The limits of forensic evidence

Ten years ago, the US National Academy of Sciences published a report that was critical of the scientific foundations of many forms of forensic evidence. It called for “more and better research”. Ten years on, *Significance* is publishing a special collection of articles that reflect on the report’s impact, the progress that has been made, and the challenges that remain for evaluating and interpreting evidence. First, Brian Tarran highlights the critical role played by statisticians in supporting the development of forensic science.

Twenty-eight years. That is how long Steven Mark Chaney spent in prison, deprived of freedom for something he did not do. Chaney was convicted of a 1987 murder, in part on the basis of the testimony of forensic dentists, which linked him to the crime through a bite mark found on the skin of the victim. The court heard that Chaney’s teeth were a “perfect match”, and that there was a “one to a million” chance that someone other than Chaney was the source of the mark (bit.ly/2EAYRt0).

The jury was convinced of his guilt, but Chaney maintained he was innocent. He remained in prison until October 2015, when his conviction was overturned. Then, in December of last year, the Texas Court of Criminal Appeals found Chaney “actually innocent” of the crime. What changed in those 28 years?
The short answer is, the science did. In its decision, the Court of Criminal Appeals found that the “body of scientific knowledge underlying the field of bitemark comparisons has evolved since [Chaney’s] trial in a way that contradicts the scientific evidence relied on by the State at trial”, and that “testimony of the sort given at Chaney’s trial is now known to be scientifically unsupportable because it ‘went too far’” (bit.ly/2EAYRt0).

Central to the arguments against the bite-mark evidence was a 2009 report by the National Academy of Sciences (NAS), called *Strengthening Forensic Science in the United States: A Path Forward*. It is a report that has done much to change the conversation around forensic science in the USA.

Scientific scrutiny

The release 10 years ago this February of the NAS report was “a pivotal moment”, says Sarah Chu, a senior advisor on forensic science policy for the Innocence Project, which works to exonerate those who are wrongly convicted of crimes. It was “a consensus report ... by an august scientific body”, says Chu, and it served to draw widespread attention to the fact that there was a worrying lack of science underpinning many types of forensic science evidence, thus limiting the inferences and conclusions that could be made.

“Much forensic evidence – including, for example, bite marks and firearm and toolmark identifications – is introduced in criminal trials without any meaningful scientific validation, determination of error rates, or reliability testing to explain the limits of the discipline,” said the report. On bite marks specifically, it said: “Although the majority of forensic odontologists are satisfied that bite marks can demonstrate sufficient detail for positive identification, no scientific studies support this assessment, and no large population studies have been conducted” – for example, the kind of studies that might help to determine the uniqueness (or not) of human bite marks.

Issues discussed in the NAS report were many and varied. It highlighted that “only nuclear DNA analysis has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between an evidentiary sample and a specific individual or source”; that for techniques like hair analysis, “No scientifically accepted statistics
exist about the frequency with which particular characteristics ... are distributed in the population”, and that there appeared to be “no uniform standards on the number of features ... which ... must agree before an examiner may declare a ‘match’.”

The report also drew attention to the limits of fingerprint identifications, long viewed as “exact means of associating a suspect with a crime scene print”. As it explained: “The question is less a matter of whether each person’s fingerprints are permanent and unique – uniqueness is commonly assumed – and more a matter of whether one can determine with adequate reliability that the finger that left an imperfect impression at a crime scene is the same finger that left an impression (with different imperfections) in a file of fingerprints.”

Karen Kafadar, professor and chair of the Department of Statistics at the University of Virginia, was one of those involved in the preparation of the report. She writes on page 19 of this magazine that, despite the criticisms, “the authors [of the 2009 NAS report] did not say that forensic methods are invalid, unreliable, or incapable of providing useful information”. “On the contrary,” she says, “the Committee simply described the state of the peer-reviewed published research on these methods at that time, based on the type of scientific scrutiny that would be done in any area of science.”

Biting the bullet

None of the issues raised by the NAS report came as much of a surprise to the scientists and statisticians who had been involved in forensic science work in the years beforehand. Many similar criticisms had been made previously. In 2003, for example, in a piece titled “Forensic science: Oxymoron?”; Science editor-in-chief Donald Kennedy wrote: “It’s not that fingerprint analysis is unreliable. The problem, rather, is that its reliability is unverified either by statistical models of fingerprint variation or by consistent data on error rates. Nor does the problem with forensic methods end there. The use of hair samples in identification and the analysis of bullet markings exemplify kinds of ‘scientific’ evidence whose reliability may be exaggerated when presented to a jury.”

What the NAS report did was to build on such criticisms, bring them together and turn a major spotlight on the problems, says Alicia Carriquiry, distinguished professor of statistics at Iowa State University.

Carriquiry first became involved in forensic science in 1998, when she and Hal Stern were asked by the Federal Bureau of Investigation (FBI) to study bullet lead analysis, a method used for identifying suspects behind shootings. Examiners would compare the chemical composition of lead in bullets recovered from a crime scene with lead in bullets found in the possession of a suspect, and if the chemical compositions of the two samples of lead were indistinguishable, according to some thresholds determined by the FBI, then the examiner would say that the two bullets came from the same box of ammunition.

It did not take long for the statisticians to see the flaws in this approach. “Hal and I looked at this data and we said, ‘Wait a minute, you are forgetting something called the probability of a coincidental match. It’s not just that the chemical composition is indistinguishable, you need to think about how many other bullets would have an indistinguishable chemical composition but come from different boxes [of ammunition].’”

Several years later, in 2004, the National Research Council published a report, Forensic Analysis: Weighing Bullet Lead Evidence. Carriquiry and Stern’s study is among the references; on the report committee sat fellow statisticians Kafadar and Clifford Spiegelman. In summary, the committee found that the production of bullets involves “compositionally indistinguishable volumes of lead” (CIVL) and that each of these volumes of lead can produce thousands to millions of bullets – meaning “multiple people would be expected to have ammunition with the same lead composition”.

The committee determined that bullet lead analysis was sufficiently reliable to support testimony that bullets from the same CIVL “are more likely to be analytically indistinguishable than bullets from different CIVLs”, and that examiners could testify that having two bullets that are analytically indistinguishable “increases the probability that two bullets came from the same CIVL, versus no evidence of match status”. But the FBI decided to stop using this evidence type thereafter.

Statistical thinking

Scientific scrutiny of bullet lead and bite-mark analysis has resulted in a number of exonerations, as has a review of FBI hair analysis. “One of the really important things [that emerged from] the FBI hair review was the principle that probabilistic statements matter and that testimony needed to be statistically sound,” says Sarah Chu of the Innocence Project. “It raised consciousness across the forensic science community about what it means to say something is ‘unique’ or is a ‘match’.”

Chu stresses “just how important the contribution of statisticians has been in the last 10 years”. “They have moved the research community to consider these issues,” she says. “They’ve engaged forensic scientists in the actual research, they’ve educated courts, they are helping set standards for testimony. They are educating the criminal justice system on what validity and reliability mean.”

Many of these efforts at research and education have arisen as a result of the 2009 NAS report and its call for “more and better research”. Carriquiry, Stern and Kafadar, for instance, make up three-quarters of the leadership team of the Center for Statistics and Applications in Forensic Evidence (CSAFE), which was funded by the National Institute of Standards and Technology, in direct response to the NAS recommendations, to conduct research to “quantify the uncertainty, limitations and errors associated with human factors; pattern evidence, such as fingerprints and bullet marks; and digital evidence, such as data from cellphones and computers”.

Some of CSAFE’s work is described in the articles that follow, alongside wider discussion of the statistical concerns relating to forensic science. On page 16, Kafadar highlights the role of objective measures in increasing consistency and reducing uncertainty in the evaluation of evidence. Then, on page 21, Stern and others consider ways to determine the reliability and
expressed as a decimal fraction), then it tells in favour of the alternative. If the ratio is exactly 1, then the evidence is neutral. The likelihood ratio thus expresses the value of the evidence for the propositions considered. 

Ian Evett, a statistician and expert in the interpretation of scientific evidence, and a former member of the now-defunct UK Forensic Science Service, considers the likelihood ratio to be “the foundation of forensic science inference”. Put simply, it provides a way for the trier of fact – the judge or jury – to update their prior belief in a suspect’s guilt or innocence after considering the evidence. Bayes’ theorem gives the method to do that: multiply the prior odds of guilt by the likelihood ratio to produce posterior odds. 

According to the forensic statistician Colin Aitken, the introduction of Bayes’ theorem represents “the major contribution by statistics to the development of forensic science”. He says: “Bayes’ theorem provides a rigorous approach to the measurement of the effect (value) of evidence in updating the prior beliefs in the prosecution and defence propositions. The theorem has enabled the rarity of the evidence and the similarity of control evidence (evidence whose source is known) and recovered evidence (evidence whose source is not known and which may be from the same source as the control evidence) to be evaluated together on a continuous scale.”

The use of likelihood ratios in court is “a battle that has been won in the UK and Europe”, says Evett. Indeed, on page 42 of this issue, Gillian Tully, the Forensic Science Regulator for England and Wales, tells how likelihood ratios form a cornerstone of her draft standard for the interpretation of evidence – a standard that was developed with the input of statisticians and members of the Royal Statistical Society’s Statistics and Law Section. 

But in the USA, the forensic science community remains to be convinced. Says Carriquiry: “Forensic scientists have been very reluctant – I’d say, dead-set against – admitting that this likelihood ratio framework is a reasonable way to evaluate evidence, because they know we are still not close to being able to compute likelihood ratios for many types of evidence – pattern evidence, for example – and they wonder what in the world they are supposed to do in the meantime.”

The sticking point seems to be knowing what probabilities to assign to the prosecution proposition (the numerator) and the defence proposition (the denominator). As Mejia et al. explain on page 27, different denominators “could yield drastically different estimates” of the likelihood ratio. Ideally, they write, “the denominator should be estimated from population reference data”. 

However, others would say that the absence of population reference data is not an argument against using likelihood ratios. Consider this from Robertson et al.: “The fact that one has imprecise or incomplete data does not negate ordinary logical relationships such as those expressed by Bayes’ theorem. To estimate the area of a rectangular field, it is still logical to multiply the estimates of length and breadth, even if they are only estimated by eye … [T]he structure of likelihood
ratios reflects the logical structure of the argument and that is not affected by the quality of data available.”

Evett, meanwhile, takes issue with the idea that “improving” forensic science “is just a matter of getting good databases and using them statistically to remove subjectivity and replace it with objectivity”.

“Subjectivity is central to forensic science inference,” he says. “If we have a relatively simple situation where we have a database which can give us numbers, even if we use that data and use the cleverest statistical modelling to assign weights of evidence based on that database, overarching all of that is the judgement of the scientist in relation to the relevance of the database and the relevance and limitations of the model. There is no such thing as a perfectly relevant database – and, of course, all models are wrong.”

Knowledge calibration
To Evett, the judgement, knowledge and understanding of the forensic scientist is paramount – but it should not be “unfettered knowledge”, he says. Rather, knowledge should be managed effectively through training and calibration. At a 2015 Royal Society discussion meeting, Evett put forward the argument that “the reliability of scientists’ opinions should be judged, at least in part, by their performance under controlled conditions”.6

In that way, he says, “when the expert goes to court, the expert doesn’t say, ‘Believe me because I’ve been doing this job for 40 years’; they can present a documented set of studies in which they have participated, and this shows what their performance is like”. He cites the “ground-breaking work” carried out in the field of handwriting comparison over many years in Australia and New Zealand under the leadership of Found and Rogers.7

Studies akin to these have shown the performance of experts to be mixed. In the area of bite-mark analysis, for example, the Washington Post reports (wapo.st/2EBQ5fh): “In 1999, members of the American Board of Forensic Odontology met to test their technique, attempting to match four bite marks to just seven dental models (as opposed to comparing a single bite mark against an ‘open population’ of thousands, even millions, as most forensic dentists … must do). They turned up false positives 63.5 percent of the time, according to the National Research Council. A 2001 study put the rate of false identifications lower – between 12 and 22 percent – but still too high to be considered conclusive. And [in 2014], a study of forensic odontologists found that the analysis couldn’t even determine which marks were bite marks.”

Let us imagine that this sort of performance data was known about the experts who testified in Steven Mark Chaney’s trial for murder. Would the outcome have changed had the jury known that dental experts sometimes (perhaps often) match the wrong sets of teeth to the wrong bite marks? Might Chaney have avoided almost three decades in jail as a result?

That is impossible to know for certain. However, the court that exonerated Chaney last year concluded that the case against him “would have been incredibly weakened” if we knew then what we know now about bite-mark evidence. All that a jury might take from such evidence, the court said, is that if the victim was bitten by someone in the days before the murder, “that person could be Chaney (or anyone else on the planet whose dentition has not been excluded)”. A likelihood ratio, based on reliable knowledge available at the time, might be of the order of 1, which is entirely uninformative.

As demonstrated by the case of Steven Mark Chaney, the 2009 NAS report forced many in the US forensic science community to confront what they know about forensic science evidence and to question what they can say about it. But, 10 years on from the NAS report, “we still do not know what we do not know”, says senior United States Circuit Judge Harry Thomas Edwards.

“We need better scientific studies and standards to shape the work of forensic practitioners and regulate the admission of forensic evidence.” he says, in a statement published by the Innocence Project (bit.ly/2EZ8Xe1). “This means that more top scientists must engage in research on forensic methods and appear in court to explain the evidence. This will allow judges to better understand forensic evidence and to more clearly and accurately instruct jurors on the limits of the evidence.”

References