Received: 17 December 2020 | Accepted: 13 August 2021

DOI: 10.1111/rssa.12747

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Royal Statistical Society
Series A
Statistics
in Society

ORIGINAL ARTICLE

Testing for calibration discrepancy of reported likelihood ratios in forensic science

Jan Hannig[1,2] | Hari Iyer[2]

[1]University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

[2]National Institute of Standards and Technology, Gaithersburg, Maryland, USA

Correspondence
Jan Hannig, University of North Carolina at Chapel Hill, 330 Hanes Hall, Chapel Hill, North Carolina 27599-3260, USA.
Email: jan.hannig@unc.edu

Funding information
Special Programs Office, National Institute of Standards and Technology; US National Science Foundation, Grant/Award Number: IIS-1633074, DMS-1916115 and DMS-2113404

Abstract

The use of likelihood ratios for quantifying the strength of forensic evidence in criminal cases is gaining widespread acceptance in many forensic disciplines. Although some forensic scientists feel that subjective likelihood ratios are a reasonable way of expressing expert opinion regarding strength of evidence in criminal trials, legal requirements of reliability of expert evidence in the United Kingdom, United States and some other countries have encouraged researchers to develop likelihood ratio systems based on statistical modelling using relevant empirical data. Many such systems exhibit exceptional power to discriminate between the scenario presented by the prosecution and an alternate scenario implying the innocence of the defendant. However, such systems are not necessarily well calibrated. Consequently, verbal explanations to triers of fact, by forensic experts, of the meaning of the offered likelihood ratio may be misleading. In this article, we put forth a statistical approach for testing the calibration discrepancy of likelihood ratio systems using ground truth known empirical data. We provide point estimates as well as confidence intervals for the calibration discrepancy. Several examples, previously discussed in the literature, are used to illustrate our method. Results from a limited simulation study concerning the performance of the proposed approach are also provided.

KEYWORDS
Bayes factor, calibration, forensic statistics, generalized fiducial inference, likelihood ratio

1 | INTRODUCTION

The forensic science community is increasingly moving towards providing expert forensic opinion in the form of a numerical summary called a *likelihood ratio* (*LR*). This view is supported by many recognized advisory bodies in the United Kingdom and in Europe. See, for instance, Aitken et al. (2010) and Willis et al. (2015). In the United States, rules concerning admissibility of expert evidence in legal proceedings are discussed in the US Federal Rules of Evidence, Rule 702 (https://www.law.cornell.edu/rules/fre/rule_702, accessed on 06/08/2021) which was most recently amended in the year 2000 in response to Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993) thus establishing the *Daubert* standard. Similar admissibility criteria have been adopted in England and Wales. See, in particular, section 19.4 (f), (g), and (h) in the Criminal Procedure Rules 2020, UK Statutory Instruments 2020 No. 759 (L. 19) (https://www.legislation.gov.uk/uksi/2020/759/article/19.4, accessed on 06/08/2021). These criteria focus on the reliability of the expert testimony, especially opinion testimony, and interpret reliability to mean 'trustworthy', see the Law Commission Report (2011), page 7, para 1.27 (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/229043/0829.pdf, accessed on 06/08/2021).

Recognizing that *LR*s are personal (Biedermann, 2013; Franck & Gramacy, 2020; Kadane, 2020; Lindley, 2013), they are to be regarded as opinion testimony given in numerical form. In particular, it is typical for an expert providing an opinion using *LR* to explain an *LR* value of x as 'the findings are x times more likely under the prosecution proposition than under the defence proposition'. However, triers of fact need to be provided whatever information is available that can be used by them to judge the reliability of such assessments.

Lindley (1977) considered the problem of deciding whether trace materials such as glass fragments or paint chips, recovered from a crime scene, have come from a known source. He first addressed this problem in the case where measurements made on the trace material are univariate. More specifically, suppose X and Y denote the measurements from the crime scene material and from a known source respectively. Suppose H_p is the proposition that X comes from the same source that produced Y and H_d denotes the complement of H_p. Noting that the posterior odds on H_p, after examining the information provided by X and Y, are obtained by multiplying the prior odds on H_p by the factor

$$\frac{P[X, Y \mid H_p]}{P[X, Y \mid H_d]} \tag{1}$$

Good (1950) referred to this factor as a ratio of likelihoods as did Lindley (1977). The notation $P[X, Y|A]$ stands for the probability of observing X and Y given that the proposition A is true. The quantity in Equation (1) is in fact a Bayes Factor since it is the ratio of the posterior odds for H_p to the prior odds for H_p (Kass & Raftery, 1995). In forensic circles this factor is often referred to as a likelihood ratio (*LR*) and we will use that terminology here. The logarithm of *LR* (to the base 10 or to any other base) is interpreted as the weight of evidence in support of H_p provided by X and Y.

Although the notion that *LR* should be used as the value of evidence appears in Good (1950), it seems that much of the recent literature on the development of quantitative methods for forensic evidence assessment has been inspired by Lindley (1977). In particular, considerable work has been done on the assessment of evidential value of DNA samples from crime scenes using *LR*s. Aided by the scientific advances in our understanding of human genetics and in our measurement capabilities of even very minute quantities of DNA, many crime laboratories currently report the strength of DNA evidence using *LR*s.

Another forensic discipline where *LR* inspired methods have been proposed is that of fingerprint comparisons. Neumann et al. (2012) developed a likelihood ratio model using the minutiae configurations on fingermarks. They conducted studies to evaluate the discriminating power of their *LR* model.

Zadora et al. (2013) describe various statistical models for calculation of likelihood ratios in situations where the evidence consists of univariate or multivariate measurements. The disciplines where their models are found to be useful include paint comparisons, ink comparisons, glass fragment comparisons and comparisons of other trace material. *LR* models are also used in forensic speaker identification and identification of illegal chemicals, for example, drugs.

Table 1 provides selected references from each discipline to emphasize the diversity of the discipline areas where the *LR* framework is being used or is being advocated for adoption. In some cases the quantitative methods advocated are not *LR*s in any sense but *scores* that have good power to discriminate between H_p-true scenarios and H_d-true scenarios.

In some disciplines *LR* assessment is made using algorithms and software. We refer to these as *LR* systems. In other disciplines the *LR* assessments are sometimes made using subjective probabilities. See, for instance, the discussion on page 16, Willis et al. (2015). In principle, our proposed methodology is applicable equally well to either situation. However, ground truth known validation data are hard to come by in the case of subjectively assessed probabilities and likelihood ratios. As a result, the reliability of such evidential value assessments is difficult to judge empirically. The methodology presented here can be readily applied to *LR* systems such as those used in probabilistic genotyping for DNA mixtures, forensic speaker recognition and other disciplines where *LR* systems are being developed.

TABLE 1 Selected references discussing the use of *LR*, or systems inspired by *LR* theory, in various forensic disciplines

DNA Single Source & Mixtures	Evett and Weir (1998), Butler (2014), Buckleton et al. (2021), Nic Daéid et al. (2017)
Fingerprints	Neumann et al. (2012), Swofford et al. (2018), Leegwater et al. (2017), Morrison and Stoel (2014)
Footwear	Park (2018), Park and Carriquiry (2020), Venkatasubramanian et al. (2021a,b)
Glass fragments	Lindley (1977), Park (2018), Park and Tyner (2019), Curran et al. (1997a,b), Curran et al. (1999), Zadora et al. (2013)
Fiber	Evett et al. (1987), Morgan (2014), Causin et al. (2004)
Drugs	Bolck et al. (2015)
Speaker recognition	Ramos (2007), Enzinger (2016), Morrison et al. (2020)
Handwriting & Authorship	Chen et al. (2018), Bozza et al. (2008), Saunders et al. (2010), Martire et al. (2018)
Firearms & Toolmarks	Bunch and Wevers (2013), Song et al. (2018), Kerkhoff et al. (2013), Dong et al. (2019)

One encounters two kinds of *LR* assessments in the literature—feature-based *LR*s and score-based *LR*s. Several authors have pointed out the deficiencies of score-based *LR*s and have questioned their suitability for use in forensic settings (Bolck et al., 2015; Neumann, 2020; Neumann & Ausdemore, 2020; Neumann et al., 2020). In any case, our procedure applies equally well for exploring the reliability of either approach.

2 | LITERATURE REVIEW

Many previous authors have contributed to the topic of calibrated probabilities and calibrated likelihood ratios. A key publication discussing the accuracy of probability assessments is by DeGroot and Fienberg (1983) where the authors consider the empirical evaluation of probability assessments using the context of weather forecasters for illustrative purposes. They showed that the overall inaccuracy of probability assessments can be decomposed into a component due to lack of calibration and a component due to lack of refinement or discrimination ability.

Morrison (2013) discussed the use of logistic regression to calibrate uncalibrated *LR*s. Morrison and Poh (2018) proposed the use of shrunk *LR*s and Bayes factors as a way of avoiding overstatement of strength of evidence. Vergeer et al. (2016) noted that *LR* values outputted by many *LR* systems are based on extrapolation and discussed a strategy for applying a cap on the reported *LR* values.

Diagnostic checks of calibration accuracy of *LR*s produced by algorithmic *LR* systems have been around for a long time. For instance, it follows from the definition of a likelihood ratio that the expected value (when it exists) of likelihood ratios corresponding to H_d-true instances must be one. This fact has led to an empirical diagnostic procedure of checking the sample average of *LR* values known to correspond to H_d true scenarios. However, this is only a necessary condition but not sufficient and therefore can be misleading in the sense that failure to detect lack of calibration using this diagnostic does not imply actual lack of calibration. Vergeer et al. (2021) discuss a number of metrics that have been previously considered as diagnostic checks of the calibration accuracy of *LR* systems. In addition, they introduce a new metric, called `devPAV`, to assess calibration accuracy and compare their metric with other metrics in the literature with respect to their ability to differentiate between well-calibrated systems and ill-calibrated systems.

Measures of overall calibration accuracy of *LR*s were introduced in Brümmer and Du Preez (2006) and further investigated in Ramos and Gonzalez-Rodriguez (2013) and Ramos et al. (2018). However, as an anonymous referee has pointed out, these authors did not consider the effect of sampling variability in the proposed metrics, thereby leaving the question open as to whether the observed deviations from being well calibrated are *real* or simply a consequence of sampling variability based on the particular sample chosen for training models or for checking the calibration status of the system.

Morrison et al. (2021) discuss, in the context of forensic voice comparison, but relevant to other forensic disciplines as well, the empirical validation of *LR* systems. They emphasize the need for using *LR* systems that are well calibrated. In Appendix C.2 of their paper they explain what is expected from a well-calibrated *LR* system in terms of the metric C_{llr} (Brümmer and Du Preez, 2006) and also in terms of what a *Tippett plot* (Meuwly, 2001) should look like. However, these requirements are only necessary but not sufficient for a system to be well calibrated. In fact, they say 'For well-calibrated systems, C_{llr} values lie in the range 0 to approximately 1'. They then also say 'A C_{llr} value less than 1 does not necessarily imply that the system is well calibrated; miscalibration may be apparent in the Tippett plot'.

A careful review of the literature reveals that there is a need for statistical methodology that can provide a clear answer to the question:

- When an *LR* system produces an *LR* value equal to *x*, to what extent might this value be an overstatement or an understatement of the value of evidence in light of the available empirical validation data? That is, to what extent can the validation data support the *LR* produced by the system?

In this paper we provide such a methodology and demonstrate, using published examples, how our method can be used to answer the above question and contrast it with what information is obtainable from previously proposed metrics.

3 | GENERAL PROBLEM STATEMENT IN THE CONTEXT OF SOURCE IDENTIFICATION

A pair of mutually exclusive and exhaustive propositions (hypotheses) are considered. One proposition, denoted H_p (prosecution proposition) states that the person or object of interest is the source (or, in some cases, one of the sources) of the crime scene evidential material. The other proposition, denoted H_d (defence proposition or alternative proposition) states that the person or object of interest is not the source of the crime scene evidential material (but another person or another object is). It is envisioned that the trier of fact (TOF) will assess the odds

$$\frac{P[H_p \mid E, I]}{P[H_d \mid E, I]}$$

where *E* denotes forensic expert's findings and *I* denotes background information available prior to introduction of *E* (case context and other evidence introduced in court prior to the particular evidence *E* under discussion). Bayes Rule says

$$\frac{P[H_p \mid E, I]}{P[H_d \mid E, I]} = \frac{P[E \mid H_p, I]}{P[E \mid H_d, I]} \times \frac{P[H_p \mid I]}{P[H_d \mid I]}$$

that is, posterior odds is the product of $LR = \frac{P[E \mid H_p, I]}{P[E \mid H_d, I]}$ and prior odds. Since prior odds are not within the purview of a forensic expert it has come to be accepted that a forensic expert makes, often with the assistance of empirically derived statistical models, an assessment of *LR* and includes it in the casework report.

LR assessments by different forensic experts can be, and often are, quite different. This can be due to a variety of reasons. Some of these are listed below.

1. **What features have been extracted from the evidential material?** Different *LR* systems may be based on different feature sets. For instance, some *LR* software for analysing DNA evidence make use of only peak locations from the electropherograms whereas others make use of peak locations as well as peak heights.
2. **Differences in model choice:** Even when statistical models are developed for the same set of features, different *LR* systems may use different statistical models.

3. **Data used during model development:** Even when the same set of features and the same model families are used, the models implemented in different *LR* systems may have been trained using different sets of data that differ with respect to size and how representative they are of any particular application scenario.

The user is then left with the task not only of assessing the adequacy of any candidate model being considered for casework use but also of comparing competing models. For a discussion of these and other related issues the reader is referred to Lund and Iyer (2017), Gelman and Hennig (2017), and Young (2018).

It is well known that the overall performance of an *LR* system is a function of two components: (1) the ability of the system to discriminate between H_p-true scenarios and H_d-true scenarios, and (2) how well calibrated the system is. See, for instance, Brümmer and Du Preez (2006), Zadora et al. (2013), and Ramos et al. (2018). Discrimination power of any system, *LR* or otherwise, is described conveniently by the receiver operating characteristic (ROC) for that system and the associated area under the ROC plot (AUC). In this paper we focus on the question of whether or not the *LR* offered by the expert is calibrated. The expert is saying that *E* is *LR* times more likely under H_p than under H_d. The expert's assessment is often aided by empirical data but subjective assessments of likelihood ratios are also acceptable to many forensic experts and forensic institutes (Aitken et al., 2010; Willis et al., 2015). In either case, how can one assess, using empirical validation data, the reliability of the process applied by the expert to make his/her assessment of the value of evidence?

3.1 | Calibration property of *LR* systems

We first remind the reader what is meant by the question 'is the *LR* system well calibrated?

Let z be equal to 1 if H_p is true and equal to 0 if H_d is true. Suppose the TOF's value for prior odds for H_p is θ, that is, the TOF's prior probability that H_p is true is $\theta/(1+\theta)$. Suppose the expert is using the pdf η to model his/her uncertainties regarding a random variable X (evidence) conditional on $z = 1$ and the pdf ψ to model uncertainties conditional on $z = 0$. Having observed a realization x of X, and in the absence of information regarding z, an expert is providing the TOF his/her likelihood ratio $r(x)$. So

$$r(x) = \frac{\eta(x)}{\psi(x)}.$$

Suppose the TOF uses the expert's *LR* as the TOF's *LR* to calculate the posterior probability $\pi = P[z = 1|r(X) = r]$ as

$$\pi = \frac{r(x)\theta}{r(x)\theta + 1}.$$

This posterior probability is calibrated, according to the TOF, if and only if the TOF's probability for the event $z = 1$, conditional on $r(X) = r$, equals π.

Suppose the TOF's uncertainties about $r(X)$ are characterized by $g(r)$ when given $z = 1$ and $f(r)$ when given $z = 0$. Then the TOF's probability that $z = 1$ given $r(X) = r$ is equal to

$$P[z=1\,|\,r(X)=r]=\frac{\frac{\theta}{\theta+1}g(r)}{\frac{\theta}{\theta+1}g(r)+\frac{1}{\theta+1}f(r)}=\frac{\theta g(r)}{\theta g(r)+f(r)}.$$

It is easy to see that this is equal to π if and only if $\frac{g(r)}{f(r)}=r$. That is, r is calibrated according to the TOF if and only if the TOF's *LR* for r (the expert's *LR*) is equal to r. We state this in the following lemma.

Lemma 1 *The expert's* LR *is calibrated with respect to the TOF's probabilities if and only if the TOF's* LR *of the expert's* LR *is equal to the expert's* LR. *That is,*

$$\frac{g(r)}{f(r)}=r\ \text{ for all }\ r\ \text{ in }\ (0,\infty). \tag{2}$$

For the record, we also state the following lemma whose proof follows from considering any strictly proper scoring rule and calculating its expectation under the TOF's probabilities for X.

Lemma 2 *The TOF's posterior probabilities calculated using his/her prior and his/her LR for H_p versus H_d (using x) will match the posterior probabilities calculated by the TOF using his/her prior and the expert's LR, in every instance, if and only if the TOF's LR and the expert's LR agree for every outcome x of X. If not, the TOF's use of the expert's LR is suboptimal from the perspective of the TOF.*

When algorithms output *LR* values and we do not get to see the underlying x values (x could be any vector of features), the only other information we can obtain are *LR* values from the algorithm under conditions with $z=1$ and *LR* values under conditions with $z=0$. In such a situation, the TOF does not have a way of comparing his/her *LR* with the *LR* from the algorithm. However, the TOF can specify g and f discussed above (derived from empirical data, or in any other manner) and hence investigate whether or not Equation (2) holds.

An alternative, frequentist, interpretation of Lemma 1 is as follows. Let us assume that there are two streams producing *LR* values, the stream associated with $z=1$ generates values according to $g(r)$, while the stream associated with $z=0$ generates values according to $f(r)$. The TOF is presented with *LR* values randomly selected from one of the streams with odds θ, that is, a fraction $\theta/(\theta+1)$ of the *LR*s in front of the TOF are from the source $z=1$. If the TOF restricts his/her attention to only *LR*s equal to r, then the posterior odds for *LR*s coming from source $z=1$ becomes $r\theta$ if and only if Equation (2) is satisfied. Again, the TOF can investigate whether or not Equation (2) holds using empirical data.

3.2 | Empirical assessment of *LR* calibration accuracy

Suppose we are interested in the calibration accuracy of a particular *LR* system. Assume that we have likelihood ratios computed using this system, for a collection of samples representative of case work for which we know ground truth. That is, some of the likelihood ratios are a result of a comparison of a crime sample with a reference sample from a source known to be the same source as the one responsible for the crime sample (i.e. H_p true) and other likelihood ratios are from a comparison of a crime sample with a reference sample from a source known to be different from the source responsible for the crime sample (i.e. H_d true). For DNA evidence, a large data set of H_p-true samples, that includes single source samples and mixtures of two to five individuals, is publicly available from the Laboratory for Forensic Technology Development and Integration (LFTDI). It is generally referred to as the PROVEDIt data set (Alfonse et al., 2018). Efforts are being made to create such publicly available data sets in other forensic disciplines as well, see, for example, CSAFE (2017).

Recall that $g(t)$ and $f(t)$ are the densities of likelihood ratios collected under the prosecution ($z = 1$) and defence ($z = 0$) hypotheses respectively. If the offered likelihood ratios are correctly calibrated, Equation (2) should be satisfied.

A commonly used test of the validity of Equation (2) is based on the average value of LR computed under the defence hypothesis ($z = 0$) (Good, 1950; Taylor et al., 2015):

$$\begin{aligned} E_f\left(r(X)\right) &= \int_0^\infty rf(r)\,dr = \int_0^\infty \frac{g(r)}{f(r)} f(r)\,dr = \int_0^\infty g(r) I_{\{f(r)>0\}}\,dr \\ &= 1 - P_g\left(r(X) = \infty\right). \end{aligned} \tag{3}$$

The last equality follows from the fact that if $r < \infty$ is in the support of g it has to be in the support of f, or else (2) would not hold. The advantage of Equation (3) is that Markov's inequality (Ross, 2014) then implies that only a small fraction of LR's arising from H_d-true situations will be large, that is, $P_f(r(X) \geq r) \leq r^{-1}$.

In practice, an average of LR values computed for the test data in the H_d-true cases is used to estimate $E_f(r(X))$, and compared to 1 as a diagnostic for LR calibration (Taylor et al., 2015). The issue with this approach is that the second moment

$$\begin{aligned} E_f\left(r(X)^2\right) &= \int_0^\infty r^2 f(r)\,dr = \int_0^\infty \left(\frac{g(r)}{f(r)}\right)^2 f(r)\,dr \\ &= \int_0^\infty \frac{g(r)}{f(r)} g(r) I_{\{f(r)>0\}}\,dr = E_g\left(r(X) I_{\{r(X)<\infty\}}\right) \end{aligned}$$

is usually very large or even infinite, see Appendix A for an example. This implies that the average of tested LR values can change dramatically from sample to sample, depending on whether some very large LRs are included. This results in extremely high uncertainty in verifying Equation (3). Furthermore, Equation (3) is only a necessary condition for Equation (2) to hold but it is not sufficient. It is therefore a weak diagnostic for checking the calibration of the LR system. For these reasons, a more nuanced approach is needed.

Consider the survival functions $S_g(s) = \int_s^\infty g(t)\,dt$ and $S_f(s) = \int_s^\infty f(t)\,dt$. Integration by parts shows that the above equation is equivalent to

$$S_g(a) - S_g(b) = aS_f(a) - bS_f(b) + \int_a^b S_f(x)\,dx \tag{4}$$

for any a and b such that $0 < a < b < \infty$. Let $L(a, b) = S_g(a) - S_g(b)$, $R(a, b) = aS_f(a) - bS_f(b) + \int_a^b S_f(x)\,dx$, and define the *interval-specific calibration discrepancy* $d_{(a,b)}(S_g, S_f)$, for the interval (a, b) by

$$d_{(a,b)}(S_g, S_f) = \log_{10}(L(a, b)) - \log_{10}(R(a, b)), \tag{5}$$

with the understanding that if both $L(a, b) = R(a, b) = 0$ we set $d_{(a,b)}(S_g, S_f) = 0$. Note that $L(a, b)$ is the probability of observing a value of *LR* in the interval (a, b) in H_p true situations. If the *LR* system is calibrated, this probability should also equal $R(a, b)$. If $L(a, b)$ is smaller than $R(a, b)$ it means that the value of evidence is being overstated since the actual number of *LR*s that fall in the interval (a, b) is smaller than what would be expected if $g(t) = tf(t)$ is true. Likewise, if $L(a, b)$ is larger than $R(a, b)$ the value of evidence is being understated. The value of $d_{(a,b)}(S_g, S_f)$ quantifies the average degree of overstatement or understatement on the interval (a, b) using the logarithmic scale. Negative values of $d_{(a,b)}(S_g, S_f)$ imply overstatement of the value of evidence for *LR* values in the interval (a, b) and positive values understate it, for example, $d_{(10,100)}(S_g, S_f) = -2$ means that, on average, the *LR* values between 10 and 100 have been overstated by 10^2 in favour of the prosecution hypothesis.

Given any sequence $0 < a_1 < \cdots < a_k < \infty$, we construct simultaneous confidence intervals for

$$d(S_g, S_f) = (d_{(a_1,a_2)}(S_g, S_f), \ldots, d_{(a_{k-1},a_k)}(S_g, S_f))^\top \tag{6}$$

using method based on generalized fiducial inference (Cui & Hannig, 2019; Hannig et al., 2016). If the confidence bounds for a particular interval level discrepancy $d_{(a_{j-1},a_j)}$ excludes zero we have evidence that *LR*s in the interval (a_{j-1}, a_j) are not well calibrated, and the bounds of the confidence interval quantify how much the *LR*s in this interval overstate or understate the value of evidence, on average, relative to the empirical data.

Details regarding how these intervals are obtained and their theoretical justification are provided in Appendix C. In particular, we prove Theorem 1 showing that generalized fiducial distribution correctly describes the uncertainty in estimating interval-specific calibration discrepancies. Moreover, the fiducial pointwise and simultaneous confidence intervals for $d(S_g, S_f)$ have asymptotically correct coverage and are therefore theoretically justified for examining calibration discrepancy of LRs.

A graphical summary of our quantitative assessment of calibration discrepancy is provided in the form of *calibration discrepancy plots*, for example, Figure 1. On the horizontal axis are the reported log-likelihood ratio intervals corresponding to the verbal equivalents as suggested in Willis et al. (2015). This default choice of intervals can be changed to accommodate user's needs. On the vertical axis we show the calibration discrepancy $d_{(a,b)}(S_g, S_f)$ explained above. The horizontal red line indicates zero discrepancy, that is, perfect calibration. The blue curve is the point estimate of the calibration discrepancy obtained as the median of the fiducial distribution. The uncertainty in estimating the calibration discrepancy is indicated by the black and cyan lines. The black lines are 95% pointwise confidence intervals, and the cyan lines are 95% simultaneous confidence bounds. The black lines are useful when a calibration of a single reported *LR* is being assessed, while the cyan lines are relevant for overall assessment of the *LR* system.

The intervals in the calibration discrepancy plots where the red line is outside the confidence bounds correspond to reported likelihood values that are ill calibrated. If the red line is above the band, the corresponding reported likelihood ratios are overstated (favouring H_p) by an amount that can be read from the calibration discrepancy plot. Similarly, if the red line is below the bounds, the corresponding reported likelihood ratios are understated (favouring H_d). For instance, Figure 1 shows that the reported likelihood ratios in the range of 10^4 to 10^5 are overstating the evidence by at least a factor of 10^2. Tight confidence bounds that include zero indicate good calibration, for example, the simulation example in Figure 15. A very wide interval that includes zero indicates inability to make definitive judgement about calibration of certain ranges of LRs based on the validation data used, for example, reported LRs in the range of 10^4 to 10^5 in Figure 9. An interesting

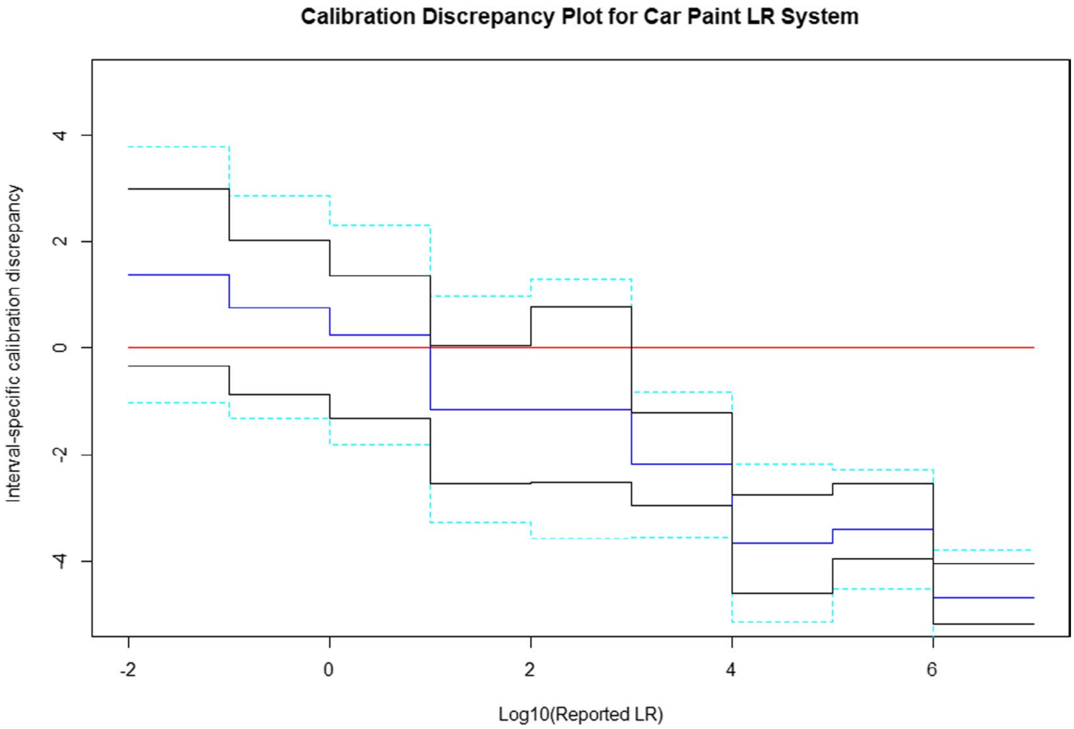

FIGURE 1 Fiducial calibration discrepancy plot for *likelihood ratio* system: Car paint data of Example 1

example is in Figure 16 where a lack of calibration is conclusively established, but the magnitude of the discrepancy is only about $10^{0.5} \doteq 3$ in favour of H_d over the *LR* ranges available, which many users might judge to be of little practical concern.

4 | EXAMPLES

We consider four examples, three of which are discussed in the book by Zadora et al. (2013). The calibration status of the *LR* systems discussed in these examples has been investigated previously and it is well known that some of these systems are poorly calibrated (Zadora et al., 2013). Our use of these examples is to compare and contrast the type of information available from empirical cross-entropy (*ECE*) plots versus information provided by the proposed calibration discrepancy plots.

4.1 | Car paint example

Paint samples were obtained from 36 different cars, see Example 4.4.3.4 in Zadora et al. (2013). Each paint sample was divided into three portions and each portion was analysed and their elemental compositions were determined. Thus there are three replicate measurements for each paint sample of each of eight organic compounds. These compounds are denoted by M2E, MST, TOL, BMA, M2P, MMA, I16 and Styrine. Logarithms to the base 10 of the peak areas from pyrograms of ratios of each of the first seven organic compounds to that of Styrine are used as the response variables (i.e. the procedure has a seven-dimensional response). The reader should refer to Zadora et al. (2013) for details.

Zadora et al. (2013) computed same-source likelihood ratios and different-source *LR*s based on the above data in order to assess false-positive and false-negative rates and also for assessing the quality of the *LR*s. They did this using each of the seven response variables individually as well as combining them to arrive at a single composite *LR* by multiplying together the individual univariate *LR*s. We use this composite *LR* from known same-source comparisons and known different-source comparisons to illustrate our method for checking *LR* calibration discrepancy. Zadora et al. (2013) use the empirical cross-entropy approach for assessing *LR* quality. This is also discussed in their book. The data are available from the website supporting the book using the following link.

https://media.wiley.com/product_ancillary/06/04709721/DOWNLOAD/BSC_files.zip

This *LR* system has good discrimination performance with the area under the ROC curve equal to 0.982 (see Figure B1 in the appendix). Next, we check how well calibrated this *LR* system is by constructing a generalized fiducial confidence band for $d_{(a,b)}(S_g, S_f)$ for different intervals (a, b). The result is shown in Figure 1.

The blue line (step function) represents the pointwise median values of $d_{(a,b)}(S_g, S_f)$ as a function of the reported $\log_{10}(LR)$. The black lines above and below the blue line define the pointwise 95% fiducial confidence intervals for $d_{(a,b)}(S_g, S_f)$ as a function of the reported $\log_{10}(LR)$. The dashed lines in cyan, above and below the blue line, define a 95% simultaneous confidence band for $d_{(a,b)}(S_g, S_f)$. The horizontal red line represents perfect calibration, that is, $d_{(a,b)}(S_g, S_f) = 0$ at all reported $\log_{10}(LR)$ values.

The median line as well as the confidence bands are described by step functions because we chose the following intervals as the intervals of interest.

$$(-2, -1), (-1, 0), (0, 1), (1, 2), (2, 3), (3, 4), (4, 5), (5, 6), (6, 7).$$

These intervals are chosen to reflect ‘order of magnitude bins’ and roughly correspond to the ENFSI verbal equivalent scale (Willis et al., 2015). For instance, if the reported *LR* is 12,000 ($\log_{10}(LR) = 4.079$), then based on the black lines, the discrepancy is roughly between −3 and −5 with 95% confidence. That is, the *LR* is possibly being overstated by a factor whose value is between 10^3 and 10^5, that is, between 1000 and 100,000.

Zadora et al. (2013) (see Chapter 6, section 6.6.2.3) investigate the accuracy of the *LR* system for the car paint example by constructing an ECE plot of the *LR* values from the validation data set (their Figure 6.18). We used the R package `comparison` (authored by Dr. David Lucy, Lancaster University) to recreate that plot which is shown in Figure 2. The figure shows three curves. The red curve is the plot of ECE versus $\log_{10}$(Prior Odds), the blue curve is the plot of ECE versus *LR* values adjusted by the PAV algorithm (where PAV stands for *pooled adjacent violators*) that applies a monotonic transformation to the *LR* values to reduce the calibration error. As pointed out by Zadora et al. (2013), the overall accuracy of the *LR* values is affected by discrimination power (investigated using ROC plots) and calibration error. The lack of agreement between the red curve and the blue curve is an indication of severe calibration loss. The dotted curve represents a noninformative *LR* system which always returns a value of 1.

Like the calibration discrepancy plot, the ECE analysis reveals that the car paint *LR* system is poorly calibrated. However, our approach gives a more easy to understand and informative answer since it is estimating the degree to which the *LR* system is overstating or understanding the strength of evidence over different *LR* ranges.

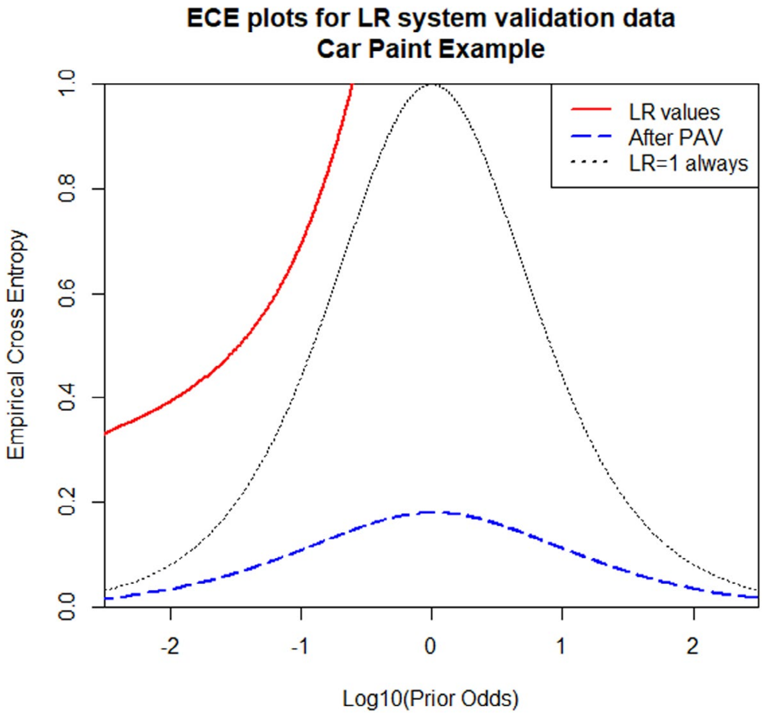

FIGURE 2 Empirical cross-entropy plot for *likelihood ratio*: Car paint data of Example 1. See also figure 6.18 in Zadora et al. (2013)

4.2 | Glass fragments example

This example is discussed in the book by Zadora et al. (2013) (see Chapter 4, section 4.4.6). Twelve fragments of glass were obtained from each of 200 glass objects and each fragment was subjected to an elemental analysis using scanning electron microscopy electron diffraction (SED-EDX). Eight elemental concentrations were measured—Na, Mg, Al, Si, K, Ca, Fe and O. Base 10 logarithms of the ratios of the first seven elemental concentrations to the concentration of O (Oxygen) formed the response variables in the analysis. Zadora et al. (2013) used a graphical model approach and kernel density estimation for computing the same-source *LR*s and different-source *LR*s. The reader should consult their book for further details. The reader is also referred to Chapter 6 of their book for details on using ECE to assess discrimination, calibration, and overall accuracy. Here we focus directly on an assessment of how well calibrated this particular *LR* system is. The discrimination potential for this *LR* system is quite good as described by the ROC plot in Figure B2 in the appendix. The area under the empirical ROC curve is 0.958. The distribution of $\log_{10}(LR)$ values for mated and nonmated cases is shown in the form of violin plots in Figure 3. Although the distributions are well separated for the most part, we can see that there are some very large $\log_{10}(LR)$ values for the nonmated cases.

Next we check how well calibrated this system is by examining the fiducial confidence band for calibration discrepancy using Figure 4. It appears that the system is understating the evidence for $\log_{10}(LR)$ values less than 1 (*LR* values less than 10) whereas it appears to be overstating the evidence for $\log_{10}(LR)$ values between 1 and 2 (*LR* between 10 and 100) and between 2 and 3 (*LR* between 100 and 1000). The confidence band ends after $\log_{10}(LR)$ equal to 3 because we do not have enough data to compute estimates beyond this point without making assumptions about

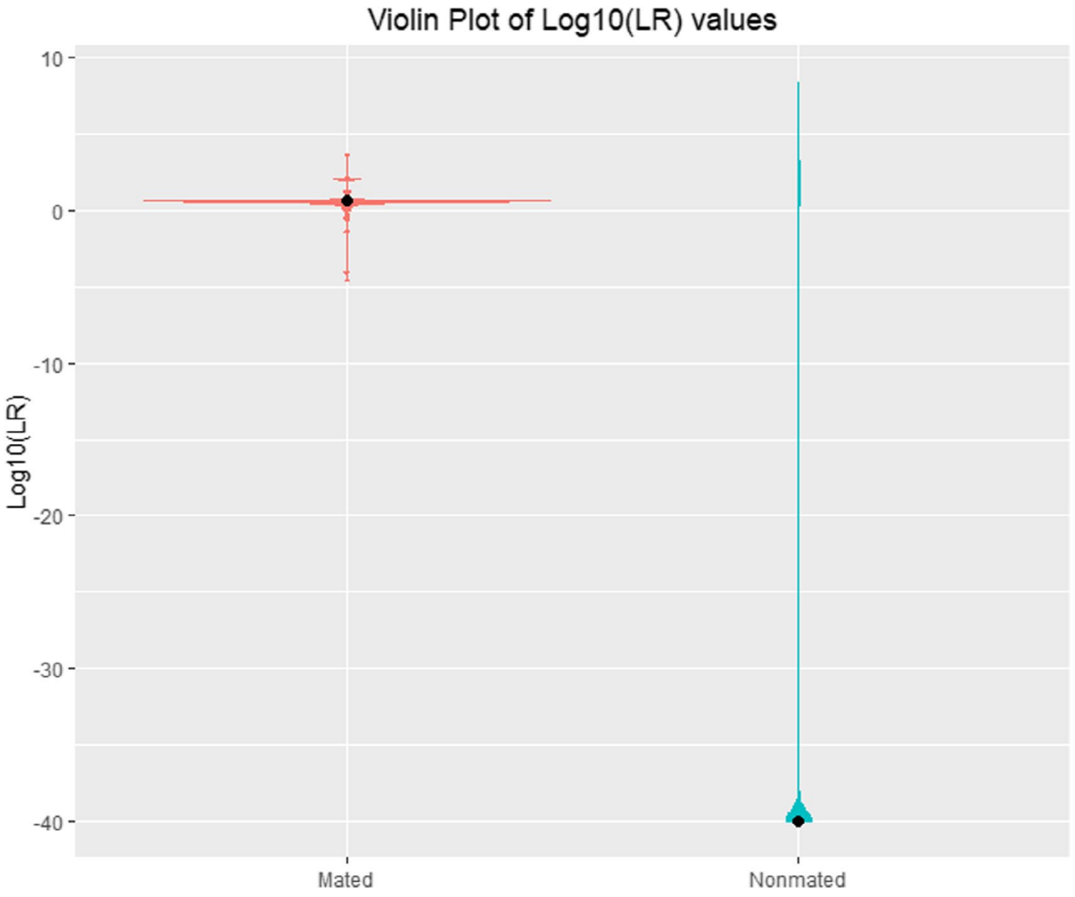

FIGURE 3 Violin plots for mated and nonmated $\log_{10}(LR)$ values for glass data

the tail behaviour of the mated and nonmated $\log_{10}(LR)$ distributions. In particular, there are not enough data to make any useful statements regarding how well calibrated the system is for values of $\log_{10}(LR)$ greater than 3.

An ECE plot for the glass LR system validation set is shown in Figure 5. This plot is constructed using the R package `comparison`. The calibration error in this example is not as pronounced as in the paint example as the red curve and the blue curve are closer to each other here than in the previous example. However, the ECE plot does not provide information regarding regions of LR values where evidence is overstated and regions where evidence is understated. The fiducial calibration discrepancy plot is able to provide such information.

4.3 | Comparison of inks

This example is from Chapter 4 (section 4.4.1) of Zadora et al. (2013) where they report results of experiments conducted to compare 40 different inks using microspectrophotometry with diode detector (MSP-DAD for short). The reader is referred to their book for details. The measured data are three chromaticity coordinates, denoted by x, y, z, that sum to one. Ten measurements were made for each ink. The authors use these data to illustrate the development of LR models for strength of evidence assessments relative to the following competing propositions.

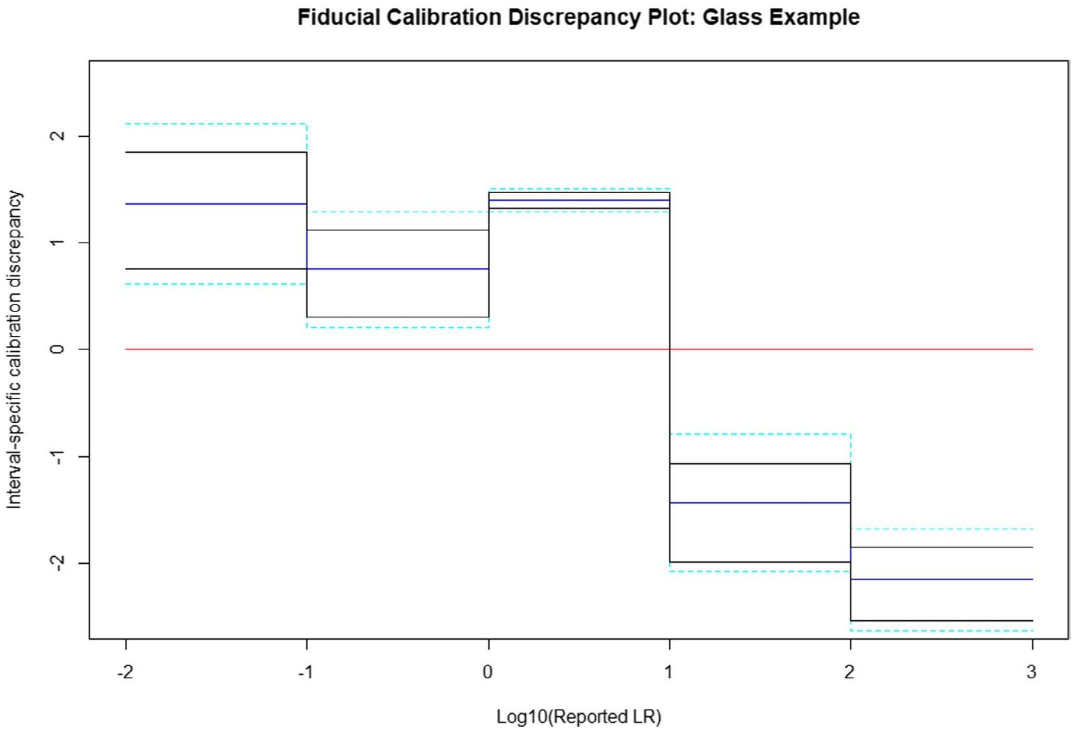

FIGURE 4 Fiducial calibration discrepancy plot for *likelihood ratio* system: Glass data of Example 2

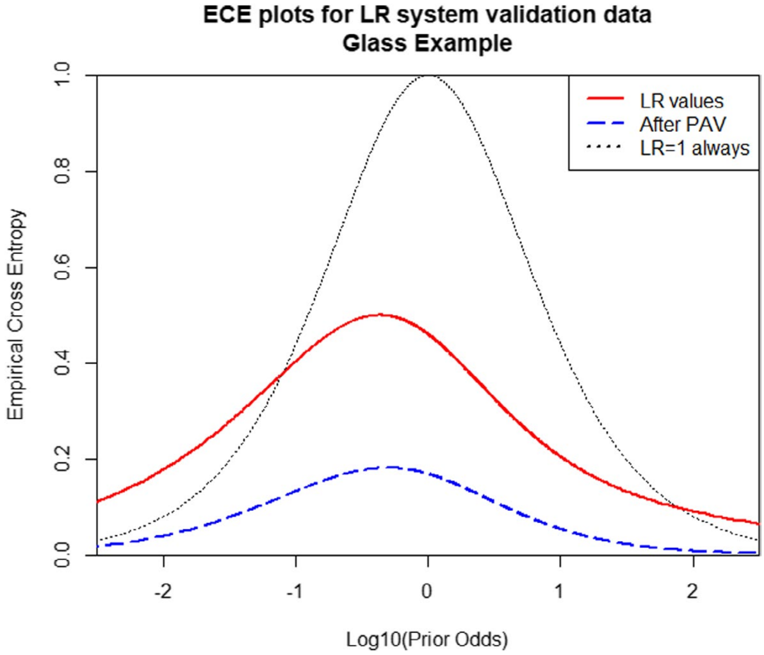

FIGURE 5 Empirical cross-entropy plot for glass data graphical model *likelihood ratio* system. See Zadora et al. (2013) Chapter 4, section 4.4.6 for details

1. H_p: The two written notes being compared used the same ink, versus
2. H_d: The two written notes being compared used different inks

The authors carried out 40 same-source comparisons and 780 different-source comparisons and obtained the corresponding *LR* values. The area under the empirical ROC curve (see Figure B3 in the appendix) for these data is 0.94 which indicates good discrimination power.

Figure 6 shows the fiducial calibration discrepancy plot. The *LR* system for ink comparisons seems to be much better calibrated than the *LR* systems in the previous examples. The zero discrepancy line (red) is either within the confidence bands or just misses the confidence band over the range of $\log_{10}(LR)$ values shown in the plot.

Figure 7 shows the ECE plot for the Ink *LR* system validation set. This plot confirms our findings using the fiducial calibration discrepancy plot. Although the effect of sampling variability is not shown in the ECE plots, the closeness of the red curve to the blue curve suggests that the point estimate of the calibration error is much smaller in this example than in the previous examples.

4.4 | Fingerprint *LR* system

We now consider data from the larger of the two empirical studies reported by Neumann et al. (2012) where the authors investigated the suitability of an *LR* system for fingerprint comparisons using ground truth known scenarios involving different numbers of minutiae available for the

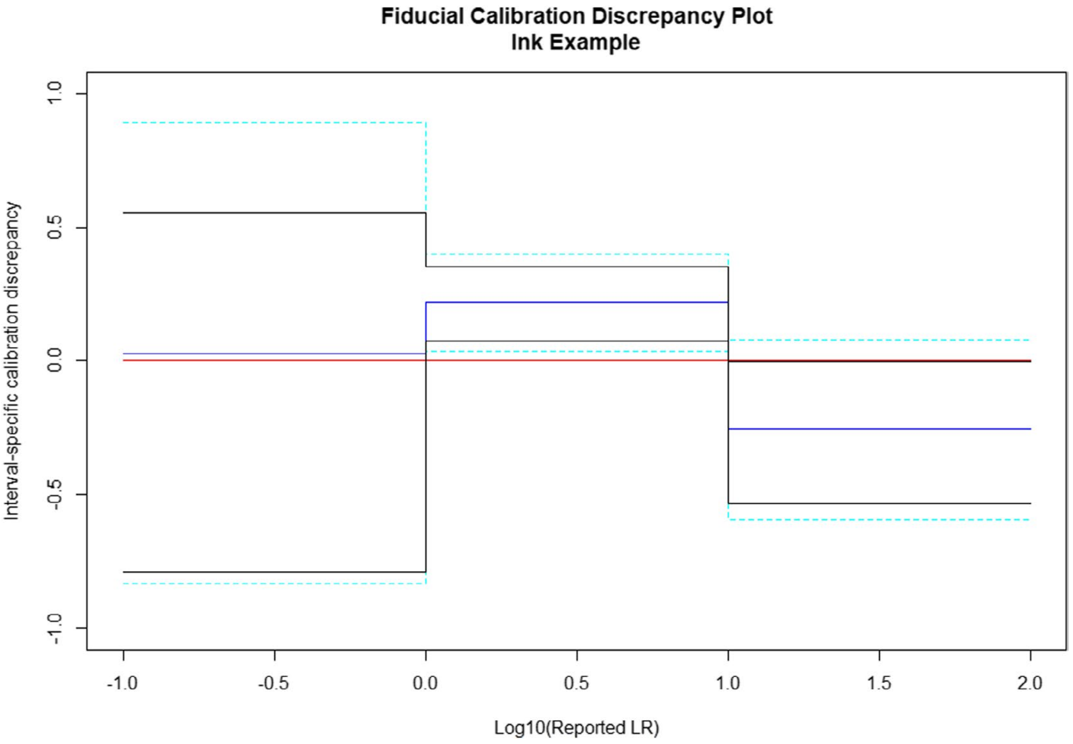

FIGURE 6 Fiducial calibration discrepancy plot for ink data

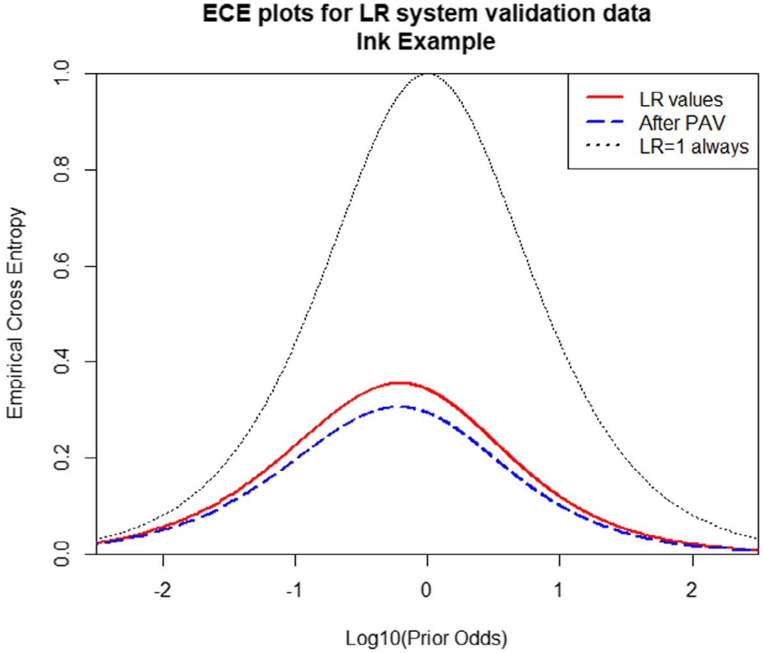

FIGURE 7 Empirical cross-entropy plot for ink data

comparison of a questioned fingermark (Q) to a reference fingerprint (R). The reader is referred to their paper for details.

Figure 8 shows boxplots of $\log_{10}(LR)$ values grouped by ground truth (H_p-true or H_d-true) and by number of minutiae available for comparison. See also figure 5 in Neumann et al. (2012). The H_p-true LR values (red) show an increasing trend as the number of minutiae increase but the H_d-true LR values (green) show no trend. The plot also reveals that the discrimination power of this LR system increases with increasing number of minutiae. See also Table B1 in the appendix.

Next we construct a fiducial calibration discrepancy plot for the LR system using all available H_p-true and H_d-true data, pooled across the entire range of available number of minutiae. This is shown in the Figure 9. This plot suggests that the fingerprint LR system is overstating the strength of evidence over the LR range of 1/100 to 10,000. Outside of this range we do not have enough data to evaluate the calibration discrepancy with any confidence. Thus, although this fingerprint LR system has excellent discrimination power, it would be desirable to reduce its calibration discrepancy. For instance, one might search for monotonic transformations of the LR output from this system that would reduce the calibration discrepancy. As is well known, this process will not affect the discrimination power of the system. One such approach has been discussed by Morrison (2013) in the context of these fingerprint data.

Figure 10 shows an ECE plot for the fingerprint LR system. This plot supports the finding from the fiducial calibration discrepancy plot since there is a noticeable improvement in the ECE after applying a monotonic transformation (PAV algorithm, see Chapter 6 of Zadora et al., 2013). However, as stated earlier, one is unable to deduce the degree of possible understatement or overstatement from the ECE plot. Furthermore, the effect of sampling variability on the ECE plot and related metrics is not available.

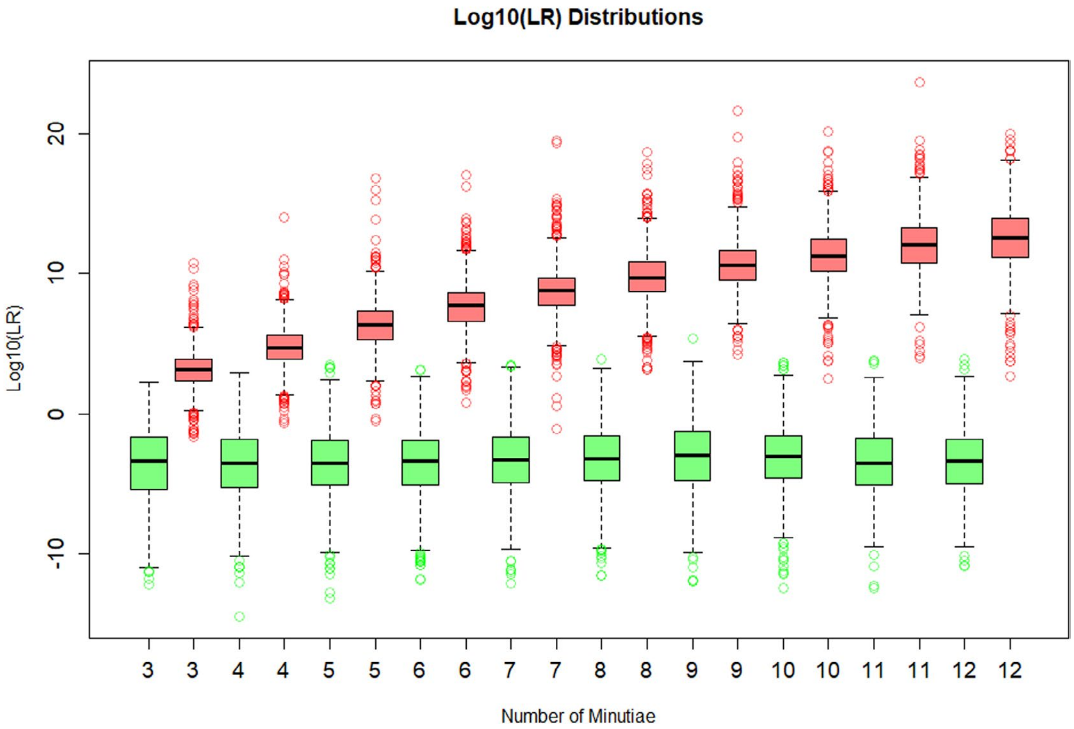

FIGURE 8 Boxplots of $\log_{10}(LR)$ values for H_p-true (Red) and H_d-true (Green) comparisons

5 | A SIMULATION EXPERIMENT

We now illustrate how our proposed procedure performs in a typical modelling situation using a ground-truth known scenario. Development of *LR* systems typically begin with features that are expected to have information for discriminating between H_p and H_d. The joint distribution of the features is modelled using ground-truth known empirical data and one obtains a (fitted) probability density η for H_p-true cases and ψ for H_d-true cases. If x is the observed feature value (may be a vector) then the *LR* is computed by

$$LR = \frac{\eta(x)}{\psi(x)}.$$

To the extent that η and ψ do not correctly describe the feature distributions we can expect a loss in calibration. Loss in discrimination power occurs when the feature vector x does not capture all discriminating information (i.e. it is an insufficient statistic) in the unprocessed raw data.

To illustrate this phenomenon of calibration loss we consider the following simple scenario. Suppose we have a one-dimensional feature-value x that is used to discriminate between H_p and H_d and suppose its true probability functions, under H_p and H_d, respectively, are $\eta(x) = Gamma(x; 10, 2)$ and $\psi(x) = Gamma(x; 1, 2)$ where

$$Gamma(x;a,b) = \frac{1}{\Gamma(a)b^a} x^{a-1} e^{-x/b}, \quad x, a, b > 0.$$

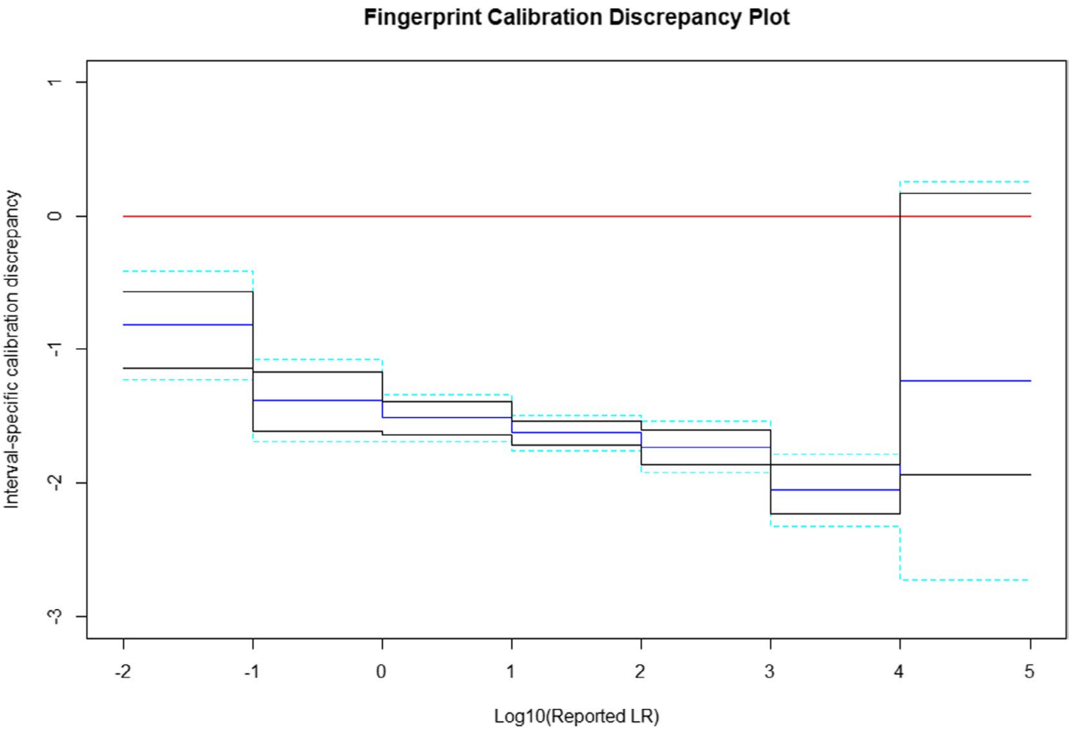

FIGURE 9 Fiducial calibration discrepancy plot for the fingerprint data

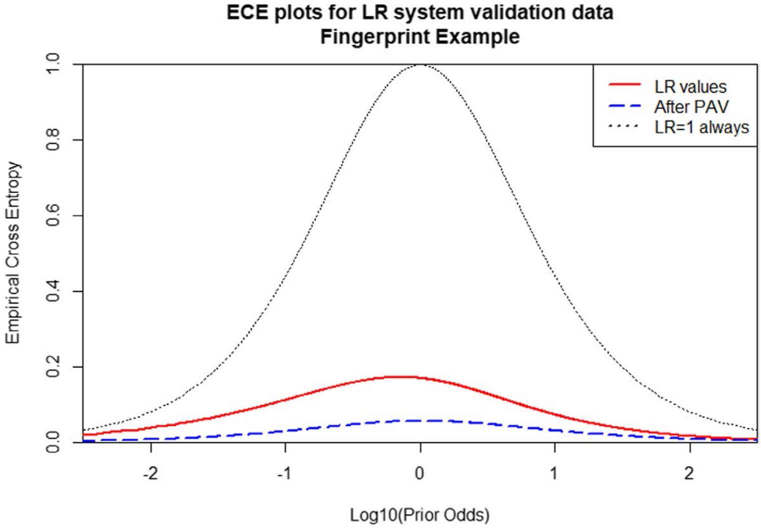

FIGURE 10 Empirical cross-entropy plots for Neumann Fingerprint *likelihood ratio* system

In practice, the data generating mechanisms are unknown and one searches for suitable distributions to describe the observed feature values and uses these fitted distributions to develop likelihood ratios. In most applications it is very rare that large quantities of ground-truth known data are available for both H_p-true and H_d-true cases. Fitting distributions to data using smaller numbers of samples has two consequences: (1) The resulting *LR* system may suffer a noticeable calibration loss, and (2) the fiducial calibration discrepancy check will exhibit substantial uncertainty. We illustrate these points by using 100 feature values known to have come from H_p-true situations and 5000 feature values known to have come from H_d-true situations (it is generally the case that we have access to more H_d-true data than H_p-true data). Next, we fit normal distributions to the logarithms (base 10) of the feature values. The $\log_{10}$(feature-value) histograms and fitted densities are shown in Figure 11.

Figures 12 and 13, respectively, show the resulting fiducial calibration discrepancy plots when likelihood ratios are computed using the true distributions and when using the fitted distributions. When the true distributions are used to compute likelihood ratios, we note that the point estimates of log(discrepancy) (solid blue line) are very close to the red line (perfect calibration) whereas, when using the fitted distributions, the blue line deviates noticeably from the red line, with the confidence band excluding the red line in some intervals, showing calibration loss (there still may be calibration loss where the confidence band includes the red line).

Note that the fiducial calibration discrepancy plot does not provide any information outside a certain range where sample values are available from both H_p-true and H_d-true scenarios. This is because there is no direct information concerning calibration performance outside this range and any statements regarding calibration would be an extrapolation based on assumptions that cannot be empirically verified.

One would expect the calibration loss to decrease as the sample size used in modelling increases. To illustrate this phenomenon we use 100,000 H_p-true feature values and 100,000 H_d-true

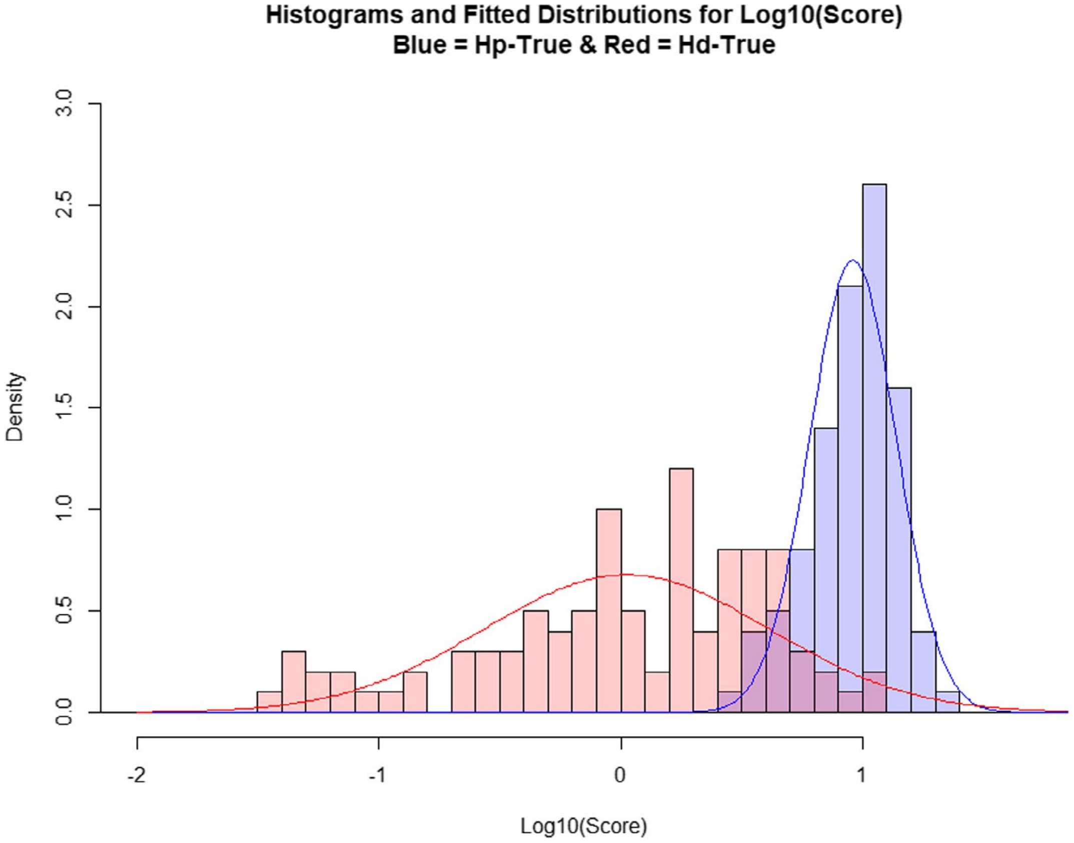

FIGURE 11 Histograms of 100 sample feature values from the H_p-true distribution (blue) and 5000 sample feature values from the H_d-true distribution (red)

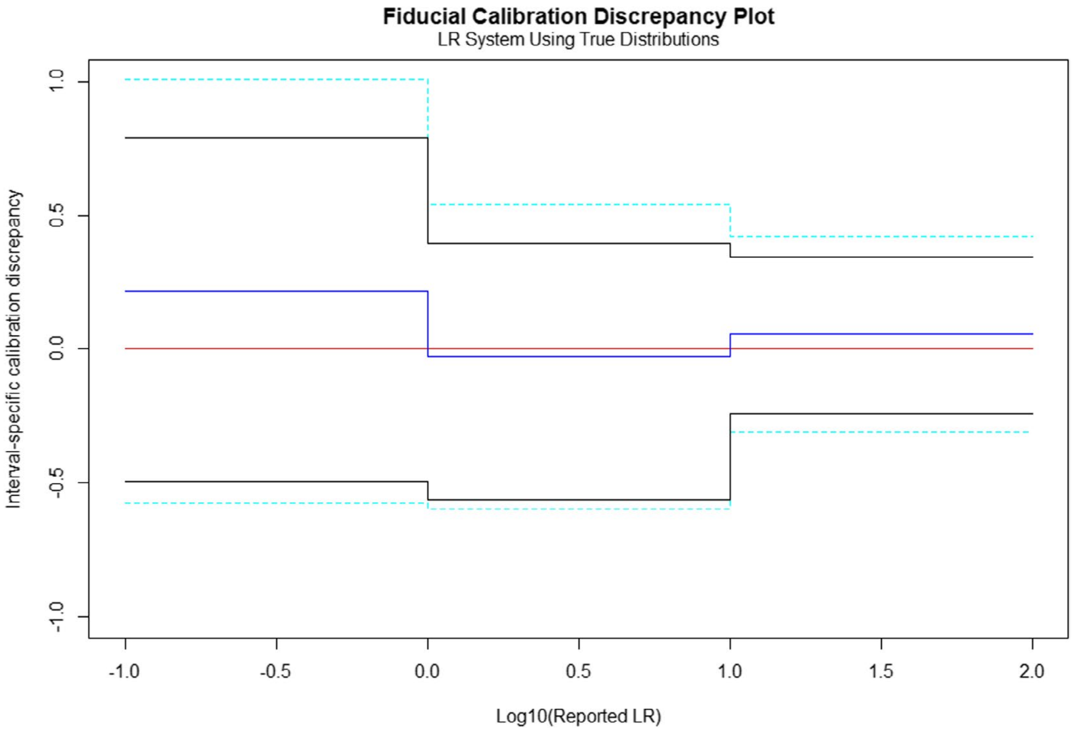

FIGURE 12 Fiducial calibration discrepancy plot for the *likelihood ratio* system using actual feature-value distributions with 100 samples from H_p-true and 5000 samples from H_d-true scenarios

feature values for model fitting. The $\log_{10}$(feature-value) histograms and the fitted densities of their corresponding normal approximation are shown in Figure 14. The fits are far from perfect. However, keep in mind that we are using 100,000 samples from each scenario here whereas, in most applications, models are fit using far fewer empirical data and this magnitude of lack of fit would not be discernible.

Again, we compute likelihood ratios using the true Gamma model first. Figure 15 shows the fiducial calibration discrepancy plot for these *LR*s. We notice that this *LR* system is well calibrated since the confidence interval for log(discrepancy) is tight and includes the zero discrepancy line (in red).

Next, we compute *LR* using the normal approximations and compute the fiducial calibration discrepancies. The results are shown in Figure 16. The plot demonstrates that calibration discrepancy indeed exists although it is not substantial. Reported likelihood ratios appear to understate the strength of evidence by a factor between 2 and 3. The point estimates of log(discrepancy) in this illustration are nearly the same as what we obtained earlier, when using smaller sample sizes, but the confidence bands are tighter as expected.

We also generated ECE plots for this example. Figure 17 shows the ECE plot for the *LR* system which uses the correct feature-value distributions and Figure 18 shows the ECE plot for the system based on fitted distributions. We can see that, in either situation, the discrimination power is very good but the second system, based on fitted distributions, does exhibit a small, but noticeable, calibration discrepancy. This result is consistent with our findings from the fiducial calibration discrepancy plots.

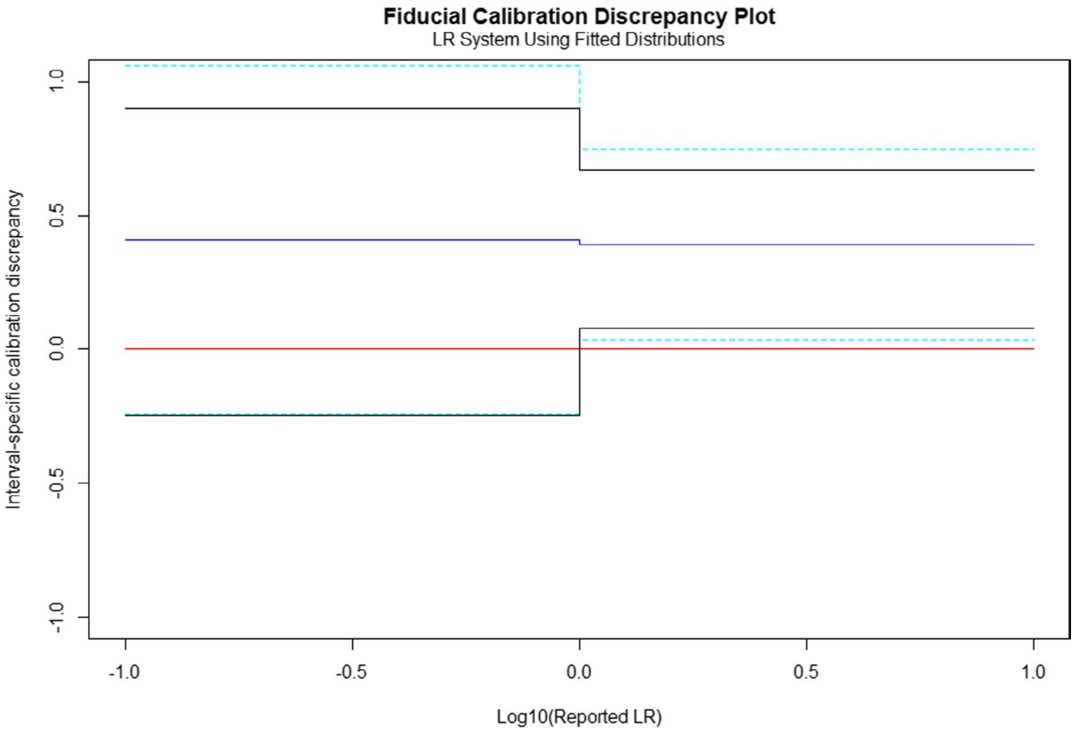

FIGURE 13 Fiducial calibration discrepancy plot for the *likelihood ratio* system using fitted feature-value distributions with 100 samples from H_p-true and 5000 samples from H_d-true scenarios

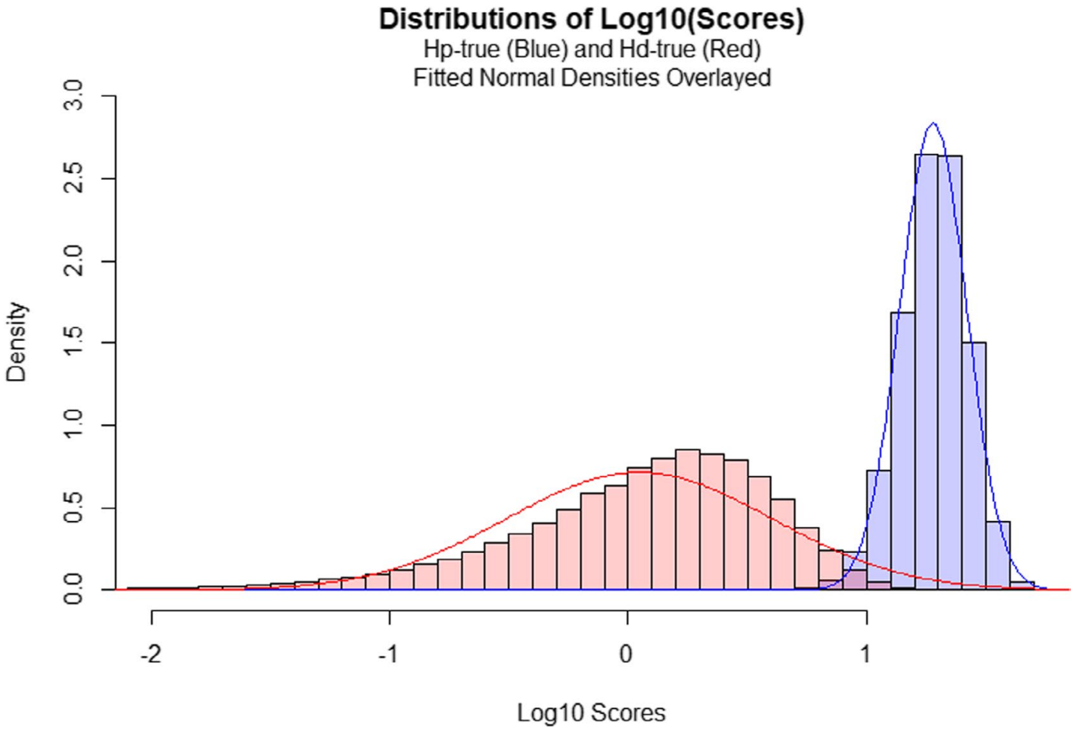

FIGURE 14 Logarithms (base 10) of the feature values and corresponding fitted normal distributions

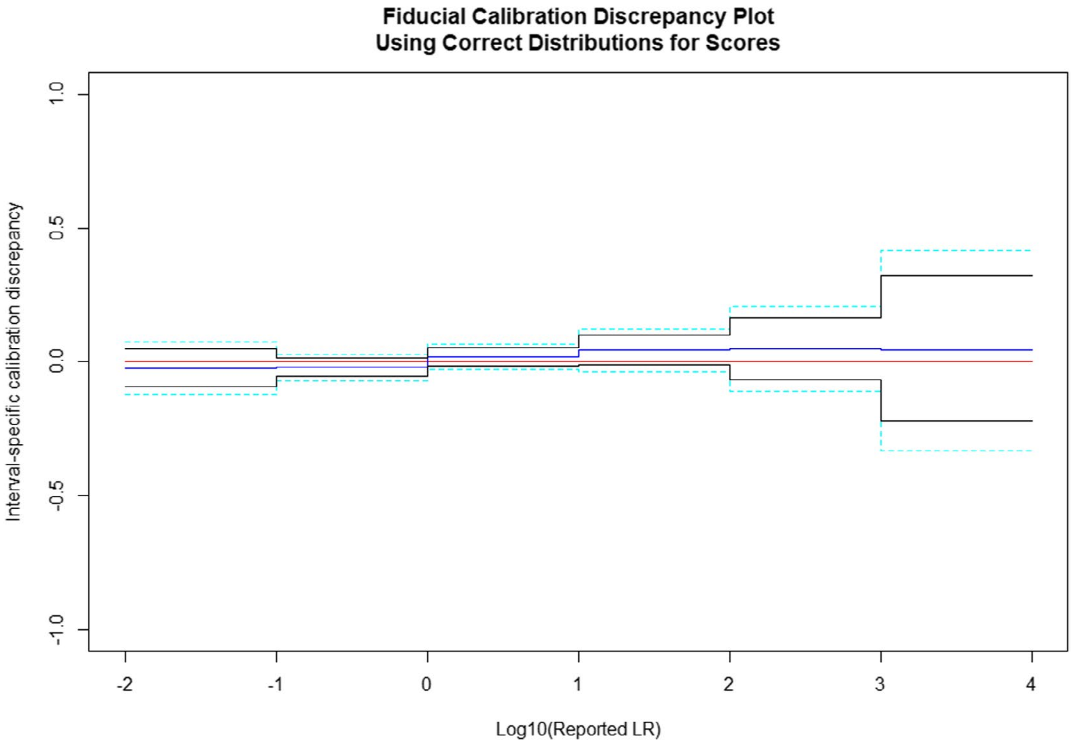

FIGURE 15 Calibration discrepancy plot for the simulated example when *likelihood ratio* values are computed from the true feature-value distributions η and ψ using 100,000 samples from each of η and ψ

5.1 | Repeated sampling verification

In order to examine the trustworthiness of the fiducial confidence intervals for calibration discrepancy, we simulate independent sets of feature values using the H_p and H_d scenarios of the previous section. Again, we consider a small data set containing 100 feature values generated under H_p-true scenarios and 5000 feature values under H_d-true scenarios. We also use a large data set containing 100,000 H_p and H_d-true feature values each. For the small data scenario as well as the larger data scenario we produce a fiducial calibration discrepancy plot and record whether each of the confidence intervals includes zero.

Tables 2 and 3 contain results for *LR*s computed using the true gamma distributions and estimated normal distributions respectively. We report the coverage, the percentage of intervals including zero, and its simulation margin of error for each pointwise (shown in black) and simultaneous (shown in cyan) confidence intervals. Only bins that had observed data during simulation are reported. Because each of the intervals in the fiducial calibration plot is nominally a 95% confidence interval, we expect about 95% of the calibration plots to wholly include zero when we use the true gamma distribution to produce *LR*s (Table 2). When *LR*s are computed using the estimated normal distribution (Table 3), they are technically not well calibrated and the closer the coverage is to zero, the more power we have to detect the miscalibration.

There is a subtle difficulty caused by the fact that the *LR* values covered by the calibration discrepancy plots can differ between data sets. Consequently, the coverage of the smallest and largest $\log_{10}(LR)$ bin is computed based on significantly fewer observations, especially for the small data set. This is not an issue when considering the simultaneous interval. We observe that overall the

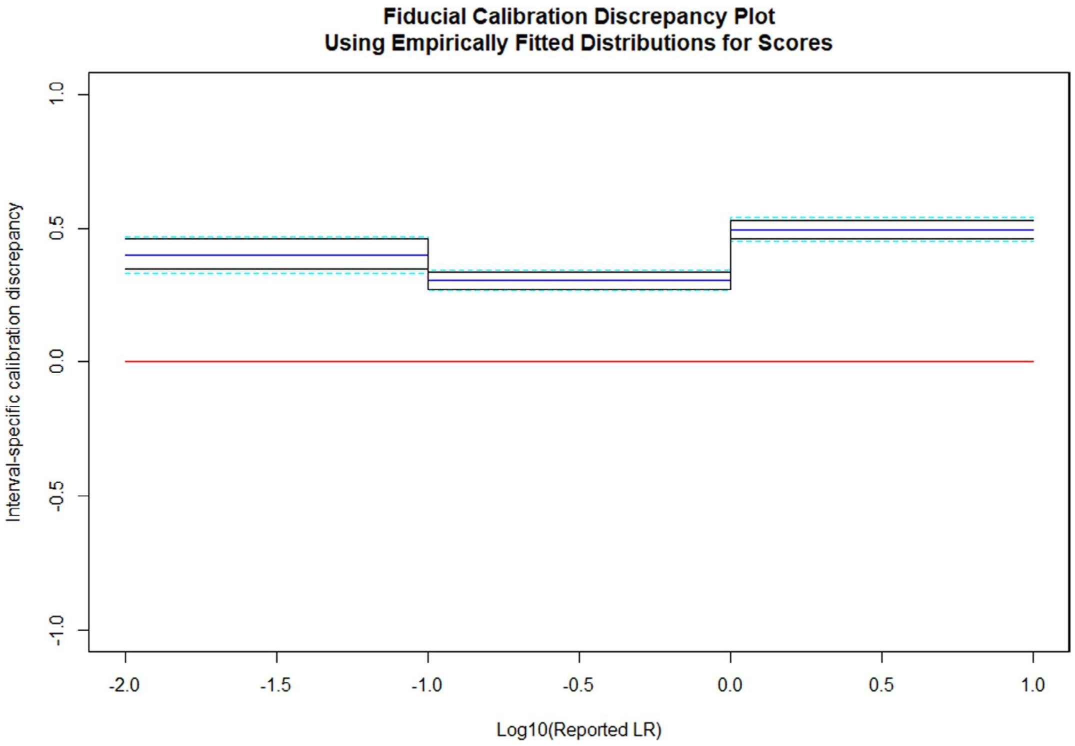

FIGURE 16 Fiducial calibration discrepancy plot for the *likelihood ratio* system using fitted distributions

fiducial calibration plot is itself well calibrated, and we recommend the simultaneous confidence band (shown in cyan) to be used when checking calibration of *LR* systems.

6 | DISCUSSION AND CONCLUSIONS

In this paper we considered the problem of quantifying the strength of evidence provided by a set of observations (x) in favour of one proposition or hypothesis (H_p) relative to its complementary proposition or hypothesis ($H_d = H_p^c$). We noted that there is widespread acceptance for using the likelihood ratio $LR = Pr[x|H_p]/Pr[x|H_d]$ as a measure of the strength of the evidence. In many applications LR is primarily used to make a decision to either believe H_p is true or to believe H_d is true. This is often implemented using a suitable chosen threshold LR_0 such that H_p will be believed to be true if the computed LR value exceeds LR_0. The threshold value is chosen to appropriately balance the trade-offs between falsely believing H_p to be true and falsely believing H_d to be true. This is done by assigning possibly different costs (or penalties) to the false-positive and false-negative errors and minimizing the expected cost, taking into account prior beliefs regarding the probability of H_p and H_d. In traditional statistics these are called classification problems or decision problems. In such applications what matters is the power of the procedure to discriminate between the two scenarios in a manner that leads to minimum expected cost due to misclassification errors.

In forensic science applications, the expert who makes strength of evidence assessments does not make any decisions regarding the truth of H_p or of H_d. Rather, he/she will report the assessed value of LR. The receiver of this information will then process the expert's report or testimony along with other evidence to make decisions regarding the truth of H_p or H_d.

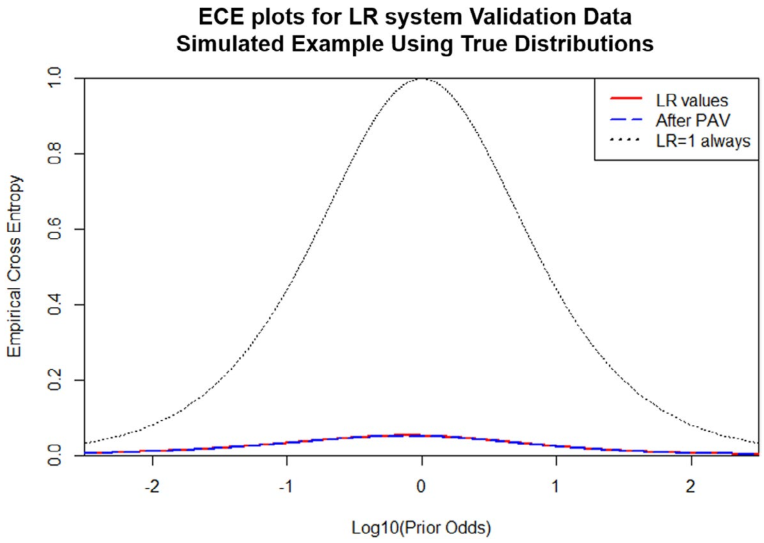

FIGURE 17 Empirical cross-entropy plots for the *likelihood ratio* system using actual feature-value distributions

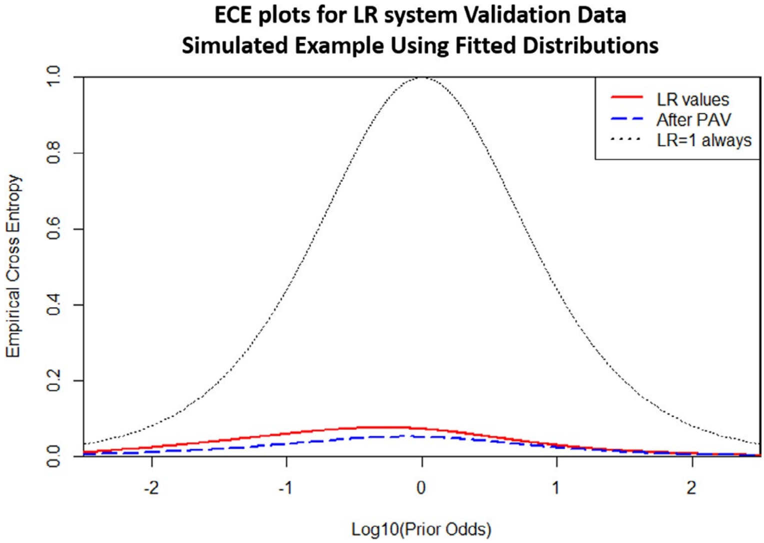

FIGURE 18 Empirical cross-entropy plots for the *likelihood ratio* system using fitted feature-value distributions

TABLE 2 Empirical coverage (in percentage) of the fiducial calibration discrepancy plots for *LR*s computed from the true gamma distributions

$\log_{10}(LR)$	(−2, −1)	(−1, 0)	(0,1)	(1,2)	(2,3)	(3,4)	(4,5)	Simultaneous
Small	88(4)	95(1)	94(1)	94(1)	94(2)	96(4)	50(71)	96(1)
Large	96(1)	96(1)	95(1)	94(1)	95(1)	92(2)	92(11)	95(1)

Notes: The numbers in parentheses are twice the standard errors. Numbers close to 95% show good coverage.

TABLE 3 Empirical coverage (in percentage) of the fiducial calibration discrepancy plots for *LR*s computed using the estimated normal distributions

$\log_{10}(LR)$	(−2, −1)	(−1,0)	(0,1)	Simultaneous
Small	81(3)	91(1)	27(2)	26(3)
Large	0(1)	0(1)	0(1)	0(1)

Notes: The numbers in parentheses are twice the standard errors. Small numbers show good power.

Typically, the value of *LR* will be communicated to the jury or other stakeholders in the following manner:

- The evidence is *LR* times more probable if H_p is true than if H_d is true.

See, for instance, Willis et al. (2015), SWGDAM (2020) and Bright and Coble (2019). Therefore it becomes important that the expert's process for assessment of *LR* be empirically checked to determine whether this process is reliable. The problem we addressed in this paper is this:

- How can we empirically check how reliable are such statements from experts, or assessments by algorithms used by experts? In particular, when an *LR* system produces an *LR* value equal to *x*, to what extent might this value be an overstatement or an understatement of the value of evidence in light of the available empirical validation data?

To answer this question we developed a measure of discrepancy between reported *LR*s and the empirically supported values for *LR*s and, using the theory of generalized fiducial inference, provided an approach for obtaining confidence bands for the calibration discrepancy. We also demonstrated how to visually display this information in the form of a fiducial calibration discrepancy plot. We provided several examples, most of them from the open literature, to illustrate the use of our proposed method. Other authors (e.g. Zadora et al., 2013) have discussed assessment of *LR* performance using the notion of ECE with particular attention to discrimination and calibration as two separate components contributing to overall performance. We compared our approach with this alternative approach of examining ECE plots and noted that, while the ECE approach can alert us to lack of calibration when it occurs, it does not directly tell us the extent to which evidence is being understated or overstated when the *LR* system is not well calibrated. Our approach does this. Moreover, our approach also takes into account sampling variability whereas the previously published methods do not. We therefore believe that fiducial calibration discrepancies and corresponding plots provide valuable additional tools for assessing the calibration status of *LR* systems.

Although we discussed empirical checking of the calibration performance of *LR* models in the context of *source identification* it is important to keep in mind that, in actual casework, the question of source identification will make sense only after its relevance is demonstrated. This is so because there can be many innocent reasons for the presence of evidential material at the crime scene originating from the person of interest. For instance, DNA from a person may be found at a crime scene but that does not mean that the person of interest was present at the crime scene when the crime event occurred. The person may have visited the crime scene either before or after the crime event. In such circumstances, relaying to the TOFs a likelihood ratio addressing the source of the evidence, in the absence of other supporting evidence that make the source question relevant, could be viewed as potentially having a prejudicial effect on the triers of fact. Although we did not address such issues in this paper, these and other considerations are of great practical importance to the proper functioning of the criminal justice system since the life and liberty of individuals are at stake.

ACKNOWLEDGEMENTS

The authors are thankful to Dr. Cedric Neumann for providing us with the fingerprint data. We also thank the anonymous referees for their valuable suggestions. Special thanks to one of the referees who raised many important issues and also provided numerous constructive suggestions. We believe these have resulted in substantial improvements to the paper.

DISCLAIMER

The views expressed in this paper are those of the authors and do not reflect the official position or policies of the National Institute of Standards and Technology.

ORCID

Jan Hannig http://orcid.org/0000-0002-4164-0173

REFERENCES

Aitken, C., Roberts, P. & Jackson, G. (2010) *Fundamentals of probability and statistical evidence in criminal proceedings: Guidance for judges, lawyers, forensic scientists and expert witnesses*. Royal Statistical Society. Available at: https://www.maths.ed.ac.uk/~cgga/Guide-1-WEB.pdf [Accessed 4th September 2021].

Alfonse, L.E., Garrett, A.D., Lun, D.S., Duffy, K.R. & Grgicak, C.M. (2018) A large-scale dataset of single and mixed-source short tandem repeat profiles to inform human identification strategies: Provedit. *Forensic Science International: Genetics*, 32, 62–70.

Biedermann, A. (2013) Your uncertainty, your probability, your decision. *Frontiers in Genetics*, 4, 148. https://doi.org/10.3389/fgene.2013.00148

Bolck, A., Ni, H. & Lopatka, M. (2015) Evaluating score-and feature-based likelihood ratio models for multivariate continuous data: Applied to forensic MDMA comparison. *Law, Probability and Risk*, 14, 243–266.

Bozza, S., Taroni, F., Marquis, R. & Schmittbuhl, M. (2008) Probabilistic evaluation of handwriting evidence: Likelihood ratio for authorship. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57, 329–341.

Bright, J.-A. & Coble, M. (2019) *Forensic DNA profiling: A practical guide to assigning likelihood ratios*. Boca Raton, FL: CRC Press.

Brümmer, N. & Du Preez, J. (2006) Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20, 230–275.

Buckleton, J.S., Bright, J.-A. & Taylor, D. (Eds.) (2021) *Forensic DNA evidence interpretation*, 2nd edition, Boca Raton, FL: CRC Press.

Bunch, S. & Wevers, G. (2013) Application of likelihood ratios for firearm and toolmark analysis. *Science & Justice*, 53, 223–229.

Butler, J.M. (2014) *Advanced topics in forensic DNA typing: Interpretation*. New York, NY: Academic Press.

Causin, V., Schiavone, S., Marigo, A. & Carresi, P. (2004) Bayesian framework for the evaluation of fiber evidence in a double murder—a case report. *Forensic Science International*, 141, 159–170.

Chen, X.-H., Champod, C., Yang, X., Shi, S.-P., Luo, Y.-W., Wang, N. et al. (2018) Assessment of signature handwriting evidence via score-based likelihood ratio based on comparative measurement of relevant dynamic features. *Forensic Science International*, 282, 101–110.

CSAFE (2017) Forensic science data portal. Available at: https://forensicstats.org/data/ [Accessed 4th September 2021].

Cui, Y. & Hannig, J. (2019) Nonparametric generalized fiducial inference for survival functions under censoring. *Biometrika*, 106, 501–518.

Curran, J., Triggs, C., Almirall, J., Buckleton, J. & Walsh, K. (1997a) The interpretation of elemental composition measurements from forensic glass evidence: I. *Science & Justice*, 37, 241–244.

Curran, J., Triggs, C., Almirall, J., Buckleton, J. & Walsh, K. (1997b) The interpretation of elemental composition measurements from forensic glass evidence: II. *Science & Justice*, 37, 245–249.

Curran, J., Buckleton, J. & Triggs, C. (1999) Commentary on Koons, RD, Buscaglia J. The forensic significance of glass composition and refractive index measurements. J Forensic Sci 1999; 44(3): 496–503. *Journal of Forensic Science*, 44(6), 1324–1325.

DeGroot, M.H. & Fienberg, S.E. (1983) The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32, 12–22.

Dong, F., Zhao, Y., Luo, Y., Zhang, W. & Li, Y. (2019) Objective evaluation of similarity scores derived by Evofinder® system for marks on bullets fired from Chinese Norinco QSZ-92 pistols. *Forensic Sciences Research*, 1–7.

Enzinger, E. (2016) Implementation of forensic voice comparison within the new paradigm for the evaluation of forensic evidence. PhD dissertation, University of New South Wales, Sydney, New South Wales.

Evett, I.W. & Weir, B.S. (1998) *Interpreting DNA evidence: Statistical genetics for forensic scientists*. Sunderland, MA: Sinauer Associates.

Evett, I., Cage, P. & Aitken, C. (1987) Evaluation of the likelihood ratio for fibre transfer evidence in criminal cases. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 36(2), 174–180.

Franck, C.T. & Gramacy, R.B. (2020) Assessing Bayes factor surfaces using interactive visualization and computer surrogate modeling. *The American Statistician*, 74(4), 359–369.

Gelman, A. & Hennig, C. (2017) Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180, 967–1033.

Good, I.J. (1950) *Probability and the weighing of evidence*. London: Charles Griffin.

Hannig, J., Iyer, H., Lai, R.C. & Lee, T.C. (2016) Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association*, 111, 1346–1361.

Kadane, J.B. (2020) *Principles of uncertainty*, 2nd edition, Boca Raton, FL: CRC Press.

Kass, R.E. & Raftery, A.E. (1995) Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795.

Kerkhoff, W., Stoel, R., Mattijssen, E. & Hermsen, R. (2013) The likelihood ratio approach in cartridge case and bullet comparison. *AFTE J*, 45(3), 284–289.

Leegwater, A.J., Meuwly, D., Sjerps, M., Vergeer, P. & Alberink, I. (2017) Performance study of a score-based likelihood ratio system for forensic fingermark comparison. *Journal of Forensic Sciences*, 62, 626–640.

Lindley, D.V. (1977) A problem in forensic science. *Biometrika*, 64(2), 207–213.

Lindley, D.V. (2013) *Understanding uncertainty*. Chichester: John Wiley & Sons.

Lund, S.P. & Iyer, H.K. (2017) Likelihood ratio as weight of forensic evidence: A closer look. *Journal of Research of National Institute of Standards and Technology*, 122, 1–32. https://doi.org/10.6028/jres.122.027

Martire, K.A., Growns, B. & Navarro, D.J. (2018) What do the experts know? Calibration, precision, and the wisdom of crowds among forensic handwriting experts. *Psychonomic Bulletin & Review*, 25, 2346–2355.

Meuwly, D. (2001) Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique. Phd dissertation, University of Lausanne, Lausanne, Switzerland.

Morgan, S.L. (2014) Evaluation of statistical measures for fiber comparisons: Interlaboratory studies and forensic databases. National Institute of Justice Washington, DC.

Morrison, G.S. (2013) Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, 45, 173–197.

Morrison, G.S. & Poh, N. (2018) Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios/Bayes factors. *Science & Justice*, 58, 200–218.

Morrison, G.S. & Stoel, R.D. (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.

Morrison, G., Enzinger, E., Ramos, D., González-Rodríguez, J. & Lozano-Díez, A. (2020) Statistical models in forensic voice comparison. In: Banks, D.L., Kafadar, K., Kaye, D.H. & Tackett, M. (Eds.) *Handbook of forensic statistics*. Boca Raton, FL: Chapman and Hall/CRC, pp. 451–497.

Morrison, G.S., Enzinger, E., Hughes, V., Jessen, M., Meuwly, D., Neumann, C. et al. (2021) Consensus on validation of forensic voice comparison. *Science & Justice*, 61(3), 299–309.

Neumann, C. (2020) Defence against the modern arts: The curse of statistics: Part I—FRStat. *Law, Probability and Risk*, 19, 1–20.

Neumann, C. & Ausdemore, M. (2020) Defence against the modern arts: The curse of statistics—part II: 'score-based likelihood ratios'. *Law, Probability and Risk*, 19, 21–42.

Neumann, C., Evett, I. & Skerrett, J. (2012) Quantifying the weight of evidence from a forensic fingerprint comparison: A new paradigm. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175, 371–415.

Neumann, C., Hendricks, J. & Ausdemore, M. (2020) Statistical support for conclusions in fingerprint examinations. In: Banks, D.L., Kafadar, K., Kaye, D.H. & Tackett, M. (Eds.) *Handbook of forensic statistics*. Boca Raton, FL: Chapman and Hall/CRC, pp. 277–324.

Nic Daéid, N., Rafferty, A., Butler, J., Chalmers, J., McVean, G. & Tully, G. (2017) *Forensic DNA analysis: A primer for courts*. The Royal Society.

Park, S. (2018) Learning algorithms for forensic science applications. Ph.D. thesis, Iowa State University.

Park, S. & Carriquiry, A. (2020) An algorithm to compare two-dimensional footwear outsole images using maximum cliques and speeded-up robust feature. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 13(2), 188–199.

Park, S. & Tyner, S. (2019) Evaluation and comparison of methods for forensic glass source conclusions. *Forensic Science International*, 305, 110003.

Ramos, D. (2007) Forensic evaluation of the evidence using automatic speaker recognition systems. Ph.D. thesis, Universidad autónoma de Madrid.

Ramos, D. & Gonzalez-Rodriguez, J. (2013) Reliable support: Measuring calibration of likelihood ratios. *Forensic Science International*, 230, 156–169.

Ramos, D., Franco-Pedroso, J., Lozano-Diez, A. & Gonzalez-Rodriguez, J. (2018) Deconstructing cross-entropy for probabilistic binary classifiers. *Entropy*, 20(3), 208. https://doi.org/10.3390/e20030208

Ross, S.M. (2014) *Introduction to probability models*. New York, NY: Academic Press.

Saunders, C., Hepler, A., Davis, L. & Buscaglia, J. (2010) Estimation of likelihood ratios for forensic handwriting analysis. *Science & Justice*, 1, 32.

Song, J., Vorburger, T.V., Chu, W., Yen, J., Soons, J.A., Ott, D.B. et al. (2018) Estimating error rates for firearm evidence identifications in forensic science. *Forensic Science International*, 284, 15–32.

SWGDAM (2020) Scientific working group on DNA analysis methods: Recommendations of the SWGDAM ad hoc working group on genotyping results reported as likelihood ratios. Available at: https://1ecb9588-ea6f-4feb-971a-73265dbf079c.filesusr.com/ugd/4344b0_dd5221694d1448588dcd0937738c9e46.pdf [Accessed 4th September 2021].

Swofford, H., Koertner, A., Zemp, F., Ausdemore, M., Liu, A. & Salyards, M. (2018) A method for the statistical interpretation of friction ridge skin impression evidence: Method development and validation. *Forensic Science International*, 287, 113–126.

Taylor, D., Buckleton, J. & Evett, I. (2015) Testing likelihood ratios produced from complex DNA profiles. *Forensic Science International: Genetics*, 16, 165–171.

van der Vaart, A.W. (1998) *Asymptotic statistics*, vol. 3 of *Cambridge series in statistical and probabilistic mathematics*. Cambridge: Cambridge University Press.

Venkatasubramanian, G., Hegde, V., Lund, S.P., Iyer, H. & Herman, M. (2021a) Quantitative evaluation of footwear evidence: Initial workflow for an end-to-end system. *Journal of Forensic Sciences*. https://doi.org/10.1111/1556-4029.14802

Venkatasubramanian, G., Hegde, V., Padi, S., Iyer, H. & Herman, M. (2021b) Comparing footwear impressions that are close non-matches using correlation-based approaches. *Journal of Forensic Sciences*, 66, 890–909.

Vergeer, P., van Es, A., de Jongh, A., Alberink, I. & Stoel, R. (2016) Numerical likelihood ratios outputted by LR systems are often based on extrapolation: When to stop extrapolating? *Science & Justice*, 56, 482–491.

Vergeer, P., van Schaik, Y. & Sjerps, M. (2021) Measuring calibration of likelihood-ratio systems: a comparison of four metrics, including a new metric devPAV. *Forensic Science International*, 321, 110722.

Willis, S., McKenna, L., McDermott, S., O'Donell, G., Barrett, A., Rasmusson, B. et al. (2015) *ENFSI guideline for evaluative reporting in forensic science*. European Network of Forensic Science Institutes. Available at: https://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf [Accessed 4th September 2021].

Young, C. (2018) Model uncertainty and the crisis in science. *Socius*, 4, 2378023117737206.

Zadora, G., Martyna, A., Ramos, D. & Aitken, C. (2013) *Statistical analysis in forensic science: Evidential value of multivariate physicochemical data*. Chichester: John Wiley & Sons.

How to cite this article: Hannig J, Iyer H. Testing for calibration discrepancy of reported likelihood ratios in forensic science. *J R Stat Soc Series A*. 2021;00:1–35. https://doi.org/10.1111/rssa.12747

APPENDIX A

EXAMPLE

We present an example of an *LR* system that produces likelihood ratios whose distribution has infinite variance when H_d is true. Suppose X is a random variable with density

$$\eta(x) = \frac{1}{4x^{3/4}} \quad 0 < x \leq 1$$

when H_p is true, and density

$$\psi(x) = \frac{3}{4x^{1/4}} \quad 0 < x \leq 1$$

when H_d is true. The likelihood ratio in favour of H_p when x is observed is

$$r = r(x) = \frac{1}{3x^{1/2}} \quad 1/3 \leq r < \infty.$$

The density of $R = R(X)$ under H_p is

$$g(r) = \frac{1}{2\sqrt{3}r^{3/2}} \quad 1/3 \leq r < \infty$$

and under H_d is

$$f(r) = \frac{1}{2\sqrt{3}r^{5/2}} \quad 1/3 \leq r < \infty.$$

Note that $\frac{g(r)}{f(r)} = r$ and $E_f[R] = \int_{1/3}^{\infty} \frac{1}{2\sqrt{3}r^{3/2}} = 1$. Furthermore, it is easily verified that

$$E_f[R^2] = \int_{1/3}^{\infty} r^2 f(r)\, dr = \infty.$$

Finally we point out that the survival functions corresponding to the densities f and g, respectively, are given by

$$S_f(r) = \frac{1}{3\sqrt{3}r^{3/2}} \quad \frac{1}{3} \leq r < \infty$$

and

$$S_g(r) = \frac{1}{\sqrt{3}r^{1/2}} \quad \frac{1}{3} \leq r < \infty.$$

APPENDIX B

DISCRIMINATION PERFORMANCE OF THE *LR* SYSTEM

B.1 | Car paint data

Figure B1 gives the receiver operating characteristic (ROC) plot of the reported *LR* values for the car paint data. The area under the ROC plot (AUC) is 0.982 indicating strong discriminating power for the *LR* system. If an *LR* value is selected randomly from the mated distribution and another *LR* value is selected randomly from the nonmated distribution then the (estimated) probability is 0.982 that the mated *LR* will be larger than the nonmated *LR*.

B.1 | Glass data

The discrimination potential for this *LR* system is quite good as described by the ROC plot in Figure B2. The area under the empirical ROC curve is 0.958.

The unusual shape of the ROC plot here is due to a few nonmated cases that yielded very high *LR* values. The overall discrimination performance would normally be judged to be quite good as the AUC value is 0.958. However, this example illustrates why AUC value alone is not sufficient to judge performance. It is undesirable to have many nonmated *LR* values taking on very high values. Of course, this will be detected by examining calibration accuracy.

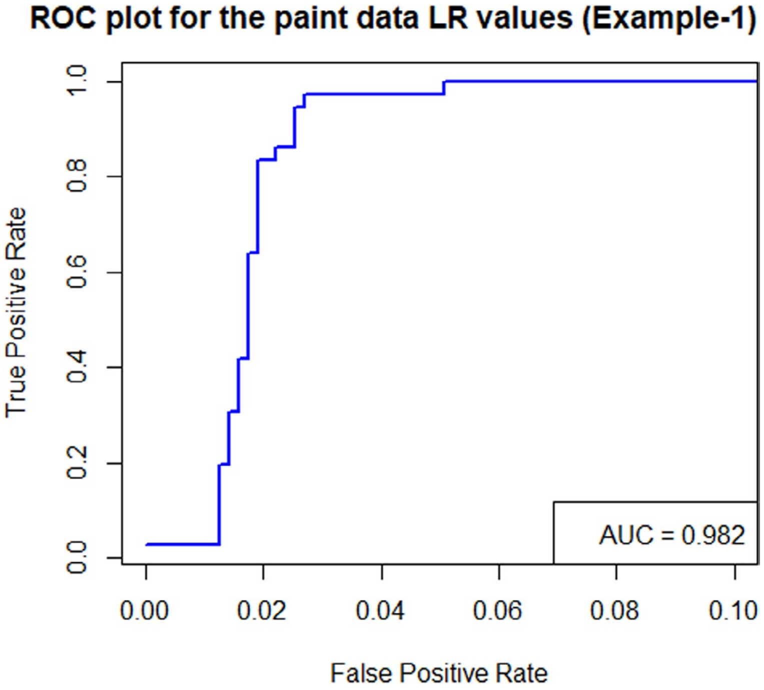

FIGURE B1 Receiver operating characteristic plot for *likelihood ratio*: Car paint data of Example 1

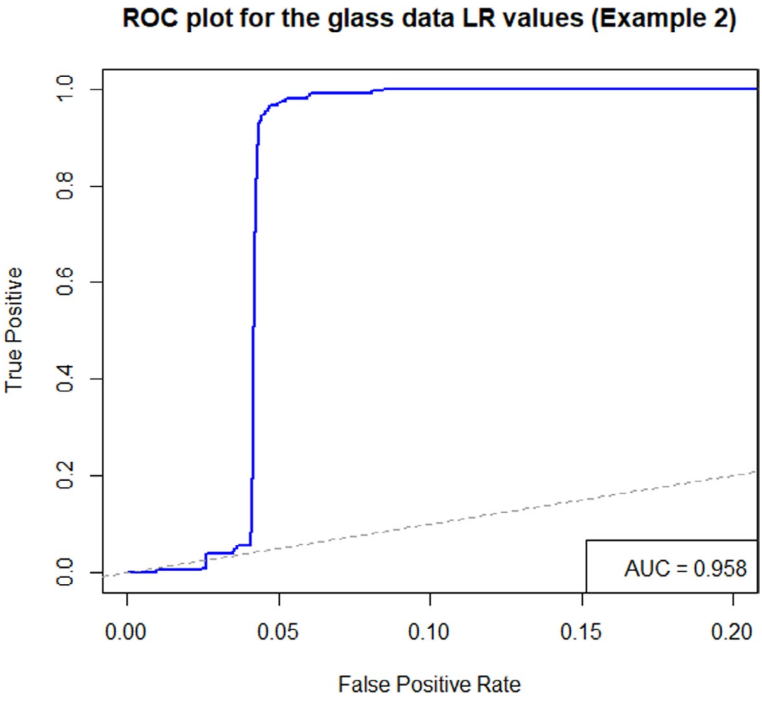

FIGURE B2 Receiver operating characteristic plot for *likelihood ratio*: Glass data of Example 2. See also Figure 6.18 in Zadora et al. (2013)

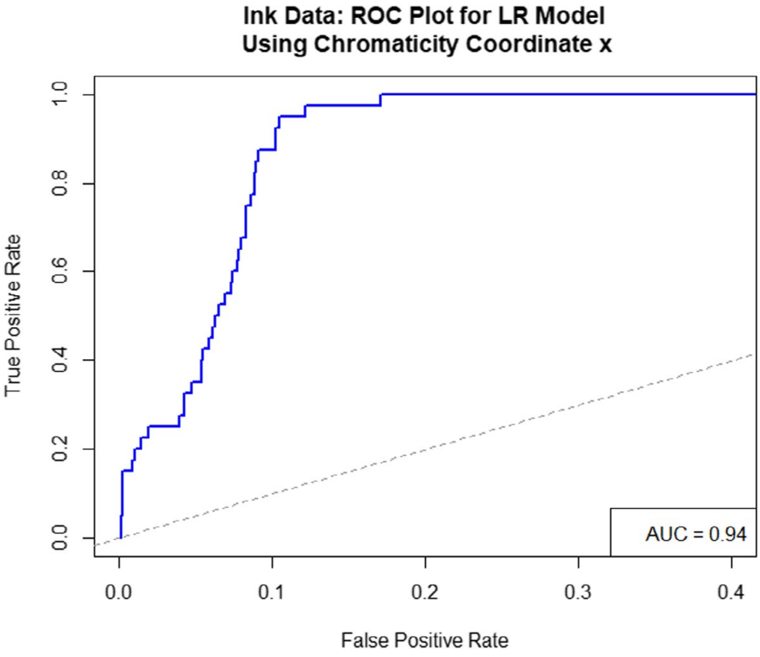

FIGURE B3 Receiver operating characteristic plot for glass data graphical model *likelihood ratio* system. See Zadora et al. (2013) Chapter 4, section 4.4.6 for details

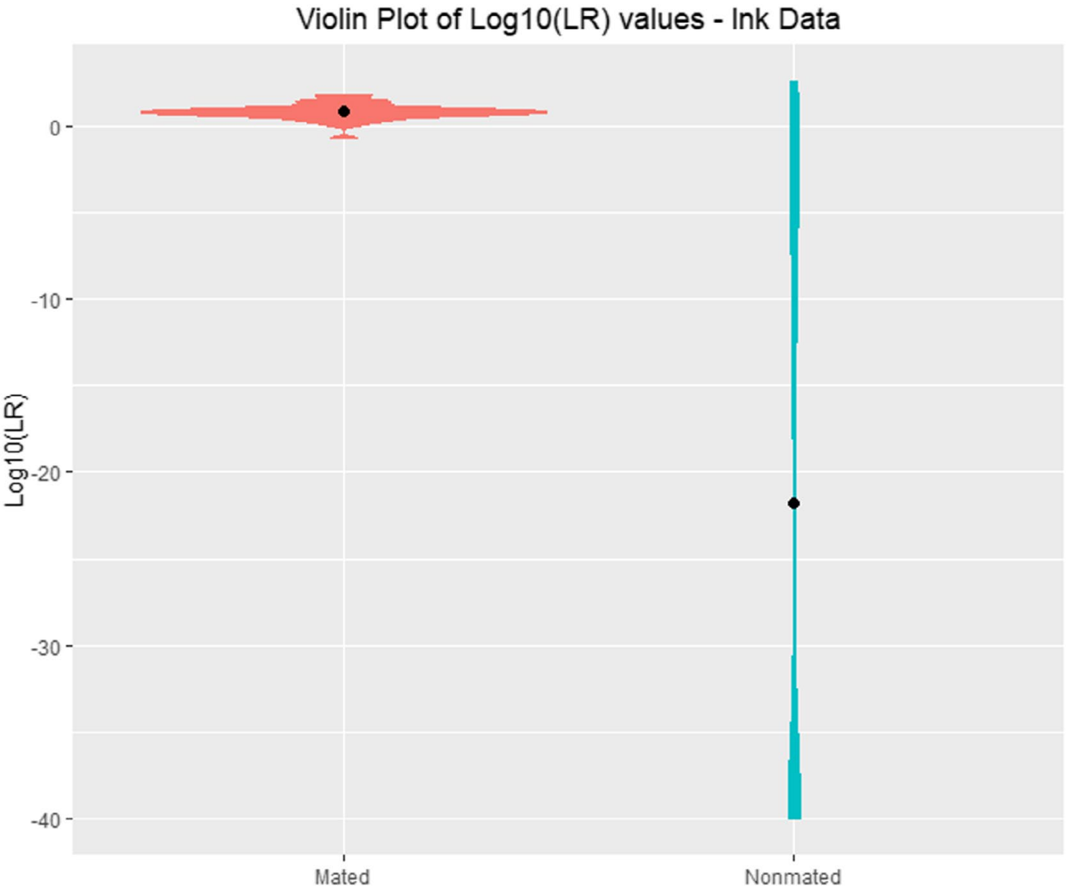

FIGURE B4 Violin Plot for ink data

B.3 | Ink data

The empirical ROC curve for the *LR* system for the Ink data is shown in Figure B3. Again, the *LR* system shows good discrimination power. A violin plot shows the distributions of $\log_{10}(LR)$ values for mated and nonmated cases. See Figure B4.

B.4 | Fingerprint data

Table B1 shows how the area under the ROC curve (AUC) changes as a function of the number of minutiae available for comparison. This confirms the increasing trend in discrimination power that was seen in Figure 8 as the number of minutiae increases.

TABLE B1 Area under the ROC curve (AUC) as a function of number of minutiae available for the comparison

Number of Minutiae	3	4	5	6	7
AUC	0.9957317	0.9993912	0.9996826	0.9999484	0.9997780
Number of Minutiae	8	9	10	11	12
AUC	0.9999951	0.9999975	0.9999921	1.0000000	0.9999746

APPENDIX C

NONPARAMETRIC FIDUCIAL CONFIDENCE INTERVALS FOR CALIBRATION DISCREPANCY: THEORETICAL BACKGROUND

In this section we explain how generalized fiducial inference is used to quantify uncertainty in estimating $d(S_g, S_f)$ defined in Equation (6). Following the discussion in Cui and Hannig (2019) we derive a generalized fiducial distribution (GFD) for the survival function.

Let $S^{-1}(u) = \inf\{x \in \mathbb{R} : S(x) \leq u\}$ be the usual inverse of the survival function. Consider the data generating algorithm

$$X_i = S^{-1}(U_i), \quad i = 1, \ldots, n, \tag{C1}$$

where U_i are mutually independent, Uniform(0, 1) random variables. A GFD is obtained by inverting the data generating algorithm, see Hannig et al. (2016) for detailed discussion. In this context, inverting a GFD means that after observing $\mathbf{x} = (x_i)_{i=1,\ldots,n}$ and generating values of $\mathbf{u}^\star = (u_i^\star)_{i=1,\ldots,n}$ from the standard uniform distribution, we need to find all survival functions that satisfy (C1). To be more precise, denote the inverse image of Equation (C1)

$$Q_{\mathbf{x}}(\mathbf{u}^\star) = \bigcap_{i=1}^{n} \{S^* : S(x_i - \epsilon) > u_i^\star \geq S(x_i) \text{ for any } \epsilon > 0\}. \tag{C2}$$

Notice that $Q_{\mathbf{x}}(\mathbf{u}^\star) \neq \emptyset$ if the order of $\mathbf{u}^\star$ matches the reverse order of $\mathbf{x}$, in which case $Q_{\mathbf{x}}(\mathbf{u}^\star)$ contains infinitely many survival functions. For the purposes of this work we will use a representative survival function $S^* \in Q_{\mathbf{x}}(\mathbf{u}^\star)$ that is a linear spline. To generate fiducial samples S_r^*, $r = 1, \ldots, m$, of S^*, we sample m independent copies of $\mathbf{u}_\mathbf{i}^\star$, where the distribution of $\mathbf{u}_\mathbf{i}^\star$ is i.i.d Uniform(0,1) conditional on matching the reverse order of $\mathbf{x}$. These fiducial samples than can be used for inference.

To construct confidence intervals for the vector $d(S_g, S_f)$ we first obtain fiducial samples $S_{g,r}^\star$ and $S_{f,r}^\star$, $r = 1, \ldots, m$, using the observed LRs that were collected under H_p-true and H_d-true scenarios respectively. Then we substitute these sampled survival functions into Equation (5). The resulting fiducial sample $d(S_{g,r}^\star, S_{f,r}^\star)$, $r = 1, \ldots, m$ is then used to obtain approximate confidence interval for the unknown $d(S_g, S_f)$ using sets of fiducial probability $1 - \alpha$, that is, containing $m(1 - \alpha)$ fiducial samples $d(S_{g,r}^\star, S_{f,r}^\star)$.

This approach is justified by the following Bernstein–von Mises type result that guarantees that the fiducial pointwise and simultaneous confidence interval for $d(S_g, S_f)$ will have asymptotically correct coverage and is therefore theoretically justified for examining calibration discrepancy of LRs.

Theorem 1 *Let us assume that* $1 > \mathrm{S_g}(\mathrm{a}_1) > \cdots > \mathrm{S_g}(\mathrm{a_k}) > 0$ *and* $1 > \mathrm{S_f}(\mathrm{a}_1) > \cdots > \mathrm{S_f}(\mathrm{a_k}) > 0$ *and the observed* LR*s are independent of each other. Denote the number of observed* LR*s under* $\mathrm{H_p}$*-true and* $\mathrm{H_d}$*-true scenarios as* $\mathrm{n_g}$, $\mathrm{n_f}$, *the corresponding empirical survival functions as* $\widehat{S}_g$, $\widehat{S}_f$, *and samples from the fiducial survival functions* $S_g^\star$, $S_f^\star$ *respectively. Finally let* $\mathrm{n} = min(\mathrm{n_g}, \mathrm{n_f})$ *and assume* $\lim_{n\to\infty} n/n_f = p_f$ *and* $\lim_{n\to\infty} n/n_g = p_g$. *Then as* $\mathrm{n} \to \infty$

$$\sqrt{n}(d(\widehat{S}_g, \widehat{S}_f) - d(S_g, S_f)) \xrightarrow{\mathcal{D}} N(0, \Sigma_{g,f}),$$

*and conditionally on the observed LR*s

$$\sqrt{n}(d(S_g^\star, S_f^\star) - d(\widehat{S}_g, \widehat{S}_f)) \xrightarrow{\mathcal{D}} N(0, \Sigma_{g,f}) \quad a.s.$$

Proof In this proof we will work with distribution functions rather than survival functions, that is, $F(s) = 1 - S_f(s)$, $G(s) = 1 - S_g(s)$. Donsker's theorem (van der Vaart, 1998) implies that

$$n_g(\widehat{G} - G) \xrightarrow{\mathcal{D}} B_g, \quad n_f(\widehat{F} - F) \xrightarrow{\mathcal{D}} B_f \quad \text{in } \mathbb{D}[0, \infty),$$

where B_g, B_f are independent mean zero Gaussian process with covariance

$$EB_g(s)B_g(t) = G(s)(1 - G(t)), \quad EB_f(s)B_f(t) = F(s)(1 - F(t)) \text{ when } s \le t.$$

It follows from Theorem 2 of Cui and Hannig (2019) that conditionally on the observed LR*s*

$$n_g(G^\star - \widehat{G}) \xrightarrow{\mathcal{D}} B_g, \quad n_f(F^\star - \widehat{F}) \xrightarrow{\mathcal{D}} B_f \quad \text{in } \mathbb{D}[0, \infty) \quad \text{a.s.}$$

Because of independence the convergence also happens jointly, that is,

$$\sqrt{n}\left(\begin{pmatrix} \widehat{G} \\ \widehat{F} \end{pmatrix} - \begin{pmatrix} G \\ F \end{pmatrix}\right) \xrightarrow{\mathcal{D}} \begin{pmatrix} p_g B_g \\ p_f B_f \end{pmatrix}, \quad \text{and } \sqrt{n}\left(\begin{pmatrix} G^\star \\ F^\star \end{pmatrix} - \begin{pmatrix} \widehat{G} \\ \widehat{F} \end{pmatrix}\right) \xrightarrow{\mathcal{D}} \begin{pmatrix} p_g B_g \\ p_f B_f \end{pmatrix} \quad \text{a.s.}$$

To complete the proof we need to use functional delta method (Theorems 20.8, 23.9 of van der Vaart, 1998).
To this end notice that Equation (5) is equivalent to

$$d_{(a,b)}(G, F) = \log_{10}(G(b) - G(a)) - \log_{10}\left(bF(b) - aF(a) - \int_a^b F(s)\, ds\right).$$

The Hadamard's derivative of $d_{(a,b)}$ at (G, F) for $U,\ W \in \mathbb{D}_0$ is

$$d'_{(a,b),G,F}(U,W) = \frac{1}{\log(10)}\left(\frac{U(b)-U(a)}{G(b)-G(a)} - \frac{bW(b)-aW(a)-\int_0^1 W(s)\,ds}{bF(b)-aF(a)-\int_a^b F(s)\,ds}\right).$$

Consequently,

$$\sqrt{n}(d(\widehat{G},\widehat{F})-d(G,F)) \xrightarrow{\mathcal{D}} (d'_{(a_1,a_2),G,F}(p_gB_g,p_fB_f),\dots,d'_{(a_{k-1},a_k),G,F}(p_gB_g,p_fB_f))^\top.$$

Similarly, conditionally on the observed *LR*s

$$\sqrt{n}(d(G^\star,F^\star)-d(\widehat{G},\widehat{F})) \xrightarrow{\mathcal{D}} (d'_{(a_1,a_2),G,F}(p_gB_g,p_fB_f),\dots,d'_{(a_{k-1},a_k),G,F}(p_gB_g,p_fB_f))^\top, \text{ a.s.}$$

This proves the theorem.