Science and Justice 56 (2016) 374–379

Contents lists available at ScienceDirect

Science and Justice

journal homepage: www.elsevier.com/locate/scijus

Position paper

What should a forensic practitioner's likelihood ratio be?☆

Geoffrey Stewart Morrison [a,b,*,1], Ewald Enzinger [a]

[a] Morrison & Enzinger, Independent Forensic Consultants, Vancouver, British Columbia, Canada & Corvallis, Oregon, USA
[b] Department of Linguistics, University of Alberta, Edmonton, Alberta, Canada

ARTICLE INFO

Article history:
Received 11 May 2016
Accepted 12 May 2016

Keywords:
Likelihood ratio
Precision
Reliability
Verbal scale

ABSTRACT

We argue that forensic practitioners should empirically assess and report the precision of their likelihood ratios. Once the practitioner has specified the prosecution and defence hypotheses they have adopted, including the relevant population they have adopted, and has specified the type of measurements they will make, their task is to empirically calculate an estimate of a likelihood ratio which has a true but unknown value. We explicitly reject the competing philosophical position that the forensic practitioner's likelihood ratio should be based on subjective personal probabilities. Estimates of true but unknown values are based on samples and are subject to sampling uncertainty, and it is standard practice to report the degree of precision of such estimates. We discuss the dangers of not reporting precision to the courts, and the problems with an alternative approach which instead reports a verbal expression corresponding to a pre-specified range of likelihood ratio values. Reporting precision as an interval requires an arbitrary choice of coverage, e.g., a 95% or a 99% credible interval. We outline a normative framework which a trier of fact could employ to make non-arbitrary use of the results of forensic practitioners' empirical calculations of likelihood ratios and their precision.

1. Introduction

We assume the reader is familiar with the introduction to the current special issue [1]. The present position paper is a contribution to the debate as to whether forensic practitioners should calculate and report the precision of the likelihood ratios which they present to the courts. We discuss differences in philosophical understanding as to the nature of the forensic practitioner's likelihood ratio. We argue that the forensic practitioner's likelihood ratio should be an empirically calculated estimate of a true but unknown value, and that the precision of the estimated value should be empirically calculated and reported. We argue against the use of likelihood ratios which are directly the result of subjective judgement. We also argue against the use of multi-level ordinal scales consisting of verbal expressions and associated ranges of likelihood ratio values. Finally, we outline a normative framework which could allow the trier of fact to make principled non-arbitrary use of the reported imprecision of forensic practitioners' likelihood ratios.

2. Different philosophical understandings as to what constitutes the forensic practitioner's likelihood ratio

As outlined in the introduction to the current special issue [1], a number of researchers and practitioners believe that forensic practitioners should empirically calculate and report the precision of the likelihood ratios which they present in court. Papers espousing this position and proposing concrete procedures for calculating precision include Chakraborty et al. [2], Balding [3], Weir [4], Curran et al. [5], Curran [6], Morrison et al. [7], Beecham & Weir [8], Curran & Buckleton [9], Morrison et al. [10]; Morrison [11], Nordgaard & Höglund [12], Hancock et al. [13], Alberink et al. [14], Zhang et al. [15], Taylor et al. [16], Kaplan Damary et al. [17]. Others, for example Taroni et al. [18,19], argue that it is not appropriate to calculate the precision of a likelihood ratio since a likelihood ratio is a personal belief which has no true value to be estimated. We find Sjerps et al. [20] counterarguments to the latter position convincing (see also Kaplan Damary et al. [17]), and we believe that the likelihood ratio which a forensic practitioner presents to the court should not be a personal belief. Once the forensic practitioner has stated what they understand to be the relevant circumstances of the case and the prosecution and defence hypotheses they have adopted, including the definition of the relevant population, and they have stated the type of measurements they will make on the known-origin sample, the questioned-origin specimen, and a sample of the relevant population, their task is to calculate an estimate of a likelihood ratio which has a true (but unknown) value. Without these specifications there is no true likelihood ratio value to be estimated. If one were to change

☆ This paper is part of the Virtual Special Issue entitled: Measuring and Reporting the Precision of Forensic Likelihood Ratios, [http://www.sciencedirect.com/science/journal/13550306/vsi], Guest Edited by G. S. Morrison.

* Corresponding author.
E-mail address: geoff-morrison@forensic-evaluation.net (G.S. Morrison).

[1] This author is also the Guest Editor for the special issue. The present paper was written before the Guest Editor handled any of the other submissions, and the Editor in Chief had editorial responsibility for the present paper.

http://dx.doi.org/10.1016/j.scijus.2016.05.007
1355-0306/

the specifications then the question being asked and answered would change. A likelihood ratio is the answer to a specific question, and changing the question will lead to a different likelihood ratio with a different true value. For example, a likelihood ratio based on the widths of objects will not only have a different value compared to a likelihood ratio based on the lengths of objects, it will be the answer to a different question.[2]

If a forensic practitioner's task were to assign personal probabilities to arrive at a likelihood ratio which does not depend on specifications including the types of measurements to be made, such a likelihood ratio could not be said to be an estimate of a true but unknown value. In contrast, we believe that the task of the practitioner is to calculate an estimate of the true but unknown value of the likelihood ratio given a specified relevant population and one or more specified types of measurement. This disagreement in belief as to what constitutes the forensic practitioner's likelihood ratio is a philosophical one. It cannot be resolved empirically. We believe that this philosophical disagreement is the fundamental source of dispute between those who believe one should calculate the precision of likelihood ratios and those who believe that one should not. Although both groups may use the term "likelihood ratio", they are referring to different things.

The strict terminological issue could perhaps be resolved by maintaining the distinction between the terms "likelihood ratio" and "Bayes factor": A likelihood ratio is calculated using sample data and maximum likelihood estimates of parameters, whereas a Bayes factor is calculated using posterior predictive probability distributions derived by combining sample data with prior probability distributions and integrating over nuisance parameters. The debate could therefore be rephrased as: Should forensic practitioners calculate estimates of likelihood ratios and their precision versus should they calculate Bayes factors? (The wider debate would also include a third option: Should forensic practitioners assign likelihood ratio values based directly on subjective judgement rather than on explicit calculations using quantitative data?) The term "likelihood ratio" is, however, often used as a cover for both the more restrictive definition of likelihood ratio and for Bayes factor (we will generally use it as a cover term in the present paper). Also, note that the debate is not actually about whether to use frequentist calculations based only on sample data, or Bayesian calculations that take account of both sample data and prior probability distributions. Both sides can advocate Bayesian approaches, albeit different Bayesian approaches. Sjerps et al. [19] argue in favour of calculating and reporting Bayesian posterior probability distributions rather than point-value Bayes factors, whereas Brümmer & Swart [21] argue in favour of the latter.

3. Why it is important to calculate and report the precision of the forensic practitioner's likelihood ratio

If one believes that the forensic practitioner's task is to calculate an estimate of a likelihood ratio which has a true but unknown value, since those calculations will be based on sample data, not oracular knowledge of population distributions, the calculated likelihood ratio value will be influenced by sampling uncertainty. Since the calculated value has uncertainty it should be accompanied by a measure of that uncertainty, i.e., a measure of its precision. This is standard practice (e.g., International Organization for Standardization [22], United Kingdom Accreditation Service [23]) and has been explicitly recommended for forensic science (e.g., National Research Council [24]).

Not to report the precision of a likelihood ratio value could be highly misleading to the court. For example, if the forensic practitioner's best estimate of the likelihood ratio is 10, but this estimate is produced by a system (an implementation of a method) which when tested under conditions reflecting those of the case is found to have a 95% credible interval of plus or minus two orders of magnitude, then the 95% credible interval in this case would be 1/10 to 1000. Knowing this, a judge at an admissibility hearing might reasonably decide that the evidence is not sufficiently reliable to warrant admission, or the trier of fact might reasonably decide that the likelihood ratio value is not substantially different from 1 and that the evidence should therefore not change their belief as to the relative probabilities of the prosecution and defence hypotheses.[3] Even if the forensic practitioner's best estimate of the likelihood ratio were relatively extreme compared to the bound of its credible interval closest to the neutral value of 1, e.g., a likelihood ratio of 1000 with a 95% credible interval of plus or minus one order of magnitude and a lower bound of this interval at 100, the trier of fact might choose to be conservative, and, based on the reported precision, they might use a likelihood ratio value closer to 1 than the forensic practitioner's best estimate, e.g., they may choose to use 100 rather than 1000. If the forensic practitioner did not report precision, then the court would be denied the information necessary to make a reasonable decision on admissibility or on what might constitute a reasonable degree of conservatism. A trier of fact might then be misled into assigning a more extreme strength of evidence to the forensic practitioner's likelihood ratio conclusion than they would have done had they known about the precision of the system used to calculate that likelihood ratio. Alternatively, a trier of fact may be incredulous as to the apparent degree of precision of a likelihood ratio reported as a point value. This might lead them to use a more conservative strength of evidence than if they had actually been presented with the results of an empirical assessment of the precision of the system, e.g., they could choose to use a likelihood ratio of 10 when the forensic practitioner's best estimate was 1000 and the unreported lower limit of the 95% credible interval would have been 100.

The discussion in the preceding paragraph raises a distinction to be made between the likelihood ratio reported by the forensic practitioner and the likelihood ratio actually used by the trier of fact. These will not necessarily have the same value. The effective likelihood ratio that the trier of fact employs, i.e., the extent by which they update their beliefs with respect to the relative probabilities of the competing prosecution and defence hypotheses, will likely depend on the trier of fact's assessment of how much they trust what the forensic practitioner reports. For example, if a practitioner reports a likelihood ratio of 1000 and their appearance and manner instil confidence, the trier of fact might use an effective likelihood ratio of 1000, but if the practitioner's appearance and manner do not instil confidence the trier of fact might be less trustful of what the practitioner reports and use an effective likelihood ratio of 100 instead. Supplying the trier of fact with empirical information about the precision of the system used to calculate the forensic practitioner's likelihood ratio would hopefully lead to the trier of fact choosing their effective likelihood ratio on a more relevant basis than the forensic practitioner's appearance and manner.

[2] For this reason, it is important that the forensic practitioner clearly explain to the judge at an admissibility hearing and to the trier of fact at trial what the question is that the forensic practitioner has set out to address. Only with a clear understanding of the question can the judge or trier of fact decide whether the forensic practitioner has set out to answer a question that is of interest to the court, and only with a clear understanding of the question can the trier of fact understand the forensic practitioner's answer to that question.

[3] The *Daubert* ruling states that "The focus, of course, must be solely on principles and methodology, not on the conclusions that they generate" [William Daubert et al. v Merrell Dow Pharmaceuticals Inc., 1993, 509 US 579, Section II C]. We think, however, that the value of the strength of evidence could play a role in deciding upon admissibility. When the likelihood ratio is 10 and the 95% credible interval is plus or minus two orders of magnitude, the testimony could be ruled inadmissible because if the likelihood ratio is not substantially different from 1 it may be that it will not "help the trier of fact… to determine a fact in issue" [Federal Rule of Evidence 702(a) as amended Apr. 17, 2000, eff. Dec. 1, 2000; Apr. 26, 2011, eff. Dec. 1, 2011]. In contrast, this would not be of concern if the value of the likelihood ratio were extreme (far from 1) compared to the width of its credible interval. Note that our discussion relates to the absolute value of the likelihood ratio compared to its precision, not to the absolute value alone. A likelihood ratio of 2 with a high degree of precision could still help the trier of fact to determine a fact in issue.

We do believe that the trier of fact's prior odds, likelihood ratio, and posterior odds will be based on subjective personal probabilities. These probabilities will be based on the trier of fact's beliefs, and these beliefs will form the basis for the trier of fact's ultimate decision in the case. Some of these beliefs will be informed by quantified strength of evidence statements presented by forensic practitioners, but other beliefs will be based on non-forensic evidence which is not accompanied by quantification of its strength. In contrast, as we previously stated, we do not believe that forensic practitioners' likelihood ratios should be based on personal probabilities. The mantra of "probabilities are personal" (Berger et al. [25]) could be (mis)used to justify the presentation of likelihood ratios which are entirely or primarily based on forensic practitioners' untested subjective judgements. In such cases, relevant data, quantitative measurements, and statistical models would play no part or play only a secondary supportive role. This has led to what we consider to be justified criticism. Risinger [26], for example, asks what the warrant is for such subjective likelihood ratios. We believe that the likelihood ratio framework is the logically correct framework for the evaluation of forensic evidence, but we do not believe that logical correctness by itself is sufficient. As warrant, we believe that a forensic practitioner's likelihood ratio should be based on relevant data, quantitative measurements, and statistical models, that the degree of accuracy and precision (validity and reliability) of the system used to calculate the likelihood ratio should be empirically assessed under conditions reflecting those of the case under investigation, and that what the forensic practitioner reports as strength of evidence should be directly the output of the statistical model (see Morrison [11], Morrison & Stoel [27], Morrison [28], Enzinger & Morrison [29], Enzinger et al. [30]).

If a subjective-judgement approach were empirically tested and found to have an acceptable degree of accuracy and precision, we would argue for its admission; however, subjective-judgement approaches are practically difficult to empirically test since it may take a human expert a substantial amount of time to perform each test trial. In contrast, once it has been trained, a system based on data, automated measurements, and statistical models can rapidly and cheaply perform thousands of test trials (the only limitation being the availability of suitable test data). A system for which the output is primarily or directly the result of subjective judgement is also potentially highly susceptible to cognitive bias (Found [31]).

A system based on relevant data, quantitative measurements, and statistical models still involves subjective judgements made by the forensic practitioner, but these are judgements about matters such as what constitute appropriate relevant data to train and test the system. These subjective judgements are far removed from the final output of the system, and the remainder of the process is objective. Systems based on relevant data, quantitative measurements, and statistical models are therefore intrinsically resistant to cognitive bias. They are also much more transparent and replicable than systems based primarily on subjective judgement.

If a forensic practitioner uses a system based on relevant data, quantitative measurements, and statistical models, and empirically tests the performance of that system under conditions reflecting those of the case under investigation, the performance testing should include an empirical assessment of precision. In such circumstances, what the forensic practitioner should report is the numeric likelihood ratio value output by the system and a numeric indicator of the precision of that value, for example: From measurements and calculations based on the known sample and questioned specimen and a sample of data representative of the relevant population and with conditions reflecting those of the known sample and questioned specimen in this case, my best estimate of the strength of evidence is 1000 (the evidence is 1000 times more probable if the prosecution hypothesis were true than if the defence hypothesis were true). Based on empirical tests of the reliability of the system I used to make these calculations, tests made using data representative of the relevant population and reflecting the conditions of the known sample and questioned specimen in this case, I am 99% certain that the strength of evidence is at least 100 (the evidence is at least 100 times more probable if the prosecution hypothesis were true than if the defence hypothesis were true).

4. Why multi-level scales with verbal expressions and associated ranges of likelihood ratio values should not be used

We believe that it would make no sense to take the numeric output of a system based on relevant data, quantitative measurements, and statistical models and, rather than report it directly along with an empirically derived indicator of its precision, instead convert it to a verbal expression with an associated pre-determined range of likelihood ratio values. Ordinal scales with sets of verbal expressions and associated ranges of likelihood ratios are popular (e.g., Champod & Evett [32], Association of Forensic Science Providers [33], Nordgaard et al. [34], Bunch & Weavers [35], Willis et al. [36]), and avoiding giving a false sense of precision has been stated as justification for the use of such scales:

> that kind of precision is rarely realistic in forensic science and the scale is no more than a guide to the judgement of the scientist. … the assessment of the evidence is not numerically precise
>
> (Evett et al. [37] p. 236)

> precise figures may not be available and the verbal conventions enable the expert to express the relative strength of the evidence in words.
>
> (Robertson et al. [38] p. 447)

> presentation of a numerical value does present a false impression of exactitude and the presentation of a confidence interval presents difficulties of interpretation. A verbal scale provides an interpretation which is relatively easy to understand by a layman and yet is not misleadingly precise.
>
> (Aitken in the discussion on Neumann et al. [39] p. 396)

We are opposed to the use of multi-level scales with verbal expressions and associated likelihood ratio ranges for a number of reasons, some of which we discuss here.

The likelihood-ratio ranges in these scales are arbitrary, and they suffer from cliff edge effects. A calculated likelihood ratio value of 999 is not meaningfully different from a calculated likelihood ratio value of 1001, but one will be converted to "more probable" and the other to "much more probable".

More substantially, expressions such as "slightly more probable", "more probable", "appreciably more probable", "much more probable", "far more probable" and "exceedingly more probable" (Willis et al. [36]) are vague and will be interpreted differently by different individuals. Their meaning can only be made explicit via reference to their associated ranges of likelihood ratio values, e.g., 2–10, 10–100, 100–1000, 1000–10,000, 10,000–1,000,000, 1,000,000+, and an explanation of what a likelihood ratio is, what it means, and how it normatively should be used. Transparency is essential and an explanation has to be given, the use of a verbal expression instead of a numeric value obfuscates rather than clarifies.

Potential ambiguity of meaning stems from the pre-specification of a universal set of verbal expressions and likelihood ratio ranges, without consideration of how the trier of fact is likely to interpret these given the particular circumstances of individual cases to which the scale is applied. Enzinger & Morrison [29] describe an analysis based on the circumstances of a real forensic voice comparison case in which the reported likelihood ratio value was 4.40 with 99% certainty that it was at least 3.77. This would be converted to a likelihood ratio range of 2–10 and a verbal expression of the evidence being "slightly more probable" if hypothesis *A* were true than if hypothesis *B* were true. If the

relevant population specified in the alternative hypothesis (hypothesis *B*) were large, e.g., on the order of thousands or millions of speakers, this information could legitimately be relevant for deciding on prior odds, but we think it might also influence the trier of fact's interpretation of the expression "slightly more probable". The trier of fact might interpret "slightly more probable" as corresponding to a substantially larger effective likelihood ratio than 4.40, or even substantially larger than 10. An effective likelihood ratio of 44, for example, would make little dent in prior odds of one in a million, and although it would be an order of magnitude greater than the calculated value of 4.4 and well beyond the upper 99th percentile from the reliability tests (5.16), a value of 44 could still reasonably be equated to "slightly more probable" in this context. In this example the trier of fact would be in danger of inferring a greater strength of evidence from the verbal expression "slightly more probable" than would be justified by the calculated likelihood ratio value and its precision. In the actual case, the circumstances were such that the hypotheses were that the speaker was one of two speakers, either speaker *A* or speaker *B*. This is information which could legitimately be relevant for deciding on prior odds, but we think it might also influence the trier of fact's interpretation of the expression "slightly more probable". If the priors were even, a trier of fact might perceive a likelihood ratio value of 4.4 as being stronger than what they would infer from "slightly more probable". They might reasonably think that "slightly more probable" covered a range of above 1 to about 2, and they might think that 4.4 corresponded to "much more probable". In this example the trier of fact would be in danger of inferring a lesser strength of evidence from the verbal expression "slightly more probable" than would be justified by the calculated likelihood ratio value and its precision. The solution would be not to use verbal expressions at all, and instead directly report the calculated value of the likelihood ratio and its precision.

Even if a calculated credible interval is one order of magnitude (plus or minus half an order of magnitude), as are most of the scale ranges in the example above, conversion of the likelihood ratio value to one of the ranges on the scale may be misleading, e.g., a likelihood ratio of 900 with a calculated credible interval[4] of 285 to 2846 would be converted to a range of 100 to 1000, but so would a likelihood ratio of 200 with a calculated credible interval of 63 to 632. The conversion in the first example would arguably under-represent the strength of evidence, and in the second over-represent it. The effect of a logarithmic scale may also not be immediately apparent, a calculated likelihood ratio of 550, linearly in the middle of the 100 to 1000 range, would actually have a one order of magnitude credible interval of 174 to 1739. Also, there is no reason why a calculated credible interval should turn out to be one order of magnitude, and conversion to a one order of magnitude range could under or over represent the empirically calculated precision of the system. For example, a likelihood ratio of 316 with a one order of magnitude (plus or minus half an order of magnitude) credible interval would have an interval of 100 to 1000, but if the credible interval were half an order of magnitude (plus or minus quarter an order of magnitude) it would be the narrower interval of 178 to 562, and if it were two orders of magnitude (plus or minus one order of magnitude) it would be the wider interval of 32 to 3162, yet these would all be converted to a range of 100 to 1000.

We therefore see no justification for using such a scale if the forensic practitioner has calculated a likelihood ratio using relevant data, quantitative measurements, and statistical models, and is able to empirically assess the validity and reliability of that system using data representative of the relevant population and reflective of the conditions of the case under investigation.

Even if the scales are intended to be used by practitioners who assign strength of evidence on the basis of subjective judgement and they wish to convey that their judgements have a degree of uncertainty, we think it would be more transparent if practitioners actually state a numeric range of values for what they believe to be the probability of the evidence given the prosecution hypothesis and a numeric range of values for what they believe to be the probability of the evidence given the defence hypothesis, and then calculate and report a range of likelihood ratios based on the ranges they gave for its numerator and denominator (e.g., Facey & Davis [40]). This would be more transparent and open up to inspection the process by which the practitioner arrives at their strength of evidence statement and their degree of uncertainty with respect to its value. The practitioner would then likely have to justify their choices of ranges of values for the numerator and denominator. This approach would also force practitioners to actually use the likelihood ratio framework, and would prevent them from directly picking a value on an ordinal scale.[5]

In general, we are opposed to the presentation of untested likelihood ratios which are based directly on subjective judgement. We feel these lack warrant, and so would still be unhappy with them even if they were presented in the form outlined in the previous paragraph. If empirical testing can be performed, however, calibration can also be performed, and calibration and testing can provide warrant. Calibration data should be distinct from test data, but initial steps in the procedure for calibrating a system are the same as for testing a system: The system is presented with a large number of trials. The system outputs a response for each trial. The calibrator compares the output with knowledge about which hypothesis was true for each trial. If, for example, over hundreds or thousands of empirical calibration trials it is found that when a practitioner declares "identification" the ratio of hits to false alarms (correct identifications versus false identifications) is 100:1, then when they declare "identification" in a case the likelihood ratio associated with them stating that conclusion if it were true versus if it were false is 100 (assuming that the calibration data reflected the conditions of that forensic case). Subjective likelihood ratios on a continuous or an ordinal scale can also be empirically calibrated using statistical models (e.g., Lindh & Morrison [41], Ramos et al. [42]). Such empirically calibrated likelihood ratios are not directly subjective judgements, they are based on data (the output of the subjective judgement process) and statistical models. Completing a sufficient number of calibration and test trials may, however, be practically difficult if a substantial amount of time is needed to make each subjective judgement.

5. How a trier of fact could make use of information about the precision of forensic practitioners' likelihood ratios

A trier of fact who always used the bound of a credible interval closest to the neutral value of 1 would end up being very conservative, probably more conservative than they anticipated. For example, if the trier of fact always used a 90% bound (the forensic practitioner reported that they were 90% certain that the likelihood ratio was more extreme than this bound), and they combined the likelihood ratios from two different pieces of evidence, the resulting conservatism for the combination would not be 90% but 99%. Also, any particular percentage for a credible interval or a conservative bound is arbitrary, there is no principled reason why one should use 90%, 95%, 98%, 99%, or any other value. A better way for a trier of fact to handle imprecision would be for them to consider the distributions of the likelihood ratios in the forensic practitioner's assessment of the reliability of the system. A trier of fact could also have uncertainty in their belief as to prior odds.[6] That

[4] This could be a 95% or 98% credible interval range, or any other percentage. It does not matter for this example as long as the same percentage is used throughout.

[5] Note that the idea of presenting a range of likelihood ratio values because the practitioner's subjective judgements are imprecise is at odds with the idea that one should report a point value and should not calculate the precision of a likelihood ratio because a likelihood ratio is based on personal probabilities. Although both approaches may be described as using subjective probabilities, they are philosophically opposed.

[6] The value of the prior odds is whatever the trier of fact believes it to be, it is not an estimate of a true but unknown value. With prior odds based on probabilities that are subjective personal beliefs, however, we do not believe that the trier of fact should be forced to have a belief that is a point value, we accept that they may have a fuzzy belief.

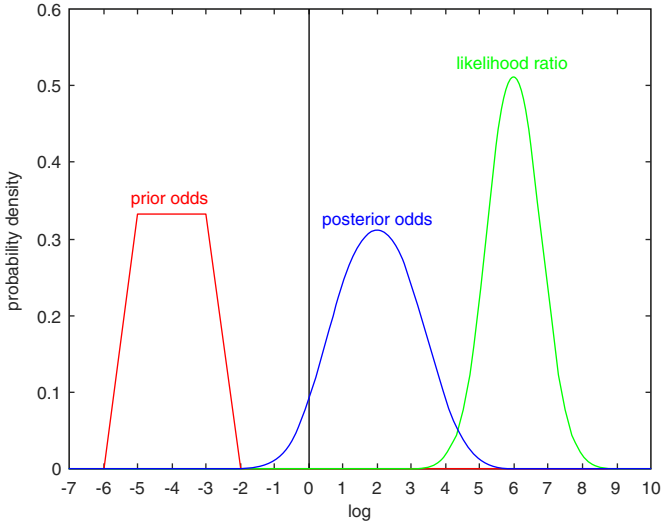

Fig. 1. Example of calculation of a posterior odds distribution using a prior odds distribution and a likelihood ratio distribution.

uncertainty need not be normally distributed, Fig. 1 shows a graphical representation in which the trier of fact's prior odds are high between 10^{-5} and 10^{-3}, fall rapidly outside this range, and are zero for less than 10^{-6} and greater than 10^{-2}. The forensic practitioner presents a likelihood ratio and the results of tests of the precision of the system which produced that likelihood ratio, and the resulting distribution is normal with a mean of 10^{6} (the best estimate of the value of the forensic practitioner's likelihood ratio) and a standard deviation of 0.765 on a base 10 logarithmic scale (the 95% credible interval is plus or minus 1.5 orders of magnitude, i.e., $10^{4.5}$ [31,623] to $10^{7.5}$). Note that the scale on the *x* axis of Fig. 1 is logarithmic. To combine prior odds and likelihood ratios on a linear scale they are multiplied together, on a logarithmic scale they are added. If both distributions were normal, they could be analytically added.[7] Since one of the distributions in this example is not normal, we used numerical convolution.[8]

Several things should be noted: The distribution of the posterior odds lies between that of the prior odds and that of the likelihood ratio. The distribution of the posterior odds is wider than both the distribution of the prior odds and the distribution of the likelihood ratio, i.e., the degree of uncertainty of the posterior odds is greater than that of either the prior odds or the likelihood ratio. Taking the current posterior odds as new prior odds and combining them with another likelihood ratio distribution will result in new posterior odds with even greater uncertainty. The greater the uncertainty of the likelihood ratio, the greater the uncertainty of the resulting posterior odds. If the bulk of the likelihood ratio distribution is, however, to the right of the prior odds distribution, then the posterior odds distribution will shift to the right relative to the prior odds distribution. In general, a series of likelihood ratios with wide distributions which straddle a log likelihood ratio of 0 (likelihood ratio of 1), will lead towards a log posterior odds distribution which has a wide distribution centred close to 0. A series of likelihood ratios with narrow distributions which are all far away in the same direction from a log likelihood ratio of 0 (all substantially higher, or all substantially lower) will lead to a log posterior odds distribution which is not as wide as in the previous example and which is centred far from 0.[9]

What would ultimately matter for the trier of fact, however, is what proportion of the log posterior odds distribution is greater than 0 and what proportion is less than 0. In the example shown in Fig. 1, 96% of the log posterior odds distribution is greater than 0. If, before the beginning of the presentation of evidence the trier of fact had decided that their decision threshold would be 95%, then they would now conclude in favour of the prosecution hypothesis, but if they had decided on a threshold of 99% they would conclude in favour of the defence hypothesis. We do not expect the courts to adopt this framework any time soon, if ever, but we present it as a normative rational decision making framework which takes imprecision into account.

One could object that the sort of framework we propose here is unnecessarily complicated (see Lindley [43] pp. 239–240). A coherent rational decision maker could simply use the expected values of the prior odds and likelihood ratio, obtain a point value for the posterior odds (in our example above $10^{-4} \times 10^{6} = 10^{2} = 100$), and compare that to a pre-specified threshold.[10] We would agree if the likelihood ratio under discussion were the trier of fact's likelihood ratio, but our aim was to describe a framework that would allow the trier of fact to make principled non-arbitrary use of the reported imprecision of the forensic practitioners' likelihood ratios. Of course, we are asking the trier of fact to adopt the forensic practitioner's likelihood ratio distribution as is, rather than only to adopt as is the forensic practitioner's best estimate of their likelihood ratio without any consideration of imprecision. Even if the sort of framework we propose here is more complicated than absolutely necessary for making coherent rational decisions, we believe that it has the practical advantage that it makes imprecision explicit and provides a principled way to deal with it, and we think it would therefore be perceived as more open and transparent by the courts. As Taylor et al. [16] p. 57 observe:

> [In] the cross examination process … any uncertainty or doubt is often explored at length. … Furthermore, … it is an accepted practice in adversarial systems that all reasonable uncertainty is conceded to the defendant. … Almost any measurement in science has an associated measure of uncertainty. Well prepared lawyers correctly investigate this avenue of questioning.

6. Conclusion

Despite their popularity, we think that ordinal scales which associate pre-specified verbal expressions with pre-specified ranges of likelihood ratio values are a very bad solution to the precision problem. We are opposed to the presentation of likelihood ratios based directly on untested uncalibrated subjective judgement. Presenting a point-value Bayes factor, rather than a likelihood ratio plus its precision, is attractive in the simplicity of the presentation of a single value which is conservative in that if the amount of training data is small (and sampling variability would likely be large) the calculated value will tend towards the neutral likelihood ratio value of 1. If, however, we take multiple samples from the same population, calculate Bayes factors, and find that each sample results in a different Bayes factor value, we think the courts would still be interested in that degree of variability.

Forensic scientists should evaluate and quantify strength of evidence in a manner which they believe to be logically correct and for which they can provide suitable warrant. Without compromising the latter, we also need to be concerned about transparency and how best to communicate our strength of evidence conclusions to the court.

[7] $\mu_{A+B} = \mu_A + \mu_B$, $\sigma^2{}_{A+B} = \sigma^2{}_A + \sigma^2{}_B$ (assuming zero covariance).

[8] A MATLAB script that the performs the calculations is available at http://geoff-morrison.net/#MorEnz2016_SCIJUS_VSI_pos.

[9] If the log likelihood ratio distributions were narrow but some substantially higher and some substantially lower than 0, this would lead to a log posterior odds distribution which is not as wide as in the first example and which is centred close to 0.

[10] Also, in our statement that what would ultimately matter for the trier of fact is the proportion of the log posterior odds distribution greater than 0 versus the proportion less than 0, we have ended up back at a posterior odds point estimate.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] G.S. Morrison, Special issue on measuring and reporting the precision of forensic likelihood ratios: introduction to the debate, Sci. Justice 56 (2016) 371–373.
[2] R. Chakraborty, M.R. Srinivasan, S.P. Daiger, Evaluation of standard error and confidence intervals of estimated multilocus genotype probabilities and their implications in DNA, Am. J. Hum. Genet. 52 (1993) 60–70.
[3] D.J. Balding, Estimating products in forensic identification using DNA profiles, J. Am. Stat. Assoc. 90 (1995) 839–844.
[4] B.S. Weir, Genetic Data Analysis II, Sinauer, Sunderland, MA, 1996.
[5] J.M. Curran, J.S. Buckleton, C.M. Triggs, B.S. Weir, Assessing uncertainty in DNA evidence caused by sampling effects, Sci. Justice 42 (2002) 29–37, http://dx.doi.org/10.1016/S1355-0306(02)71794-2.
[6] J.M. Curran, An introduction to Bayesian credible intervals for sampling error in DNA profiles, Law Prob. Risk 4 (2005) 115–126, http://dx.doi.org/10.1093/lpr/mgi009.
[7] G.S. Morrison, T. Thiruvaran, J. Epps, Estimating the precision of the likelihood-ratio output of a forensic-voice-comparison system, in: H. Cernocký, L. Burget (Eds.), Proceedings of Odyssey 2010: The Language and Speaker Recognition Workshop, Czech Republic, International Speech Communication Association, Brno 2010, pp. 63–70.
[8] G.W. Beecham, B.S. Weir, Confidence interval of the likelihood ratio associated with mixed stain DNA evidence, J. Forensic Sci. 56 (2011) S166–S171, http://dx.doi.org/10.1111/j.1556-4029.2010.01600.x.
[9] J.M. Curran, J.S. Buckleton, An investigation into the performance of methods for adjusting for sampling uncertainty in DNA likelihood ratio calculations, Forensic Sci. Int. Genet. 5 (2011) 512–516, http://dx.doi.org/10.1016/j.fsigen.2010.11.007.
[10] G.S. Morrison, C. Zhang, P. Rose, An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system, Forensic Sci. Int. 208 (2011) 59–65, http://dx.doi.org/10.1016/j.forsciint.2010.11.001.
[11] G.S. Morrison, Measuring the validity and reliability of forensic likelihood-ratio systems, Sci. Justice 51 (2011) 91–98, http://dx.doi.org/10.1016/j.scijus.2011.03.002.
[12] A. Nordgaard, T. Höglund, Assessment of approximate likelihood ratios from continuous distributions: a case study of digital camera identification, J. Forensic Sci. 56 (2011) 390–402, http://dx.doi.org/10.1111/j.1556-4029.2010.01665.x.
[13] S. Hancock, R. Morgan-Smith, J.S. Buckleton, The interpretation of shoeprint comparison class correspondences, Sci. Justice 52 (2012) 243–248, http://dx.doi.org/10.1016/j.scijus.2012.06.002.
[14] I. Alberink, A. Bolck, S. Menges, Posterior likelihood ratios for evaluation of forensic trace evidence given a two-level model on the data, J. Appl. Stat. 40 (2013) 2579–2600, http://dx.doi.org/10.1080/02664763.2013.822056.
[15] C. Zhang, G.S. Morrison, F. Ochoa, E. Enzinger, Reliability of human-supervised formant-trajectory measurement for forensic voice comparison, J. Acoust. Soc. Am. 133 (2013) EL54–EL60, http://dx.doi.org/10.1121/1.4773223.
[16] D. Taylor, J.-A. Bright, J.S. Buckleton, J.M. Curran, An illustration of the effect of various sources of uncertainty on DNA likelihood ratio calculations, Forensic Sci. Int. Genet. 11 (2014) 56–63, http://dx.doi.org/10.1016/j.fsigen.2014.02.003.
[17] N. Kaplan Damary, M. Mandal, N. Levin, E. Izrael, Calculation of likelihood ratios for gunshot residue evidence — statistical aspects, Law Prob. Risk. http://dx.doi.org/10.1093/lpr/mgw001
[18] F. Taroni, S. Bozza, A. Biedermann, C.G.G. Aitken, Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio, Law Prob. Risk 15 (2016) 1–16, http://dx.doi.org/10.1093/lpr/mgv008.
[19] F. Taroni, S. Bozza, A. Biedermann, C.G.G. Aitken, Rejoinder to Sjerps et al. and Nordgaard, Law Prob. Risk 15 (2016) 31–34, http://dx.doi.org/10.1093/lpr/mgv007.
[20] M.J. Sjerps, I. Alberink, A. Bolck, R.D. Stoel, P. Vergeer, J.H. van Zanten, Uncertainty and LR- to integrate or not to integrate, that's the question, Law Prob. Risk 15 (2016) 23–29, http://dx.doi.org/10.1093/lpr/mgv005.
[21] N. Brümmer, A. Swart, Bayesian calibration for forensic evidence reporting, Proceedings of Interspeech 2014, pp. 388–392 (http://www.isca-speech.org/archive/interspeech_2014/i14_0388.html).
[22] International Organization for Standardization, ISO/IEC 17025:2005 General Requirements for the Competence of Testing and Calibration Laboratories, International Organization for Standardization, Geneva, Switzerland, 2015.
[23] United Kingdom Accreditation Service, UKAS M3003 The Expression of Uncertainty and Confidence in Measurement, United Kingdom Accreditation Service, Feltham, UK, 2012.
[24] National Research Council, Strengthening Forensic Science in the United States: A Path Forward, National Academies Press, Washington, DC, 2009 (http://www.nap.edu/catalog.php?record_id=12589).
[25] C.E.H. Berger, J. Buckleton, C. Champod, I.W. Evett, G. Jackson, Evidence evaluation: a response to the court of appeal judgment in R v T, Sci. Justice 51 (2011) 43–49, http://dx.doi.org/10.1016/j.scijus.2011.03.005.
[26] D.M. Risinger, Reservations about likelihood ratios (and some other aspects of forensic 'Bayesianism'), Law Prob. Risk 12 (2013) 63–73, http://dx.doi.org/10.1093/lpr/mgs011.
[27] G.S. Morrison, R.D. Stoel, Forensic strength of evidence statements should preferably be likelihood ratios calculated using relevant data, quantitative measurements, and statistical models — a response to Lennard (2013) Fingerprint identification: how far have we come? Aust. J. Forensic Sci. 46 (2014) 282–292, http://dx.doi.org/10.1080/00450618.2013.833648.
[28] G.S. Morrison, Distinguishing between forensic science and forensic pseudoscience: testing of validity and reliability, and approaches to forensic voice comparison, Sci. Justice 54 (2014) 245–256, http://dx.doi.org/10.1016/j.scijus.2013.07.004.
[29] E. Enzinger, G.S. Morrison, Mismatched distances from speakers to telephone in a forensic-voice-comparison case, Speech Comm. 70 (2015) 28–41, http://dx.doi.org/10.1016/j.specom.2015.03.001.
[30] E. Enzinger, G.S. Morrison, F. Ochoa, A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case, Sci. Justice 56 (2016) 42–57, http://dx.doi.org/10.1016/j.scijus.2015.06.005.
[31] B. Found, Deciphering the human condition: the rise of cognitive forensics, Aust. J. Forensic Sci. 47 (2015) 386–401, http://dx.doi.org/10.1080/00450618.2014.965204.
[32] C. Champod, I.W. Evett, Commentary on Broeders (1999) Some observations on the use of probability scales in forensic identification, Forensic Ling. 7 (2000) 238–243.
[33] Association of Forensic Science Providers, Standards for the formulation of evaluative forensic science expert opinion, Sci. Justice 49 (2009) 161–164, http://dx.doi.org/10.1016/j.scijus.2009.07.004.
[34] A. Nordgaard, R. Ansell, W. Drotz, L. Jaeger, Scale of conclusions for the value of evidence, Law Prob. Risk 11 (2012) 1–24, http://dx.doi.org/10.1093/lpr/mgr020.
[35] S. Bunch, G. Weavers, Application of likelihood ratios for firearm and toolmark analysis, Sci. Justice 53 (2013) 223–229, http://dx.doi.org/10.1016/j.scijus.2012.12.005.
[36] S.M. Willis, L. McKenna, S. McDermott, G. O'Donell, A. Barrett, B. Rasmusson, A. Nordgaard, C.E.H. Berger, M.J. Sjerps, J.J. Lucena-Molina, G. Zadora, C.G.G. Aitken, L. Lunt, C. Champod, A. Biedermann, T.N. Hicks, F. Taroni, ENFSI Guideline for Evaluative Reporting in Forensic Science, European Network of Forensic Science Institutes, Wiesbaden, Germany, 2015 (http://enfsi.eu/sites/default/files/documents/external_publications/m1_guideline.pdf).
[37] I.W. Evett, G. Jackson, J.A. Lambert, S. McCrossan, The impact of the principle of evidence interpretation on the structure and content of statements, Sci. Justice 40 (2000) 233–239, http://dx.doi.org/10.1016/S1355-0306(00)71993-9.
[38] B. Robertson, G.A. Vignaux, C.E.H. Berger, Extending the confusion about Bayes, Modern Law Rev. 74 (2011) 444–455, http://dx.doi.org/10.1111/j.1468-2230.2011.00857.x.
[39] C. Neumann, I.W. Evett, J. Skerrett, Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm, J. R. Stat. Soc. Ser. A 175 (2012) 371–415, http://dx.doi.org/10.1111/j.1467-985X.2011.01027.x.
[40] O.E. Facey, R.J. Davis, Re: Expressing evaluative opinions; a position statement, Sci. Justice 51 (2011) 212, http://dx.doi.org/10.1016/j.scijus.2011.06.001.
[41] J. Lindh, G.S. Morrison, Forensic voice comparison by humans and machine: forensic voice comparison on a small database of Swedish voice recordings, in: W.-S. Lee, E. Zee (Eds.), Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, China, Organizers of ICPhS XVII at the Department of Chinese, Translation and LinguisticsCity University of Hong Kong, Hong Kong, China 2011, pp. 1254–1257.
[42] D. Ramos, J. Franco-Pedroso, J. González-Rodríguez, Calibration and weight of the evidence by human listeners. The ATVS-UAM submission to NIST Human Aided Speaker Recognition 2010, Proceedings of the International Conference on Speech Signal Processing, Czech Republic, IEEE, Prague 2011, pp. 5908–5911, http://dx.doi.org/10.1109/ICASSP.2011.5947706.
[43] D.V. Lindley, Understanding Uncertainty, Wiley, Hoboken, NJ, 2006.