software for mixture interpretation [27]. Mixtures should be gauged against profiles of true donors (i.e., ground truth known trials) to test the sensitivity of the software whereas a comparison to non-contributors is necessary to test its specificity. Where previous interpretation resulted in an inclusion of a person of interest, one should expect an inclusionary likelihood ratio for the same profile using the software under validation; deviations should be discussed in the validation report.

Determination of the limits of the software is important to establish the types of profiles that are suitable for handling by the laboratory. It is acceptable to manipulate the input data so as to create challenging profiles with the desired properties to test.

Probabilistic software, especially for low-level DNA mixtures, may allow a laboratory to widen the scope of their casework in terms of the type of evidence handled. However, there may also be a temptation to submit all complex mixtures to particularly versatile software. Therefore, the community is reminded of a previous recommendation of the DNA Commission [20] that is still valid:

(Gill et al., 2006, Recommendation 8): If the alleles of certain loci in the DNA profile are at a level that is dominated by background noise, then a biostatistical interpretation for these alleles should not be attempted.

Recommendation 13

The laboratory should determine whether the results produced by the software are consistent with the laboratory's previously validated interpretation procedure if the data and/or method exist.

In general, known samples are used as part of the internal validation and the results from previous validation exercises (for example, a simple spreadsheet to calculate kinship statistics for parent-child trios) should be compared to the output of new software. One would expect the results of the different procedures to be sufficiently similar.

3.2. Standard operating procedure development

Recommendation 14

In addition to the user manual, the laboratory should develop standard operating procedures based upon the internal validation data outlining the types of cases and data to which the software can be applied, the source of population allele frequencies, the testing of one or more propositions, reporting, and how software updates are performed regularly.

The SOP for any laboratory should take into account both the developmental and internal validations. They should guide end users on when and how to use the software and when it should not be used. The latter can be achieved by providing explicit guidance on the limitations of the software. The SOP should be detailed enough to ensure consistent use of the software across the laboratory. It is important to note with both kinship analysis [21] and forensic evidence evaluation [20], the construction of clearly defined hypotheses (propositions) is critical, and the key assumptions underlying the computational process will affect the final interpretation of the output [28–30].

Prior to training laboratory staff on new SOPs, the instructions should be tested on a controlled data set to verify that workflow laid out by the SOPs performs as expected.

Software bug-fix releases should be installed with priority according to a plan as part of the SOP. The laboratory should define a general policy on software updates and upgrades in terms of validation and personnel responsibilities.

3.3. Training and education

Recommendation 15

The laboratory should develop and follow a policy or procedure for the training of software end users in the laboratory.

Training laboratory personnel on the use of bio-statistical software is mandatory and must include a range of cases and require a competency test as a qualifying exam. In addition, the DNA Commission recommends that basic training on likelihood ratios and proposition building should be an integral part of the professional qualification of forensic geneticists.

The training policy should outline the prerequisite competencies for an examiner using the software. For example, if the software requires manual elements such as removal or recognition of artifacts, or for mixture software the assignment of a number of contributors, then these are prerequisite competencies. For each competency mandated for the examiner using the software, the exact learning outcome, the examination strategy, and the expectations required to pass the exam should be defined.

Additional proficiency testing and continuous competency monitoring of the software users is also recommended. The ISFG encourages the participation of external collaborative exercises such as proficiency testing workshops and interlaboratory studies [31,32] to develop a "community of users".

Recommendation 16

The DNA Commission encourages the forensic community to establish a public repository of typing results from adjudicated casework covering a wide range of kinship cases and mixture samples including different challenging scenarios like low-level mixtures and related contributors. The data need to be in a universal, useful file format. The repository should be governed by a neutral organization providing equal access to all interested international parties.

Mock or case-like samples may be a useful alternative for the repository. Meta-data associated with the submitted profiles should include relevant information such as the kit used, PCR

cycle conditions, the separation polymer used, the CE system electrophoretic injection parameters, and any other relevant information about the sample.

The DNA Commission envisions the repository to become a rich resource for both, the initial testing of new software and continuing training programs. For example, a set of candidate family reference data from NIST [33] available at <http://www.cstl.nist.gov/biotech/strbase/kinship.htm> was used by one laboratory to confirm the concordance between a kinship software program and algebraic calculations verified by a spreadsheet program [34]. Likewise, the Biomedical Forensic Sciences program at Boston University (USA) has developed a training website (<http://www.bu.edu/dnamixtures/>) with a variety of single-source and mixture profiles for testing and training.

3.4. Additional guidance on software usage and application

Cosmetic modifications such as a change in the graphical interface of the program, or changes in the reporting format, may not require developmental validation but should be subjected to additional tests to ensure that the changes do not affect the interpretation of the software output. This may be achieved by running a range of identical cases before and after the changes, followed by comparative reviewing of the output. Core changes to the implemented algorithms should be subjected to additional developmental validation prior to their release.

In addition to supporting internal laboratory validation, it is recommended that software providers or developers, together with laboratories and other stakeholders, create Supporting information targeted towards the legal community. This information shall be made up such that it allows end users to successfully debate the scientific merits of the software in admissibility hearings and court cases. In jurisdictions employing an adversarial system, this should include a defense access policy.

If the cost to purchase the software is prohibitive, access, at reasonable or no cost, to an executable version of the software for use in a particular case, along with sufficient support that the defense could realistically run the software with some understanding should be provided. If alternative validated software using similar, scientifically sound and widely accepted algorithms is available, then the defense scientist may use this different software to analyze the case in question. There may be examples where the analysis of one and the same evidence with different software produces statistical output that may lead to differing conclusions. This could possibly cause confusion in the legal system although it should not be interpreted as one software being “better” than the other. It is important instead that the end users understand the underlying assumptions, models, and limitations of the software used.

Acknowledgements

The authors would like to thank Dr. Bruce Weir (University of Washington) and June Guinness (UK Forensic Science Regulation Unit) for their input and helpful discussions. Points of view in this document do not necessarily represent the official position or policies of the U.S. Department of Commerce. Certain commercial equipment, instruments, or materials are identified in this paper in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the U.S. Department of Commerce nor does it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

References

- [1] C.H. Brenner, Symbolic kinship program, *Genetics* 145 (1997) 535–542.
- [2] J.A. Riancho, M.T. Zarrabeitia, A Windows-based software for common paternity and sibling analyses, *Forensic Sci. Int.* 135 (2003) 232–234.
- [3] J. Drábek, Validation of software for calculating the likelihood ratio for parentage and kinship, *Forensic Sci. Int. Genet.* 3 (2009) 112–118.
- [4] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, B.W. Duceman, Validating TrueAllele[®] DNA mixture interpretation, *J. Forensic Sci.* 56 (2011) 1430–1447.
- [5] A. Röck, J. Irwin, A. Dür, T. Parsons, W. Parson, SAM: String-based sequence search algorithm for mitochondrial DNA database queries, *Forensic Sci. Int. Genet.* 5 (2011) 126–132.
- [6] K. Slooten, Match probabilities for multiple siblings, *Forensic Sci. Int. Genet.* 6 (2012) 466–468.
- [7] A.A. Mitchell, J. Tamariz, K. O’Connell, N. Ducasse, Z. Budimlija, M. Prinz, T. Caragine, Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in, *Forensic Sci. Int. Genet.* 6 (2012) 749–761.
- [8] P. Gill, H. Haned, A new methodological framework to interpret complex DNA profiles using likelihood ratios, *Forensic Sci. Int. Genet.* 7 (2013) 251–263.
- [9] D. Kling, A.O. Tillmar, T. Egeland, Familias 3—extensions and new functionality, *Forensic Sci. Int. Genet.* 13 (2014) 121–127.
- [10] R. Puch-Solis, T. Clayton, Evidential evaluation of DNA profiles using a discrete statistical model implemented in the DNA LiRa software, *Forensic Sci. Int. Genet.* 11 (2014) 220–228.
- [11] C.D. Steele, M. Greenhalgh, D.J. Balding, Verifying likelihoods for low template DNA profiles using multiple replicates, *Forensic Sci. Int. Genet.* 13 (2014) 82–89.
- [12] R.G. Cowell, T. Graverson, S.L. Lauritzen, J. Mortera, Analysis of forensic DNA mixtures with artefacts, *J. R. Stat. Soc. Ser. C: Appl. Stat.* 64 (2015) 1–48.
- [13] G. Dørum, D. Kling, C. Baeza-Richer, M. García-Magariños, S. Sæbø, S. Desmyter, T. Egeland, Models and implementation for relationship problems with dropout, *Int. J. Legal Med.* 129 (2015) 411–423.
- [14] D. Taylor, J. Buckleton, I. Evett, Testing likelihood ratios produced from complex DNA profiles, *Forensic Sci. Int. Genet.* 16 (2015) 165–171.
- [15] Ø. Bleka, G. Storvik, P. Gill, EuroForMix: an open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts, *Forensic Sci. Int. Genet.* 21 (2016) 35–44.
- [16] IEEE Standard for System and Software Verification and Validation, in *IEEE Std 1012–2012 (Revision of IEEE Std 1012–2004)*, pp. 1–223, May 25 2012; 10.1109/IEEESTD.2012.6204026.
- [17] IEEE Standard for Software and System Test Documentation, in *IEEE Std 829–2008*, pp. 1–150, July 18 2008; 10.1109/IEEESTD.2008.4578383.
- [18] General Principles of Software Validation; Final Guidance for Industry and FDA Staff, U.S. Department Of Health and Human Services – Food and Drug Administration—Center for Devices and Radiological Health & Center for Biologics Evaluation and Research. (2002) available at: <http://tinyurl.com/m39b2g>.
- [19] N. Morling, R.W. Allen, A. Carracedo, H. Gead, F. Guidet, C. Hallenberg, W. Martin, W.R. Mayr, B. Olaisen, V.L. Pascali, P.M. Schneider, Paternity Testing Commission of the International Society of Forensic Genetics: recommendations on genetic investigations in paternity cases, *Forensic Sci. Int.* 129 (2002) 148–157.
- [20] P. Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayr, N. Morling, M. Prinz, P.M. Schneider, B.S. Weir, DNA Commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures, *Forensic Sci. Int.* 160 (2006) 90–101.
- [21] D.W. Gjerston, C.H. Brenner, M.P. Baur, A. Carracedo, F. Guidet, J.A. Luque, R. Lessig, W.R. Mayr, V.L. Pascali, M. Prinz, P.M. Schneider, N. Morling, ISFG: recommendations on biostatistics in paternity testing, *Forensic Sci. Int. Genet.* 1 (2007) 223–231.
- [22] P. Gill, L. Gusmão, H. Haned, W.R. Mayr, N. Morling, W. Parson, L. Prieto, M. Prinz, H. Schneider, P.M. Schneider, B.S. Weir, DNA Commission of the International Society of Forensic Genetics: recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods, *Forensic Sci. Int. Genet.* 6 (2012) 679–688.
- [23] Scientific Working Group on DNA Analysis Methods (SWGDM), Validation Guidelines for DNA Analysis Methods (2012), pp. 1–13. http://media.wix.com/ugd/4344b0_cbc27d16dcb64fd88cb36ab2a2a25e4c.pdf.
- [24] J.A. Bright, I.W. Evett, D. Taylor, J.M. Curran, J. Buckleton, A series of recommended tests when validating probabilistic DNA profile interpretation software, *Forensic Sci. Int. Genet.* 14 (2015) 125–131.
- [25] Scientific Working Group on DNA Analysis Methods (SWGDM), Guidelines for the Validation of Probabilistic Genotyping Systems (2015), pp. 1–12. http://media.wix.com/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf.
- [26] D. Taylor, Using continuous DNA interpretation methods to revisit likelihood ratio behaviour, *Forensic Sci. Int. Genet.* 11 (2014) 144–153.
- [27] P. Gill, J. Curran, C. Neumann, A. Kirkham, T. Clayton, J. Whitaker, J. Lambert, Interpretation of complex DNA profiles using empirical models and a method to measure their robustness, *Forensic Sci. Int. Genet.* 2 (2008) 91–103.
- [28] J. Buckleton, J.A. Bright, D. Taylor, I. Evett, T. Hicks, G. Jackson, J.M. Curran, Helping formulate propositions in forensic DNA analysis, *Sci. Justice* 54 (2014) 258–261.

- [29] T. Hicks, A. Biedermann, J.A. de Koeijer, F. Taroni, C. Champod, I.W. Evett, The importance of distinguishing information from evidence/observations when formulating propositions, *Sci. Justice* 55 (2015) 520–525.
- [30] S. Gittelsohn, T. Kalafut, S. Myers, D. Taylor, T. Hicks, F. Taroni, I.W. Evett, J.-A. Bright, J. Buckleton, A practical guide for the formulation of propositions in the Bayesian approach to DNA evidence interpretation in an adversarial environment, *J. Forensic Sci.* 61 (2016) 186–195.
- [31] A.R. Thomsen, C. Hallenberg, B.T. Simonsen, R.B. Langkjær, N. Morling, A report of the 2002–2008 paternity testing workshops of the English speaking working group of the International Society for Forensic Genetics, *Forensic Sci. Int. Genet.* 3 (2009) 214–221.
- [32] L. Prieto, H. Haned, A. Mosquera, M. Crespillo, M. Alemañ, M. Aler, F. Ivarez, C. Baeza-Richer, A. Dominguez, C. Doutremepuich, M.J. Farfán, M. Fenger-Grøn, J. M. García-Ganivet, E. González-Moya, L. Hombreiro, M.V. Lareu, B. Martínez-Jarreta, S. Merigioli, P. Milans Del Bosch, N. Morling, M. Muñoz-Nieto, E. Ortega-González, S. Pedrosa, R. Pérez, C. Solís, I. Yurrebaso, P. Gill, EuroforGen-NoE collaborative exercise on LRmix to demonstrate standardization of the interpretation of complex DNA profiles, *Forensic Sci. Int. Genet.* 9 (2014) 47–54.
- [33] K.L. O'Connor, S.P. Myers, E.L.R. Butts, C.R. Hill, J.M. Butler, P.M. Vallone, Candidate reference family data: a tool for validating kinship analysis software, 21st International Symposium On Human Identification, San Antonio, TX, October 11–14, 2010 http://www.cstl.nist.gov/biotech/strbase/pub_pres/Promega2010_OConnor.pdf. <http://www.cstl.nist.gov/biotech/strbase/kinship.htm>.
- [34] D.A. Azevedo, G.R.B. Souza, I.H.E.F. Silva, L.A.F. Silva, Genetic kinship analysis: a concordance study between calculations performed with the software Familias and algebraic formulas of the American Association of Blood Banks, *Progress in Forensic Genetics 14: Proceedings of the 24th International ISFG Congress* (2011) e186–e187.

IN THE COURT OF COMMON PLEAS OF ALLEGHENY COUNTY, PENNSYLVANIA

CRIMINAL DIVISION

COMMONWEALTH OF PENNSYLVANIA)
)
 v.) CC 201307777
)
 MICHAEL ROBINSON,)
)
 Defendant)

ORDER OF COURT

AND NOW, to-wit, this 7th day of December, 2015, having considered testimony, exhibits, and arguments presented, this Court hereby DENIES Defendant's Discovery Motion to the extent it requests production of True Allele Casework software source code.

BY THE COURT:


Honorable Jill E. Rangos

IN THE COURT OF COMMON PLEAS OF ALLEGHENY COUNTY, PENNSYLVANIA
CRIMINAL DIVISION

COMMONWEALTH OF PENNSYLVANIA)
)
 v.) CC 201307777
)
 MICHAEL ROBINSON,)
)
 Defendant)

MEMORANDUM ORDER

AND NOW, to-wit, this 4th day of February, 2016, this Court hereby DENIES Defendant's "Application Pursuant to Title 42 Pa.C.S.A. Section 702(B), Interlocutory Orders, for Amendment to Include Certification of the Interlocutory Discovery Order Issued on December 7, 2015." This Court denied Defendant's discovery request for the "source code" for Cybergenetics TrueAllele Casework System, which was used to test a bandana recovered from the crime scene which the Commonwealth alleges belongs to Defendant. This source code is the intellectual property of Cybergenetics.

Pa. R. Crim. P. 573 states that a trial court may permit discovery of items which are material, reasonable and in the interests of justice, and Defendant asserts that his request for the source code has met this criteria. However, "[e]vidence is material only if there is a reasonable probability that, had the evidence been disclosed to the defense, the result of the proceeding would have been different. A 'reasonable probability' is a probability sufficient to undermine confidence in the outcome." *Pennsylvania v. Ritchie*, 480 U.S. 39, 57 (1987). Since materiality requires that the material sought must be outcome-determinative (*See also Commonwealth v. Tharp*, 101 A.3d 736, 748 (Pa. 2014)), Defendant must establish that production of the source

FILED
16 FEB -4 PM 1:34

code is a linchpin to undermining the Commonwealth's case as it pertains to the DNA evidence on the bandana.

In support of its assertion, Defendant alleges that TrueAllele's reliability cannot be evaluated without the source code. The Pennsylvania Superior Court, in *Commonwealth v. Foley*, 38 A.3d 882 (Pa. Super. 2012) (*en banc*), disagreed. The *Foley* court discussed whether TrueAllele testing was admissible pursuant to *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923) and in so doing found that TrueAllele was not "novel" science. *Foley* addressed the issue of assessing the reliability of TrueAllele without the production of the source codes and determined that scientists could validate the reliability of TrueAllele without the source code. *Id.* at 889-90. In addition, the *Foley* court noted that the trial court had "[found] Dr. Perlin's methodology [to be] a refined application of the "product rule," a method for calculating probabilities that is used in forensic DNA analysis." *Foley*, 38 A.3d at 888. The Superior Court noted that evidence based on the product rule previously has been deemed admissible under *Frye*. *Id.*, citing *Commonwealth v. Blasioli*, 713 A.2d 117, 1118 (Pa. 1998).

As the defense has argued that *Foley* is not controlling on the question of materiality of the source code, this Court held a two day hearing and considered expert testimony and argument. After considering the testimony, this Court determined that the source code is not material to the defendant's ability to pursue a defense.

Moreover, release of the source code would not be reasonable under Pa. R. Crim. Pro. 573 (A). Dr. Mark Perlin, founder of Cybergenetics, stated in his April 2015 Declaration that disclosure of the source code would cause irreparable harm to the company, as other companies would be able to copy the code and potentially put him out of business. (Commonwealth's Supplemental Answer to Motion for Discovery, Exhibit 1, "Declaration of Mark W. Perlin, April

2015” para. 54-55) An order requiring Cybergenetics to produce the source code would be unreasonable, as release would have the potential to cause great harm to Cybergenetics. Rather than comply, Dr. Perlin could decline to act as a Commonwealth expert, thereby seriously handicapping the Commonwealth’s case.

42 Pa.C.S. § 702(b) states that if the trial court believes the interlocutory order “involves a controlling question of law as to which there is substantial ground for difference of opinion and that an immediate appeal from this order may materially advance the ultimate termination of the matter, it shall so state in such order.” This Court is not of the opinion that the discoverability of the source code for Cybergenetics’ TrueAllele Casework system involves a controlling issue of law to which a substantial ground for a difference of opinion exists. Defendant alleges that the Honorable Jeffrey A. Manning’s ruling in the *State of California v. Martell Chubbs* creates a substantial ground for a difference of opinion. However, in that case J. Manning merely enforced a subpoena *duces tecum* ordering Dr. Perlin to appear in California with the documents subject to the subpoena but he left the ultimate disposition of the discovery request to the California court. Ultimately, the California Superior Court did not require Cybergenetics to produce the source code.¹ Further, J. Manning, in another pending matter involving a discovery request for the TrueAllele source code, declined² to read his ruling in *Chubbs* as controlling or contradictory and deferred to this Court for a ruling on the issue of the discoverability of source code. Similarly, the Honorable Edward J. Borkowski, without a hearing, quashed a subpoena *duces tecum* requesting production of the TrueAllele source code in another case pending in this Court.³

¹ 2015 WL 139069 (Unpublished Opinion)

² *Commonwealth v. Chelsea Arganda and Chester White*, CC# 2013-17748 and CC# 2013-17753.

³ *Commonwealth v. Wade*, CC# 2014-04799.

Reviewing *Foley* and *Chubb*, as well as the pretrial proceedings of record in other matters pending before my colleagues in the Criminal division of the Court of Common Pleas of Allegheny County, and taking into consideration the briefs and arguments of the parties, this Court finds no reason to certify its December 7, 2015 Discovery Order for Interlocutory Appeal.

BY THE COURT:


Honorable Jill E. Rangos

IN THE SUPERIOR COURT OF COWETA COUNTY

STATE OF GEORGIA

STATE OF GEORGIA

V.

CASE NO. 2017-CR-618

MONTE BAUGH,
THADDEUS HOWELL,

DEFENDANTS.

FILED
CLERK OF SUPERIOR COURT
COWETA COUNTY, GA
2019 MAR 22 PM 2:12

ORDER

This case is before the court regarding the state's intent to present evidence at trial of DNA analysis using TrueAllele® software. The Defendants in the case have moved to exclude this evidence arguing that it does not meet the standard for the admission of scientific evidence set out in *Harper v. State*, 249 Ga. 519 (1982) and subsequent cases. The state has opposed that motion and has moved the court to take judicial notice that this evidence has reached a state of scientific certainty sufficient to admit it under *Harper* without a hearing.

After consideration of the issue, the court denied the state's motion to take judicial notice based on the relative novelty of TrueAllele evidence and the absence of its prior use in this court. The court then held an evidentiary hearing on the admissibility of the TrueAllele DNA evidence on March 11, 2019. After conducting the hearing and considering the evidence presented, the record of the case and arguments of counsel, the court hereby finds that the TrueAllele DNA evidence does meet the *Harper* standard, will be admissible in this case and makes the following findings of fact and conclusions:

DNA Evidence in Georgia

DNA evidence has been routinely admitted in the State of Georgia for decades. As the manner of DNA analysis has evolved over time, Georgia courts have kept up with this evolution by continuously assessing the reliability and validity of any significant advancements in DNA analysis.

DNA evidence's admissibility was first addressed by the Georgia Supreme Court in the landmark decision of *Caldwell v. State*, 260 Ga. 278 (1990). In *Caldwell*, the Georgia Supreme Court first recognized the reliability and admissibility of DNA evidence involving the use of restriction fragment length polymorphism analysis ("RFLP"). Thereafter, advances in DNA analysis led to the development of a new technique of DNA analysis involving the using of polymerase chain reaction ("PCR") as part of the process of extracting, amplifying, and profiling a DNA sample in preparation for making DNA profile comparisons. *Redding v. State*, 219 Ga. App. 182 (1995). Since that time, PCR has continually been recognized as a valid and reliable form of creating DNA profiles for comparison, even as PCR based DNA analysis was applied to different forms of DNA. *Thrasher v. State*, 261 Ga. App. 650 (2003) (holding that PCR based DNA analysis is accepted as valid in Georgia); *Shabazz v. State*, 265 Ga. App. 64 (2004) (affirming the trial court's admission of Y-STR DNA analysis from PCR generated DNA profiles); *Vaughn v. State*, 282 Ga. 99 (2007) (affirming the admission of mitochondrial DNA (mtDNA) analysis results at trial).

The Role of TrueAllele Software in DNA Analysis

Dr. Mark Perlman, the creator of TrueAllele software, provided expert testimony which included an explanation as to how the long-established procedures involving PCR

that have been used in the preparation of DNA profiles for comparison purposes are still used today. TrueAllele does not change in any manner this established and reliable process of generating DNA profiles. Rather, TrueAllele now offers the ability to analyze such DNA profiles using a computer - a task traditionally performed by a human analyst.

Traditionally, PCR generated DNA profiles have been compared by human analysts using the long-standing statistical association technique known as the Random Match Probability (“RMP”) based on peak height thresholds. These data thresholds are most suitable for analyzing a simple DNA profile involving a single contributor. Dr. Perlin explained how human analysts are limited in their ability to apply thresholds to a complex DNA profile involving a mixture of DNA formed from multiple contributors.

The threshold-based Combined Probability of Inclusion (“CPI”) statistical association analysis of a DNA mixture often results in an “inconclusive” result. This is because humans tend to lack the extraordinary time and mathematical ability needed to analyze the complicated possibilities involved in attempting to unsort the mixture. This is where TrueAllele comes in.

How TrueAllele Software Functions

TrueAllele is a probabilistic genotyping software that analyzes DNA evidence using a mathematical model based on Bayesian statistical analysis and the Markov chain Monte Carlo algorithm. This probabilistic analysis includes a careful consideration of DNA’s known biological and PCR properties, and the prevalence of certain DNA variants in the population.

TrueAllele operates by initially analyzing a DNA mixture¹ that was obtained from a piece of *physical evidence*². In analyzing particular locations of DNA in this mixture, TrueAllele considers the overlapping DNA components present from each contributor's DNA. These overlapping components are termed alleles. Alleles may be visualized as peaks of varying heights and locations on an electropherogram. TrueAllele considers, in part, that each individual contributor to the DNA mixture contributes two alleles at any given location. An individual's two alleles at any location is called that individual's genotype.

Deconvolution of a mixture of DNA involves assessing the entire group of alleles present at a particular location of the DNA mixture and considering the likelihood of different possibilities of sorting and pairing the alleles into separated genotypes. Taking certain known biological principles into consideration, TrueAllele is able to determine which proposed configurations of genotypes are more likely. For example, since a genotype is composed of two alleles (one received from the mother and one received from the father), when analyzing a DNA mixture, it is expected that the two alleles forming an individual's genotype will be present in equal amounts represented on the electropherogram. With a number of these biological principles factored in, TrueAllele considers very many possible assortments of pairs of alleles and then determines the probability of each proposed configuration (or genotype). TrueAllele assesses the possible genotypes and assigns a probability that reflects the likelihood the proposed genotype correctly explains the DNA mixture.

¹ Although TrueAllele's functionality is unique in its ability to analyze DNA mixtures, it's functionality also can apply to non-complex single contributor DNA profiles.

² A suspect's DNA is not a part of this initial analysis.

Once every possible genotype has been objectively assigned a probability corresponding to the likelihood that the proposed genotype belongs to one of the contributors, TrueAllele subsequently compares the suspect's genotype to the corresponding genotype which was previously inferred. Where the suspect's genotype corresponds with the inferred genotype, the previously determined probability is obtained.

This probability that is associated with the suspect's genotype is then divided by the probability of a random person in the population having the same genotype. This final consideration of the prevalence of the particular genotype in the population helps provide context for assessing whether it is just a coincidence the suspect's genotype is present or whether it is more likely present because the suspect actually contributed it. The result of this completed analysis is a match statistic referred to as the likelihood ratio ("LR"). The LR reflects the likelihood of a DNA match between the evidence occurring because the suspect actually contributed their DNA to the mixture versus the probability of a match existing by mere coincidence.

The aforementioned procedure is repeated on a number of different locations of the DNA mixture (typically 15 to 25 locations). The LR's determined for each of these locations are then multiplied together to obtain a final LR that reflects the strength of a match with the suspect out of consideration of all of these locations in the DNA mixture. This final LR may be reported, as it was in the instant case, as "A match between the firearm grip (Item 13) and Monte Baugh, Jr. (Item 21) is: 3.02 million times more probable than a coincidental match to an unrelated African American person; 305 million times more probable than a coincidental match to an unrelated Caucasian

person, and 67.6 million times more probable than a coincidental match to an unrelated Hispanic person.”

TrueAllele is Reliable

There is no genuine controversy as to the validity and reliability of TrueAllele’s method of analysis. To the contrary, computer analysis of uncertain data using probability modeling is the scientific norm. The reliability of the mathematical concepts TrueAllele uses are not at issue. Bayesian Statistics have been used since the 1700’s, and the Markov Chain Monte Carlo algorithm is a well-established algorithm used since the 1950’s. The PCR generated DNA profiles TrueAllele analyzes are the same profiles analyzed by other methods of admissible DNA analysis that have existed for decades.

Cybergenetics thoroughly tests its software before it is released. Over thirty five validation studies have been conducted by Cybergenetics and other groups to establish the reliability of the TrueAllele method and software. Seven of these studies have been published in peer-reviewed scientific journals, for both laboratory-generated and casework DNA samples.

In the “peer-review” process, scientists describe their research methods, results and conclusions in a scientific paper, which they submit to a journal for publication. An editor at the journal has, at a minimum, two independent and anonymous scientists in the field read the paper, assess its merits, and advise on the suitability of the manuscript for publication. The paper is then accepted, rejected, or sent back to the authors for revision and another round of review.

A “laboratory-generated” validation study uses data that has been synthesized in a DNA laboratory, and is of known genotype composition. The State provided four published TrueAllele papers of this type for this Court to consider.³

A “casework” validation study uses DNA data exhibiting real-world issues developed by a crime laboratory in the course of their usual casework activity. The State provided three published TrueAllele papers of this type.⁴

Conducting such validations is consistent with the FBI’s 2010 Scientific Working Group on DNA Analysis Methods (SWGDM) interpretation guidelines. TrueAllele complies with the 2015 SWGDAM validation guidelines for probabilistic genotyping systems. Regulatory bodies in New York and Virginia have had independent scientists review validation studies before they granted approval for their state crime laboratories to use TrueAllele for casework.

Validation studies concerning TrueAllele assessed and recognized its reliability in the areas of reproducibility, specificity, and sensitivity. (State’s Exhibits 7 and 11).

Reproducibility speaks to the consistency of the results of the analysis. As Dr. Perlin explained, and as was demonstrated by the validations studies, the LR’s produced

³ (1) Perlin, MW, Sinelnikov, A. **An information gap in DNA evidence interpretation.** *PLOS ONE*. 2009;4(12): e8327; (2) Ballantyne J, Hanson EK, Perlin MW. **DNA mixture genotyping by probabilistic computer interpretation of binomially-sampled laser captured cell populations: combining quantitative data for greater identification information.** *Science & Justice*. 2013;52(2): 103-14; (3) Perlin MW, Hornyak J, Sugimoto G, Miller K. **TrueAllele genotype identification on DNA mixtures containing up to five unknown contributors.** *Journal of Forensic Sciences*. 2015;60(4):857-868; (4) Greenspoon SA, Schiermeier-Wood L, and Jenkins BC. **Establishing the limits of TrueAllele Casework: a validation study.** *Journal of Forensic Sciences*. 2015 ;60(5): 1263-1276.

⁴ (1) Perlin MW, Legler MM, Spencer CE, Smith JL, Allan WP, Belrose JL, Duceman BW. **Validating TrueAllele™ DNA mixture interpretation.** *Journal of Forensic Sciences*. 2011 ;56(6): 1430-1447; (2) Perlin MW, Belrose JL, Duceman BW. **New York State TrueAllele Casework validation study.** *Journal of Forensic Sciences*. 2013 ;58(6): 1458-66; (3) Perlin MW, Dormer K, Hornyak J, Schiermeier-Wood L, and Greenspoon S. **Casework on Virginia DNA mixture evidence: computer and manual interpretation in 72 reported criminal cases.** *PLOS ONE*. 2014;9(3): e92837.

from successive runs of TrueAllele tend to all be within a factor of 100, a reasonable margin given that TrueAllele's match statistics can range into numbers upwards of sextillion (1 followed by 21 zeroes).

Sensitivity measures the extent to which a mixture interpretation method identifies the correct person as a contributor, and Specificity measures the extent to which a mixture interpretation method does not misidentify someone as a contributor. In this context, the validation studies demonstrated how the LR for a known non-contributor is nearly never greater than 999. Thus, great reliability exists in LR's which are greater.

TrueAllele analysis also results in a predictable LR. As the amount of a contributor's DNA in a mixture increases, so does the LR in a predictable manner. (State's exhibits 8 and 9).

TrueAllele's Widespread Acceptance

TrueAllele has been used in approximately 688 criminal cases, with Cybergeneics expert witness testimony given in approximately 85 trials. TrueAllele results have been reported in 43 of the 50 states.

Courts accepting TrueAllele evidence include California, Florida, Indiana, Louisiana, Maryland, Massachusetts, Michigan, Nebraska, New Hampshire, New York, Ohio, Pennsylvania, South Carolina, Tennessee, Texas, Virginia, Washington, the United States Federal Courts (Eastern District of Virginia), United States Marine Corps, Northern Ireland, and Australia.

Over 10 crime laboratories have purchased the TrueAllele system for their own in-house use, and 8 labs are on-line with their validated systems, including the GBI Crime

Lab. These crime laboratories issue their own TrueAllele reports, and give expert witness testimony at trial about their TrueAllele results.

TrueAllele was used to identify human remains in the World Trade Center disaster, comparing 18,000 victim remains with 2,700 missing people. Both prosecutors and defenders use TrueAllele for determining DNA match statistics. TrueAllele is also used by innocence projects and for post-conviction relief. TrueAllele's reliability has been confirmed in appellate precedent in Pennsylvania.⁵

TrueAllele has been admitted into evidence after opposition challenges in nineteen courts in multiple states, including recently in Georgia after a Harper hearing. Jurisdictions that have admitted TrueAllele results after analyzing its reliability include California, Florida, Georgia, Indiana, Louisiana, Massachusetts, Nebraska, New York, Ohio, Pennsylvania, South Carolina, Tennessee, Virginia, Washington, Northern Ireland and Australia.

Nineteen admissibility decisions in the United States are: People of California v. Dupree Langston, Kern County (Kelly-Frye), BF139247B, January 10, 2013; State of Florida v. Lajayvian Daniels, Palm Beach County (Frye), 2015CF009320AMB, October 31, 2018; State of Indiana v. Randal Coulter, Perry County (Daubert), 62C01-1703-MR-192, August 2, 2017; State of Indiana v. Dionisio Forest, Vanderburgh County (Daubert), 82D03-1501-F2-566, June 3, 2016; State of Indiana v. Daylen Glazebrook, Monroe County (Daubert), 53C02-1411 -F 1-1066, February 16, 2018; State of Indiana v. Malcolm Wade, Monroe County (Daubert), 53C02-1411-F3-1042, August 3, 2016; State of Louisiana v. Chattel Chesterfield and Samuel Nicolas, East Baton Rouge Parish

⁵ See Commonwealth v. Foley, 47 A.3d 882 (Pa. Super. 2012).

(Daubert), 01 13-0316 (II), November 6, 2014; State of Louisiana v. Harold Houston, Jefferson Parish (Daubert), 16-3682, May 19, 2017; Commonwealth of Massachusetts v. Heidi Bartlett, Plymouth County (Daubert), PLCR2012-00157, May 25, 2016; State of Nebraska v. Charles Simmer, Douglas County (Daubert), CR16-1634, February 2, 2018; People of New York v. John Wakefield, Schenectady County (Frye), A-812-29, February 11, 2015; State of Ohio v. Maurice Shaw, Cuyahoga County (Daubert), CR-13-575691, October 10, 2014; State of Ohio v. David Mathis, Cuyahoga County (Daubert), CR-16-61 1539-A, April 13, 2018; Commonwealth of Pennsylvania v. Kevin Foley, Indiana County (Frye), 2012 PA Super 31, No. 2039 WDA 2009, Superior Court affirmed February 15, 2012; State of South Carolina v. Jaquard Aiken, Beaufort County (Jones), 20121212-683, October 27, 2015; State of Tennessee v. Demontez Watkins, Davidson County (Daubert), 2017-C-1811, December 17, 2018; Commonwealth of Virginia v. Matthew Brady, Colonial Heights County (Spencer-Frye), CR11000494, July 26, 2013; State of Washington v. Emanuel Fair, King County (Frye), 10-109274-5 SEA, January 12, 2017; State of Georgia v. Thaddus Nundra, Ronnie McFadden, and Louis Ousley (Harper), 18-CR-134, January 29, 2019.

DR. PERLIN IS CREDIBLE

Dr. Perlin testified or has been called to court as an expert witness more than fifty times in fifteen state courts as well as military and federal courts. Dr. Perlin reviewed his credentials, summarized in his curriculum vitae admitted as State's Exhibit 1, and the Court declared him an expert in DNA evidence interpretation, TrueAllele, and the field of software engineering. Dr. Perlin first walked the court through the science of DNA analysis and the processes TrueAllele uses to calculate LR's, using slide shows, which is included in the record as State's Exhibit 3. Dr. Perlin then testified about how

TrueAllele had been tested and used a second slide presentation as he described the validation process and explained the sensitivity, specificity, and reproducibility of TrueAllele also included on State's Exhibit 4.

Availability to Test the Reliability of the TrueAllele Method

Cybergenetics provides opposing experts the opportunity to review the TrueAllele process, examine results, and ask questions. This review can be done in Cybernetics's Pittsburgh office, or through an Internet Skype-like meeting. Cybergenetics regularly explains the system, and the results obtained in a case, to both prosecution and defense.

This introduction to the TrueAllele method, the case data, and the application of the method to the data, is a logical first step. The TrueAllele method is inherently objective, since the computer determines evidence genotypes without any knowledge of the comparison reference genotypes. Hence, there is no possibility of examination bias when determining genotypes from the DNA data. Match statistics, whether inclusionary or exclusionary, are calculated only afterwards by comparing evidence genotypes with reference genotypes. TrueAllele's reliability was established on the evidence in this case. The report and its supporting case packet admitted by the State of Georgia in this case described the system's sensitivity, specificity and reproducibility on the DNA evidence. The case packet gives the data and parameter inputs used in running the program in the case. The packet also includes a case-specific mini-validation study of reported TrueAllele match statistics, measuring match specificity by comparison with non-contributor genotypes. (State's Exhibit 5)

Dr. Perlin testified thirty-seven validation studies have been conducted on TrueAllele either by Cybergenetics, independent crime labs, or collaboration of both; studies, twenty-three are internal validation studies. (State's Exhibits 7 and 11)

Seven of thirty-seven studies have been published in peer-reviewed journals—the first published in 2009. Six of the seven published studies were authored or co-authored by Dr. Perlin. The 2016 PCAST Report states, “it is completely appropriate for method developers to evaluate their own methods”, while noting that “establishing scientific validity also requires scientific evaluation by other scientific groups that did not develop the method.”⁶ Here, although the majority of the publications have been by Cybergenetics, other entities have also reviewed TrueAllele's method.⁷

Dr. Perlin further testified TrueAllele abides by quality assurance standards established by the FBI, as well as guidelines issued by the Scientific Working Group on DNA Analysis Methods (herein “SWGDM”). In 2015, SWGDAM issued guidelines specifically for validation of probabilistic genotyping systems like TrueAllele abides by today.⁸

Dr. Perlin testified sophisticated computer programs solve problems with a hundred dimensions, and TrueAllele uses Markov chain Monte Carlo (MCMC) computing, one of the oldest and well-adopted methods, dating back to the 1950s.⁹ Dr. Perlin testified the MCMC algorithm is considered one of the ten most widely used in computer science.

⁶ 2016 Report on Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, President's Council of Advisors on Science and Technology (PCAT) Report, at 93.

^{7,7} See S. Greenspoon, L. Schiermeir-Wood & B. Jenkins, Establishing the Limits of TrueAllele Casework: A Validation Study, 60 Journal of Forensic Science, 1263 (2015).

⁸ See also State's Exhibit 15 binder titled “Method Reports”

⁹ See also State's Exhibit 20 binder titled “Other Papers”

TrueAllele's Visual User Interface (VUIer™) tool uses MATLAB programming language, which Dr. Perlin described as a standard, and widely relied upon and accepted programming language.

Bayesian methods update belief (i.e., probability) based on evidence. Before seeing evidence (e.g., scientific data), one begins with initial beliefs about hypotheses. Informative evidence changes those beliefs. Bayes wrote his mathematical rule 250 years ago, and modern computing has broadly applied it to the natural and social sciences. A forensic hypothesis is that someone was at a crime scene; Bayes rule weighs DNA evidence to assess that hypothesis.¹⁰

To permit any interested expert witnesses to take a closer look at how TrueAllele software is coded to implement its analysis, Dr. Perlin explained that approximately two years ago he agreed to disclose TrueAllele's source code under specific conditions. (State's Exhibit 12). Dr. Perlin testified the defense in this case did not accept the offer nor has anyone else. Moreover, Cybergenetics offers free cloud-based TrueAllele testing to defense experts.

Dr. Perlin testified the mathematics underlying TrueAllele comply with the SWGDAM guidelines and recommendations. He provided a document that described the TrueAllele methods with both statistical equations and plain English. (State's Exhibit 20). Dr. Perlin further testified TrueAllele has a known error rate under a

¹⁰ Dale J. Poirier, The Growth of Bayesian Methods in Statistics and Economics Since 1970, Bayesian Analysis (2006), which is included in the binder admitted into evidence as State's Exhibit 20; Matthew Richey, The Evolution of Markov Chain Monte Carlo Methods, Math. Assoc. of America. (May 2010), which is also included in the binder admitted into evidence as State's Exhibit 20; See, e.g. Sho Manab, et al., Development and validation of open-source software for DNA mixture interpretation based on a quantitative continuous model, PLOS One (Nov. 2017) (printout included in the binder admitted into evidence as State's Exhibit 20).

fraction of 1%, and the calculation for a false positive in this case was included on the Cybergenetics Report. He explained false-positive error rates are stratified by the strength of the match statistic; he demonstrated with data on the slides, that when a match statistic, or LR, is up to a hundred, the error rate is one in a million, but by the time TrueAllele gets a match statistic of a thousand, no false positives were seen in the study. In comparison to other genotyping methods used and admitted before, such as the Modified Combined Probability of Inclusion (CPI), TrueAllele has a far lower error rate.

Conclusion

The Court finds TrueAllele software satisfies the *Harper* standard. The procedure or technique in question, TrueAllele's method of probabilistic genotyping and DNA analysis, has reached a scientific stage of verifiable certainty and "rests upon the laws of nature". There has been substantial peer review of the subject matter. Validation studies have been conducted that recognize TrueAllele's reliability. The error rate for TrueAllele's manner of probabilistic genotyping is much less than that of other genotyping methods the Courts have already deemed scientifically reliable, such as CPI and modified CPI.

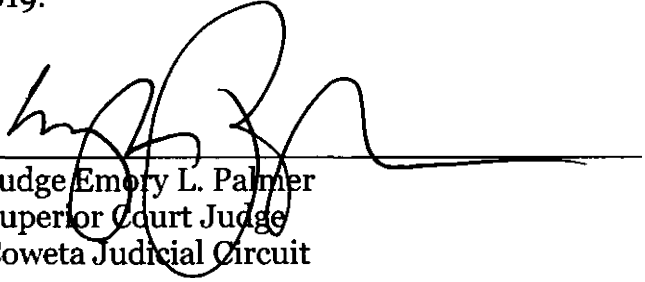
The trial court makes this determination from evidence presented to it at hearing in the form of expert testimony from Dr. Perlin. The Trial Court also bases its determination on all the exhibits and treatises submitted on behalf of the State as shown in the record, including the rationales of other jurisdictions and in Decatur County, Georgia. (State's Exhibits 1 – 27A).

Based on all the evidence presented, this Court finds the TrueAllele analysis was performed in an acceptable manner in this case, that TrueAllele software is capable of

producing reliable results, and the testimony of either Dr. Perlin or Jennifer Hornyak concerning these results would substantially assist the trier of fact in understanding the evidence. The criticisms raised by the defense go towards the weight of the evidence, not admissibility.

For the reasons set forth above, the Court finds the TrueAllele analysis scientifically reliable, and the testimony concerning the TrueAllele's results are admissible at trial. The Trial Court finds that the State has met its burden under Harper. This matter remains scheduled for trial on April 29, 2019.

IT IS SO ORDERED.



Judge Emory L. Palmer
Superior Court Judge
Coweta Judicial Circuit