2157118X.4.1.R3 Peer Reviewed
Received 31 May 2014
Received in revised form 7 July 2014
Accepted 7 July 2014
Available online 28December 2014

On Limiting the Use of Bayes in Presenting Forensic Evidence

Norman Fenton [a], ***Martin Neil*** [b]

a Professor and Director of Risk and Information Management Research Group (Queen Mary University of London) and CEO Agena Ltd. Email: norman@eecs.qmul.ac.uk.

b Computer Science and Statistics (Queen Mary University of London) and CTO Agena Ltd. Email: martin@eecs.qmul.ac.uk.

Abstract A 2010 UK Court of Appeal Ruling (known as "R v T") asserted that Bayes theorem and likelihood ratios should not be used in evaluating forensic evidence, except for DNA and 'possibly other areas where there is a firm statistical base'. The potential impact of this ruling is enormous and it has drawn fierce criticism from expert witnesses, academics and lawyers, who have identified various weaknesses and fallacies in the ruling. This paper focuses on the strategic and cultural challenges that the ruling raises to ensure that the role of Bayes is better understood and exploited in the presentation of forensic evidence. We provide a simple unifying way of describing all probabilistic forensic 'match' evidence; this enables us to easily identify and avoid the kind of common misunderstandings and fallacies that have afflicted probabilistic reasoning about evidence, including especially why it is irrational to assume that some forensic evidence is 'statistically sound' whereas other less established forensic evidence is not. But these misunderstandings are not restricted to lawyers, since we show that both forensic scientists and even Bayesian experts have consistently failed to include all relevant information in their evidence, such as error probabilities, and this applies to DNA as much as any other forensic science. We also show that there are severe limits of the extent to which the results of Bayes can be presented in purely intuitive terms; we show that the scope in forensics is even narrower than previously assumed. Hence, there are two major challenges facing the opponents of the R v T ruling: First, there must be much greater awareness of the need to improve Bayesian forensic arguments (before they are even presented in court) in order to avoid the common errors and omissions that are made. Second, there must be a radical rethink on the strategy for presenting the results of Bayesian arguments in court. Resorting to the formulas and calculations in court is a dead end strategy since these will never be understood by most lawyers, judges and juries, but the intuitive presentations simply do not scale up. Ultimately this means getting the lay observers to 'accept' that they need only question the prior assumptions that go into the Bayesian calculations and not the accuracy or validity of the calculations given those assumptions. Bayesian networks may provide a suitable mechanism for performing these calculations.

Keywords: Forensic science, Bayes, Evidence.

1 Introduction

Proper use of probabilistic reasoning has the potential to improve dramatically the efficiency and quality of the entire criminal justice system. Bayes theorem is a basic rule, akin to any other proven maths theorem, for updating the probability of a hypothesis given evidence. Probabilities are either combined by this rule, or they are combined wrongly. Yet, the Court of Appeal in the case of R v T [1] ruled that the use of formulas to calculate probabilities and reason about the value of evidence was inappropriate in the area of footwear evidence. It regarded the forensics of footwear matching as 'unscientific' and not having a sufficiently 'firm statistical base' in contrast to DNA forensics. Specifically, Points 86 and 90 of the ruling respectively assert:

"..*We are satisfied that in the area of footwear evidence, no attempt can realistically be made in the generality of cases to use a formula to calculate the probabilities. The practice has no sound basis*".

" *It is quite clear that outside the field of DNA (and possibly other areas where there is a firm statistical base)*

this court [1] *has made it clear that Bayes theorem and likelihood ratios should not be used*"

Given its potential to change the way forensic experts analyse and present evidence in court, experts have been understandably quick to publish articles criticising the ruling. At the time of writing there have already been at least four such excellent articles [12],[29],[31],[32] that provide a detailed analysis of the case and ruling. These papers recognise that there were weaknesses in the way the expert presented the probabilistic evidence (in particular not making clear that likelihood ratios for different aspects of the evidence were multiplied together to arrive at a composite likelihood ratio), but nevertheless express deep concern about the implications for the future presentation by experts of forensic evidence. The papers recognise positive features in the ruling (notably that experts should provide full transparency in their reports and calculations) but they provide compelling arguments as to why the main recommendations stated above are problematic. For example, [32] uses the following analogy of likelihood ratio calculations with area calculations:

Saying the expert should not use this 'mathematical formula' to assess the composite likelihood ratio is like saying that if one is just estimating by eye the area of a field, one is not allowed to multiply estimates of its width and length together. Clearly it is the correct procedure: there is no uncertainty in the relationship between length, width, and area, only in their values. If the Court were to say that the expert was not to use a logical procedure, rather than a 'mathematical formula', the flaw in its reasoning would be obvious.

1 The judge is actually referring to the Court of Appeal ruling in the case of Adams, which is mentioned in Point 89. ■

The authors in [32] also conclude that:

..the Court has not understood the difference between assessments of the probability of a proposition and of the strength of evidence for the proposition; the second is a confusion between uncertainty in the values of the variables and uncertainty in their relationship in a mathematical formula. The fact that variables cannot be precisely expressed does not affect the validity of the relationships described by the formula.

The authors in [31] highlighted the inconsistency in the ruling which, on the one hand rejects the use of Bayes and likelihood ratio calculations, while on the other hand insists on full transparency of all calculations. They ask:

..how could such an injunction ever be enforced on forensic scientists … The best that might be imagined would be a policy of "don't ask, don't tell", whereby experts formulated their conclusions according to their good faith understanding of scientific protocol but carefully concealed their "deviant" probabilistic reasoning from legal scrutiny.

On a similar theme the authors in [11] assert that:

…the evaluation of evidence for a court of law is not just a matter of "using likelihood ratios" but one of working to a set of principles that are founded on logic. To deny scientists the contemplation of the likelihood ratio – whether quantitative or qualitative – is to deny the central element of this logical structure

Clearly, as pointed out in [32], the ruling in [1] exhibits misunderstandings of some fundamental ideas of probabilistic reasoning and even includes instances of the fallacy of the transposed conditional, despite the dozens of papers and even rulings about it over many years. That such errors should continue to be made routinely by members of the legal profession (see also [19] for other recent examples) indicates that we (meaning the community of experts in probabilistic reasoning) have failed to communicate our arguments effectively where it matters most. In Section 2 we explain the challenges that this failing poses for expert witnesses and Bayesians. The rest of the paper addresses the challenge and is structured as follows:

- In Section 3 we introduce a hypothetical forensic 'science' in order to present the core ideas of forensic match evidence in a simple unifying way. This enables us to explain in very simple terms the Bayesian approach and to expose not just the fundamental misunderstandings in the R v T ruling, but also a number of key issues that have been missed in previous discussions.
- In Section 4 we use the generic example to highlight the irrationality of the core message in the R v T ruling (namely that there can be a clear distinction between forensic methods that are or are not 'statistically sound' and different allowed reasoning applied).
- While Sections 3 and 4 expose the weaknesses in the R v T ruling, Section 5 explains why, in many ways, the ruling is perfectly understandable, since we show that forensic probabilistic evidence is usually presented in a confusing -, and often incorrect - way. In particular, forensic scientists and even Bayesian experts typically ignore (or do not properly articulate) the potential for testing errors (false positives and false negatives).
- Hence in Section 6 we show that, when the potential for testing errors is included (as

it should be) this introduces significant complexity even in very simple cases. The key point is that, even in the simplest case, it is unrealistic ever to expect the associated Bayesian argument to be understood by lay people. We explain how the use of Bayesian network models may potentially address this problem.

- Finally, in Section 7 we present the grand challenge that Bayesians need to address before Bayes can ever take (what Bayesians feel should be) its rightful central position in legal reasoning.

2 The main challenges for expert witnesses and Bayesians

While the various papers on the ruling in [1] have done a fine job analysing in depth the weaknesses contained therein, there should be no doubt that the ruling is a damning indictment of the community of experts and academics who recognise the central importance of Bayesian reasoning for evidence evaluation. Despite some twenty-five years of work explaining the power and relevance of Bayes to the law, (resulting in several hundred academic publications and dozens of textbooks) the actual impact on legal practice has been minimal.

This failure must be attributed to our inability to communicate the core ideas in such way that they are accepted as a standard tool of the trade rather than as they are perceived now by much of the legal profession: an exotic, somewhat eccentric method to be wheeled out for occasional specialist appearances whereupon a judge or lawyer will cast doubts on, and even ridicule, its integrity (hence ensuring it is kept firmly locked in the cupboard for more years to come).

To address the problem we need to communicate the core ideas more effectively to both forensic scientists and lawyers. Specifically, we need to ensure that:

a. both the forensic scientists and lawyers know when Bayesian reasoning should be used.

b. the forensic scientists are able to properly articulate the assumptions required for a Bayesian analysis.

c. both the forensic scientists and lawyers know the difference between the assumptions required for the analysis (which will generally be disputed) and the Bayesian calculations that determine the conclusions based on the assumptions (which must not be disputed).

d. before evidence is used, the forensic scientists are able to perform the Bayesian calculations correctly and efficiently. The scale of this problem has been massively underestimated, and as we shall explain in this paper, can only be resolved by more widespread acceptance of the use of tools.

e. the forensic scientists (and ultimately the lawyers themselves) are able to present the results of Bayesian reasoning about evidence in a way that is understandable to jurors and other lawyers. This is the most difficult challenge of all since, ultimately it will only be achieved once it is accepted that we do not actually have to reason in court about the results of the Bayesian calculations themselves (i.e. the calculations are accepted in the same way as we might accept the results of using a calculator for long division[19]).

f. likelihood ratios (or some suitable graphical/verbal equivalent representation) are used as a standard means for stating the value of evidence (individually and in combination).

To see the extent of how and why we have failed to meet the above objectives we need only look at the range of relevant textbooks:

- There are two standard textbooks, [24] and [27], for forensic science training. Despite its apparently encyclopaedic coverage, [27] contains nothing at all on Bayes and only some basic high school material on statistics such as graphs and bar charts. The book [24] does contain a very brief introduction to Bayes and the likelihood ratio right at the end, but without attempting to link it in any way to the core material of the book (so that it appears as an afterthought, out of context).
- There is one standard book, [37], aimed at forensic scientists presenting evidence in court. Until its latest 2010 edition, this book did not contain any mention of Bayes, likelihood ratios, or even probability, and so failed to consider such basic issues as random match probability and the probability that tests may have less than perfect accuracy (more encouragingly, the new 2010 edition does contain a chapter on trace and contact evidence [14] that includes a discussion of the Bayesian approach).
- There are several excellent books that focus on the statistical and probabilistic aspects of forensic evidence. These include [8], [10], [11], [13], [18], [21], [28], [29]. These books cover exactly the right material in depth, and they also include introductory material on Bayes. However, they are most suited for people with a statistical or mathematical background (who wish to find out in detail how to properly reason with forensic evidence) rather than practicing forensic scientists lawyers. So, for example, even those that are considered the most accessible to non-

experts, namely [8], [18], [29], make extensive use of formulas and hence require a significant level of mathematical sophistication. The books also tend to focus on the details of specific types of forensics (especially DNA).

- There are no suitable relevant books we are aware of that are specifically targeted at lawyers. The closest would be populist books on probability and risk, such as [22] and [23], but these do not address the issue of evidence presentation.

In [19] [20] we argued that it was a mistake to assume that any kind of Bayesian formulas - such as those used in the case of R v Adams (and shown in Figure 1) could be presented to lawyers and juries no matter how 'simple' they appeared to statisticians.

In the relevant text books and papers discussed above the best approaches start with visual explanations of a very simply instance of Bayes (using, for example, tree diagrams with frequentist versions of the probabilities). However, for reasons we will explain in Section 3 below, these visual approaches do not scale up meaningfully in any realistic situation. It is at this point that the various authors normally resort to the formulas instead; hence, this is the point that most forensic scientists and lawyers never get beyond.

3 Clarifying the notions of 'forensic match' and common fallacies

To help readers understand that there is a simple unifying way to present any kind of forensic 'match' evidence we use a hypothetical (but not unreasonable) example of a completely new forensic science, which we call 'stature matching'. This avoids the problem of getting distracted by the details and biases of specific areas (such as shoeprint matching or DNA matching). This approach will enable us to expose numerous common misunderstandings about the meaning of match evidence and that, contrary to what the judge ruled in [1] (and indeed what forensic many experts assume), it is inappropriate to assume that certain methods are inherently 'scientifically sound' and others are not.

3.1 A new, but typical, forensic science: Stature matching

Our 'new' forensic science is called "stature matching". Stature matching assumes that, for any person, we can measure the following features:

- Sex (male, female)
- Height (in centimetres)
- Waistline ((in centimetres)

So each person has their own stature profile such as:

- (male, 131, 65)

The 'science' of stature matching is the ability to determine a person's stature profile accurately. They can do this either directly by observing and measuring the person or indirectly from an image of the person. If, for example, CCTV captures the image of a man at the scene of a crime (we can think of the image as a 'trace' left by the man) then stature matching scientists might determine that the trace has the following stature profile:

- (male, 132, 64)

A real person is said to be a 'match' to the stature profile of the trace if the following criteria are satisfied:

- Sex of the person = sex of the trace stature profile
- Height of the person differs from height of the trace stature profile by less than 2 centimetres
- Waistline of the person differs from waistline of the trace stature profile by less than 2 centimetres

So, for example, four different people with respective stature profiles

- (male, 132, 64)
- (male, 132, 64)
- (male, 131, 65)
- (male, 132, 65)

would all be considered to be a 'match' to the stature profile (male, 132, 64), whereas people with the following stature profiles would not be considered a match:

- (male, 135, 65) - this 'fails' on height
- (female, 132, 65) – this 'fails' on sex

Every branch of forensic matching that is based on some properties of people [2] (be it DNA, fingerprint, blood type, shoe-print, earprint, Gait, voice, ….and any other type not yet invented) is based on the same underlying principles as stature matching:

Specifically:

- Every person has a 'profile' (defined by the area of forensics) that can be
- measured by some defined procedure.
- In certain circumstances a person leaves a 'trace' (or 'print') of this profile
- In certain circumstances we can measure the profile of the trace that was left.
- There is a criterion for

$$V=\frac{Pr(H_p|E,I_1,I_2)}{Pr(H_d|E,I_1,I_2)}=\frac{Pr(E|H_p)}{Pr(E|H_d)}\times\frac{Pr(I_1|H_p)}{Pr(I_1|H_d)}\times\frac{Pr(I_2|H_p)}{Pr(I_2|H_d)}\times\frac{Pr(H_p)}{Pr(H_d)}$$

Fig 1. Typical Bayesian likelihood ratio calculation. Far too complex for lay people to understand

2 Other types of forensic match analysis, such as glass, fibres, pollen etc, are not concerned with attributes of people and do not exactly fit the same framework. ■

determining whether a trace profile matches the profile of a person.

The first simple (but extremely) important observation to make about forensic matching is that (in contrast to widely held assumptions) there is no definitive means for considering a forensic matching method to be 'scientific' or not. Most people assume that DNA is 'scientific' because the measurement and matching criteria and protocols are objective and reasonably standardised (in contrast to those that are widely assumed to be 'non-scientific' like gait analysis, face mapping, and fingerprinting). Yet, our new stature matching method is at least as scientific as DNA in this respect. For example, in stature matching we insist on always measuring the three specific values (sex, height, and waistline) and never any others; we can always assume that the height and waistlines are measured without clothes or shoes, and we always include the 2cm error margins for the match. There is no fundamental reason why any forensic method cannot in principle be made 'scientific'.

The second extremely important observation to make about forensic matching is the following (see [14], [26], [34] for a comprehensive discussion of this issue):

- A 'match' never means a unique identification of a person.

This is important because the assumption of uniqueness is a common fallacy arising in DNA, fingerprint, and many other areas of forensics. For example, in R v Kempster, EWCA Crim 975 [3] the ruling includes the following assertion about earprint evidence:

It is clear … that ear-print comparison is capable of providing information which could identify the person who has left an ear-print on a surface.

This assertion is highly misleading. In fact, when we find a 'match' (be it for stature matching, earprint matching, DNA or any of the areas of forensics discussed above) all we can conclude is that within the agreed criteria, the person's profile is the same as the profile of the trace. To equate this notion with 'identification' is always flawed.

An expert in stature matching could, in court, present the information about a match as follows:

"*I am absolutely certain that the stature profile of the trace found at the scene is a match of the defendant's stature profile.*"

Instead, the common error made by experts is to assert the following:

"*I am absolutely certain the stature profile trace found at the scene is that of the defendant*"

Indeed, this was exactly the error made by the expert witness on earprint evidence in R -v- Dallagher, EWCA Crim 1903 [4]. The judge consequently rejected the entire earprint evidence as inadmissible. While the judge's ruling was understandable in this particular case it would be extremely dangerous to interpret this as meaning that, unless a 'match' is the same as an 'identification', then match evidence can never be admissible. For not only would this rule all future earprint evidence as inadmissible, it would also rule as inadmissible every area of forensic match evidence.

3.2 Understanding the Bayesian approach to match evidence

In the simplest use of forensic match evidence in legal cases we assume that a person has left a trace at a particular location. Then we have the following (continuing with the stature matching example):

- Source profile: This is the stature profile of the trace found at the location.
- Target profile: This is the stature profile of a particular person believed (normally called the defendant) who some believe may have been the one who left the trace.

Let us, for the time being, make a massive simplification (it turns out that it is ONLY for this restrictive case that a simple explanation of Bayes is possible). We will assume that our stature testing is perfect. So, someone with type (male, 131, 65) will always be tested to be of type (male, 131, 65) and someone who is not type (male, 131, 65) will never be tested to be of type (male, 131, 65) .

With the above assumptions our typical simple forensic case amounts to the following:

- Prosecution hypothesis (H): "The target is the source" (i.e. the defendant is the person who left the trace at the scene).
- Defence hypothesis (not H): "The target is not the source" (i.e. a person other than the defendant left the trace at the scene).
- Evidence E1: The source profile type is known, say to be of type (male, 132, 64). For simplicity and generality we shall refer to a particular profile as type X.
- Evidence E2: Target profile matches the source profile (i.e. both have type X).

From an evidential perspective, the 'value' of the evidence is therefore completely determined by the following two pieces of (probabilistic) information:

1. 'Defence likelihood': How likely are we to see the evidence if the defence hypothesis is true. In other words how likely is it that the source and target (defendant) are both of type X, if the target was not the source.

   With the above simplistic assumptions, the defence likelihood is represented by the single branch (H false, E1 true, E2 true) in Figure 2. Suppose *m* is the proportion of people in the population who have type X. This is sometimes called the

frequency (of the particular type) or the *random match probability* (of the particular type). So, the defence likelihood is equal to m^2.

2. 'Prosecution likelihood': How likely are we to see the evidence if the prosecution hypothesis is true. In other words how likely is it that the source and target are both of type X if the target is the source.

With the above simplistic assumptions, the prosecution likelihood is represented by the single branch (H true, E1 true, E2 true) in Figure 2. Hence, the prosecution likelihood is simply equal to *m* (because our testing is perfect the target is certain to be of type X if the target is the source).

So, if the random match probability *m* is equal to 1 in a 100, then the prosecution likelihood is 100 times greater than the defence likelihood. In fact, we are 100 times more likely to observe the evidence if the prosecution hypothesis is true than if the defence hypothesis is true.

The likelihood ratio (the prosecution likelihood divided by the defence likelihood) is simply the mathematical formalism that expresses exactly this intuitive information.

The likelihood ratio is very well-suited to the legal context because it enables us to evaluate the impact of the evidence without having to specify what our prior belief is in the prosecution or defence hypothesis. What Bayes theorem additionally tells us is that, whatever our prior odds were for the prosecution hypothesis, the result of seeing the evidence is such that those odds are multiplied by the likelihood ratio[3]:

Posterior odds = Likelihood ratio × Prior odds

So, according to Bayes, if we started off assuming that the odds in favour of the defence hypothesis were 1000 to 1, then the 'correct' revised belief once we see the evidence is that the odds still favour the defence, but only by a factor of 10 to 1:

	Prior odds		*Likelihood ratio*		*Posterior odds*
Prosecutor	1	×	100	=	1
Defence	1000		1		10

And if we started off assuming that the odds in favour of the defence hypothesis were 4 to 1, then the 'correct' revised belief once we see the evidence is that the odds now favour the prosecution by a factor of 25 to 1:

	Prior odds		*Likelihood ratio*		*Posterior odds*
Prosecutor	1	×	100	=	25
Defence	4		1		1

3 Note the following (which we will assume later): If we assume that the prior odds are 'evens' i.e. 50:50 then the posterior odds will be the same as the likelihood ratio. Also odds can easily be transformed into probabilities: specifically, if the odd are x to y for hypothesis H over not H then the probability of H is x/(x+y) and the probability of not H is y/(x+y). So odds of 100 to 1 in favour of H means the probability of H is 100/101 and the probability of not H is 1/101. ■

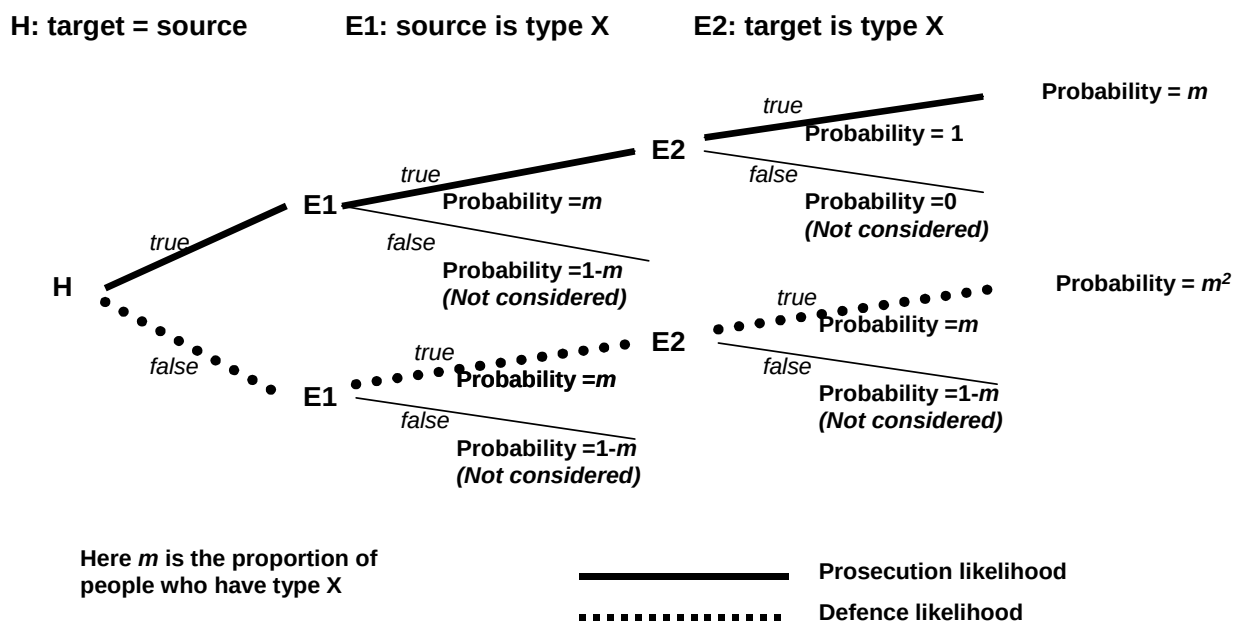


Fig 2. Determining the possible scenarios and likelihoods in simple case

But why should we accept that Bayes is the 'correct' interpretation? The standard way to convince lay people that Bayes is correct is to consider examples (often called the 'Island' example) like the following:

> ***Example 1:*** Suppose that, in addition to the defendant, it is known that another 1,000 other people were in the vicinity of the crime scene[4] – see Figure 3. Then our prior assumption, i.e. what we should assume before any evidence has been presented, is that any one of these other people is just as likely to be the person who left the trace as the defendant. So the prior odds are 1000 to 1 in favour of the defence hypothesis (or equivalently the probability that the defence hypothesis is true is 1000/1001). Since the random match probability is 1/100, we expect about 10 of the other 1000 people to have the type X. So, once we observe the evidence (defendant is type X) we can rule out all other people, except those 10, as having possibly left the trace. So, after observing the evidence the defendant and 10 others remain as possibilities. So the revised odds are now 10 to 1 in favour of the defendant (or equivalently the probability that the defence hypothesis is true is now 10/11). So, although the odds still favour the defence hypothesis the odds have swung by a factor of 100 (the likelihood ratio) towards the prosecution hypothesis.
>
> If we change the number of people we start with the odds still always swing by a factor of 100 (the likelihood ratio) towards the prosecution hypothesis. So, if there were 500 other people then we expect about 5 to have the same stature type. So the prior odds, which in the case are 500 to 1 in favour of the defence, drop to 5 to 1 after observing the evidence

If there were just 10 other people then the use of population diagrams such as in Figure 3 to represent Bayes becomes difficult because, in this case, the expected number of people who match is a fraction (one tenth) of a person. From a mathematical perspective this is not a problem: the prior odds are 10 to 1 in favour of the defence. After the evidence there is just 1/10 of another person other than the defendant. So the odds are now 10 to 1 in favour of the prosecution hypothesis. The swing is still a factor of 100 toward the prosecution. But this example shows that, even with the most simplistic assumptions we have made the standard explanation of Bayes and likelihood ratios may not be easily understandable to lay people. Because many types of forensic science (such as DNA) have very low match probabilities, it is inevitable that we have to consider 'fractions' of people if we adopt this approach. The trick to gaining acceptance from lay people is therefore to use hypothetical examples that do not involve fractions, and then explain that exactly the same method works no matter what the actual match probabilities are.

3.3 Exposing some common misunderstandings

Before tackling the core problem of what constitutes 'statistically sound' evidence it worth noting that the framework we have provided makes it easy to expose three common misunderstandings in probabilistic reasoning about evidence:

When likelihood ratios can and cannot be multiplied

The practice of multiplying likelihood ratios was explicitly criticised in [1]. The error in the ruling was the failure to understand and

4 In the standard 'Island problem' presentation it is assumed that the crime was committed on an island and that, in the absence of evidence, all residents are equally likely suspects. ■

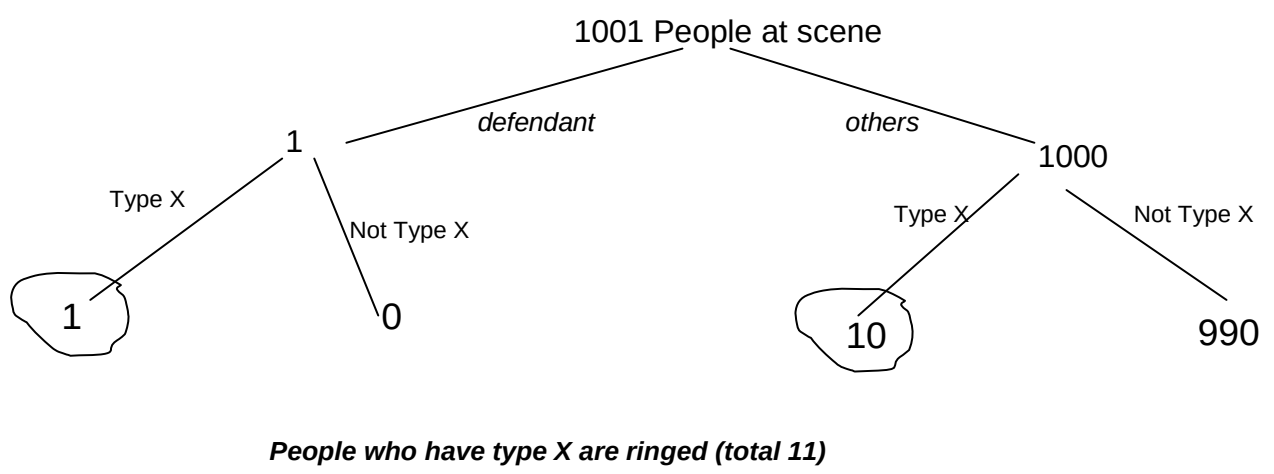


Fig 3. Bayes calculation explained visually

distinguish between the circumstances when multiplying likelihood ratios was and was not the correct thing to do.

When there are two pieces of independent evidence then multiplying likelihood ratios is the only correct way to reason about the impact of the combined evidence. To see why, suppose, that in addition to a match of the defendant's stature profile, we also discover a match of hair colour; the defendant and the person at the scene have brown hair. Suppose that the random match probability for brown hair is 1 in 5. Then the evidence in this case is that the stature profile and the hair profile of the defendant both match that of the person who left the trace (in the form of a CCTV image) at the scene. Since stature and hair colour can be considered independent, the probability of seeing both matches given that the defendant was not the person who left the print is the product of the two random match probabilities, i.e. 1/500. Hence the likelihood ratio is now 500. Assuming there are 1000 other people who were at the scene, it follows that 10 of these is likely to have the same stature profile as the defendant and of those 10 two are likely to have the same hair colour as the defendant. This means the odds in favour of the defence hypothesis have come down from 1000 to 1 to 2 to 1. That is a factor of 500, which is equal to the product of the two likelihood ratios (100 times 5).

So, when two pieces of evidence are genuinely independent it would, contrary to the ruling in [1], be irrational not to multiply the likelihoods - even for such 'unscientific' forensics as stature matching and hair colour.

However, the ruling against multiplying likelihood ratios is perfectly justified if the defence was unable to demonstrate that the underlying pieces of evidence were independent. If, for example, instead of hair colour we chose 'weight' it would certainly be wrong to conclude that weight was independent of stature. In such circumstances there are standard, but different, Bayesian calculations that need to be used (we have to consider explicitly the probability of one piece of evidence given the other). But such a scenario already puts us into the realms of problem complexity beyond which it is reasonable – or even possible – to perform manual calculations that lay people would be able to understand intuitively.

Fallacy of the transposed conditional.

This occurs when the defence likelihood, i.e. the probability of seeing the evidence given the defence hypothesis, is wrongly assumed to be equivalent to the probability of the hypothesis given the evidence.[5]

So, suppose we know that the defence likelihood is 1/100. By wrongly assuming this is the same as the probability of the hypothesis given the evidence, a prosecutor might state

- "The probability the defendant was not at the scene given this match evidence is 1 in 100"

In fact, if our prior was 1000 to 1 in favour of the defence hypothesis (as in Example 1 above) it turns out that what should have been stated was:

- "The probability the defendant was not at the scene given this match evidence is 10 in 11"

The danger of reading too much into very low match probabilities

For DNA the probability is normally presented as being so low (for example, 1 in 2 billion) that is it as 'good as' equal to zero[6] and hence a match is (wrongly) considered as a unique identification. In the case of fingerprints the situation is even worse, since there is still a strong assumption by many that a match is, by definition, a unique identification (i.e. the random match probability is assumed to be equal to zero).

Recent research, such as [16], has exposed this fallacy for fingerprint evidence and this was best exemplified by the dramatic Mayfield case [6] where a fingerprint match was subsequently discovered not to be that of the defendant. Primarily on the basis of this instance of a known match 'error', a State of Maryland Court subsequently ruled that fingerprint evidence was not admissible in a totally unrelated murder case [7]. If that way of thinking was applied to DNA or any other type of forensic evidence, then any example of a 'match' in which the person deemed matching was NOT the one who left the 'print', would be justification for rejecting as inadmissible the whole of that field of forensic evidence.

4 The irrational notion of 'statistically sound' evidence

Having dealt with some of the misunderstandings and fallacies in rulings such as [1] we now turn to the most critical and challenging misunderstanding that lies at the heart of the ruling: the assumption that the random match probability is 'statistically sound' for some areas of forensic science and not others. We again expose the weakness of this assumption by using our hypothetical stature matching example.

For any forensic science the match probability is based on some database of profiles. For our new science of stature matching we therefore need a database of people's stature profiles. For a particular

5 So, using the language of statisticians P(H|E) is wrongly assumed to be equal to P(E|H) hence why it is referred to as transposing the conditional.

6 This is especially true of the FBI in the US. In the UK the Forensic Science Service no longer assumes this, although lay people and many lawyers do. ■

profile, say (male, 132, 64), we simply count the frequency of profiles in the database that would be classified as a match to this profile. So this would include profiles like: (male, 132, 64), (male, 131, 65), (male, 132, 65), etc. If there are 1000 such matches in a database of 100,000 then we can say that the random match probability is 1 in a hundred, or equivalently 0.01.

The 'reliability' of the database could, of course, be questioned on numerous grounds such as the following:

- If the crime happened in the UK and the database comes from the USA then it may not be representative; perhaps people in the UK are smaller
- If it is known that the person who left the print was definitely a man then perhaps we should we consider only a database containing stature profiles of men.
- If it is known that the person who left the print was definitely Caucasian, then perhaps we should consider only a database containing stature profiles of Caucasians.
- If we are '90%' sure that the person who left the print walked with a limp, then perhaps we should consider only a database in which 90% of the stature profiles belong to people who walk with a limp
- Etc.

Clearly even if we were able to change to a more 'representative' database or restrict the existing database to people with the relevant criteria (and normally this is not possible because the database will only contain the stature profiles and few other details) the random match probability will also change. Hence it is impossible to assume that there is a truly objective random match probability (what makes a measurement objective or subjective is the supposed level of rigour of the measurement instrument). But, all of these issues are inevitable for any database for any area of forensic science. In other words there are no objective criteria by which our stature matching database could be ruled as any less 'reliable' than the most sophisticated DNA database[7]. This fictional example exposes exactly the kind of questions that need to be asked about any forensic database (including DNA databases), but which rarely are. Indeed, as described in [13] and [17], because of very different databases and different assumptions about how to use them, DNA experts in the UK and the USA report very different random match probabilities for the same person (often many orders of magnitude different such as one in a billion compared to one in a trillion). These differences, even when the probabilities are so low, matter greatly as we have already shown (and matter even more when we factor in the possibility of testing errors as we show in the next section).

Contrary to what was argued in [1] the 'statistical base' for determining the defence likelihood in stature matching is no less well defined than it is it for DNA. In fact it is actually much easier to get a relevant database, easier to do the matching, and easier to explain to a jury precisely what the match probability means. The match probabilities are as well defined (in fact less subjective) than those in the 'mature' science of DNA.

The 'scientific' quality or maturity of the type of forensic science being considered is therefore irrelevant as far as the statistical argument is concerned. The level of 'scientific' or 'statistical' quality is certainly not synonymous with very low defence likelihood figures. This point is important because there is a misconception that DNA evidence is scientific because it produces very low defence likelihood figures, while earprint or footprint evidence is less scientific because it rarely produces very low defence likelihoods. The value for the defence likelihood actually has nothing to do with the reliability of the data.

What matters is that in all cases of a match (whether it be DNA, fingerprinting, footprints, earprints, stature matching or anything else) the expert should be obliged to present the random match probability (possibly as a range) along with a statement about the limitations of the underlying data. For example,

- "The probability of finding this match in a person who was NOT the one who left their stature print at the scene is between one in a thousand and one in two thousand. This figure is arrived at from a database of 100,000 stature profiles of which 150 match the print at the scene."

The defence likelihood is inevitably a statement of subjective probability, as is any statement involving uncertainty.

So, given that there is no rational basis for declaring DNA 'statistics' as more 'scientific' than any other type of forensic match evidence, the prohibition from using likelihood ratios and Bayes on all but "DNA (and possibly other areas where there is a firm statistical base)" [1] makes no sense. The only consistent strategy would be to either allow its use for all forensic match evidence or to ban it for all (including DNA).

Clearly our argument is that the former should apply. To support this

7 It is, of course, important to note that the databases that provide a basis for the frequency statistics for DNA cases are far more comprehensive than for most other areas of forensic science, and this is presumably what the Judge in R v T was recognising. However, that does not alter the fact that DNA is not inherently more or less scientific than other areas of forensics currently lacking extensive databases. ■

we can point to the examples we have already provided where Bayes provided the correct results that match our intuition. But an even more convincing argument is to show that banning it for all arguments would mean that we would have to reject all statistical analysis as the following example should make clear:

- ***Example Case 1:*** A man is charged with a gaming offence, specifically that he was using a rigged coin when taking bets on whether the coin he was tossing comes up Tails. The defence hypothesis is that it was a fair coin. The prosecution hypothesis is that the coin was double-headed (so the punters were always sure to lose). The evidence E is that the coin landed as Heads on 9 out of 9 plays.

The point about this example is that the evidence is not only purely statistical, but that the statistics involved – coin tossing – allow us to use classic frequentist analysis and hence avoid any debates between Bayesian and non-Bayesian statisticians. Thus, everybody will certainly agree on the following:

- The defence likelihood is 1/512 (a half to the power of 9) because that is the probability of seeing 9 out 9 Heads given that the coin is fair. This analogous to the random match probability in a forensic case.
- The prosecution likelihood is is equal to 1, because that is the probability of seeing 9 out 9 Heads given that the coin is double-headed.

It is clear that the evidence favours the prosecution hypothesis more than the defence hypothesis. Moreover, the likelihood ratio of 512 can be proved to be the 'correct' factor in favour of the prosecution hypothesis. For, suppose that before the game was played a double-headed coin was added to a bag of 1000 coins that were known to be fair. Suppose also that the coin played in the game was selected randomly from this bag. Then before we see the evidence the odds must favour the defence hypothesis by a factor of 1000 to 1 (these are just the odds of selecting the double-headed coin). We know that there is a 1 in 512 chance of tossing 9 out of 9 heads in a fair coin. So, having seen the evidence the odds are 1000 to 512 (i.e. about 2 to 1) that the coin chosen was a fair coin. So, the evidence increases the odds in favour of the prosecution hypothesis by a factor of 512, but the defence hypothesis is still more likely. Hence, any rational juror should not convict the defendant on the basis of this evidence alone. Think of it this way: The chance of getting 9 Heads in 9 tosses of a fair coin (defence hypothesis) is still higher than the chance of selecting the one double headed coin from a bag of 1001 coins (prosecution hypothesis).

There is no dispute, therefore, that in the above hypothetical legal case the use of likelihood ratios and Bayes leads to the undisputedly correct conclusion. There is no 'statistical doubt' at all. Why is this important if the case is purely hypothetical? The answer can be gleaned by changing the assumptions very slightly. The assumption that a 'fair' coin has a probability of ½ of landing on Heads is a simplification. Even if we have no reason to believe there are double-headed coins in circulation the actual frequency of heads tossed in all coins in circulation is not a number that can be practically determined, and even if we had a very large datatabase of coins and toss results on them, it would certainly not be exactly equal to ½. These minor additional assumptions of reality, already shift us out of the 'purely statistical' scenario. Do these changes mean that our approach to evaluating evidence using likelihood ratios is no longer valid? Of course not. The exact same methods apply. All that has changed is our confidence in the original assumptions. We counter this uncertainty not by declaring the calculus of probability as invalid but by either stating our uncertainty clearly up front or using ranges instead of exact values.

All evidence in any case ultimately has a 'statistical basis'. The 'soundness' of the statistical basis is a spectrum where examples like that of case 1 above just happen to sit firmly at the 'soundest' end. The rationale for the ruling in [1] is not just that that there is some point at the opposite end of the spectrum at which the use of likelihood ratios become inappropriate, but that most types of forensic match evidence are even further beyond this point of the spectrum. Readers may yet be unconvinced that the minor change in the example already discussed is insufficient to push the example beyond this point, but surely the following leaves no doubt.

- ***Example Case 2:*** This case is the same as case 1, except for the fact there is no possibility that the coin was double-headed because the defendant clearly showed the coin to have a head and tail before tossing it. The prosecution hypothesis here is simply that the coin is 'biased' – i.e. will in the long run produce a greater ratio of Heads than Tails. It is still an offence to knowingly use such a coin. It is not known exactly what this bias is, but it is known that a magic shop in the area was selling special coins that looked real but were biased. These coins were all made with a different weighting and all that can be said with reasonable certainty was that the range of Heads 'bias' in these coins was between 0.6 and 0.7. The prosecution hypothesis is that the defendant used one of the

coins from this magic shop.

The evidence of 9 Tails in this example case has less 'statistical soundness' than the evidence of a stature match (or indeed any type of forensic match) in the previous section. Yet, it is easy to see that the use of likelihood ratios can be applied just as rationally in this example as in example case 1. Specifically:

The defence likelihood is the probability of seeing 9 Heads in 9 tosses given that the coin is fair. We cannot assume that the probability of tossing a Head on a fair coin is exactly ½. If we have a database of what are believed to be fair coins in which the lowest frequency of heads is 0.495 and the highest frequency is 0.505 then we could consider a range for the defence likelihoods using these as assumptions that are respectively least and most favourable to the defence hypothesis. So the least favourable is 0.00178 (that is 0.495 to the power of 9) and the most favourable is 0.002136 ((that is 0.505 to the power of 9).

The prosecution likelihood is the probability of seeing 9 out of 9 Heads given that the coin is biased. Here we have an infinite number of different prosecution hypotheses corresponding to every potential number between 0.6 and 0.7. Taking just the two extremes as those being respectively least and most supportive of the prosecution hypothesis we end up with respective prosecution likelihoods of 0.01 (that is 0.6 to the power of 9) and 0.04 (that is 0.7 to the power of 9) .

Despite the 'unscientific' nature of the evidence, we can conclude that, with the assumptions that most favour the prosecution, the likelihood ratio is 22.5 (0.04 divided by 0.00178), while with the assumptions that most favour the defence the likelihood ratio is 4.7 (0.04 divided by 0.00178). So, despite the clear lack of 'statistically' sound evidence, we can rationally conclude that the odds in favour of the prosecution hypothesis have increased by a factor of between 4.7 to 22.5. Indeed that is the only rational conclusion to make.

If the evidence made by either an expert or a member of the jury does not lead to the conclusion that the evidence supports the prosecution hypothesis by a factor of at least 4.7 to 1, assuming the most optimistic defence assumptions, then such a conclusion is irrational. If, as the ruling in [1] suggests, the use of likelihood ratios to explain the impact of this kind of evidence was not allowed in court, then the jury would be expected to do their own reasoning. This would mean, for example, that it would be acceptable to conclude that the evidence actually supported the defence by a factor of 100 to 1 if that is what their own 'method' led them to conclude.

Having, hopefully, countered the argument against using Bayes for 'non-scientific' statistical evidence, we next return to the crucial issue of why Bayesian reasoning has failed to make an impact on 'non-scientific' forensic match evidence.

5 Moving to more realistic assumptions: why the R v T ruling was understandable

Recall that the assumption of perfect testing accuracy, used so far in our forensic match evidence examples, means that:

- Someone with type X will always be tested to be of type X. This means that there is *zero probability of false negatives*:
- Someone who is not type X will never be determined to be of type X. This means that there is *zero probability of false positives*:

In the case of stature matching neither of these assumptions is at all realistic, as they would require all of the following to hold:

- Stature traces (taken either from the crime scene or taken directly from the defendant) are always 'perfect' (so, for example, there is no possibility that distortion of the photographic/video evidence is such that the person's height could be determined to be 136 centimetres as opposed to 132 centimetres).
- The process of analysing the stature trace is infallible (so, for example, it is impossible for one stature expert to determine from a photo that the person is a man and for a different stature expert to determine from the same photo that the person has is a woman.
- Stature prints can never be tampered with before they are examined by the expert.
- A person's stature profile can never change (so, for example, if their waistline was 65 centimetres at the time they made the print, then when they are subsequently tested their waistline will inevitably be within 2 centimetres of 65 centimetres).

But these assumptions (especially the first three) are even more dubious in the case of DNA evidence than in the case of stature matching. If any of these statements is not true then neither the false negative probability nor false positive probability will be zero.

Yet, while it is accepted that random match probabilities need to be 'statistically sound' the same is never demanded of the probabilities of false positives and false negatives. Indeed, in many analyses they are simply (but wrongly) assumed to be zero, while in others (including DNA analyses) they are simply stated as subjective estimates. This prompted the authors in [36] to ask pointedly:

- "Why are the two possible sources of error in DNA testing treated so differently? In particular, why is it

considered essential to have valid, scientifically accepted estimates of the random match probability but not essential to have valid, scientifically accepted estimates of the false positive probability?"

The authors in [36] provide a strong argument on why it is just as critical to include the false positive probability as the random match probability. However, their omission of the case for the false negative probability (presumably because they only consider the scenario where there have been positive tests for both the source and target) is itself an oversight. Even assuming that both tests are positive, the Bayesian reasoning still requires us to know the probability of a true positive (which is equal to one minus the probability of a false negative, as shown in Table 1). The calculations in [36] assume that the true positive probability is 1 (and hence the false negative is 0). This is unrealistic. By assuming the more realistic assumption of non-zero false negative probability we allow for the scenario in which it is possible that some other suspect with profile type X was never considered because they were wrongly tested as not being type X.

It follows that, as soon as we drop our assumption about 'perfect testing' (as in practice we surely must), then the notion of a sound 'statistical base' for DNA compared with other types of forensic evidence becomes even more blurred than we previously explained, since there is no 'statistically sound' base for determining the error probabilities in DNA testing. If anything it will surely be easier and more objective to determine the values and exact causes of false positive and false negative errors for stature matching than it would be for DNA. It would also be easier to explain to a jury precisely what these errors are.

It should be clear now, conceptually, that there is no more justification for using the probabilities that arise from DNA as there is in using the probabilities that arise from just about any other type of forensic match evidence.

However, it turns out as we show in the next section, that as soon as we incorporate the potential for testing error in a Bayesian argument things become complex. It is not clear, for example, that these issues were properly addressed for the footwear evidence that was the subject of the R v T ruling, and this possibly makes the judge's lack of trust in the transparency and accuracy of the results of the Bayesian analysis more understandable.

6 The problem with scaling up Bayesian arguments

In [35] the authors state:

- "The best argument for the application of Bayesian theory in forensic science is to show that the theory agrees with personal intuitions, when inference problems are simple and intuitions are reliable, and that it helps to go beyond them, when problems become more complicated and intuitions are not so reliable."

This is exactly the strategy we have suggested. The problem with this strategy is that as soon as we recognise that the false positive and false negative probabilities may not be zero, the 'simple' problem actually becomes very difficult to explain using the intuitive, tree-diagram approach. In fact, although several authors have tried it, we are not aware of the problem being presented correctly in any way *other* than by using the formulaic approach. And, even then, the presentations fail to include the false negative probability. The net effect is that, unless people are prepared to understand the formulas they will ***not*** be able to see that the theory agrees with personal intuitions even in the 'simple' problem case. This goes some way to explaining why the basic misunderstandings discussed in Section 3 persist in the law.

To explain what the problem really is and how we might solve it, let us review the relevant information we have to consider for any forensic match case when the testing cannot be assumed to be perfect:

- ***Prosecution hypothesis (H1):*** "The target is the source" (unchanged)
- ***Defence hypothesis (not H1):*** "The target is not the source" (unchanged)
- ***Evidence E1:*** "The source profile is tested to be of type X" (note: we can no longer assume the source profile actually is type X)
- ***Evidence E2:*** "The target profile is tested to be of type X (note: we can no longer assume the target profile actually is type X)

Because of the probability of false positives we cannot assume from the above evidence that either the source or the target have type X. Instead these assertions are also unknown hypotheses:

- ***Source type hypothesis (H2):*** "The source profile really is type X" (true or false)
- ***Target type hypothesis (H3):*** "The target profile really is type X" (true or false)

What we have, therefore, is a problem involving five 'variables' H1,

Table 1. Error probabilities

Actual Type	Not X	Not X	X	X
Test result	Not X	X	X	Not X
	(True negative)	(False positive)	(True positive)	(False negative)
Probability	1-*u*	*u*	1-*v*	*v*

H2, H3, E1, E2 which can all be true or false (in order to do the necessary Bayesian reasoning). But this means there are 32 different scenarios representing the different possible true/false combinations (although some are 'impossible' and some are not observed, such as false values for the evidence). We can show this in a tree diagram - Figure 4 - but of course it is now far more complex than before; possibly too complex for lay people to understand.

Even when we ignore the impossible branches and all the scenarios in which the evidence E1 and E2 is false, we are left with six scenarios that need to be incorporated in the likelihood calculations:

- ***Scenario 1*** (this is the 'normal' prosecution scenario) in which H1, H2, H3, E1 and E2 are all true. This scenario has probability $m(1-v)^2$
- ***Scenario 2*** (this is an often ignored prosecution scenario) in which H1 is true (the target is the source) but the target is not actually type X. Both the test of the target and source, however, incorrectly result in an X. This scenario has probability $(1-m)u^2$.
- ***Scenario 3*** (this is the 'normal' defence scenario) in which the tests are correct but the match is coincidental. This scenario has probability $m^2(1-v)^2$.
- ***Scenario 4*** this is the defence scenario in which the target is incorrectly tested to be type X. This scenario has probability $m(1-m)(1-v)u$.
- ***Scenario 5*** this is the defence scenario in which the source is incorrectly tested as type X. This scenario has probability $(1-m)mu(1-v)$.
- ***Scenario 6*** this is an often ignored defence scenario in which both the source and target are wrongly tested to be X. This scenario has probability $(1-m)^2u^2$.

The prosecution likelihood is the sum of the probabilities for scenarios 1 and 2, while the defence likelihood is the sum of the probabilities for

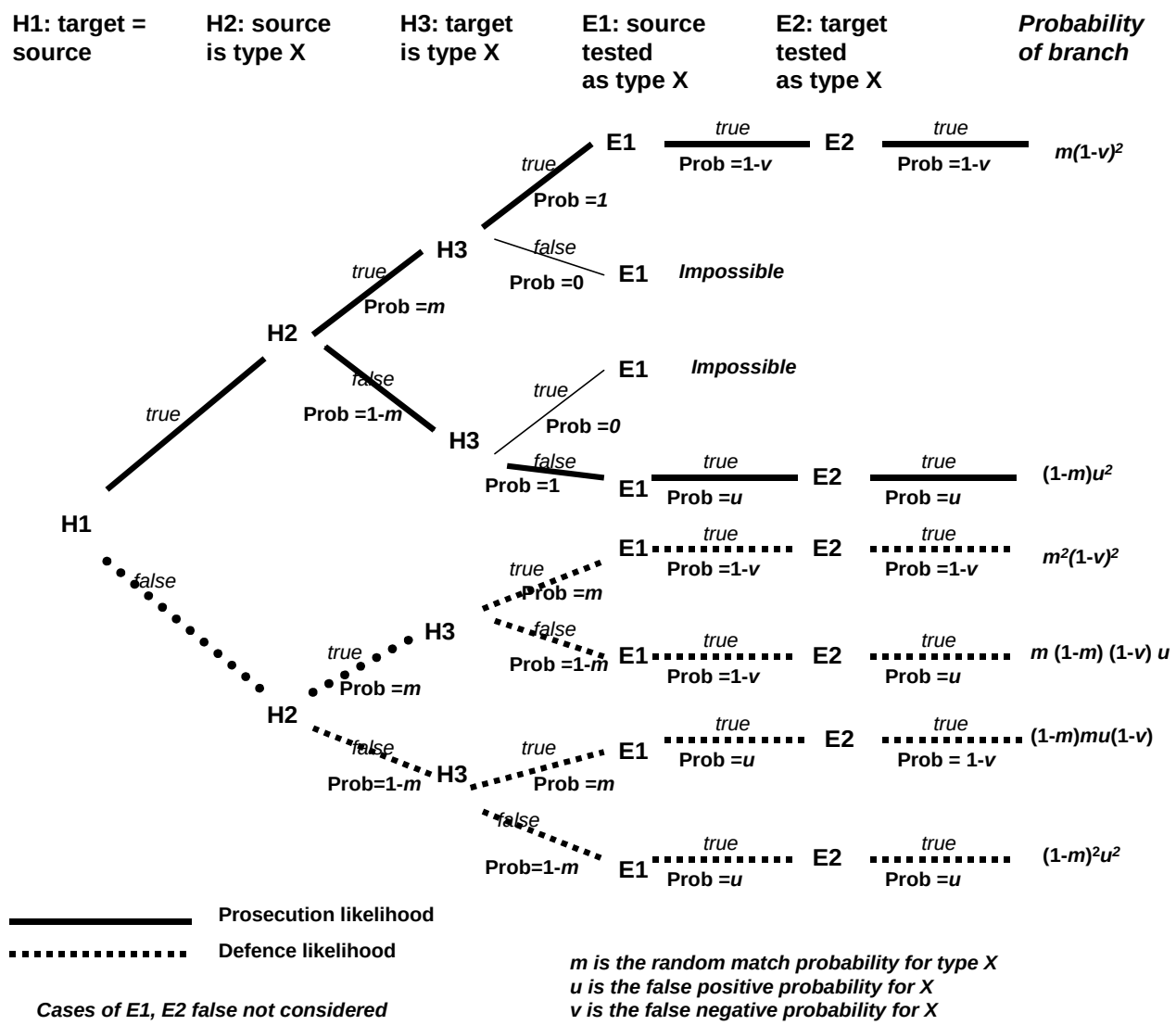


Fig 4. Bayes calculation explained visually (but this time possibly too complex to understand)

scenarios 3, 4, 5, and 6.

The problem is that the likelihoods, and hence also the resulting likelihood ratio, are not sufficiently 'simple and intuitive' to ensure that people can check they 'agree with personal intuition' (which is why it is not even worth the effort here of going through the motions). Resorting to the Bayes formulas, of course, only makes things much worse.

The example also shows that, even for experienced Bayesians, it can be difficult to model the problem in this way and difficult to perform the calculations (as we argued earlier, we have not previously seen a full solution of this problem taking into account both error probabilities). And this example still has many simplifications: it assumes that all three probabilities (random match, false positive, false negative) are all 'point' values, whereas in practice they would be uncertain distributions [11]; it assumes that all variables have just two possible values (true and false); it assumes that there is just one print; and it assumes the only evidence is the match evidence. When we include further aspects of reality (especially including multiple, related pieces of evidence) the possibility for producing the correct Bayesian calculations manually (with or without formulas) – let alone being able to explain them to a lay person – are non-existent.

In our view the best way to minimise this problem is to use Bayesian networks (as explained in [19] [25] [35]). By exploiting assumptions of independence between variables, a Bayesian network (BN) model is typically compact and efficient, since it avoids the problem we saw above whereby we had to consider all possible combinations of variable values (statisticians express this formally by saying that 'it is not necessary to consider the full joint probability distribution').

A BN (see Figure 5) is a graphical model that shows the dependency relationships between the unknown variables of interest (each variable is represented by a node in the graph).

In addition to the graphical structure we define, for each node, a probability table that defines the probability values for the node given the different combinations of parent states. For example, the probability table for node "target tested as X" simply encodes the error rates as shown in Table 2.

It is much easier to build and run this model with the relevant information than it is to either construct a tree as before or to produce the necessary formulas. Once built we can enter evidence and get the calculations immediately as shown in Figure 6 (this shows the results using a standard BN tool). Here we actually compare the results under two different sets of assumptions:

- In a) we encoded the assumption of perfect testing

Table 2. Probability table for node "target tested as X"

Target is type X	False	True
Target tested as X (False)	1-*u*	*v*
Target tested as X (True)	*u*	1-*v*

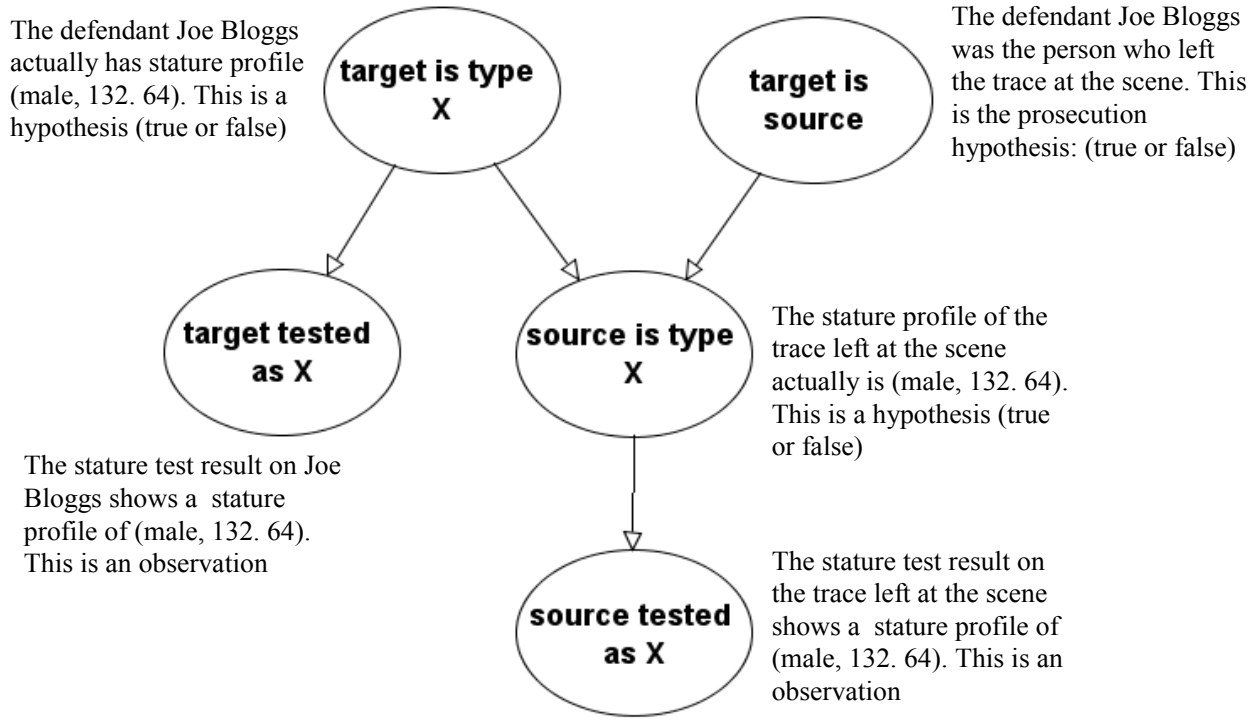


Fig 5. Bayesian network solution to the problem (with an example showing what the nodes would mean for a specific stature matching case)

accuracy (i.e. u and v are both set to zero).

- In b) we encoded the assumption that u (false positive) is 0.1 and v (false negative) is 0.01.

Although in both cases we assume the same match probability (1/100) and the same prior (50:50)[8] for the prosecution hypothesis ("target is source") the difference is quite dramatic. Although the evidence is identical in both cases, in the former t posterior odds[9] are 100 to 1 in favour of the prosecution hypothesis, whereas in the latter the posterior odds[10] are only 65 to 35 (i.e. about 2 to 1) in favour of the prosecution hypothesis.

Not only does the BN remove the need for performing the difficult Bayesian calculations manually, but its graphical representation is easy for a lay person to understand. We are not, however, suggesting that the BN model is what should be presented court. It should be used for pre-trial analysis of the evidence by forensic experts, preferably using different scenarios for the different ranges of match probabilities and error probabilities. The model structure should be agreed between legal teams and forensic experts on both sides. All that should be presented in court are clear statements of the prior assumptions being used (the match probabilitity, and error probabilities) and the results of the calculations under the different assumptions.

A detailed history of BNs in legal reasoning, along with proposed mechanisms for using them in practice can be found in [19] and [20].

7 Conclusions and recommendations

The ruling in R v T displayed some fundamental misunderstandings, including assertions that can be shown to be either illogical or irrational. However, the presentation of the Bayesian argument and likelihood ratios in the original case was both inadequate and inaccurate, as it has been in many similar cases. We have argued that this may be, in large part, due to the continued failure of the statistical community to provide the necessary support to forensic scientists and lawyers. That fundamental probabilistic reasoning should have therefore been discredited in the R v T ruling is hard for statisticians to take but, even in our view as Bayesians, was totally understandable.

If statisticians continue to believe that the way to explain their arguments in legal reasoning is by using first principle calculations and formulas, then the future for Bayes in the law is doomed.

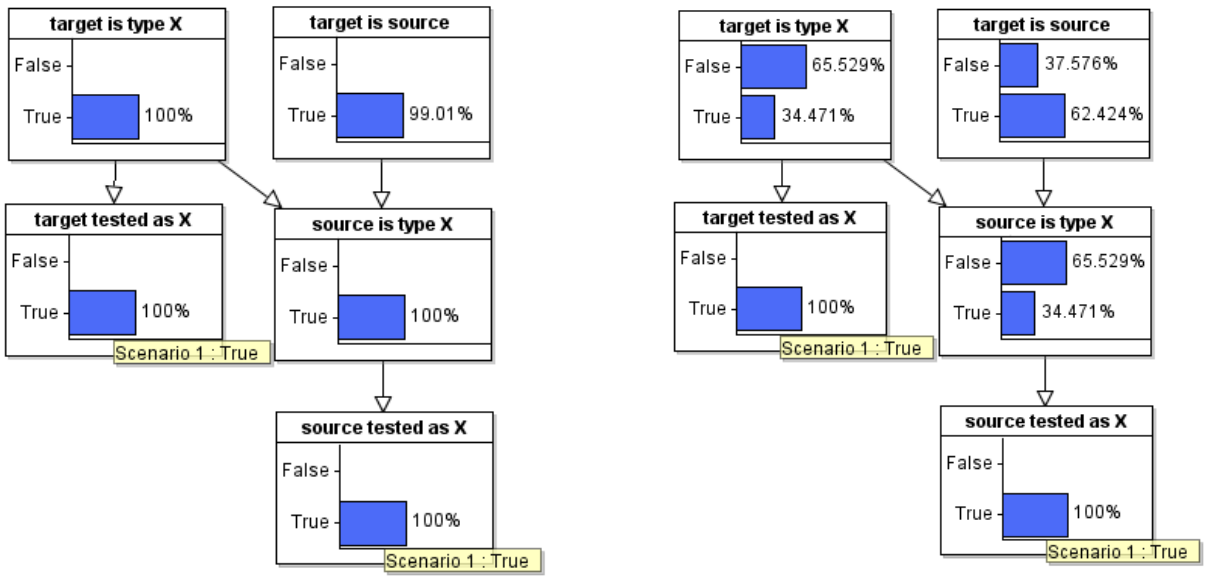
The challenge over the next few years is to get to the situation whereby everybody in the legal system understands the difference between

a. the genuinely disputable assumptions

8 Recall that, by assuming a 50:50 prior, we know that the posterior odds are equal to the likelihood ratio..

9 The likelihood ratio is 100, meaning equivalently the probability the prosecution hypothesis is true is 100/101 = 99.01%)

10 The likelihood ratio is 65/35, meaning equivalently the probability the prosecution hypothesis is true is 65%). ■



a) Impact of evidence when error probabilities are assumed to be zero

b) Impact of evidence when false positive rate is 0.1 and false negative is 0.01

Fig 6. Comparing the different impact of the evidence when we assume different error rates (in both cases the match probability is 1/100 and the prior probability for "target is source" is ½)

that go into a probabilistic argument; and

b. the Bayesian calculations required to compute the conclusions based on the different disputed assumptions.

Crucially, there should be no more need to explain the Bayesian calculations in a complex argument than there should be any need to explain the thousands of circuit level calculations used by a calculator to compute a long division. Lay people do not need to understand how the calculator works in order to accept the results of the calculations as being correct to a sufficient level of accuracy. The same must eventually apply to the results of calculations from a Bayesian analysis. The more widespread use of tools such as Bayesian networks makes this a feasible target.

However, ensuring that the distinction between a) and b) is firmly understood by lawyers is only a necessary requirement for the more widespread takeup of Bayes. There is, as yet, no significant understanding among lawyers that any legal argument can ever be couched in Bayesian terms. The challenge for statisticians is to break down this significant cultural barrier. In this challenge we also propose that the use of Bayesian network models will be useful, but any progress requires a major educational effort aimed at all levels of the criminal justice system. It requires 'buyin' from senior members of the legal profession and politicians, as well as a united front presented by the community of statisticians.

If we can meet these challenges then there is no reason why Bayes should not become a standard (possibly even the central) method for evaluating evidence in every aspect of legal reasoning.

Acknowledgements

We are indebted to the following for providing comments, corrections, relevant information, and contacts: David Balding, Sheila Bird, Tiernan Coyle, Ian Evett, Joseph Kadane, Jay Koehler, Margarita Kotti, Amber Marks, William Marsh, Geoff Morrison, Richard Nobles, David Ormerod, Mike Redmayne, David Schiff, Bill Thompson, Patricia Wiltshire.

References

[1] (2010). R v T. EWCA Crim 2439 http://www.bailii.org/ew/cases/EWCA/Crim/2010/2439.pdf
[2] R v Adams [1996] 2 Cr App R 467, [1996] Crim LR 898, CA and R v Adams [1998] 1 Cr App R 377
[3] (2008). R v Kempster, EWCA Crim 975
[4] (2002). R -v- Dallagher, EWCA Crim 1903
[5] (2007). R. v. George, EWCA Crim 2722
[6] Brandon Mayfield v. U.S.A (04cv1427)
[7] (2007). "Ruling Casts Doubts On Infallibility Of Fingerprint Evidence." from http://www.wbaltv.com/news/14433107/detail.html.
[8] Adam, C. (2010). Essential Mathematics and Statistics for Forensic Science. Chichester, John Wiley & Sons.
[9] Aitken, S. G. G. (2000). Interpretation of evidence, and sample size determination. Statistical Science in the Courtroom. J. L. Gastwirth. New York, Springer: 1-24.
[10] Aitken, C. G. G. and F. Taroni (2004). Statistics and the evaluation of evidence for forensic scientists (2nd Edition), John Wiley & Sons, Ltd.
[11] Balding, D. J. (2005). Weight-of-Evidence for Forensic DNA Profiles, Wiley.
[12] Berger, C. E. H., J. Buckleton, C. Champod, I. Evett and G. Jackson (2011). "Evidence evaluation: A response to the court of appeal judgement in R v T." Science and Justice 51: 43-49.
[13] Buckleton, J., C. M. Triggs and S. J. Walsh (2005). Forensic DNA Evidence Interpretation, CRC Press.
[14] Broeders, T. (2009). Decision-Making in the Forensic Arena. In "Legal Evidence and Proof: Statistics, Stories and Logic". (Eds H. Kaptein, H. Prakken and B. Verheij, Ashgate) 71-92.
[15] Coyle, T. (2010). Trace and Contact Evidence. Crime Scene to Court: Essentials of Forensic Science. P. C. White, Royal Society of Chemistry.
[16] Dror, I. E. and R. Rosenthal (2008). "Meta-analytically quantifying the reliability and biasability of forensic experts." Journal of Forensic Sciences 53(4): 900-903.
[17] Evett, I. W., L. A. Foreman, G. Jackson and J. A. Lambert (2000). "DNA profiling: a discussion of issues relating to the reporting of very small match probabilities." Criminal Law Review (May) 341-355.
[18] Evett, I. W. and B. S. Weir (1998). Interpreting DNA evidence : statistical genetics for forensic scientists, Sinauer Associates.
[19] Fenton, N. E. and M. Neil (2008). Avoiding Legal Fallacies in Practice Using Bayesian Networks. Seventh International Conference on Forensic Inference and Statistics. Lausanne, Switzerland
[20] Fenton, N. and M. Neil (2010). "Comparing risks of alternative medical diagnosis using Bayesian arguments." Journal of Biomedical Informatics 43: 485-495.
[21] Gastwirth, J. L., Ed. (2000). Statistical Science in the Courtroom. New York, Springer-Verlag.
[22] Gigerenzer, G. (2002). Reckoning with Risk: Learning to Live with Uncertainty. London, Penguin Books.
[23] Haigh, J. (2003). Taking Chances: Winning with Probability, Oxford University Press.
[24] Jackson, A. R. W. and J. M. Jackson (2004). Forensic Science. Harlow, Pearson.
[25] Kadane, J. B. and D. A. Schum (1996). A Probabilistic Analysis of the Sacco and Vanzetti Evidence, John Wiley & Sons.
[26] Kaye, D. H. (2009). "Identification, Individualization, Uniqueness." Law, Probability & Risk 8(2): 85-94.
[27] Langford, A., J. Dean, et al. (2005). Practical Skills in Forensic Science. Harlow, Pearson.
[28] Lucy, D. (2006). Introduction to Statistics for Forensic Scientists. Chichester, UK, John Wiley & Sons Ltd.
[29] Morrison, G. M. (2012). "The likelihood ratio framework and forensic evidence in court: a response to RvT." International Journal of Evidence and Proof 16(1).
[30] Redmayne, M. (2001). Expert Evidence and Criminal Justice, Oxford University Press.
[31] Redmayne, M., P. Roberts, C. Aitken and G. Jackson (2011). "Forensic Science Evidence in Question." Criminal Law Review (5): 347-356.
[32] Robertson, B., G. A. Vignaux and C. E. H. Berger (2011). "Extending the confusion about Bayes." The Modern Law Review 74(3): 444-455.
[33] Robertson, B. and T. Vignaux (1995). Interpreting Evidence: Evaluating Forensic Science in the Courtroom, John Wiley and Son Ltd.
[34] Saks, M. J. and J. J. Koehler (2007). "The Individualization Fallacy in Forensic Science Evidence." http://works.bepress.com/michael_saks/1
[35] Taroni, F., C. Aitken, P. Garbolino and A. Biedermann (2006). Bayesian Networks and Probabilistic Inference in Forensic Science, John Wiley & Sons.
[36] Thompson, W. C., F. Taroni and C. G. G. Aitken (2003). "How the probability of a false positive affects the value of DNA evidence." Journal of Forensic Sciences 48(1): 47-54.
[37] White, P. C., Ed. (2004). Crime Scene to Court: Essentials of Forensic Science, Royal Society of Chemistry. ■