Forensic Science International: Synergy 3 (2021) 100157

Contents lists available at ScienceDirect

Forensic Science International: Synergy

journal homepage: www.sciencedirect.com/journal/forensic-science-international-synergy

In the context of forensic casework, are there meaningful metrics of the degree of calibration?

Geoffrey Stewart Morrison [a,b,*]

[a] Forensic Data Science Laboratory & Forensic Speech Science Laboratory, Computer Science Department & Aston Institute for Forensic Linguistics, Aston University, Birmingham, UK
[b] Forensic Evaluation Ltd, Birmingham, UK

ARTICLE INFO

Keywords:
Forensic inference and statistics
Likelihood ratio
Calibration
Metric

ABSTRACT

Forensic-evaluation systems should output likelihood-ratio values that are well calibrated. If they do not, their output will be misleading. Unless a forensic-evaluation system is intrinsically well-calibrated, it should be calibrated using a parsimonious parametric model that is trained using calibration data. The system should then be tested using validation data. Metrics of degree of calibration that are based on the pool-adjacent-violators (PAV) algorithm recalibrate the likelihood-ratio values calculated from the validation data. The PAV algorithm overfits on the validation data because it is both trained and tested on the validation data, and because it is a non-parametric model with weak constraints. For already-calibrated systems, PAV-based ostensive metrics of degree of calibration do not actually measure degree of calibration; they measure sampling variability between the calibration data and the validation data, and overfitting on the validation data. Monte Carlo simulations are used to demonstrate that this is the case. We therefore argue that, in the context of casework, PAV-based metrics are not meaningful metrics of degree of calibration; however, we also argue that, in the context of casework, a metric of degree of calibration is not required.

1. Introduction

1.1. Forensic-evaluation systems should output well-calibrated likelihood-ratio values

Forensic-evaluation systems should output likelihood-ratio values that are well calibrated [1–12]. If they do not, their output will be misleading. For a well-calibrated system, the likelihood ratios of the likelihood-ratio values that it outputs will be the same as the likelihood-ratio values that it outputs (Birdsall [13] §1.2). In practice, unless one were to train and test on the same data, because of sampling variability, if one were to re-calibrate an already well-calibrated system one would expect the likelihood ratios of the likelihood-ratio values only to be approximately the same as the original likelihood-ratio values.[1]

1.2. Causes of poorly-calibrated likelihood-ratio output

If a forensic-evaluation system makes use of feature vectors (i.e., sets of measurements made on the objects of interest) that have a small number of dimensions and that have distributions that do not violate the assumptions of a parsimonious parametric statistical model, and the number of data points available for model training is large compared to the number of parameter values to be estimated, then the output of the model will be intrinsically well calibrated. In real forensic settings, however, it is common for the feature vectors to have a large number of dimensions, for the fitted models to be complex, and for the number of data points available for training to be small, thus requiring a large number of parameter values to be estimated from a limited amount of data. Classic examples of high-dimensional data and complex models can be found in forensic voice comparison [6], but, with limited data, even moderate numbers of dimensions can lead to miscalibrated results even for relatively parsimonious models; see, for example [14,15], and the commentary of [8] on the latter.

* Forensic Data Science Laboratory & Forensic Speech Science Laboratory, Computer Science Department & Aston Institute for Forensic Linguistics, Aston University, Birmingham, UK.
E-mail address: geoff-morrison@forensic-evaluation.net.

[1] We use the term "sampling variability" to mean variation between different samples drawn from the same population in the same set of conditions, not variation between different samples drawn from different populations or from different sets of conditions.

https://doi.org/10.1016/j.fsisyn.2021.100157
Received 7 May 2021; Received in revised form 29 May 2021; Accepted 8 June 2021
Available online 12 June 2021

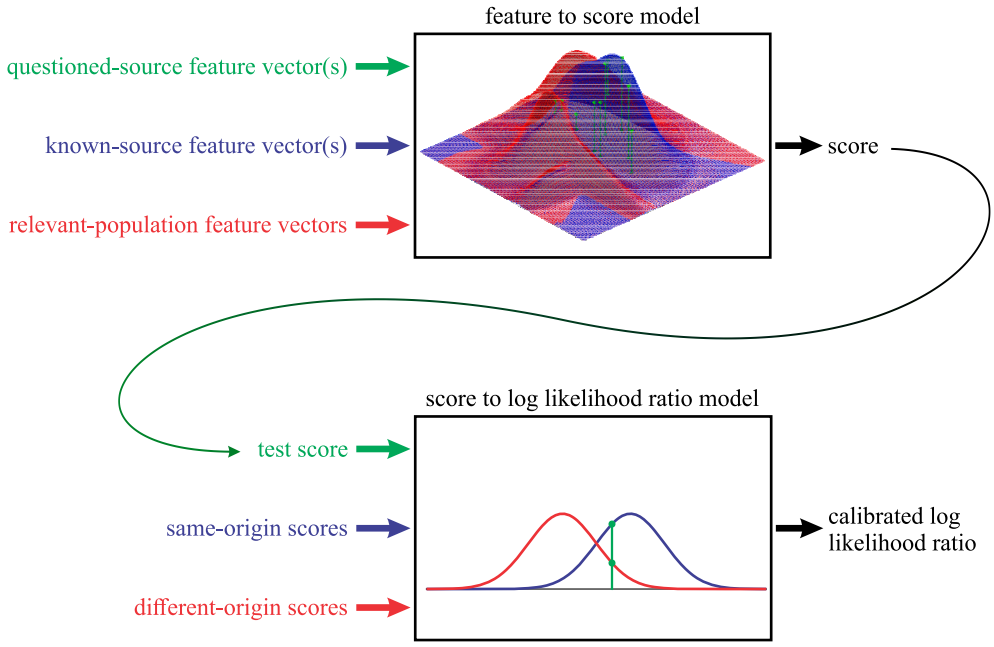

Fig. 1. Schematic of a forensic-evaluation system consisting of a feature-to-score model (a complex multidimensional model that outputs uncalibrated likelihood ratios) followed by a score-to-log-likelihood-ratio model (a parsimonious unidimensional calibration model).

1.3. How to calibrate forensic-evaluation systems

A practical solution to the problem described in the previous section is to treat the output of the model as uncalibrated likelihood ratios, and then use a second model to calibrate the output of the first model ([3, 16–18]), see Fig. 1. For simplicity, we will henceforth refer to the uncalibrated likelihood ratios output by the first-stage model as "scores", and refer to the calibrated likelihood ratios output by the second-stage model as "likelihood ratios". Also for simplicity, we will assume that the forensic problem at hand is source attribution.

The second-stage module is trained using a separate dataset from that previously used to train the first-stage model. We will henceforth refer to the second dataset as the "calibration data". Same-source and different-source pairs are constructed from the calibration data. Those pairs are input to the first-stage model which then outputs a set of same-source scores and a set of different-source scores. The second-stage calibration model is trained using those same-source and different-source scores. The scores are univariate, and a parsimonious parametric model is used as the calibration model. Hence, even with a moderate amount of calibration data, there are a relatively large number of data points available to estimate a small number of parameter values. This results in well-calibrated output.

A point to note is that the calibration model is applied to scores that are uncalibrated log likelihood ratios – the calculation of the scores has taken account of both the similarity between the members of each pair, and their typicality with respect to the relevant population. Using scores that only take account of similarity will not result in meaningful likelihood-ratio values [19–21].

Another point to note is that the calibration data must be representative of the relevant population for the case and must reflect the conditions of the questioned-source specimen and known-source sample in the case ([1,10]). If there is a mismatch between the conditions of the questioned-source specimen and known-source sample, then one member of each pair in the calibration data must reflect the conditions of the questioned-source specimen and the other member of the pair must reflect the conditions of the known-source sample. If the calibration data do not represent the relevant population for the case and do not reflect the conditions for the case, then the resulting model will miscalibrate the output. The decision as to whether the calibration data are sufficiently representative of the relevant population for the case and sufficiently reflective of the conditions for the case will be a subjective judgement made by the forensic practitioner, but this should be made transparent so that the decision can be reviewed by an independent practitioner and potentially be debated before the court ([10,22,23]).

1.4. Metrics of degree of calibration

1.4.1. Introduction

Several metrics have been proposed for measuring the degree of calibration of the output of a forensic-evaluation system.[2] Vergeer et al. [11] explored the performance of different metrics using simulated data for which the true distributions were known. Metrics based on the expected value of different-source likelihood-ratio values and the expected value of the inverse of same-source likelihood-ratio values (after Good [24]) did not perform as desired, nor did metrics based on the proportion of different-source likelihood ratios above 2 and the proportion of same-source likelihood ratios below 0.5 (after Royall [25]). We will not discuss these metrics further here. Instead, we will focus on metrics that make use of the pool-adjacent-violators (PAV) algorithm ([16,26,27]).[3]

1.4.2. $C_{\text{llr}}^{\text{cal}}$

The more established of the PAV-based metrics is $C_{\text{llr}}^{\text{cal}}$ (Brümmer & du Preez [16]). $C_{\text{llr}}^{\text{cal}} = C_{\text{llr}} - C_{\text{llr}}^{\text{min}}$, where C_{llr} is the log-likelihood-ratio cost, calculated as in Eq. (1),[4] and $C_{\text{llr}}^{\text{min}}$ is C_{llr} calculated after the log-likelihood-ratio values resulting from the validation data have been transformed using PAV. PAV is a non-parametric algorithm that, subject only to the constraint of monotonicity, shifts the log-likelihood-ratio

[2] There can be ambiguity as to whether the term "calibration" refers to the process of calibrating a system or to a property of a system, i.e., how well calibrated its output is. We will use "degree of calibration" to refer to the latter meaning.

[3] PAV is also known as isotonic regression.

[4] The form of Eq. (1) is that given in González-Rodríguez et al. [1] and thereafter widely repeated in the literature. It can be derived from Brümmer & du Preez [16] Eq. (43). C_{llr} is equivalent to the deviance statistic assuming equal priors for the two categories.

values so as to minimize C_{llr}. The same same-source and different-source log-likelihood-ratio values that are used for training PAV are themselves transformed and used to calculate $C_{\text{llr}}^{\text{min}}$.

$$C_{\text{llr}} = \frac{1}{2}\left(\frac{1}{N_s}\sum_{i}^{N_s}\log_2\left(1+\frac{1}{\Lambda_{s_i}}\right)+\frac{1}{N_d}\sum_{j}^{N_d}\log_2\left(1+\Lambda_{d_j}\right)\right) \quad (1)$$

In Eq. (1), Λ_{s_i} and Λ_{d_j} are respectively the same-source and different-source likelihood-ratio values output by the system in response to the validation data, and N_s and N_d are respectively the number of same-source and different-source likelihood-ratio values.[5] In order for the results to be meaningful in the context of the case, the validation data must be representative of the relevant population for the case and must reflect the conditions of the questioned-source specimen and known-source sample in the case, including any mismatch between them ([1, 10]). The validation data must also be separate from the calibration data (and from any other data used for training the system).[6]

1.4.3. devPAV

A novel PAV-based metric, devPAV, was introduced in Vergeer et al. [11]. For a graphical explanation of the calculation of devPAV, see Ref. [11] Fig. 2. Log likelihood ratios are calculated using the validation data. The PAV algorithm is applied to the resulting log likelihood ratios. The PAV-based log-likelihood-ratio to recalibrated-log-likelihood-ratio mapping function is plotted, with log-likelihood-ratio values on the x axis and recalibrated-log-likelihood-ratio values on the y axis. The line $y = x$ is plotted on the same axes. If the log-likelihood-ratio values output by the system were perfectly calibrated, then recalibrating them would (theoretically) result in the same values, i.e., $y = x$. Within the range from the smallest same-source log likelihood ratio to the largest different-source log likelihood ratio (the range within which the recalibrated-log-likelihood-ratio values will be finite), the area between $y = x$ and the log-likelihood-ratio to recalibrated-log-likelihood-ratio mapping function is calculated. This is achieved by stepping through adjacent pairs of log-likelihood-ratio and recalibrated-log-likelihood-ratio values, and calculating the areas of rectangles and triangles that piecewise make up the total area. The total area is then divided by the length of the range from the smallest same-source log likelihood ratio to the largest different-source log likelihood ratio. For the calculation of devPAV, both log-likelihood-ratio values and recalibrated-log-likelihood-ratio values are scaled as base-ten logarithms.

2. Argument

2.1. Introduction

The purpose of the present paper is to present the argument that, in the context of conducting casework, once a forensic-evaluation system has been appropriately calibrated, subsequently calculated PAV-based ostensive metrics of degree of calibration do not in fact provide information about degree of calibration. Instead, PAV-based ostensive metrics of degree of calibration provide information about sampling variability between the calibration and validation data and about overfitting on the validation data. Once a forensic-evaluation system has been appropriately calibrated, PAV-based metrics are not meaningful metrics of degree of calibration. The fact that they are not meaningful metrics of degree of calibration, however, is not of concern because, in the context of deciding whether a system is sufficiently well calibrated to be used for a case, a metric of degree of calibration is not required.

Note that the argument presented here relates to attempted measurement of degree of calibration of already-calibrated systems, not to measurement of degree of calibration of uncalibrated systems.

Note, also, that the argument presented here relates to the use of metrics of degree of calibration in the context of using a forensic-evaluation system in a case and presenting the results to a court (or to some other decision maker in the judicial process). It does not relate to the use of metrics of degree of calibration in the context of research and development of forensic-evaluation systems nor to selection of which of multiple systems to use. The present paper is written from the perspective of best practice for a forensic practitioner who is conducting a forensic evaluation or who is independently reviewing a report on a forensic evaluation conducted by another forensic practitioner. The present paper should be read in the context of the *Consensus on validation of forensic voice comparison* [10].

2.2. Metrics of degree of calibration are not required

An astute reader of the *Consensus on validation of forensic voice comparison* [10] may have noticed that, although it recommended that forensic-evaluation systems be well calibrated, it did not recommend that practitioners calculate and present to a court a metric of degree of calibration. In the context of a case, discussion regarding calibration should not centre around ostensive metrics of degree of calibration. Instead, it should centre around the following questions:

- Has the system been calibrated using an appropriate calibration model?[7]
- Has the calibration model been trained using appropriate data?

In order for these questions to be answerable, the forensic practitioner must describe the calibration model and the calibration data so that their appropriateness can potentially be reviewed by an independent practitioner and can potentially be debated before the court.

An example of lack of appropriate calibration in the context of a forensic-voice-comparison case is described in Morrison [28]: The questioned-speaker recording was a recording of a mobile telephone call in which the speaker of interest was distant from the telephone, and the known-speaker recording was a recording of a landline telephone call in which the speaker of interest was in a highly reverberant environment. In contrast, the forensic-voice-comparison system was trained on high-quality audio recordings, and it did not include an explicit calibration stage.

An example of a calibration model that would be inappropriate for evidential casework is described in Jessen et al. [29]: The calibration model included shifting the scores so that 10% of the different-source scores had values greater than 0. This may be appropriate in an investigative context in which one requires a 10% false-alarm rate, but, in the context of assessing strength of evidence for presentation in court, unless this accidently corresponds to the shift that minimizes C_{llr} (and for the conditions tested in Ref. [29], it did not), this procedure deliberately miscalibrates the output of the system.

An example of use of inappropriate calibration data in the context of a forensic-voice-comparison case is described in Morrison [23]: The speakers of interest on the questioned- and known-speaker recordings had West Yorkshire accents, and the questioned-speaker recordings were covert recordings made in a car. These were poor-quality recordings that included engine and traffic noise. In contrast, the calibration data were high-quality audio recordings of speakers with "standard southern British English" accents, and consisted of only one recording of each speaker (different parts of the same recordings were

[5] Readers requiring a gentler introduction to C_{llr} are referred to Ref. [6] §20.8 or [10] Appendix C.

[6] Cross-validation is often used as a means of maximizing use of limited calibration and validation data while avoiding training and testing on the same data.

[7] We define an appropriate calibration model as one that, subject to the constraints of the model, either directly or indirectly minimizes the C_{llr} for the calibration data.

used to create same-speaker pairs). For another example of use of inappropriate calibration data, see Morrison & Thompson [22] §7.[8]

These are examples in which the appropriateness of the calibration model and calibration data could be (or actually were) debated before a court. In none of these examples would a metric of degree of calibration have been of assistance.

A metric of degree of calibration might be of assistance in demonstrating that a calibration model is inappropriate, but a C_{llr} value greater than 1 or a graphical representation (such as a Tippett plot or probability-density plot) would probably be sufficient to convey gross miscalibration to a court. If an appropriate calibration model and appropriate calibration and validation data have been used, then one would not expect a C_{llr} value greater than approximately 1. If the C_{llr} value is less than 1 the system is providing useful information, therefore, *ceteris paribus*, it would be better to use that system than to use no system.

In the context of providing a critique of a forensic-evaluation report, the practitioner who is critiquing the report is unlikely to have access to the evaluation-software (including the calibration model) or the calibration and validation data used by the practitioner who conducted the evaluation (assuming they exist – all too often a critique points out that there was no calibration or validation). Hence a practitioner who is independently critiquing the report will usually not be able to generate graphics or metrics indicative of the degree of calibration of the forensic-evaluation system that was actually used. In such circumstances, all they can do is discuss whether the calibration model and the calibration and validation data were appropriate from a theoretical perspective.

A metric of degree of calibration would not be of assistance in deciding on the appropriateness of the calibration data (nor would a graphical representation of results): If it were decided that the calibration model and the calibration data were appropriate, and validation data were selected using the same criteria as were used to select the calibration data, but in reality the calibration and validation data were not appropriate, then no amount of testing using the validation data would reveal that mistake. All resulting performance metrics, including metrics of degree of calibration, would be misleading, but there would be no way of knowing that this was the case. The decision as to whether the calibration and validation data are appropriate is a pre-empirical decision.

2.3. PAV-based ostensive metrics of degree of calibration actually measure sampling variability and overfitting

Assume we have a two-stage system including a feature-to-score model then a score-to-log-likelihood-ratio model. The latter is the calibration model. The calibration model is a parsimonious model trained on a set of calibration data, and the performance of the system is tested using a set of validation data. Both the calibration data and the validation data are selected using the same criteria to decide whether they are sufficiently representative of the relevant population for the case and sufficiently reflective of the conditions for the case. In fact, it would be usual to obtain a single data set and then split it into a calibration set and a validation set, either as two completely separate sets or via cross-validation.

Either directly or indirectly, an appropriate calibration model will, subject to the constraints of the model, minimize C_{llr} for the calibration data.[9] The C_{llr} value calculated for the calibration data, however, will be the result of training and testing on the same data, and will therefore be overfitted on the calibration data and will tend to be overly optimistic with respect to the expected performance of the system when applied to previously-unseen data. Importantly, previously-unseen data include the questioned-source specimen and known-source sample in the case. Results intended to be representative of the expected performance of the system when generalized to previously-unseen data are obtained using validation data. Same-source and different-source pairs are constructed from the validation data. Those pairs are input to the first-stage model which then outputs a set of same-source scores and a set of different-source scores. These validation scores are input to the calibration model that was already trained on the calibration data. The resulting calibrated log-likelihood-ratio values derived from the validation data are used to calculate a C_{llr} value. The latter C_{llr} value represents the expected performance of the system when applied to previously-unseen data, such as the questioned-source specimen and known-source sample in the case.

If one were to take the log-likelihood-ratio values resulting from the validation data, use them to train a new calibration model and then recalibrate them using that model, one would be both training and testing on the validation data, would overfit on the validation data, and would tend to obtain overly optimistic results. If the same type of model were used for calibration and recalibration, a metric based on the difference between calibrated and recalibrated results would therefore simply capture the difference due to sampling variability between the calibration and validation data, and due to overfitting because of both training and testing on the validation data. If the recalibration model were PAV and it was both trained and tested on the validation data, then the results would be doubly overfitted. They would be doubly overfitted not just because of training and testing on the same data, but also because of the weak constraints of the non-parametric PAV algorithm.

In the description of the devPAV metric in §1.4.3 above, we wrote: "If the log-likelihood-ratio values output by the system were perfectly calibrated, then recalibrating them would (theoretically) result in the same values, i.e., $y = x$." We included the parenthetical "theoretically" because, in practice, even if the log-likelihood-ratio values were perfectly calibrated, the overfitting of PAV to real data would result in differences between the PAV-transformed log-likelihood-ratio values (the y values) and the original pre-PAV log-likelihood-ratio values (the x values).

The C_{llr} value based on the calibration data and the C_{llr} value based on the validation data will differ because of sampling variability, but this is not a problem. These two values are not compared with each other, only the latter is presented as a metric of accuracy. The same would be true for other metrics of accuracy such as false-alarm rate and miss-rate in a classification framework. The problem lies in both training and testing on the validation data, and overtraining on the validation data, then comparing a measure of the accuracy of the resulting system ($C_{\text{llr}}^{\text{min}}$) with a measure of the accuracy of the system that will actually be used in the case (the C_{llr} value based on the calibrated system and validation data). $C_{\text{llr}}^{\text{min}}$ characterizes the performance of a system that included PAV-calibration on the validation data. Since this is not the system that will actually be used to compare the questioned-source specimen and known-source sample in the case, $C_{\text{llr}}^{\text{min}}$ is not informative about the performance of the system that will actually be used in the case.

One would not usually use the non-parametric PAV as the actual calibration model because it would overfit its training data and tend not to generalize well to new data. One would usually deliberately choose a parsimonious parametric model that would be a less good fit for its training data but tend to generalize better to new data. Linear discriminant analysis (LDA) and logistic regression (LogReg) are examples of parsimonious models that could be used – they both result in a linear mapping between scores and the log likelihood ratios. A linear mapping requires the estimation of only two parameter values. LogReg is usually preferred over LDA because it does not depend on as strong assumptions – it is more robust when the data deviate from being Gaussians with the

[8] Technically, in these examples, the data were used to implement a normalization procedure, which serves a similar function but is not exactly the same as an explicit calibration model.

[9] A model that did not directly or indirectly minimize C_{llr} could potentially be appropriate in some other context, but would not be appropriate in the context of calculating likelihood ratios as expressions of strength of evidence to be used for legal-decision making.

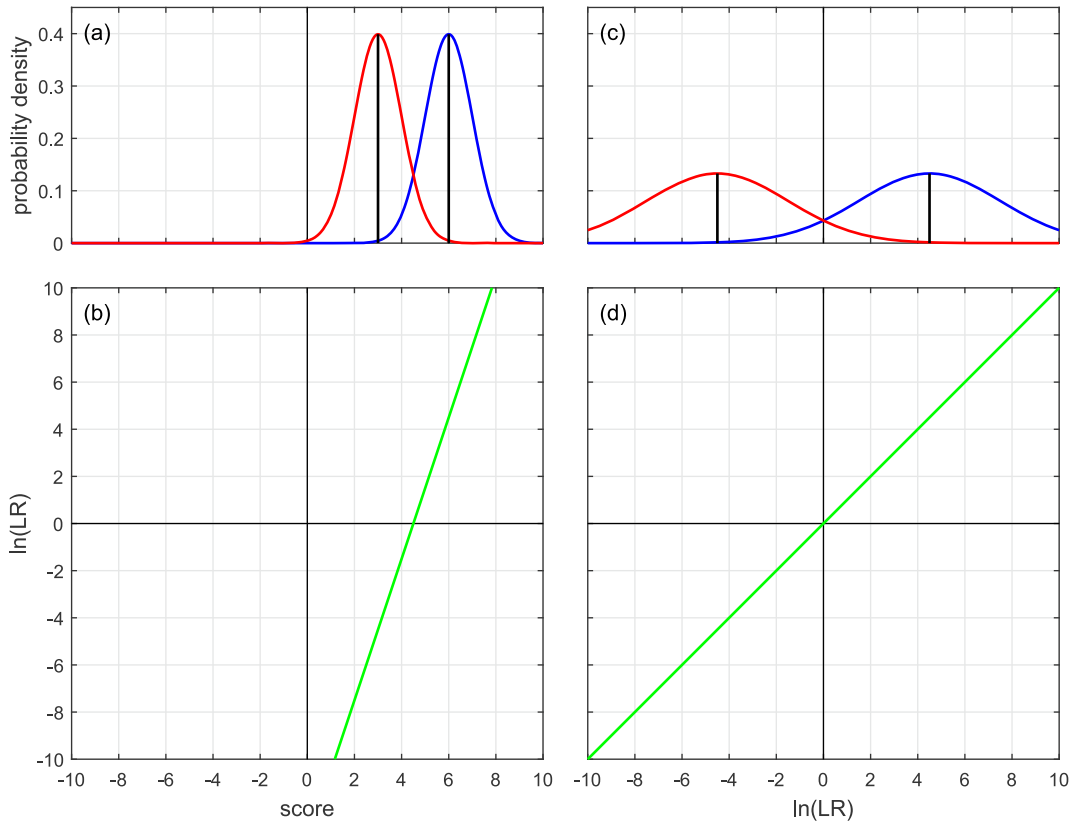

Fig. 2. (a): Monte Carlo population distributions, $\mu_d = 3$, $\mu_s = 6$, $\sigma = 1$. (b): Score-to-log-likelihood-ratio mapping function corresponding to (a). (c): Distributions of (a) after transformation using the mapping function in (b), $\mu_d = -4.5$, $\mu_s = 4.5$, $\sigma = 3$. (d): Log-likelihood-ratio-to-log-likelihood-ratio mapping function corresponding to (c).

same variance. One could potentially use non-linear, but still monotonic, models that would require estimating only a few more parameter values.

The effects of sampling variability and of overfitting would be reduced for very large data sets, but in forensic practice the amount of case-relevant data available is usually relatively small.

2.4. Conclusion

Forensic-evaluation systems should be calibrated using a parsimonious parametric calibration model trained using calibration data, and should then be tested using validation data. The calculation of PAV-based ostensive metrics of degree of calibration involves both training and testing on the validation data. Therefore, for a system that has already been calibrated using a parsimonious calibration model, what the PAV-based metrics are measuring is not degree of calibration. What they are measuring is sampling variability between the calibration and validation data, and the difference in fit between a parsimonious parametric model and an overfitted non-parametric model.

In the next section we support the theoretical argument made in the present section by presenting demonstrations based on Monte Carlo simulations.

3. Demonstrations

3.1. Introduction

Following Vergeer et al. [11], we present demonstrations based on simulated data. By specifying the population distributions, we can compare empirical results with expected results. By generating multiple Monte Carlo samples, we can explore effects due to sampling variability.

3.2. Perfectly calibrated systems

Assume a Gaussian population distribution for the same-source scores and a Gaussian population distribution for the different-source scores, and assume that the two Gaussians have the same variance.[10] Fitting an LDA model would result in a score-to-log-likelihood-ratio mapping function that is linear, i.e., has the equation $y = a + bx$, in which x is a score value (which has the form of a log-likelihood-ratio value), y is the corresponding calibrated natural-log-likelihood-ratio value,[11] and a and b are the intercept and slope. As shown in Eqs. (2)–(4), the value of the slope (b) depends on the separation of the same-source mean and the different-source mean (μ_s and μ_d) relative to their shared variance (σ^2), and the intercept (a) depends on the location of the midpoint between the same-source mean and the different-source mean.

$$y = \ln\left(\frac{\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{(x-\mu_s)^2}{-2\sigma^2}}}{\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{(x-\mu_d)^2}{-2\sigma^2}}}\right) = \ln\left(e^{\frac{(x-\mu_s)^2-(x-\mu_d)^2}{-2\sigma^2}}\right) = \frac{x^2+\mu_s^2-2x\mu_s-x^2-\mu_d^2+2x\mu_d}{-2\sigma^2}$$
$$= \frac{-\mu_s^2+2x\mu_s+\mu_d^2-2x\mu_d}{2\sigma^2} = \frac{-\mu_s^2+\mu_d^2}{2\sigma^2} + \frac{\mu_s-\mu_d}{\sigma^2}x = a+bx \quad (2)$$

[10] Usually the different-source scores would have substructure due to each feature-data point used to generate them being used in multiple pairs (source 1 v source 2, source 1 v source 3, source 2 v source 3, etc.), but for simplicity we ignore that here.

[11] Henceforth, unless explicitly stated otherwise, all references to logarithms are to natural logarithms.

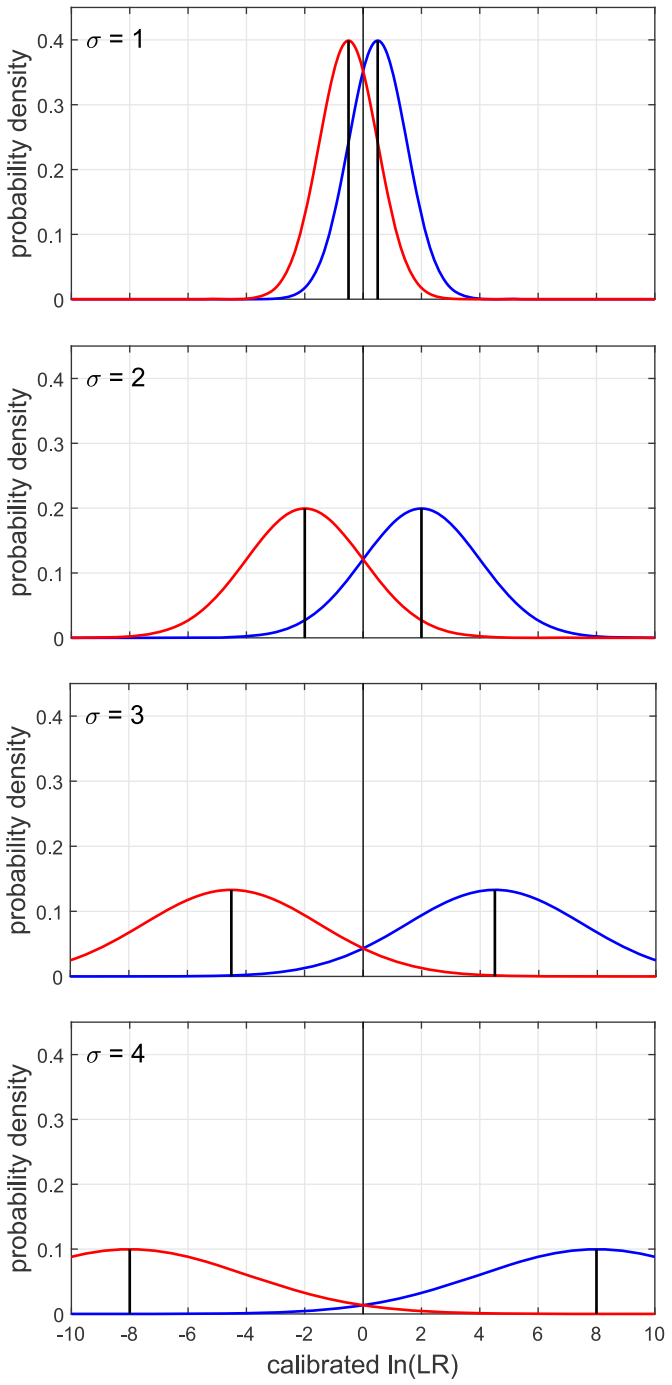

Fig. 3. Examples of $\ln(LR)$ distributions for perfectly calibrated systems with different values for σ_{cal}.

$$b = \frac{\mu_s - \mu_d}{\sigma^2} \tag{3}$$

$$a = \frac{-\mu_s^2 + \mu_d^2}{2\sigma^2} = -b\frac{\mu_s + \mu_d}{2} \tag{4}$$

If, for example, for the Monte Carlo population distributions we specify $\mu_d = 3$, $\mu_s = 6$, and $\sigma = 1$ (Fig. 2(a)), the score-to-log-likelihood-ratio mapping function will be $y = -\frac{6-3}{1^2} \times \frac{6+3}{2} + \frac{6-3}{1^2}x = -13.5 + 3x$ (Fig. 2(b)). Hence, the parameter values for the transformed distributions will be $\mu_d = -13.5 + 3 \times 3 = -4.5$, $\mu_s = -13.5 + 3 \times 6 = 4.5$, and $\sigma = 3 \times 1 = 3$ (Fig. 2(c)). If we recalibrate these values, the recalibration mapping function (the log-likelihood-ratio-to-log-likelihood-ratio function) will be $y = -\frac{4.5+4.5}{3^2} \times \frac{4.5-4.5}{2} + \frac{4.5+4.5}{3^2}x = -0 \times 1x = x$ (Fig. 2(d)), i.e., the log likelihood ratios of the calibrated log likelihood ratios equal the calibrated log likelihood ratios.

In general, once one has ascertained σ^2 (the common variance for the same-source and different-source scores) and μ_s and μ_d (the means of the same-source and different-source scores) one knows everything about the distributions of the log-likelihood-ratios of the calibrated system: The calibrated standard deviation will be $\sigma_{\text{cal}} = b\sigma = \frac{\mu_s - \mu_d}{\sigma}$,[12] and the calibrated means will be located symmetrically about 0 with a separation of σ_{cal}^2, i.e., $\mu_{\text{cal},d} = -\frac{\sigma_{\text{cal}}^2}{2}$ and $\mu_{\text{cal},s} = +\frac{\sigma_{\text{cal}}^2}{2}$ (Peterson et al. [30] §4.9; Good [24]; van Leeuwen & Brümmer [18]). Fig. 3 shows examples of $\ln(LR)$ distributions for perfectly calibrated systems with different values for σ_{cal}.

3.3. Monte Carlo simulations

Assume a situation in which the feature data for each of the calibration set and the validation set allow us to generate 50 same-source scores and 1225 different-source scores (the latter number being the size of the upper-right of a 50 × 50 matrix). Further assume that the Monte Carlo populations consist of Gaussians with means $\mu_d = 3$ and $\mu_s = 6$, both with the same standard deviation $\sigma = 1$ (Fig. 2(a)). We draw Monte Carlo samples consisting of 50 same-source scores and 1225 different-source scores. We draw one sample as a calibration set and one sample as a validation set. We use the calibration set to train a calibration model, apply the calibration model to the validation set, then calculate C_{llr} for the resulting calibrated log-likelihood-ratio values. We recalibrate the calibrated log-likelihood-ratio values, both training and testing the recalibration model on the calibrated log-likelihood-ratio values, then calculate C_{llr} for the recalibrated log-likelihood-ratio values. Hereinafter, we refer to the latter as $C_{\text{llr}}^{\text{recal}}$, which equals $C_{\text{llr}}^{\text{min}}$ if the recalibration model is PAV.[13]

We repeat this process 10,000 times, and each time:

- We compare C_{llr} for the calibrated log-likelihood-ratio values with the expected C_{llr} value given the Monte Carlo population parameters, i.e., we calculate $C_{\text{llr}} - C_{\text{llr}}^{\text{expect}}$.
- We compare C_{llr} for the calibrated log-likelihood-ratio values with the recalibrated C_{llr} value, i.e., we calculate $C_{\text{llr}} - C_{\text{llr}}^{\text{recal}}$. If the recalibration model is PAV, $C_{\text{llr}} - C_{\text{llr}}^{\text{recal}} = C_{\text{llr}} - C_{\text{llr}}^{\text{min}} = C_{\text{llr}}^{\text{cal}}$.
- We calculate devPAV, and, for comparison purposes, devLDA and devLogReg. The latter were calculated in the same way as devPAV, but, rather than using the PAV-derived likelihood-ratio-to-recalibrated-log-likelihood-ratio mapping function, the LDA- or LogReg-derived likelihood-ratio-to-recalibrated-log-likelihood-ratio mapping function was used instead.[14]

$C_{\text{llr}} - C_{\text{llr}}^{\text{expect}}$ is the perfect metric of degree of calibration, but it is not a practical metric: It can only be calculated when one has oracle

[12] In signal detection theory, $d' = \frac{\mu_s - \mu_d}{\sigma}$.

[13] We used an implementation of the PAV algorithm that returned $-\infty$ for all different-source log-likelihood-ratio values below the smallest same-source log-likelihood-ratio value and $+\infty$ for all same-source log-likelihood-ratio values above the largest different-source log-likelihood-ratio value. If small and large finite values were used instead (e.g., following Laplace's rule of succession), then when there are large separations between μ_s and μ_d (as in §3.5.4 below), most of the transformed values would be either the small finite value or the large finite value, and the value of $C_{\text{llr}}^{\text{min}}$ could be larger than the value of C_{llr}. In pilot work, this was actually the case.

[14] Part of calculating in the same way as devPAV included only calculating devLDA and devLogReg over the range between the smallest same-source log-likelihood-ratio value and the largest different-source log-likelihood-ratio value.

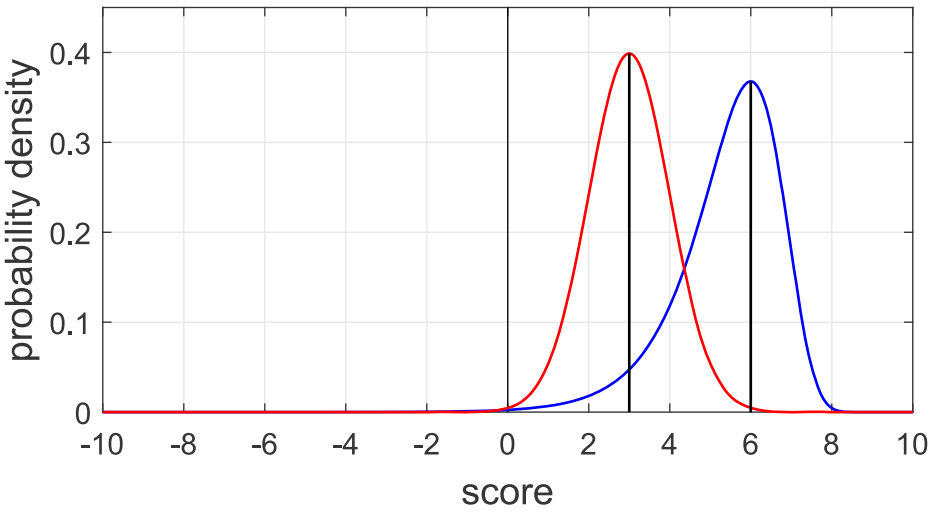

Fig. 4. Monte Carlo population distributions: a Gaussian distribution for different-source scores and a skewed distribution for same-source scores.

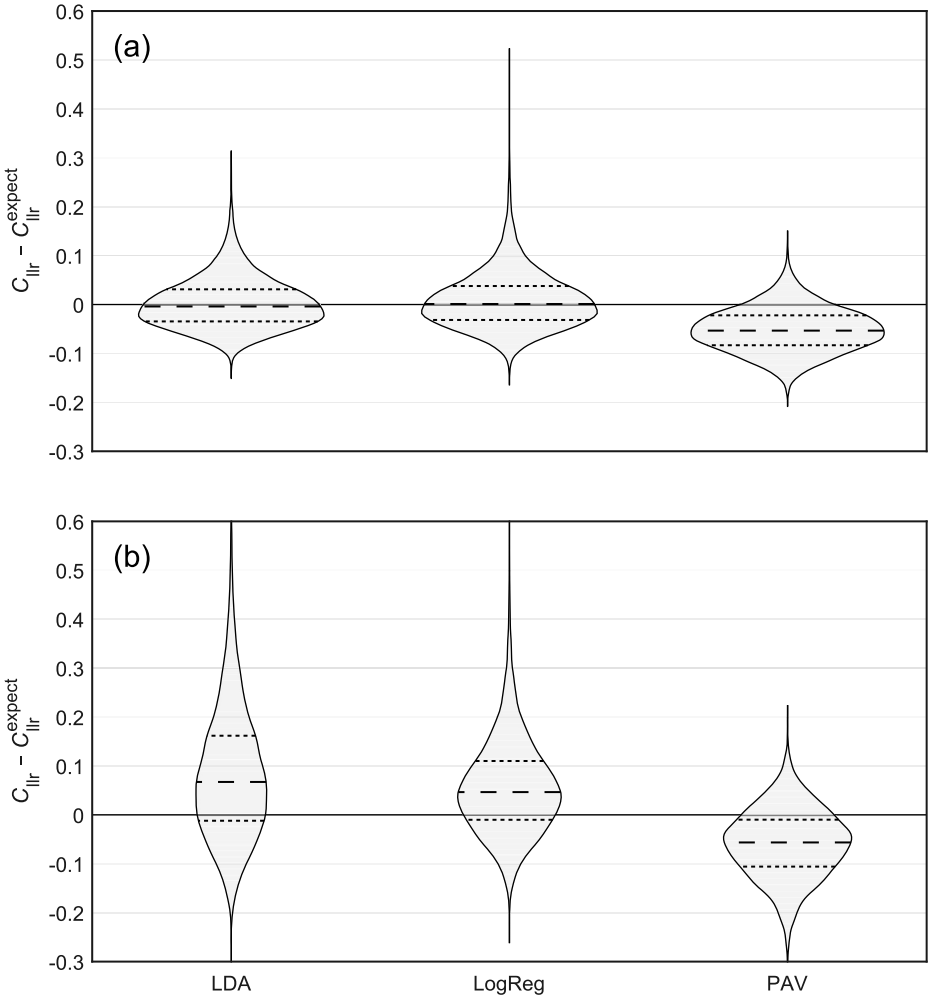

Fig. 5. Violin plots for $C_{\text{llr}} - C_{\text{llr}}^{\text{expect}}$ given: (a) Gaussian distributions for both different-source and same-source scores; (b) Gaussian distribution for different-source scores and a skewed distribution for same-source scores.

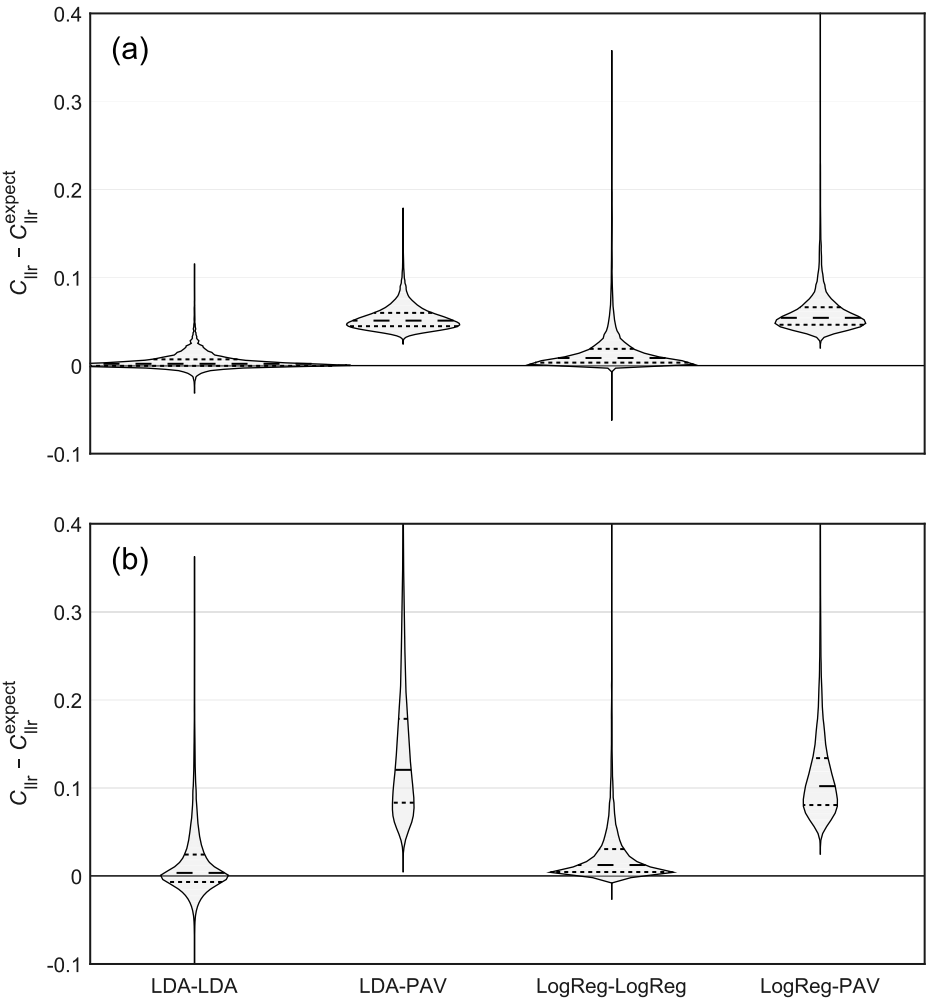

Fig. 6. Violin plots for $C_{llr} - C_{llr}^{recal}$ given: (a) Gaussian distributions for both different-source and same-source scores; (b) Gaussian distribution for different-source scores and a skewed distribution for same-source scores.

knowledge of the population distributions, i.e., in the context of Monte Carlo simulations. Rather than an analytical solution for C_{llr}^{expect} (which, depending on the population distributions, may not exist), we obtain a Monte Carlo approximation by drawing a sample of 500,000 same-source score values and 500,000 different-source score values, then for each score value we calculate the corresponding log-likelihood-ratio value given the Monte Carlo population models, and finally we calculate C_{llr} for those log-likelihood-ratio values.[15]

We compared the following four combinations of calibration and recalibration models: LDA-LDA, LDA-PAV, LogReg-LogReg, and LogReg-PAV.[16]

We repeated the entire process using different Monte Carlo population distributions: Reflecting a pattern seen for empirical score distributions (for examples, see §5 of Morrison & Poh [31]), we used a same-source distribution that has a negative skew (a heavy left tail), see Fig. 4. We generated this based on a Gumbel distribution, see Eq. (5), in which g is the probability density function for a Gumbel distribution, g^{-1} is a function that generates random numbers based on a Gumbel distribution with the specified parameter values, and μ_s and σ have the same values as previously used for the Gaussian same-source distribution.

$$x = \mu_s - g^{-1}(\nu = 0, \tau = \sigma) \tag{5}$$

$$g(x|\nu,\tau) = \frac{e^{-\left(\frac{x-\nu}{\tau} + e^{-\frac{x-\nu}{\tau}}\right)}}{\tau}$$

The Matlab code used to run these simulations is available at http://geoff-morrison.net/#no_cal_metric. The code can be modified to explore other settings, including changing the separation between the same-source and different-source distributions and changing the sample size.

3.4. Results

For the system with Gaussian distributions for both different-source and same-source scores, the value for C_{llr}^{expect} was 0.240. For the system with a Gaussian distribution for different-source scores and a skewed distribution for same-source scores, the value for C_{llr}^{expect} was 0.461.

Figs. 5–7 summarize the values for $C_{llr} - C_{llr}^{expect}$, $C_{llr} - C_{llr}^{recal}$, and devPAV/devLDA/devLogReg resulting from the Monte Carlo simulations. The figures have been formatted such that all violin plots within a

[15] Similarly, van Leeuwen & Brümmer [18] proposed calculation of C_{llr}^{expect} using numerical integration.

[16] The logistic-regression models were not regularized.

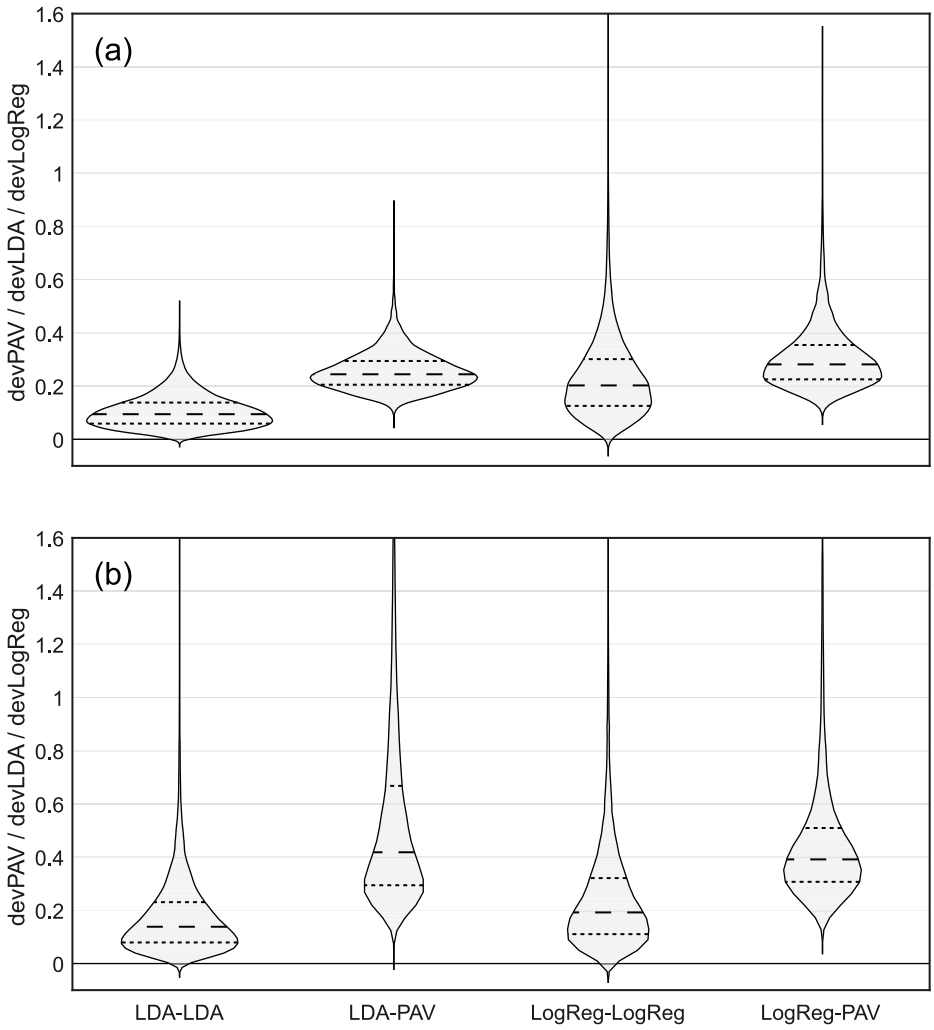

Fig. 7. Violin plots for devPAV given: (a) Gaussian distributions for both different-source and same-source scores; (b) Gaussian distribution for different-source scores and a skewed distribution for same-source scores.

figure have the same area.

3.5. Discussion

3.5.1. $C_{llr} - C_{llr}^{expect}$ results

As shown in Fig. 5(a), when both the same-source and different-source Monte Carlo population distributions were Gaussian, for both LDA- and LogReg-calibration models, C_{llr} values were centred around C_{llr}^{expect} with a slight positive skew. LDA, a parametric model whose assumptions were met by the population distributions, had a slightly tighter C_{llr} distribution than did LogReg, which was fitted using an iterative algorithm. The variation of the C_{llr} values about C_{llr}^{expect} reflects sampling variability of both the calibration data and the validation data.

Sampling variability also accounts for the spread of the C_{llr} values when the recalibration model was PAV, but those values were not centred around C_{llr}^{expect}, their distribution was substantially lower. The PAV-recalibrated C_{llr} values, i.e., C_{llr}^{min} values, tended to be lower. The reason for this is overfitting as a result of using a minimally-constrained non-parametric model that was both trained and tested on the validation data. In this example, the LDA-calibrated and LogReg-calibrated C_{llr} values were on-average closer to C_{llr}^{expect}, i.e., the LDA-calibrated and LogReg-calibrated log-likelihood-ratio values were on-average closer to the "true" log-likelihood-ratio values obtained using oracle knowledge of the true Monte Carlo population distributions, i.e., the LDA-calibrated and LogReg-calibrated log-likelihood-ratio values were better calibrated than the PAV-recalibrated log-likelihood-ratio values.

As shown in Fig. 5(b), when the different-source Monte Carlo population distribution was Gaussian but the same-source Monte Carlo population distribution was skewed, for both LDA- and LogReg-calibration models, C_{llr} values were usually higher than C_{llr}^{expect}. The results were not as well calibrated as when the models' assumptions were met by the population distributions. LogReg, which is not as sensitive to deviations from Gaussian distributions with the same variance as is LDA, had a somewhat tighter C_{llr} distribution than did LDA, i.e., LogReg-calibrated log-likelihood-ratio values where somewhat better calibrated than LDA-calibrated log-likelihood-ratio values. A model with a few more parameters to fit a non-linear (but still monotonic) mapping function would potentially lead to a better degree of calibration. Potential improvement in degree of calibration would have to be traded off against the danger of overfitting on the calibration data and then not generalizing well. As before, PAV overfitted the validation data and PAV-recalibrated C_{llr} values, i.e., C_{llr}^{min} values, tended to be lower than C_{llr}^{expect}.

The results shown in Fig. 5 indicate that the values of the metric $C_{llr}^{cal} = C_{llr} - C_{llr}^{min}$ will tend to be larger than the corresponding value for the perfect metric of degree of calibration, $C_{llr} - C_{llr}^{expect}$, i.e., the value $\left(C_{llr} - C_{llr}^{min}\right) - \left(C_{llr} - C_{llr}^{expect}\right) = C_{llr}^{expect} - C_{llr}^{min}$ will tend to be positive. This is demonstrated in the rightmost violin plots of Fig. 5, for which the values of $C_{llr}^{min} - C_{llr}^{expect}$ tended to be negative. C_{llr}^{cal} would therefore tend

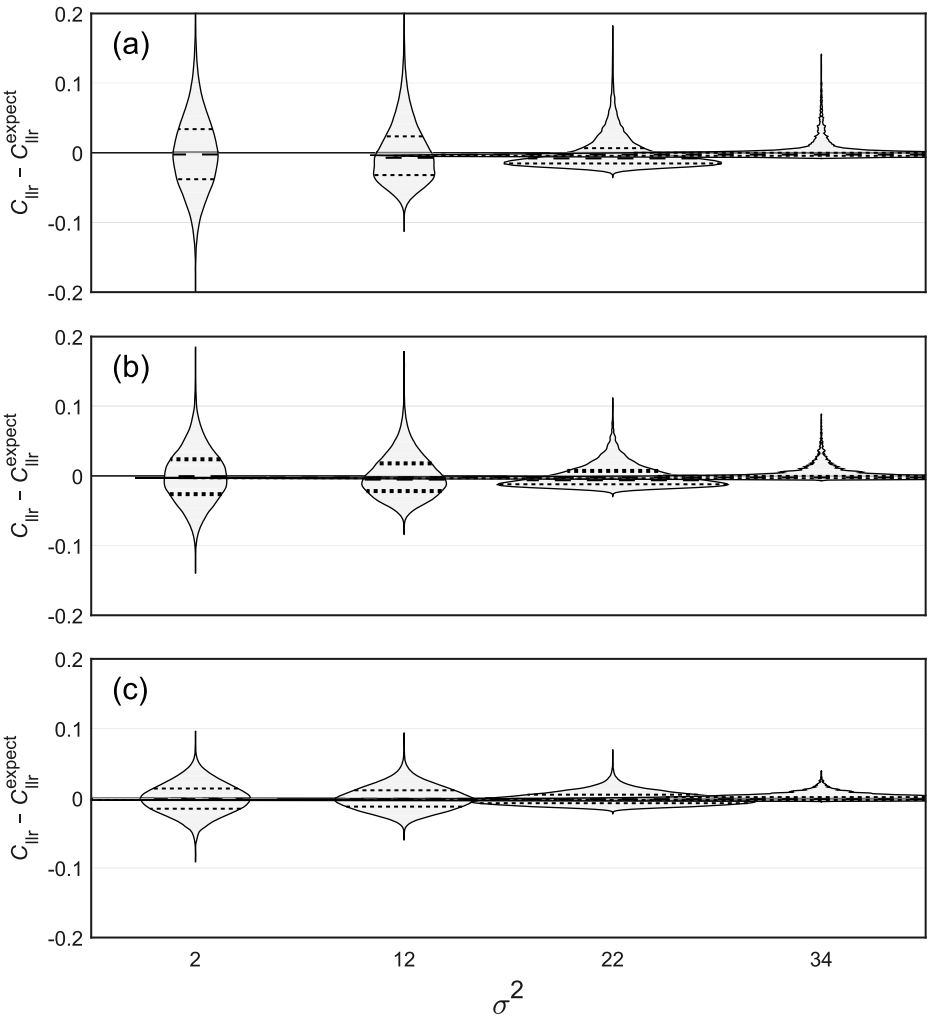

Fig. 8. Violin plots for $C_{\text{llr}} - C_{\text{llr}}^{\text{expect}}$ given a range of σ^2_{cal} values, and sample sizes of (a) 50, (b) 100, and (c) 300.

to indicate a poorer degree of calibration than is actually the case.

3.5.2. $C_{\text{llr}} - C_{\text{llr}}^{\text{recal}}$ *results*

As shown in Fig. 6, when calibrating and recalibrating using the same type of model, LDA-LDA and LogReg-LogReg, there was a spread in the distribution of the $C_{\text{llr}} - C_{\text{llr}}^{\text{recal}}$ values. This was due to sampling variability between the calibration data and the validation data. The spread was narrower in Fig. 6(a), for which the Monte Carlo population distributions met the assumptions of the LDA-LDA models (both the same-source and different-source distributions were Gaussians with the same variance). In Fig. 6(a) the distributions for LDA-LDA were slightly narrower than for LogReg-LogReg. In Fig. 6(b), for which the different-source Monte Carlo population distribution was Gaussian but the same-source Monte Carlo population distribution was skewed, the spread in the distribution of the $C_{\text{llr}} - C_{\text{llr}}^{\text{recal}}$ values for both LDA-LDA and LogReg-LogReg were wider than in Fig. 6(a). In Fig. 6(b), the spread for LogReg-LogReg was less than for LDA-LDA, LogReg being more robust to violations of the assumptions of Gaussians with the same variance.

In Fig. 6, the distributions of $C_{\text{llr}} - C_{\text{llr}}^{\text{recal}}$ for LDA-LDA were close to symmetrical about 0, with only very slight positive skew. It appears that both training and testing the LDA recalibration model on the validation data did not lead to substantial overfitting. This is an advantage of a parsimonious parametric model. For LogReg-LogReg, there was a positive skew to the $C_{\text{llr}} - C_{\text{llr}}^{\text{recal}}$ distributions. This reflects some overfitting due to both training and testing the LogReg recalibration model on the validation data.

For the LDA-LDA and LogReg-LogReg models, is $C_{\text{llr}} - C_{\text{llr}}^{\text{recal}}$ an indicator of degree of calibration? If it were, would we expect to see a shift in values in Fig. 6(b) similar to the shift in values in Fig. 5(b)? We argue that what $C_{\text{llr}} - C_{\text{llr}}^{\text{recal}}$ reflects is not degree of calibration but sampling variability between the calibration and validation data, and, more so for LogReg-LogReg, some overfitting on the validation data.

For LDA-PAV and LogReg-PAV, for which the recalibration models were PAV,. The distributions of the $C_{\text{llr}}^{\text{cal}}$ values were substantially greater than 0, and substantially greater than the $C_{\text{llr}} - C_{\text{llr}}^{\text{recal}}$ distributions for LDA-LDA and LogReg-LogReg. As previously discussed in the context of the $C_{\text{llr}} - C_{\text{llr}}^{\text{expect}}$ results, the substantially larger values for $C_{\text{llr}}^{\text{cal}}$ were due to the minimally-constrained non-parametric PAV model being both trained and tested on the validation data and overfitting the validation data. Due to this overfitting, $C_{\text{llr}}^{\text{cal}}$ tends to indicate a poorer degree of calibration than is actually the case. Due to this overfitting, we argue that $C_{\text{llr}}^{\text{cal}}$ is not a meaningful metric of degree of calibration for systems that have already been calibrated using a parsimonious parametric model.

Ferrer et al. [33] similarly observed that PAV overfitted on small data sets. Rather than calculating $C_{\text{llr}}^{\text{min}}$ using PAV, they calculated an alternative version using LogReg, i.e., the same as our $C_{\text{llr}}^{\text{recal}}$ for the LogReg-LogReg models. Data sets that are considered "small" in the automatic-speaker-recognition literature (to which Ferrer et al. [33] belongs) may be larger than case-relevant data sets typically available in forensic casework contexts.

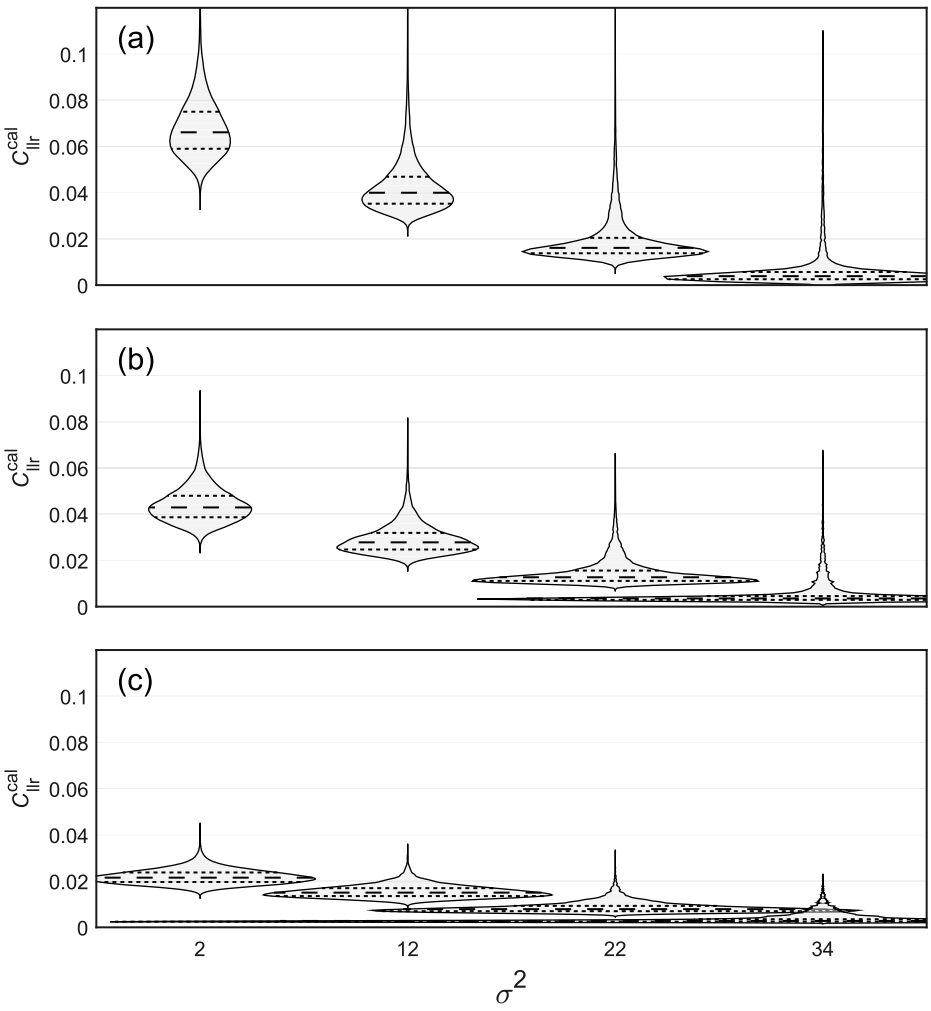

Fig. 9. Violin plots for C_{llr}^{cal} given a range of σ^2_{cal} values, and sample sizes of (a) 50, (b) 100, and (c) 300.

3.5.3. devPAV results

By design, devPAV values are greater than or equal to 0.[17] Otherwise, results in Fig. 7 show the same relative pattern as for $C_{llr} - C_{llr}^{recal}$ results in Fig. 6. In particular, as for C_{llr}^{cal} compared to $C_{llr} - C_{llr}^{recal}$ for LDA-LDA and LogReg-LogReg, the devPAV distributions (for LDA-PAV and LogReg-PAV) had larger values than for devLDA and devLogReg (for LDA-LDA and LogReg-LogReg). The larger values for devPAV were due to the minimally-constrained non-parametric PAV model being both trained and tested on the validation data, and it overfitting the validation data to a greater extend than did the parsimonious parametric models. Due to this overfitting, we argue that devPAV is not a meaningful metric of degree of calibration for systems that have already been calibrated using a parsimonious parametric model.

3.5.4. Selected Vergeer et al. (2021) results

Vergeer et al. [11] included comparison of the behaviour of different calibration metrics given perfectly calibrated Monte Carlo population distributions consisting of Gaussians with the same variance. The values used for σ^2_{cal} were 2, 12, 22, and 34, and same-source sample sizes used were 50, 100, and 300.[18] Vergeer et al. [11] did not present results from the full factorial of these combinations. We replicated this portion of Vergeer et al. [11], and our Figs. 8–10 show the full factorial of results for, $C_{llr} - C_{llr}^{expect}$, C_{llr}^{cal}, and devPAV distributions respectively. The arguments we make below could have been based on results already presented in Vergeer et al. [11], but examining the full factorial makes the pattern of results more obvious.

The C_{llr}^{expect} values for σ^2_{cal} of 2, 12, 22, and 34 were 0.710, 0.155, 0.038, and 0.007 respectively. For the perfect metric of degree of calibration, $C_{llr} - C_{llr}^{expect}$, the distributions shown in Fig. 8 were centred around 0. The spread of the distributions is due to sampling variability. As the size of the samples increased, from panel (a) through panel (c), the spread of the distributions decreased. This is the expected effect on sampling variability of increasing the sample size. As the separation between the same-source and different-source log-likelihood-ratio values increased, from left to right, the spread of the distributions also

[17] Any values in Fig. 7 or 10 that appear to be less than 0 are due to the bandwidth of the kernels used to draw the violin plots.

[18] The σ^2_{cal} values of 2, 12, 22, and 34, were encoded in Ref. [11] as $\mu_{cal,s}$ values of 1, 6, 11, and 17. Separations of 22 and 34 variance units would have produced likelihood-ratio values many orders of magnitude larger than the sample size. In the context of a case we would apply a method to avoid overstating strength of evidence (e.g. Refs. [31,34,35]), but did not do so for these simulations.

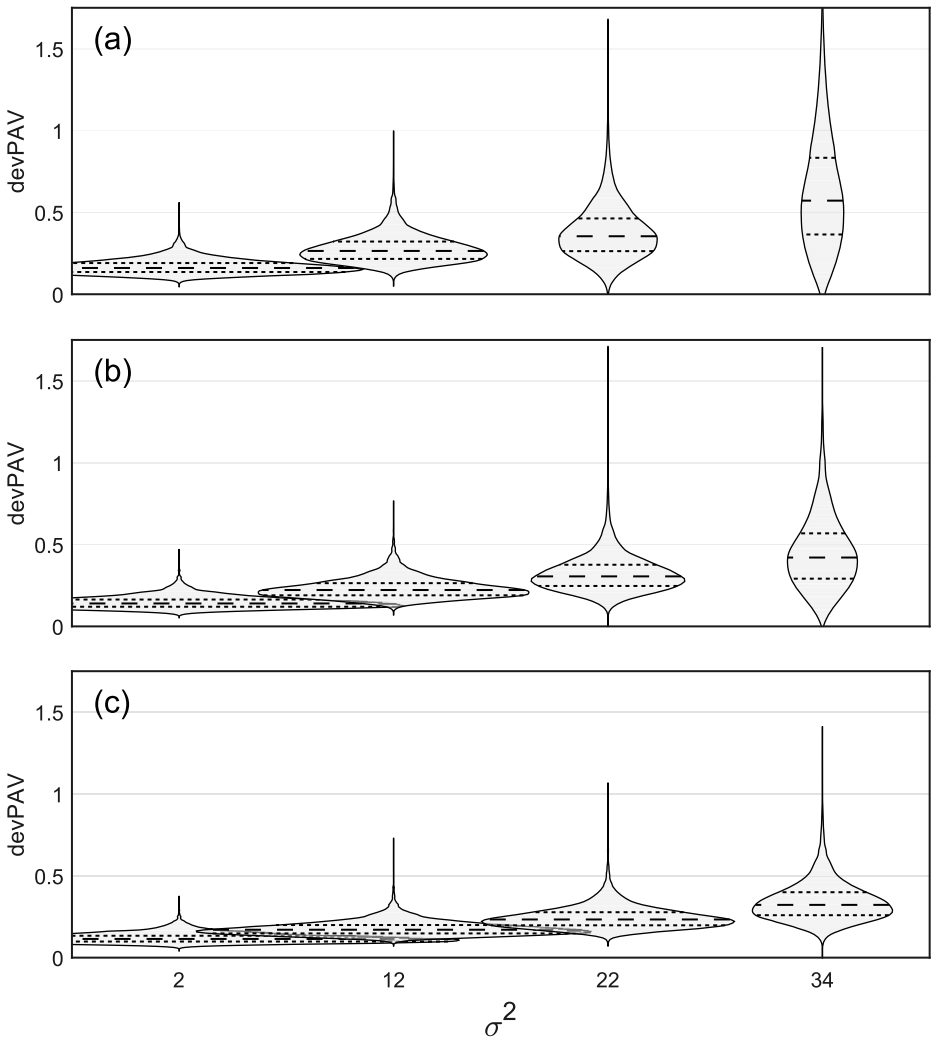

Fig. 10. Violin plots for devPAV given a range of σ^2_{cal} values, and sample sizes of (a) 50, (b) 100, and (c) 300.

decreased. This is due to the fact that as the separation between the same-source and different-source log-likelihood-ratio values increased both C_{llr} values and $C_{\text{llr}}^{\text{expect}}$ values decreased, thus the magnitude of the difference between them decreased.

The distributions of $C_{\text{llr}}^{\text{cal}}$ values shown in Fig. 9 exhibited the same pattern of spread as for the $C_{\text{llr}} - C_{\text{llr}}^{\text{expect}}$ values, but, in addition, absolute $C_{\text{llr}}^{\text{cal}}$ values decreased as sample size increased and as the separation between the same-source and different-source log-likelihood-ratio values increased.

The distributions of devPAV values shown in Fig. 10 exhibited a different pattern: As the size of the samples increased, from panel (a) through panel (c), the spread of the distributions decreased, but as the separation between the same-source and different-source log-likelihood-ratio values increased, from left to right, the spread of the distributions increased rather than decreased. Also in contrast to $C_{\text{llr}}^{\text{cal}}$ values, as the separation between the same-source and different-source log-likelihood-ratio values increased, the average devPAV values increased rather than decreased. The increase in the spread as the separation between the same-source and different-source log-likelihood-ratio values increased may be due to the fact that devPAV is only calculated for log-likelihood-ratio values in the range from the smallest same-source value to the largest different-source value. As the separation between the same-source and different-source log-likelihood-ratio values increases, this range will decrease, making the amount of data on which devPAV is calculated smaller and thus making devPAV more sensitive to sampling variability. The increase in average devPAV values as the separation between the same-source and different-source log-likelihood-ratio values increased may also be related to the decrease of the range over which it is calculated – since devPAV only has positive values, an increase in the spread of those values would be correlated with an increase in their average value.

Given that all the Monte Carlo population distributions were perfectly calibrated, a good metric of degree of calibration should have had the same average value for all the different population distributions. Because this was not the case for either $C_{\text{llr}}^{\text{cal}}$ or devPAV (across the different population distributions the median varied 27-fold for $C_{\text{llr}}^{\text{cal}}$ and 5-fold for devPAV), we argue that neither is a good metric of degree of calibration.[19]

4. Conclusion

All forensic-evaluation systems used in casework should be calibrated. If they are not intrinsically well calibrated, they should include an explicit calibration model.

[19] The comparison across the different population distributions may be somewhat unfair given that a σ^2_{cal} of 34 is rather extreme, but this was a replication of the range of values used in Vergeer et al. [11].

We have presented an argument that, in the context of casework, PAV-based ostensive metrics of degree of calibration (C_{llr}^{cal} and devPAV) are not meaningful metrics of degree of calibration for systems that have already been calibrated using a parsimonious calibration model. We have argued that, in this context, rather than measuring degree of calibration, PAV-based metrics reflect sampling variability between the calibration and validation data and overfitting on the validation data.

The fact that PAV-based ostensive metrics of degree of calibration are not meaningful metrics of degree of calibration in the context of a casework is not of concern, however, because a metric of degree of calibration is not required: A decision as to whether a calibration model is appropriate in the context of a case does not require the use of a metric of degree of calibration. $C_{llr}>1$ and graphical representations would be sufficient to indicate gross miscalibration, which will not occur if an appropriate calibration model has been used. Whether a calibration model is appropriate may (and often only can) be argued on theoretical grounds, and a decision as to whether the calibration data are appropriate is a pre-empirical decision that cannot be informed by a metric of degree of calibration.

Disclaimer

All opinions expressed in the present paper are those of the author, and, unless explicitly stated otherwise, should not be construed as representing the policies or positions of any organizations with which the author is associated.

Declaration of competing interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by Research England's Expanding Excellence in England Fund as part of funding for the Aston Institute for Forensic Linguistics 2019–2022.

The author thanks those who provided feedback on drafts of the paper and by so doing helped improve the final version.

References

[1] J. González-Rodríguez, P. Rose, D. Ramos, D.T. Toledano, J. Ortega-García, Emulating DNA: rigorous quantification of evidential weight in transparent and testable forensic speaker recognition, IEEE Trans. Audio Speech Lang. Process. 15 (2007) 2104–2115, https://doi.org/10.1109/TASL.2007.902747.

[2] D. Ramos, J. González-Rodríguez, Reliable support: measuring calibration of likelihood ratios, Forensic Sci. Int. 230 (2013) 156–169, https://doi.org/10.1016/j. forsciint.2013.04.014.

[3] G.S. Morrison, Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio, Aust. J. Forensic Sci. 45 (2013) 173–197, https://doi.org/10.1080/00450618.2012.733025.

[4] I.W. Evett, The logical foundations of forensic science: towards reliable knowledge, Philosophical Transactions of the Royal Society B 370 (2015), https://doi.org/10.1098/rstb.2014.0263 article 20140263.

[5] D. Meuwly, D. Ramos, R. Haraksim, A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation, Forensic Sci. Int. 276 (2017) 142–153, https://doi.org/10.1016/j.forsciint.2016.03.048.

[6] G.S. Morrison, E. Enzinger, D. Ramos, J. González-Rodríguez, A. Lozano-Díez, Statistical models in forensic voice comparison, in: D. Banks, K. Kafadar, D.H. Kaye, M. Tackett (Eds.), Handbook of Forensic Statistics, CRC, Boca Raton, FL, 2020, pp. 451–497, https://doi.org/10.1201/9780367527709. Ch. 20.

[7] D. Ramos, D. Meuwly, R. Haraksim, C.E.H. Berger, Validation of forensic automatic likelihood ratio methods, in: D. Banks, K. Kafadar, D.H. Kaye, M. Tackett (Eds.), Handbook of Forensic Statistics, CRC, Boca Raton, FL, 2020, pp. 143–163, https://doi.org/10.1201/9780367527709. Ch. 7.

[8] P. Vergeer, I. Alberink, M. Sjerps, R. Ypma, Why calibrating LR-systems is best practice. A reaction to "The evaluation of evidence for microspectrophotometry data using functional data analysis", Forensic Sci. Int. 314 (2020) https://doi.org/10.1016/j.forsciint.2020.110388 article 110388.

[9] Forensic Science Regulator, Codes of Practice and Conduct: Development of Evaluative Opinions (FSR-C-118 Issue 1), Forensic Science Regulator, Birmingham, UK, 2021. https://www.gov.uk/government/publications/development-of-evaluative-opinions.

[10] G.S. Morrison, E. Enzinger, V. Hughes, M. Jessen, D. Meuwly, C. Neumann, S. Planting, W.C. Thompson, D. van der Vloed, R.J.F. Ypma, C. Zhang, A. Anonymous, B. Anonymous, Consensus on validation of forensic voice comparison, Sci. Justice 61 (2021) 229–309, https://doi.org/10.1016/j.scijus.2021.02.002.

[11] P. Vergeer, Y. van Schaik, M. Sjerps, Measuring calibration of likelihood-ratio systems: a comparison of four metrics, including a new metric devPAV, Forensic Sci. Int. 321 (2021), https://doi.org/10.1016/j.forsciint.2021.110722 article 110722.

[12] R.J.F. Ypma, P.A. Maaskant - van Wijk, R. Gill, M. Sjerps, M. van den Berge, Calculating LRs for presence of body fluids from mRNA assay data in mixtures, Forensic Sci. Int.: Genetics 52 (2021), https://doi.org/10.1016/j.fsigen.2020.102455 article 102455.

[13] T.G. Birdsall, The Theory of Signal Detectability: ROC Curves and Their Character. Technical Report No. 177, Cooley Electronics Laboratory, Department of Electrical and Computer Engineering, The University of Michigan, Ann Arbor, Michigan, 1973.

[14] G.S. Morrison, Y. Kinoshita, Automatic-type calibration of traditionally derived likelihood ratios: forensic analysis of Australian English /o/ formant trajectories, in: Proceedings of Interspeech 2008 Incorporating SST 2008, 2008, pp. 1501–1504. https://www.isca-speech.org/archive/interspeech_2008/i08_1501.html.

[15] C.G.G. Aitken, Y.-T. Chang, P. Buzzini, G. Zadora, G. Massonnet, The evaluation of evidence for microspectrophotometry data using functional data analysis, Forensic Sci. Int. 305 (2019), https://doi.org/10.1016/j.forsciint.2019.110007 article 110007.

[16] N. Brümmer, J. du Preez, Application independent evaluation of speaker detection, Comput. Speech Lang 20 (2006) 230–275, https://doi.org/10.1016/j.csl.2005.08.001.

[17] D. Ramos Castro, Forensic Evaluation of the Evidence Using Automatic Speaker Recognition Systems. Doctoral Dissertation, Autonomous University of Madrid, 2007.

[18] D.A. van Leeuwen, N. Brümmer, The distribution of calibrated likelihood-ratios in speaker recognition, in: Proceedings of Biometric Technologies in Forensic Science, BTFS 2013, Nijmegen, The Netherlands, 2013, pp. 24–29.

[19] G.S. Morrison, E. Enzinger, Score based procedures for the calculation of forensic likelihood ratios – scores should take account of both similarity and typicality, Sci. Justice 58 (2018) 47–58, https://doi.org/10.1016/j.scijus.2017.06.005.

[20] C. Neumann, M. Ausdemore, Defence against the modern arts: the curse of statistics –Part II: 'Score-based likelihood ratios', Law Probab. Risk 19 (2020) 21–42, https://doi.org/10.1093/lpr/mgaa006.

[21] C. Neumann, J. Hendricks, M. Ausdemore, Statistical support for conclusions in fingerprint examinations, in: D. Banks, K. Kafadar, D.H. Kaye, M. Tackett (Eds.), Handbook of Forensic Statistics, CRC, Boca Raton, FL, 2020, pp. 277–324, https://doi.org/10.1201/9780367527709. Ch. 14.

[22] G.S. Morrison, W.C. Thompson, Assessing the admissibility of a new generation of forensic voice comparison testimony, Columbia Science and Technology Law Review 18 (2017) 326–434, https://doi.org/10.7916/stlr.v18i2.4022.

[23] G.S. Morrison, Admissibility of forensic voice comparison testimony in England and Wales, Crim. Law Rev. 2018 (1) (2018) 20–33.

[24] I.J. Good, Weight of evidence: a brief survey, in: J.M. Bernardo, M.H. DeGroot, D. V. Lindley, A.F.M. Smith (Eds.), Bayesian Statistics 2, Elsevier, 1985, pp. 249–270.

[25] R.M. Royall, Statistical Evidence: A Likelihood Paradigm, Chapman & Hall, London, 1997.

[26] M. Ayer, H.D. Brunk, G.M. Ewing, W.T. Reid, E. Silverman, An empirical distribution function for sampling with incomplete information, Ann. Math. Stat. 26 (1955) 641–647. https://www.jstor.org/stable/2236377.

[27] B. Zadrozny, C. Elkan, Transforming classifier scores into accurate multiclass probability estimates, in: Proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining, 2002, pp. 694–699.

[28] G.S. Morrison, The impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings, Forensic Sci. Int. 283 (2018) e1–e7, https://doi.org/10.1016/j.forsciint.2017.12.024.

[29] M. Jessen, J. Bortlík, P. Schwarz, Y.A. Solewicz, Evaluation of Phonexia automatic speaker recognition software under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01), Speech Commun. 111 (2019) 22–28, https://doi.org/10.1016/j.specom.2019.05.002.

[30] W.W. Peterson, T.G. Birdsall, W.C. Fox, The theory of signal detectability, Transactions of the IRE Professional Group on Information Theory 4 (1954) 171–211, https://doi.org/10.1109/TIT.1954.1057460.

[31] G.S. Morrison, N. Poh, Avoiding overstating the strength of forensic evidence: shrunk likelihood ratios/Bayes factors, Sci. Justice 58 (2018) 200–218, https://doi.org/10.1016/j.scijus.2017.12.005.

[33] L. Ferrer, M. McLaren, N. Brümmer, A speaker verification backend with robust performance across conditions, Preprint (2021). https://arxiv.org/abs/2102.01760.

[34] P. Vergeer, A. van Es, A. de Jongh, I. Alberink, R.D. Stoel, Numerical likelihood ratios outputted by LR systems are often based on extrapolation: when to stop extrapolating? Sci. Justice 56 (2016) 482–491, https://doi.org/10.1016/j.scijus.2016.06.003.

[35] R. Corzo, T. Hoffman, P. Weis, J. Franco-Pedroso, D. Ramos, J. Almirall, The use of LA-ICP-MS databases to calculate likelihood ratios for the forensic analysis of glass evidence, Talanta 186 (2018) 655–661, https://doi.org/10.1016/j.talanta.2018.02.027.