NATIONAL COMMISSION ON FORENSIC SCIENCE

Views of the Commission

Statistical Statements in Forensic Testimony

Subcommittee:	Reporting and Testimony
Date of Current Version:	19/07/16
Approved by Subcommittee:	22/08/16
Status:	Initial Draft

Comments by: Geoffrey Stewart Morrison

http://geoff-morrison.net/

Comment version: 2016-09-13a

Comments

I appreciate the direction of the draft views document and would like to express that I am supportive of what I perceive to be the goal. I find, however, that parts of the document are unclear, overly simplified, or appear to represent misconceptions. I provide these comments in the hope that they will be helpful to the Subcommittee in producing an even better document in the next draft.

The draft views document conceptualizes evaluation of evidence in terms of the probability of obtaining a “match” if the questioned-source specimen and the known-source sample had the same origin versus the probability of obtaining a “match” if they had different origins. Although this is not an incorrect characterization of a likelihood ratio, the concept of “match” is not appropriate for the type of data which occurs in many branches of forensic science: data which are continuously valued and which have intrinsic within-source variability, and/or inherent within-source variability due to transfer and/or measurement processes. A classic example is forensic voice comparison (aka forensic speaker recognition) in which the intrinsic properties if each speaker’s speech are highly variable and the speech signal is often distorted due to transmission through communications channels. The “match / non-match” approach assumes that the numerator of the likelihood ratio will be either 1 or 0. This is appropriate if the data are discrete and invariant. Data are treated as such by first-approximation DNA models, but note that there is variability in the measurement process and DNA models for continuously-valued data are now in operation. For continuously-valued data with intrinsic or inherent variability it is generally not appropriate to apply a “match” threshold to set the numerator of the likelihood ratio as either 1 or 0. It is much more appropriate for the numerator of the likelihood ratio to be a likelihood value (which is not constrained to be 1 or 0) calculated using a model trained directly on the continuously-valued data from the known-source sample. Likewise, it is much more appropriate for the denominator of the likelihood ratio to be a likelihood value calculated using a model trained directly on the continuously-valued data from a sample of the relevant population. These models can be used to estimate the probability of the continuously-valued data from the questioned-source specimen if they came from the known source versus the probability of the continuously-valued data from the questioned-source specimen if they came from some other source selected at random from the relevant population. (There are other legitimate ways to describe and calculate a forensic likelihood ratio. They are generally more complex, and for simplicity I will consistently use the formulation just given.)

In his presentation to the Commission on September 13, Hari Iyer characterized a forensic practitioner’s likelihood ratio as a subjective belief. Although there are some who argue that the forensic practitioner’s likelihood ratio should be a subjective belief (e.g., Taroni et al, 2016), I and others (e.g, Sjerps et al, 2016) argue against this. I advocate that a likelihood ratio presented by a forensic scientist be a numeric value calculated using relevant data, quantitative measurements, and statistical models (and the it be directly the value output by the statistical model); and that it not be a numeric value or verbal expression based directly on subjective judgments made by the forensic practitioner (see, Morrison, 2014; Morrison & Stoel, 2014; Morrison & Enzinger, 2016).

Footnote 3 of the draft views document states: “Some courts have held that probabilities need not be provided in such cases—the analyst may testify that the defendant could not be excluded as a contributor to a mixed DNA sample without giving a probability of exclusion or other statistic.” To state that an individual could not be excluded may be highly misleading unless information is also provided about how many other individuals from the relevant population could not be excluded. Jackson (2009) explains why it is misleading to express strength of evidence in this way.

The draft views document states that “For most of trace and pattern evidence today there is no commonly accepted probabilisitic model”. I think there is a need to emphasize that this is not a problem. Forensic scientists and forensic statisticians will apply different models to different problems (e.g., different data in different branches of forensic science) because different tools are appropriate for solving different problems. A tool which is appropriate for addressing one problem may be inappropriate for addressing another problem. Also, rapid advances are being made in statistical modeling in some branches of forensic science, and newer models with better performance will often replace older models. There is no single commonly accepted statistical model, because such a thing would be inappropriate. There may, however, be generally accepted approaches to statistical modeling. Although "general acceptance" is listed as one of the *Daubert* criteria, from a scientific perspective it is one of the other *Daubert* criteria, the need for empirical testing of validity and reliability, which is much more important. If a back-box system is demonstrated to have a sufficient degree of accuracy and precision under conditions sufficiently reflective of the conditions of the case under investigation, then the internal workings of the black-box and whether they are generally accepted should not be of concern.

The draft views document states: "All of the statistical calculations should be replicable given the data and statistical model, whereas the quantitative summary of the actual forensic evidence may vary from examiner to examiner and from laboratory to laboratory. Such measurement error should be an integral part of the expert report." It is not clear what the referent is of "such measurement error" in the last sentence. I would argue that examiner to examiner and laboratory to laboratory variably are irrelevant for the court. What the court needs to know, in order to exercise its gatekeeping role under *Daubert* or as part of considering the weight of the evidence, are the results of empirical tests of the accuracy and precision of the particular method as used by the particular examiner who performed the analysis in the current case, and those tests should be conducted under conditions which are sufficiently similar to the conditions of the current case for the results to be meaningful for the current case. *Kumho* emphasizes that the judge in an admissibility hearing should determine whether the expert's methods are valid for drawing the kinds of conclusions that the examiner drew in the case at hand. I discuss this issue in Morrison (2014).

The draft views document states: "Any recommendation on presenting explicit probabilities or likelihood ratios in light of forensic evidence might distinguish between probabilities based on some statistical model and ones said to flow from the forensic evidence itself." I do not know what is meant by "said to flow from the forensic evidence itself". Likelihood ratios calculated using relevant data, quantitative measurements, and statistical models "flow from the forensic evidence itself". Was the intention to contrast procedures based on relevant data, quantitative measurements, and statistical model versus procedures in which the conclusion as to the strength of the evidence is based directly on a subjective judgment?

The draft views document asks: "Does a forensic technician testifying in court actually compute significance probabilities or likelihood ratios, or do they come from a computer program developed by statisticians and related forensic experts?" What is intended is unclear. All non-trivial calculations of likelihood ratios will be performed by computer programs.

The draft views document asks: “If the statistical model is known to be at best an approximation, how should the probabilistic statements coming from it be viewed?” All statistical models are approximations. That is the nature of statistical models. A subjective judgment could not be any less of an approximation, and would be much less transparent and replicable than a procedure based on relevant data, quantitiative measurments, and statistical models. “The fact that [a model] is an approximation does not necessarily detract from its usefulness because all models are approximations. Essentially, all models are wrong, but some are useful.” (Box & Draper, 1987, p. 424). The relevant question is whether the model is sufficiently useful, i.e., whether it performs sufficiently well under the conditions of the case under investigation, which is something which should be empirically tested.

The draft views document asks: “What if the statistical model and method used to analyze the evidence do not admit naturally to the simplistic form of likelihood ratio increasingly favored in the forensic-science literature?” This question is based on a false premise. For didactic reasons, introductions to the likelihood ratio framework often start with simple or simplified examples, but the likelihood ratio framework is not simplistic. Additional research is needed, there remain problems to be solved and improvements to be made, but there is abundant peer-reviewed literature describing sophisticated statistical modeling techniques applied to the problems of calculating likelihood ratios for data from multiple branches of forensic science.

The draft views document asks: “How should probabilistic statements be viewed if they are not based on all of the ‘relevant evidence’?” There can be legitimate debate about what evidence is relevant; however, this question is usually asked by opponents of the use of likelihood ratios in general, or of the calculation of a likelihood ratio in a particular instance. They attempt to argue that likelihood ratios in general are deficient, or that a particular forensic analysis is deficient. The answer is the same as for the question related to statistical models being approximations: What matters is whether the probabilistic statements generated by a statistical model are useful, and this can be determined via empirical testing of the statistical model under conditions reflecting those of the case under investigation. The number of potential loci in the human genome is in the millions, but forensic DNA analyses are conducted using measurements made on a small number of loci, the number usually varies from the low teens to the low twenties depending on the particular system. The loci that are used in forensic analyses and the statistical models applied to data from those loci, are not, however, the result of arbitrary choices – they were selected via research and testing that demonstrated high degrees of accuracy and precision. The appropriate question is not whether all possible data have been used, the appropriate question is whether the performance of a system that exploits some data is good enough. If the performance of the system is sufficient, then it must have exploited sufficient relevant data.

The draft views document states that it is the view of the Commission that: “1. No one form of statistical statement is most appropriate to all forms of pattern and trace evidence”. I disagree with this view. I believe that a numeric likelihood ratio calculated on the basis of relevant data, quantitative measurements, and statistical models is the most appropriate means of quantifying and stating the strength of evidence; and that the accuracy and precision of the system calculating the likelihood ratio value must be empirically tested under conditions sufficiently similar to those of the case under investigation. For supporting arguments, see: Morrison (2014), Morrison & Stoel (2014), Morrison & Enzinger (2016).

The draft views document states that it is the view of the Commission that: “2. More importantly, the forensic expert, reporting whatever statistical quantity, needs to be able to also report on the uncertainty associated with it in some form. This might take the form of a reported interval or more typically separate statements regarding errors and uncertainties associated with the analysis of the evidence and not simply the variations in the likelihoods themselves.” These statements are too vague for me to understand exactly what the Commission is in favor of. Personally I believe that forensic practitioners should assess and report the precision of the likelihood ratio they present to the courts, but whether forensic practitioners should do this, and if so how, is actually a matter of debate. I am currently a Guest Editor for a virtual special issue on this topic in the journal Science & Justice: http://www.sciencedirect.com/science/journal/13550306/vsi Authors of all contributions published so far appear to argue that “uncertainties associated with the analysis” be either directly incorporated into the calculation of the likelihood ratio value itself, or be reported as a credible or confidence interval or an upper or lower bound associated with the calculated likelihood ratio value.

The draft views document states that it is the view of the Commission that: “3. Forensic experts should present and describe the similarities and differences in the feature sets of the questioned and known samples (the data).” I disagree with this view as stated. Forensic practitioners should assess BOTH the similarity of the questioned-source specimen with the known-source sample, AND the typicality of the questioned-source specimen with respect to a sample of the relevant population. That is they should estimate the probability of obtaining the observed properties of the questioned-source specimen if it came from the known source versus if it came from some other source selected at random from the relevant population.

The draft views document states that it is the view of the Commission that: “4. Forensic experts should not state that a specific individual or object is the source of a trace without explaining that it is possible that other individuals or objects could possess or have left a similar set of observed features.” I agree with what I interpret to be the general intent of this view, but I disagree with the view as expressed in that I believe that under no circumstances should a forensic practitioner “state that a specific individual or object is the source of a trace”. To do so would be to express an opinion as to the posterior probability of the hypothesis. Logically, a posterior probability can only be derived by combining a likelihood ratio with a prior probability. The prior probability would properly be in the mind of the trier of fact, and would take account of other evidence which may already have been presented at trial. Weighing all the evidence and reaching a decision is the task of the trier of fact, not the forensic practitioner. Unless the trier of fact tells the forensic practitioner what prior probability value to use, the forensic practitioner cannot calculate an appropriate posterior probability. It would not be appropriate for forensic practitioners to consider the other evidence in the case or to use arbitrary priors, irrespective of whether each forensic practitioner used a different prior or they all used the same prior. Likelihood ratios associated with different pieces of evidence which can reasonablely be treated as statistically independent, e.g., finger print evidence and voice evidence, can logically be combined via simple multiplication, but there is no logically correct way to combine posterior probabilities (or posterior odds, except if one ascertained what prior odds were used in calculating the posterior odds in each case so as to be able to calculate what the likelihood ratio was in each case).

In view 5A, “relevant population” should read “sample of a relevant population”.

In view 5A, “noting the uncertainties in these frequencies as estimates of the frequencies in the population” is incorrect. In frequentist terms, the frequency in the sample would be the estimate of the frequency in the population. There would be no uncertainties in the frequency in the sample, but analytical or numerical procedures could be used to obtain an estimate of the uncertainty of the estimate of the frequency in the population.

View 5B is very difficult to understand. It needs to be rewritten.

The draft views document states that it is the view of the Commission that: “6. Forensic experts should confine themselves to speaking of the weight of the evidence (the support it lends to the parties’ claims, e.g., ‘this is strongly indicative of identity – we expect to find it 10,000 times more often when it comes from the suspect than when it comes from a coincidentally matching person’).” I agree that forensic practitioners should restrict themselves to opining on the strength of the evidence given the competing hypothesis (e.g., same-origin versus different-origin hypotheses); however, “this is strongly indicative of identity" seems to me to be contrary to this principle. It appears to be a verbal expression of the posterior probability of a single hypothesis. Speaking in terms of "support" for a hypothesis also invites the prosecutor's fallacy (Thomson & Schumann, 1987). In general, I am do not favor the use of verbal expressions instead of (or in addition to) numeric likelihood ratio values, see: Morrison & Enzinger (2016).

I agree with view 7. (The question mark should be deleted.)

References

Box GEP, Draper NR (1987). *Empirical model-building and response surfaces*. Oxford: Wiley.

Jackson G (2009). Understanding forensic science opinions. Ch. 16 (pp. 419–445) in Fraser J, Williams R (Eds.), *Handbook of Forensic Science*. Cullompton, UK: Willan.

Morrison GS (2014). Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison. *Science & Justice*, 54, 245–256. http://dx.doi.org/10.1016/j.scijus.2013.07.004

Morrison GS, Enzinger E (2016). What should a forensic practitioner's likelihood ratio be? *Science & Justice*. http://dx.doi.org/10.1016/j.scijus.2016.05.007

Morrison GS, Stoel RD (2014). 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? *Australian Journal of Forensic Sciences*, 46, 282–292. http://dx.doi.org/10.1080/00450618.2013.833648

Sjerps MJ, Alberink I, Bolck A, Stoel RD, Vergeer P, van Zanten JH (2016). Uncertainty and LR: to integrate or not to integrate, that's the question. *Law, Probability and Risk*, 15, 23–29. http://dx.doi.org/10.1093/lpr/mgv005

Taroni F, Bozza S, Biedermann A, Aitken CGG (2016). Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. *Law, Probability and Risk*, 15, 1–16. http://dx.doi.org/10.1093/lpr/mgv008

Thompson WC, Schumann EL (1987). Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior*, 11, 167–187. http://www.jstor.org/stable/1393631