

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

© <2017>. This manuscript version is made available under the CC-BY-NC-ND 4.0 license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

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.

Keywords: Forensic inference, Evidence, Comparison methods, Probability, Likelihood ratio

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-focussed 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

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.

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], [13], [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 *probability*⁴ 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.

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.

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.

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

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.

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:

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 it 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)

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):

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.

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.

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 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

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

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.

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”.

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 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.
- [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.
- [3] National District Attorneys Association—NDAA, Report Entitled Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. November 16, 2016.
- [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), pp. e7-e9.
- [5] Association of Firearm and Tool Mark Examiners AFTE, Response to PCAST Report on Forensic Science. October 31, 2016.
- [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.
- [7] The International Association for Identification (IAI), IAI Response to the President's Council of Advisors on Science and Technology Report, 2016.
- [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.
- [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.
- [11] Expressing evaluative opinions: a position statement, *Sci. Justice*, 51 (2011), pp. 1-2.
- [12] R. Cook, I.W. Evett, G. Jackson, P.J. Jones, J.A. Lambert, A model for case assessment and interpretation, *Sci. Justice*, 38 (1998), pp. 151-156.
- [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 (1998), pp. 231-240.
- [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 (1999), pp. 103-111.
- [15] P.L. Kirk, The ontogeny of criminalistics, *J. Crim. Law Criminol. Police Sci.*, 54 (1963), pp. 235-238.
- [16] D.A. Stoney, What made us ever think we could individualize using statistics, *J. Forensic Sci. Soc.*, 31 (1991), pp. 197-199.

- [17] A. Biedermann, S. Bozza, F. Taroni, Decision theoretic properties of forensic identification: underlying logic and argumentative implications, *Forensic Sci. Int.*, 177 (2008), pp. 120-132.
- [18] A. Biedermann, S. Bozza, F. Taroni, The decisionalization of individualization, *Forensic Sci. Int.*, 266 (2016), pp. 29-38.
- [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 (1987), pp. 167-187.
- [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).
- [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 (2003), pp. 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 (2013).
- [27] N. Brümmer, J. du Preez, Application-independent evaluation of speaker detection, *Comput. Speech Language*, 20 (2006), pp. 230-275.
- [28] B. Robertson, G.A. Vignaux, C.E.H. Berger, *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).