Validation of forensic automatic likelihood ratio methods

Daniel Ramos, Didier Meuwly, Rudolf Haraksim, Charles E.H. Berger

1 Introduction

Forensic practice is more and more under scrutiny, both from the general public and the scientific community. Press reports about forensic examination in criminal cases regularly question the scientific foundation of forensic science and challenge the results of its analysis and interpretation [Wash15]. In 2016, a report from the US President's Council of Advisors on Science and Technology appeared, entitled "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods" [PCAST16]. This report emphasizes the importance of the validity of methods for the credibility of forensic science. In their 2017 Annual Report, the UK forensic science regulator also stressed the importance of the validity of the methods in the accreditation process:

> "*The accreditation system is predicated on organisations being: a) accountable for the quality of their work and, b) able to demonstrate through regular audit and through evidence of staff competence and method validity that they are sustainably competent to produce reliable results"* [ForSciReg18].

The forensic community is currently actively developing and implementing quality assurance, by establishing worldwide harmonized minimum quality standards. Such standards can be used to demonstrate the scope of validity, the reliability and the adequacy of the methods applied to the data collected in forensic casework. The validation and accreditation of instrumental and automatic methods used for forensic analysis is well-studied and reflected in the scientific literature, and their harmonization and standardization are already in progress at the regional level, in Europe (European Network of Forensic Science Institutes, ENFSI [ENFSI14]), in the US (Scientific Working Group for Forensic Analysis of Chemical Terrorism/Threats SWGFACT [SWGFACT05]), in Australia and New Zealand (Australian and New Zealand Policing Advisory Agency and National Institute of Forensic Science, ANZPAA NIFS and Standards Australia [ST-AU12]), and soon globally (International Organization for Standardization, ISO [ISO21043]).

This Chapter addresses the validation of automatic likelihood ratio methods for forensic evidence evaluation.

1.1 Scope

In forensic evidence evaluation practitioners assign a strength of evidence to forensic observations and analytical results, in order to address hypotheses at source or activity level [Cook98]. This assignment is based on the practitioner's assessment and increasingly on the computations of automatic likelihood ratio (LR) methods.

This chapter focuses on the validation of automatic methods[1] developed to assign a strength of evidence at source level to the analytical results originating from the comparison of distinctive features of 2 specimens: a trace or mark of an unknown source and a reference specimen of a known source [Meuwly06, Robertson16]. Usually, a trace or mark is produced under the uncontrolled conditions of a criminal activity while a reference specimen is produced under controlled and more ideal conditions.

We will review some of the performance characteristics needed to accomplish any validation process, and we will give special attention to the calibration of likelihood ratios, because of its importance and its relative novelty in forensic interpretation. Throughout this chapter, we will follow a Bayesian interpretation of probability [Lindley06], and the recent guideline for evaluative reporting in forensic science in Europe [ENFSI15].

1.2 Aim

The main aim of this Chapter is to offer guidance to forensic practitioners assessing the scope of validity and applicability of automatic likelihood ratio methods, when implementing a new and non-standard method in forensic practice. These are essential steps towards the demonstration that such a method provides results that are fit for their intended use and allow it to be accredited and used in forensic casework. The validation and accreditation of automatic forensic evaluation methods serves several purposes. Primarily, it enables the demonstration of compliance with the quality standards adopted globally [ISO17025, ILACG19], specifically the way in which specimens are handled, what methods are used and how the results are interpreted. Beyond that, scientific and transparent validation of new and non-standard forensic methods favors their acceptance within the forensic community.

[1] Currently the validation of human-based interpretation methods focuses mainly on competence assessment. In the future it is desirable that the validation also addresses performance assessment, and the methods described in this chapter are also suitable for this purpose.

Accreditation enables the legal community to recognize the methods' merits and whether or not a method works reliably under forensic conditions.

Another aim of this chapter is to elaborate on the concept of calibration as a performance characteristic for likelihood ratios. We will justify its critical importance in the validation process.

1.3 Structure

This chapter is structured as follows. In Section 2 a review of the most important standards for validation is given, along with the concepts of performance characteristics, performance metrics and validation criteria, which constitute the validation process. The approach for the measurement of performance of the methods under validation is developed in Section 3. Section 4 and 5 describe the primary and secondary characteristics used to assess the performance of automatic forensic evaluation methods. The chapter ends with our conclusions in Section 6.

2 Validation process

2.1 Standardization

The ISO/IEC 17025:2017 standard [ISO17025] is used worldwide as one of the main bases for the accreditation of forensic service providers carrying out laboratory activities, while some more specifically forensic ISO standards are currently in development [Wilson18]. In its Clause 7.2.2.1 the ISO/IEC 17025:2017 standard specifies that non-standard methods, laboratory-developed methods and standard methods used outside their intended scope or otherwise modified need to be validated.

Likelihood ratio methods used for forensic evaluation can be considered as non-standard in two aspects. Firstly because they are laboratory-developed, and secondly because they address forensic evaluation from an automatic perspective, when forensic evaluation is generally only considered as an opinion formed by a practitioner.

In its Section 7.8.7.1 the ISO/IEC 17025:2017 standard specifies that:

> "*only personnel authorized for the expression of opinions and interpretations release the respective statement*",

considering this step exclusively as a human competence.

A similar approach had already been pursued in 2010 by the Dutch Accreditation Council in its explanation of the ISO/IEC 17025:2005 standard. In its Section 3.2, the criteria to assess the competence of laboratories to express opinions and interpretations are listed as follows:

> *"(1) examining the implementation of the procedures and practices, (2) examining the adequacy of the competence criteria, (3) verifying qualifications, experience, training and knowledge of personnel, (4) examining the adequacy of mechanisms in place to monitor the competence of personnel, (5) examining reports where opinions and interpretations have been expressed, (6) examining records showing the basis on which opinions and interpretations are based, (7) using other appropriate assessment techniques"*.

A similar approach is also pursued in Section 4.8.3 of the ILAC- G19:08/2014 document [ILACG19] "*Modules in a Forensic Science Process*":

> "*personnel interpreting results shall have been assessed and deemed competent before reporting statements including interpretation and opinions of results and findings*".

In its Section 3.10 it also specifies that

> "*interpretations of results and findings shall be based on robust studies and documented procedures*",

and in its Section 3.12 it states that

> "*where software is used it shall be demonstrated as being fit for purpose. This may be a verification check of the software functionality, for example, the use of a spreadsheet to calculate values, or could be as part of the more wide reaching validation of the forensic science process in which the software is used, for example, the use of databases for matching specific characteristics*".

But neither the ISO/IEC 17025:2017 standard nor the ILAC-G19:08/2014 document explicitly consider automatic interpretation methods for forensic evaluation, or the fact that

these methods require a validation based on their performance, just as instrumental analytical methods require validation.

2.2 Validation of theoretical and empirical aspects

Validation can address the theoretical or empirical aspects of the LR method. The validation of the theoretical aspects rests upon mathematical proof or falsification. The validation of the empirical aspects, on the other hand, rests upon the acceptance or rejection of validation criteria on the basis of experimental results. This requires the definition of a validation protocol and experiments, which are used to accept or reject the method's validity, based on the chosen validation criteria.

The theoretical validation is necessary and the literature regarding the theoretical grounds for using likelihood ratio methods for forensic evaluation is already abundant [Robertson16]. On the other hand, the empirical validation of automatic likelihood ratio methods is an emerging area, for which literature has been sparse to date [Haraksim15, Ramos17, Meuwly17]. Therefore, this chapter limits its focus to the empirical validation of automatic LR methods.

In essence, the approach for the empirical validation of automatic LR methods is analogous to the one described for the empirical validation of instrumental analytical methods in the ISO/IEC 17025:2017 standard. The aim is to establish the scope of validity of the method, and to determine the operational conditions under which it meets some performance requirements or validation criteria. In its Section 7.2.2, the ISO/IEC 17025:2017 standard mentions that validation can be one or a combination of measurements of several performance characteristics, such as:

1. the calibration or evaluation of bias, precision, a systematic assessment of the factors influencing the results;
2. the evaluation of the robustness for variation of controlled parameters;
3. the comparison of results achieved with other validated methods and the evaluation of measurement uncertainty of the results, based on an understanding of the theoretical principles of the method and practical experience with the method.

The note of its Section 7.2.2.3 also provides a definition of performance characteristics:

> *"performance characteristics can include, but are not limited to, measurement range, accuracy, measurement uncertainty of the results, limit of detection, limit of*

quantification, selectivity of the method, linearity, repeatability or reproducibility, robustness against external influences or cross-sensitivity against interference from the matrix of the sample or test object, and bias."

2.3 Performance characteristics for automatic LR methods

Currently, accuracy, discrimination, calibration, robustness, monotonicity[2] and generalization have been identified as relevant to the validation performance characteristics for the assessment of automatic likelihood ratio methods [Meuwly17]. Performance metrics and graphical representations are associated with each performance characteristic for the measurement and representation of the method's performance.

Accuracy, discrimination and calibration have been defined as primary performance characteristics, as they relate directly to performance metrics and focus on desirable properties of the LR methods. They address the required behavior of the automatic LR method if it is intended to be fit for purpose. In [Meuwly17] their selection is based on the statistics literature on the evaluation of Bayesian probabilities, and in particular on the use of proper scoring rules.

Robustness, monotonicity and generalization have been identified as secondary performance characteristics. They describe how the primary characteristics behave in different conditions representing the extreme variability of forensic casework. Factors of variability are usually degrading, as e.g. data sparsity, quality of the specimens or mismatch in the conditions between training data and operational data.

2.4 Empirical validation

Empirical validation is strictly necessary before making use of a new method in practice, because of the variability and often low quality of the operational data analyzed, which may cause sound LR models to present undesirable behavior. Among the most common degrading factors are: data sparsity, high variability of the quality of specimens, a shift between the conditions of the data used for LR model training and the data captured in the different forensic scenarios, etc.

---

[2] This was previously referred to as coherence [Haraksim15], but the name was changed for the sake of clarity, and in order to avoid confusion with statistical coherence.

As a central procedure of the validation process, performance measurement requires careful definition. In particular, the performance characteristics must guarantee that the likelihood ratios are fit for purpose, and that they have desirable properties under operational conditions.

Some definitions are given here for better understanding of the rest of the chapter[3]:

- A *performance characteristic* represents the answer to the question "*What to measure?*". It is a characteristic of an LR method that is thought to have an influence on the desired or undesired behavior of a given interpretation method. For example, we want LR values that help the trier of fact to reach better decisions, and in that sense the LR values should possess the performance characteristic defined as *accuracy*[4].
- A *performance metric* represents the answer to the question "*How to measure?*". It gives a quantitative measure of a performance characteristic, usually as a scalar. For the performance characteristic defined above as accuracy, the performance metric can be implemented by the use of proper scoring rules [deGroot82, Gneiting07a] on an empirical set of likelihood ratios (see Section 1.4.1). Thus, this performance metric will yield a single number that measures accuracy: the lower this number, the better the accuracy[5], and *vice versa*.
- A *validation criterion* represents the answer to the question "*what performance is needed to regard a method as valid?*". It is defined as the decision rule to determine when a method is acceptable and fit for purpose according to a given performance characteristic. For the performance metric *accuracy* defined above (empirical average of a proper scoring rule), a possible validation criterion is a scalar threshold over the performance metric. When the metric is above the threshold, the method is not validated from the point of view of the accuracy, and *vice versa*.

2.5 Validation protocol

The validation protocol begins with a validation plan describing the experiments. This plan lists the performance characteristics considered for validation of the method and the

[3] As explained in [Meuwly17], these terms have been defined to be, as much as possible, in accordance with relevant ISO standards.

[4] Here, it can be seen that we define *accuracy* in terms of proper scoring rules, in contrast to its usual definition. See Section 1.4.1.

[5] As we will see, the average of a proper scoring rule yields a penalty, which is lower when the accuracy is better.

performance metrics and graphical representations used to assess those performance characteristics. It also describes the aim of the experiments, the data used and the validation criteria applicable. In order to get more insight into the expected performance of the method, a comparison with either the current state of the art or with a baseline method can be performed, which provides an initial set of validation criteria.

Experiments are performed in two stages, the first entails the development and validation of the method and the second the validation for varying conditions. The development and validation of the method uses a training dataset (with a known *ground truth*) to select the automatic LR method, and to refine the parameters of this method and the statistical models involved in it. The aim is to measure the primary performance characteristics of the method and to obtain the best performance with the most representative dataset for the widest possible range of conditions.

The validation of the developed method for varying conditions consists in measuring its performance on a previously unseen set of data captured under forensic conditions (with a known ground truth), using both the primary and secondary performance characteristics. The aim is to test the automatic LR method under conditions that are as similar as possible to conditions in forensic casework, and to arrive at the validation decision. If a dataset is used to assign the value of some hyperparameter, which is often the case in the method development stage, then the same dataset should not be used to estimate the performance in the validation stage. The reason is to avoid a possible inadequate generalization to new data in casework (overfitting). The validation experiments in two stages are summarized in the flowchart shown in Figure 1.

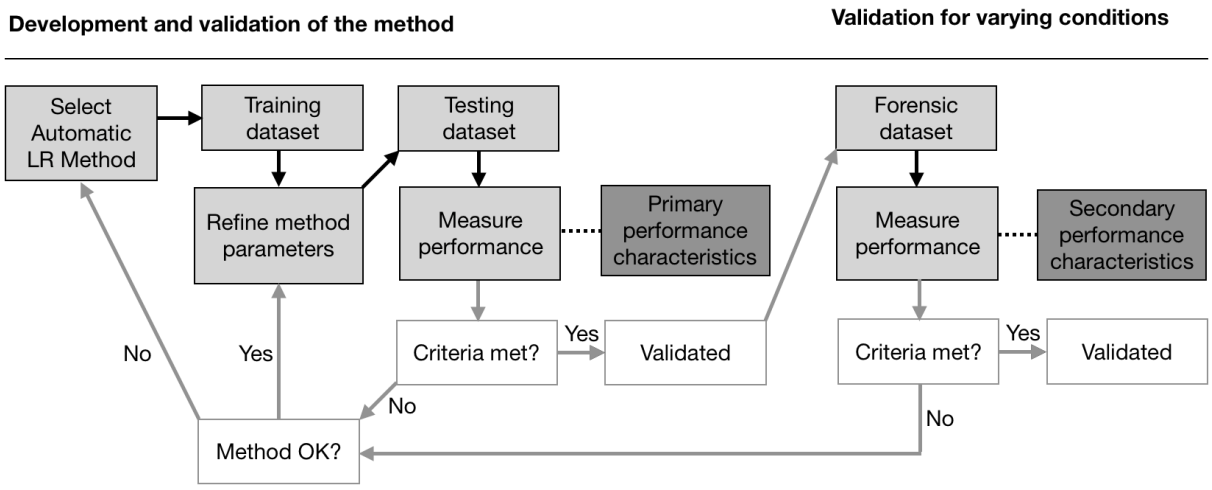

Figure 1. Diagram describing the development and validation stages of the validation process.

Table 1. Performance characteristics and examples of performance metrics and graphical representations.

Performance Characteristic	**Performance Metrics Examples**	**Graphical Representation Examples**
Accuracy	Empirical average of a proper scoring rule for a given prior probability, such as C_{llr}.	Prior-dependent representation of a proper scoring rule, such as an ECE plot.
Discrimination	Discrimination component of the empirical average of a proper scoring rule for a given prior probability, such as C_{llr}^{min} or EER.	Discrimination component of a prior-dependent representation of a proper scoring rule, such as an ECE^{min} plot or a DET plot.
Calibration	Calibration component of the empirical average of a proper scoring rule for a given prior probability, such as C_{llr}^{cal}.	Calibration component of a prior-dependent representation of a proper scoring rule, such as an ECE^{cal} plot. Also visible in the symmetry of a Tippett plot (i.e., cumulative histograms).
Robustness, Monotonicity, Generalization	Variation of primary metrics such as C_{llr} or EER, range of LR values.	Variation of primary representations such as ECE, Tippett or DET plots.

Table 2. Validation matrix for automatic likelihood ratio methods.

Performance Characteristic	Performance Metrics	Graphical Representation	Validation criteria	Experiments	Data	Results	Validation Decision
For each listed characteristic	As appropriate for characteristic	As appropriate for characteristic	According to the definition	Description of the experimental settings	Data used	+/- [%] compared to the baseline	Pass / Fail

Finally, the results of the validation experiments are summarized in a validation report, recording the decision of acceptance or rejection, depending on whether the experimental results meet the validation criteria or not. A validation decision should always be linked to a specific set of experimental conditions determining the scope of validity of the method.

The protocol for the validation of an automatic LR method is summarized in the validation matrix as shown in Table 2. Note that all the validation processes, seen as columns of the validation matrix in Table 2 apply to each of the performance characteristics (i.e., all the rows in Table 1). This might mean that a validation process could end with a "pass" validation decision for some characteristics, and with a "fail" validation decision for some others. To apply the method in casework (or not) will be the decision of the forensic science institute, but the validation report should be transparent and made public.

The guideline for validation proposed in [Meuwly17] is the first initiative in a long-term effort. It will be improved in the future, considering suggestions from others (see e.g. [Alberink17]).

An example of a validation report using development and forensic data can be found in [Ramos17]. It is linked to the necessary data used to reproduce the results, in the form of empirical sets of likelihood ratios with corresponding ground-truth labels. Interested researchers can access the data and follow the set of steps presented in this report, which can help them to proceed with the empirical validation of their own methods.

Moreover, a toolbox for performance assessment is available with the main tools necessary to generate the performance metrics and graphical representations needed to validate an LR method from an empirical set of LR values. This toolbox is freely available online [Perfevtoolbox].

3 Primary Performance Characteristics

The primary characteristics allow to define the minimum requirements that a LR method must satisfy empirically. This section begins with a description of how Bayesian probability theory has addressed this problem in other areas outside forensic science, leading to the concept of a *strictly proper scoring rule* as a function to assess the *accuracy* of probabilities. Then, the section describes the decomposition of the accuracy into discrimination and calibration, and their main properties.

The forensic evaluation of observations with regard to propositions is done by the practitioner assigning a likelihood ratio [ENFSI15], while the prior probability of propositions is the province of the trier of fact. However, the framework proposed to measure performance focuses on validation and is based on proper scoring rules, which apply to posterior probabilities and thus also depend on the prior probabilities. In a validation process, the forensic scientist must demonstrate that the LR method is valid for a wide range of prior probabilities. The validation process therefore involves testing the LR method in such a range of prior probabilities. To measure the performance of LRs, the associated posterior probabilities are assessed by using proper scoring rules.

It does not suggest that forensic evaluation and reporting should involve assigning prior probabilities, but that the LR method should be tested for a wide range of prior probabilities, as described in [Meuwly17, Ramos13b].

3.1 Performance of probabilities by proper scoring rules

The assessment of the goodness of probabilistic opinions has been the focus of extensive research in statistics since long. According to a widely accepted interpretation of probability from a Bayesian perspective, probability is personal [Lindley06], and therefore there is no such thing as a *true* probability or a *true* likelihood. One can assign a probability distribution for some uncertain observation, which might be different from the assignment made by someone else, but not necessarily better or worse. The Bayesian perspective also allows for subjective assignment of probabilities, as long as the rules of probability are respected, making it a coherent probability assignment.

Although this interpretation of probability proves to be flexible and useful, this does not mean that probabilities always lead to accurate actions when used to make decisions. For

instance: a person might have a gut feeling that it will not rain the next day, which motivates an assignment of a probability of 1% to the event "rain next day". In accordance with this probability value, it is most probable that this person will decide not to take an umbrella the next day. However, if it rains the next day, the decision not to take the umbrella is not a successful one. Nothing in this example has violated the laws of probability, nor coherence, nor the logic of decision-making, but the outcome can rationally make us question the earlier probability assessment.

This fact has motivated the assessment of the goodness of probabilities. In fact, the earliest works on this topic addressed the problem of probabilistic weather forecasting [Brier50], where an empirical criterion was proposed. Suppose a forecast on whether it will rain the next day or not is given by a forecaster every day. We denote $H \in \{H_r, H_{nr}\}$the random variable that represents one of two propositions, taking two categorical values depending on whether it rains the next day or not. For a given day d_i, the forecaster assigns a probabilistic forecast $P_i \equiv P(H = H_r|d_i)$. Then, the next day, the true value of the random variable H for day $d_i$will be known, and will be denoted the ground truth label for that day, as L_i, an observation drawn from random variable $L \in \{H_r, H_{nr}\}$. The empirical measurement of performance requires the availability of a database of past forecasts where the ground truth labels are known. Empirical measurement has been proposed in the past by the use of a so-called *proper scoring rule* (PSR), in the following way:

$$S = \frac{1}{m}\sum_{i=1}^{m} R(P_i, L_i) \tag{1}$$

where *S* is the *average PSR score* of the forecaster, and $R(P_i, L_i)$ is the strictly proper scoring rule that assigns a *penalty* to the forecast P_i depending on the value of L_i. Useful examples of proper scoring rules are the quadratic rule and the Brier rule [Brier50]:

$$R(P_i, L_i) = (1 - P_i)^2 \quad \text{if } L_i = H_r$$

$$R(P_i, L_i) = (P_i)^2 \quad \text{if } L_i = H_{nr}, \tag{2}$$

and also the logarithmic scoring rule:

$$R(P_i, L_i) = -log(P_i)^2 \quad \text{if } L_i = H_r$$

$$R(P_i, L_i) = -log(1 - P_i)^2 \quad \text{if } L_i = H_{nr}. \tag{3}$$

Figure 2 shows both examples of proper scoring rules. It can be seen that, if $L_i = H_r$, then it rained on day d_i, and therefore the proper scoring rule's penalty is lower if $P_i \equiv P(H = H_r|d_i)$ is closer to 1, and *vice-versa*. As a proper scoring rule is defined here as a penalty to a single probabilistic forecast, it penalizes forecasts P_i more when they are further from 1 while $L_i = H_r$, or further from 0 while $L_i = H_{nr}$. In other words, forecasters that will tend to assign probabilities that are closer to 1 when $L_i = H_r$and closer to 0 when $L_i = H_{nr}$will receive a better average PSR score.

This methodology to measure the performance of probabilistic forecasts also applies in the forensic evaluation context. We consider a trier of fact aiming to assign a posterior probability $P_i \equiv P(H = H_1|E)$, where H_1 is the proposition that associates the suspect with some trace at a crime scene, and E are the observations to be evaluated by the forensic examiner. As is recommended in [ENFSI15], the posterior probability is obtained from a prior probability, province of the trier of fact, and the likelihood ratio from the forensic examiner, using Bayes' theorem:

$$\frac{P(H_1|E)}{P(H_2|E)} = \frac{P(E|H_1)}{P(E|H_2)}\frac{P(H_1)}{P(H_2)} = LR\,\frac{P(H_1)}{P(H_2)}. \quad (4)$$

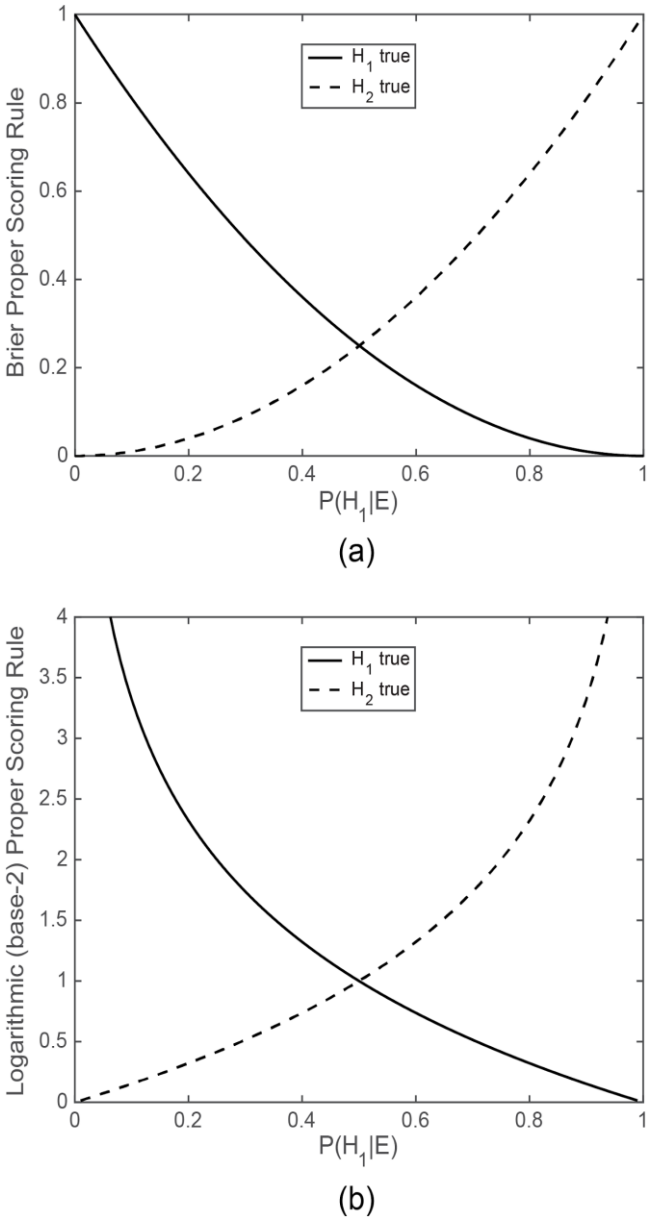

Figure 2. Examples of Proper Scoring Rules (PSR): (a) Brier, or quadratic, scoring rule. (b) Logarithmic scoring rule (base-2, although other bases could be considered, differing only by a scale factor).

If we have a database where the ground truth about the proposition for each of the observations in the database is known, we can replicate the weather forecasting performance measurement, and evaluate the posterior probabilities using proper scoring rules. The only relevant difference between the weather forecasting example and the forensic scenario is that, in the forensic evaluation scenario, we aim at measuring the performance of a method used by the forensic examiner, who only computes the likelihood ratio in Equation 4. In weather

forecasting, the database contains the forecasts, which are equivalent to posterior probabilities in forensic science. However, in the validation of forensic methods, we cannot have a database of posterior probabilities, because the prior probabilities in Equation 4 are generally not known, and in any case, they are not the responsibility of the forensic examiner. Thus, in forensic science the equivalent of the database with forecasts is a database of likelihood ratios.

We denote $H \in \{H_1, H_2\}$ the random variable, taking one of two categorical values H_1or H_2. For a given forensic case c_i, with findings E_i, the forensic examiner will use their LR method to obtain LR_i If the trier of fact would assign the prior probability, then the posterior probability $P_i \equiv P(H = H_1|E_i)$ could be obtained by Equation 4. In the database of LR values, the true value of the random variable H for each case c_i is known, and will be denoted the *ground-truth* label for that case, as $L_i \in \{H_1, H_2\}$. The empirical measure of performance for that database of cases, where posterior probabilities and ground-truth labels are known, will then be given by the average PSR score as in Equation 1.

In order to compute the value of the average PSR score in forensic evaluation, a prior must be fixed, because proper scoring rules apply to posterior probabilities. However, it is well-known that the forensic examiner has no role in assigning the prior, since it concerns evidence and information that falls outside their area of expertise. Thus, our proposed validation methodology only uses prior probabilities for measuring performance. The performance will be dependent on those prior probabilities, and indeed it is observed in practice that an LR method, which presents a better average PSR score for a given prior probability, could present a bad average PSR score for another prior probability. Therefore, the performance of the average PSR score should be measured for a wide range of prior probabilities, in order to guarantee that the LR method will perform adequately in a range of diverse realistic scenarios.

Some popular proper scoring rules have been introduced in recent literature. Among them, those based on the logarithmic scoring rule have gained popularity, more specifically the Empirical Cross-Entropy (ECE) [Ramos13a, Ramos18], and the log-likelihood ratio cost C_{llr} [Brummer06].

ECE is the average PSR score of the logarithmic strictly proper scoring rule, where the cases c_i, where H_1or H_2 is true, are combined with the information of the prior probabilities $P(H =$

$H_1)$ and $P(H = H_2)$, respectively. Thus, ECE is a measure of the average penalty that all the LRs in the set have. ECE is a prior-dependent measure, and therefore it must be represented as dependent on the prior probabilities in forensic science, since those are not the province of the forensic examiner. Therefore, ECE is better as it is lower, and the higher its value, the poorer the performance. C_{llr} is the value of ECE at the prior probabilities of 0.5, i.e. when $P(H = H_1) = P(H = H_2) = 0.5$, and it can be seen as a summarizing measure of ECE when the information about the propositions $H_1$and $H_2$is minimum (i.e., maximum prior uncertainty).

The ECE has been represented as a prior-dependent measure by means of ECE plots [Ramos13a, Ramos13b]. It therefore measures the performance of LR values as a function of the prior probabilities. The C_{llr} also has an interpretation as the average cost for all possible prior probabilities, and can be seen as a summarizing measure of ECE in terms of information theory [Brummer06, Ramos13a]. Both performance measures have a relevant information-theoretical interpretation and have been scientifically justified [Ramos18].

3.2 Discrimination and Calibration of Probabilities

An important property of proper scoring rules is that they allow a decomposition of the PSR score into two components: a refinement, discrimination, or sharpness component; and a calibration component [Murphy87, deGroot82]. Because this decomposition is additive, both of these components should be minimized to optimise the performance of the system. This performance is measured for an empirical set of posterior probabilities P_i, obtained from a database of likelihood ratios and a value of the prior probability where the performance is to be measured. In general, we will refer to this decomposition using the following notation:

$$R(P_i, L_i) = R^{min}(P_i, L_i) + R^{cal}(P_i, L_i), \tag{5}$$

where R^{min} is the discrimination component of the penalty according to the proper scoring rule, and R^{cal} its component due to calibration. This decomposition can always be considered theoretically, and in some particular cases it allows a straightforward closed-form expression (see, for example, [deGroot82]). However, this is not the case in general and some algorithmic methods have been proposed in recent literature to separate both components, the most relevant one being the Pool Adjacent Violators (PAV) algorithm [Brummer06, Fawcett07].

In a forensic evaluation using a Bayesian decision framework, posterior probabilities must differentiate between different propositions, in the sense that the observations should lead to the true proposition by increasing its probability. This property of probabilities has been dubbed as *refinement* [Murphy87, deGroot82], *sharpness* [Gneiting07b] or discrimination [Brummer06, Ramos13b]. Roughly speaking, discrimination is the property that allows to separate the sets of posterior probabilities that are obtained when one or the other proposition in the case is true. In other words, with findings E for experiments where either H_1 or H_2 is true, $P(H_1|E)$ should be higher when H_1 is true than where H_2 is true.

Thus, if we compute posterior probabilities for many cases in our development or validation database (i.e., by choosing a range of prior probabilities in our experimental set-up), the values of $P(H_1|E)$ when H_1 is true should overlap as little as possible with the values of $P(H_1|E)$ when H_2 is true. It is this relative overlap between probabilities which defines the discrimination. Popular measures of discrimination of probabilities are the Area under the Receiver Operating Characteristic curve (Area Under ROC, or AUC) [Fawcett07], the Equal Error Rate [Martin97] and the C_{llr}^{min} [Brummer06]. Their corresponding graphical representations are Receiver Operating Characteristic (ROC) curves, Detection Error Tradeoff (DET) curves and minimum Empirical Cross-Entropy (ECE^{min}) curves.

But discrimination is not enough: posterior probabilities must also be *reliable* [Ramos13b], in the sense that triers of fact can rely on them to improve their decisions on average. This means they represent the findings that are being evaluated, and the prior probability for which the performance is evaluated. However, in this Bayesian context, *reliability* has a different meaning than the classical, frequentist one: in Bayesian statistics, as mentioned earlier, probability is personal [Lindley06]. Thus, reliability does not measure closeness to a given *true* probability distribution. The most common property that Bayesian statisticians have attributed to reliable probabilities is their *calibration* [Dawid82, deGroot82], and calibration has been used in many works as a synonym of reliability [Dawid82]. According to these works, posterior probabilities should not only be more discriminating, but also more reliable (in the sense of better-calibrated). If this is the case, probabilities lead the trier of fact to make better decisions on average [Ramos13b].

Calibration can be defined in many ways, but two definitions have been commonly accepted. The first definition is perhaps the most theoretical, and the most general: a method that assigns posterior probabilities is *perfectly calibrated* if, when a posterior probability is

assigned using this method, and the method computes a posterior probability again once the previous posterior probability has been observed, it will yield the same posterior probability [vanLeeuwen13]. This has the immediate implication that, if a posterior probability is perfectly calibrated, it will perfectly represent the probability of the propositions given the findings, and allows to arrive at better decisions [Robertson16].

Another definition of calibration is more empirical, informal and practical. Imagine an empirical set of posterior probabilities $P(H_1|E)$, with many cases when H_1 and H_2 are respectively true. Imagine also a subset of posterior probabilities that lay *close* to a value k. Then, k' is computed as the proportion of those probabilities in the subset where H_1 is actually true. The closer k is to k' for all values of k, the better the calibration. For instance, a weather forecaster will have a proportion $k'$of days with rain for all the days where s/he gave the posterior probability k of rain forecast. This definition motivates the use of performance representations like the so-called empirical calibration curves [Zadrozny02, Cohen04], also known as reliability plots, where the values of k' (proportion of cases) are represented as a function of k (probabilities assigned in the empirical set) for a binning of the latter. In these representations, calibration will be better if the curve is closer to the diagonal of the plot. See, e. g., Figure 3(d).

Of course, it follows from Equation 5 that perfect calibration does not imply certainty about the propositions, because the term due to discrimination still remains in the proper scoring rule [Brummer06]. But if the discrimination remains the same, then improving the calibration improves performance.

Following this approach, the performance of probabilities is measured as the average penalty, i.e., the empirical average of the proper scoring rule penalties [deGroot82]. Thus, the improvement in the accuracy of the decisions made by the trier of fact by a reduction of the proper scoring rule penalty may not manifest for a single case, where a single evaluation of findings is performed; but it will on average over a large amount of cases. In fact, discrimination and calibration do not apply to single probabilities, or to single cases, but to averages over empirical sets.

3.3 Performance of likelihood ratios

As mentioned earlier, the validation process is aimed at the likelihood ratios given by a forensic method, not at posterior probabilities. The reason is that a forensic examiner cannot assign the prior probability, leaving that task to the trier of fact. However, LR methods that are going to be validated for their use in casework must be validated for a wide range of prior probabilities. This is because an LR method might pass the validation criterion for some prior probabilities, but not for others. There are several solutions to this issue, from the point of view of the validation process:

- The performance of likelihood ratios can be measured for a wide range of prior probabilities. Thus, LR methods that perform according to their validation criterion for the whole range of analyzed prior probabilities will be validated. This is done using prior-dependent performance representations of likelihood ratios, such as ECE plots (and the corresponding calibration and discrimination components, ECE^{min} and ECE^{cal}), ROC curves and DET plots.
- The performance of likelihood ratios can also be measured for a single, summarizing prior probability; or as an average over prior probabilities. Thus, those methods will give a single scalar measure that summarizes the performance for all the prior probabilities. This is done using e.g. the C_{llr} (with its corresponding discrimination and calibration components C_{llr}^{min}d C_{llr}^{cal}), the EER and the AUC.

The properties discussed above for probabilities easily extend to LRs, since the posterior probability is simply the product of the LR and a (fixed) prior probability. Therefore, the LR and the posterior probabilities share the same information if the prior is fixed. This can be proved using *e. g.* theorems of information theory [Cover06] related to the so-called *data-processing inequality*. Thus, an empirical set of LRs also presents the two properties determining the performance of a set of posterior probabilities: discrimination and calibration.

3.4 Properties of well-calibrated likelihood ratios

Calculating the LRs without mistake is necessary but, perhaps surprisingly, not sufficient for the numbers obtained to have the properties LRs should have. When used to update the prior probability ratio, on average LRs have to decrease the uncertainty about the hypotheses. If

this property is observed on large set of LRs, then the method is deemed to produce 'well-calibrated' LRs. A method producing 'well-calibrated' LRs would help, on average, the trier of fact to optimally reach a decision on the issue in a case. Numbers presented as LRs - from here on 'reported LRs' - can have any degree of performance. Reported LRs are not always helpful for the trier of fact, whereas calibrated LRs are, on average.

The above should make clear that anyone considering using LRs should also be interested in the performance of those LRs. This not only shows whether a trier of fact can expect to benefit from the system at all, it can also help the forensic practitioner choose between different systems. For forensic scientists that attempt to improve a system's performance, it can help to assess the improvement achieved [Berger12].

The performance of an automatic system that generates LRs is limited for a variety of reasons. It can be due to blatant mistakes, but more often the laws of probability theory are followed and other factors are limiting performance. Limiting factors can be e.g. modelling assumptions, or databases that are never perfectly representing the population of interest because of their size or nature. For example, the database may be of studio recordings of speech but the system may be used to compare voices on tapped phone calls.

LRs have a specific probabilistic interpretation: they represent strength of evidence. Apart from discrimination (the low degree of overlap of LRs under either hypothesis), the values of the LRs themselves are important. We can consider the LR reported by a system (or person), as evidence E for our own evaluation. This evidence is interpreted, as always, by considering the probability of obtaining the evidence if either hypothesis is true:

$$LR = \frac{P(E|H_1)}{P(E|H_2)} = \frac{P(reported\ LR|H_1)}{P(reported\ LR|H_2)}. \quad (6)$$

If the reported LR is equal to LR, the result of our own trusted evaluation, then the calibration of the system is ideal. Ideal calibration means that our assessment of the evidential value of the reported LR agrees with the assessment of the evidence by the system. A system with ideal calibration thus reports LRs that fully capture the evidential strength of the observation, such that this original observation can be replaced by its LR. This means that the interpretation of the original observation gives the same result as the interpretation of the reported LR. Or, in other words, for ideal calibration the LR of the reported LR is equal to the

reported LR. But if the calibration is not ideal, the reported LRs will be misleading more often than better-calibrated LRs [Dawid82, Cohen04, Brummer06, Ramos13a, Ramos18].

An empirical observation is that, if the discrimination of an empirical set of LR values increases, the strength of evidence of well-calibrated LRs also tends to increase [Ramos13b]. This agrees with common sense: an LR method with better discrimination (for instance, DNA analysis) should yield higher strength-of-evidence than an LR method with lower discrimination (for instance, forensic voice comparison). This happens if the set of LR values is well-calibrated, but not necessarily otherwise. In [vanLeeuwen13], a proof of this property is given for equal-variance Gaussian distributions of the scores of speaker recognition systems.

3.5 Examples with primary performance characteristics

In this section several examples are given of performance metrics and graphical representations related to primary performance characteristics. The following performance metrics will be used for 3 sets of LR values produced by 3 simulated systems:

- Accuracy: C_{llr} [Brummer06], ECE [Ramos13a, Ramos18].
- Discrimination: C_{llr}^{min}rummer06], ECE^{min} [Ramos13a], AUC [Fawcett07], EER [Martin97].
- Calibration: C_{llr}^{cal} [Brummer06], ECE^{cal} [Ramos13a, Ramos18].

Also, the following graphical representations are shown:

- Accuracy: ECE plot [Ramos13a, Ramos18].
- Discrimination: ROC curve [Fawcett07], DET curve [Martin97].
- Calibration: Empirical calibration plot [Zadrozny02].

It is out of the scope of this Chapter to give a thorough interpretation of these performance measures, and interested readers can find further details in the indicated references. However, a brief description of their interpretation is given in relation with performance.

The histograms in Figure 3(a) show the explicit empirical distribution of log(LR) values for either hypothesis being true. These graphical representations do not explicitly measure performance, but show the degree of overlap as a measure of discrimination. They also give some indication of the calibration of the log(LR) values, since they should be centered around

log(LR) = 0 if they are well-calibrated [vanLeeuwen13]. The cumulative version of these histograms is the Tippett plot [Meuwly00, Lucena-Molina15].

The DET and ROC graphs (Figures 3b and 3c respectively) are measures of discrimination. The closer the curves are to the top-left corner of the graph for ROC curves, and to the origin for the DET curves, the better the discrimination. The AUC is a summarizing measure of discrimination that integrates the whole ROC curve, the higher the better. Its equivalent is the C_{llr}^{min}[Brummer06, Fawcett07], for which lower values represent better discrimination. The EER is a point-summary of the discriminating power, the lower the better. However, the calibration is not visible in these measures, as can be seen from the fact that Set 1 and Set 3 present similar DET and ROC curves (Figure 3b, c), but very different calibration (Figure 4).

The empirical calibration graph in Figure 3(d) gives a measure of calibration based on its empirical definition. It shows the relation between the proportion of cases where H_1 is true and the posterior probability of H_1, and takes the proportion of cases where H_1 is true as prior probability (named *empirical prior*). The set of LR values is well-calibrated when the data points approach the diagonal. This representation does not take into account discrimination. The example shows that Set 2 has the best-calibrated set of LR values, when the DET and ROC graphs show that it is also the least discriminating (Figure 3b, c). This example also shows that discrimination and calibration are two essential and complementary performance characteristics for the validation of LR methods.

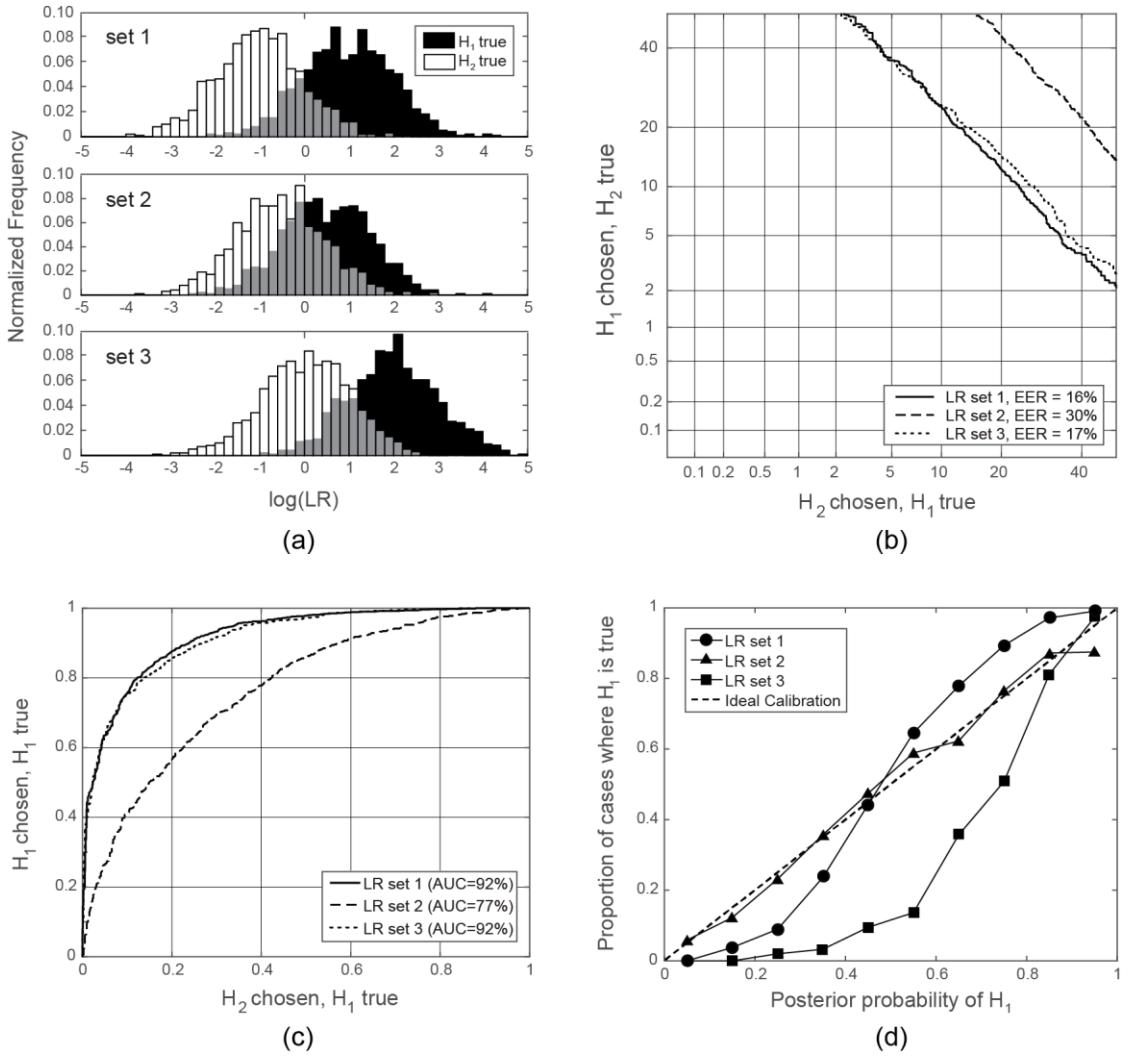

Figure 3. Some popular and useful performance metrics and graphical representations: (a) Histograms, (b) DET plots with EER, (c) ROC curves with AUC, (d) empirical calibration plots.

The ECE plots in Figure 4 give the performance as a function of prior probabilities. The accuracy (solid curve, ECE) is plotted with the loss of accuracy due to imperfect discrimination (dashed curve, ECE^{min}) and to imperfect calibration (difference between both, ECE^{cal}). The lower the corresponding ECE curve the better the performance. The plots also show the ECE of a non-informative system, such as the tossing of a coin (short-dash curve). They demonstrate that the LR method can produce results better than the non-informative system in some ranges of prior probability, while worse in other ranges. The scope of validity of the method increases with the range for which the method performs better than the level of chance. Finally, the values of C_{llr}, C_{llr}^{min} and C_{llr}^{cal} are the ECE values at the prior $\log_{10}$ (odds) value of 0 (i.e., *y*-axis).

These performance metrics and graphical representations are examples of the many that could be included in this validation framework based on proper scoring rules [Meuwly17].

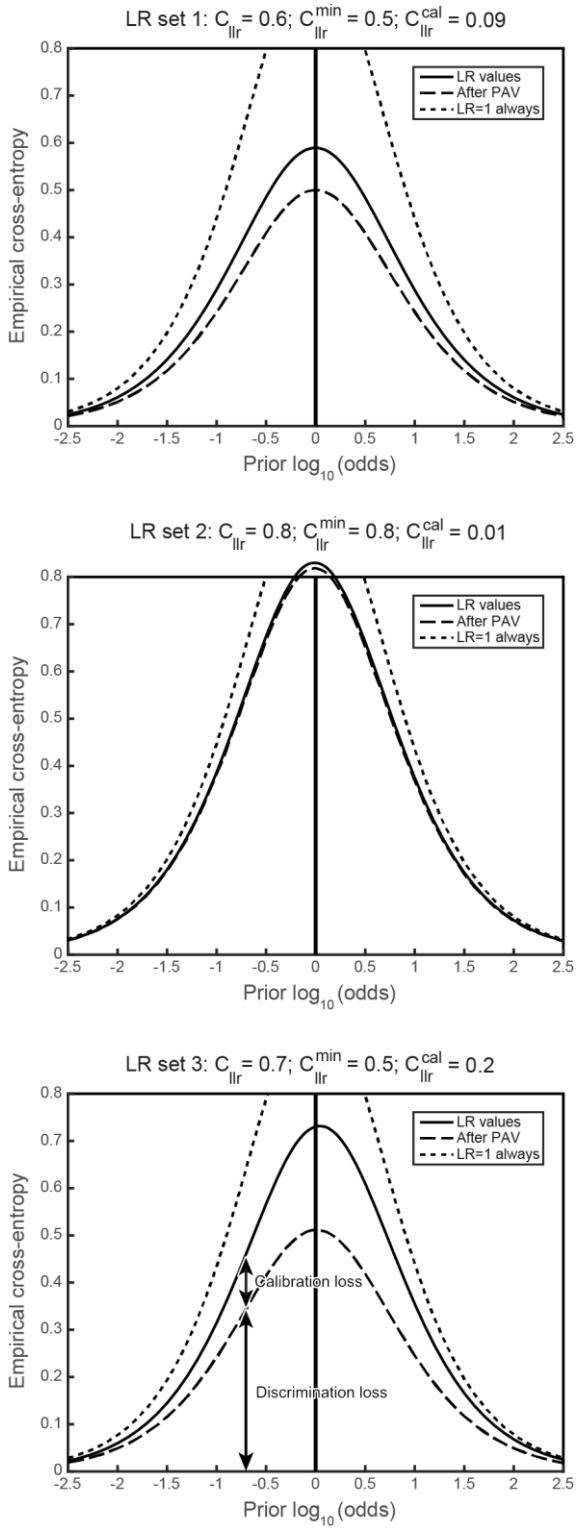

Figure 4. Empirical Cross-Entropy (ECE) plots with C_{llr}, C_{llr}^{min}d C_{llr}^{cal}, based on the same sets of LR values as Figure 3. The "After PAV" curve label refers to the ECE^{min} curve, since ECE^{min} is the ECE obtained after the application of the PAV algorithm.

4 Secondary Performance Characteristics

The secondary characteristics describe how the primary metrics behave in different situations, such as typical forensic casework conditions (e.g. differing quality of the training data and the trace material). The aim of secondary performance characteristics is to assess the performance of the LR method in forensic casework. Therefore, secondary performance characteristics are mainly assessed at the stage of validation for varying conditions (see Figure 1). However, if necessary or if possible, they could also be used in the stage of development and validation of the method.

The secondary performance characteristics are related to a single primary performance metric or a single graphical representation. Thus, we talk about e.g. the robustness of the accuracy, of the discriminating power or of the calibration.

We define the proposed secondary performance characteristics LR-based forensic evaluation methods as follows:

4.1 Robustness

The robustness of an LR method is the ability of the method to maintain the value of a primary performance metric when the data changes. For instance, Method A is more robust to a lack of data than Method B if, as the data gets more sparse, the primary performance metric of Method A degrades less than the same metric for Method B. In the LR context, robustness usually refers to the stability of the performance of LR methods to varying conditions (e.g. quality/quantity of the data).

4.2 Monotonicity

The monotonicity[6] of an LR method is defined as the ability of the method to yield LR values with better performance when increasing the intrinsic quantity/quality of the data. Examples are the number of minutiae in a fingermark or the signal-to-noise ratio in a voice recording.

4.3 Generalization

[6] In [Meuwly17, Haraksim15, Ramos17], this property has been named *coherence*. However, it has been decided to change its name in order to not confound it with the statistical coherence of subjective probabilities.

The generalization of an LR method is the ability of the method to maintain its performance for previously unseen data (even when the quality/quantity of the data is the same). An LR system for speaker identification for example, could be used for recordings of the same quality and quantity but in a different language.

5 Conclusion

Many automatic methods have been developed to compute the strength of evidence, particularly for the forensic evaluation of genetic and biometric traces. But there is currently no standard for the validation of such forensic evaluation methods. This book chapter summarizes the first steps taken in the direction of the validation of forensic automatic likelihood ratio methods. Many more steps, and the involvement of the forensic community are necessary to further develop this multidisciplinary and complex matter.

6 References

[Alberink17] I. Alberink, A. Bolck, M. Sjerps, P. Vergeer. “Comment to ‘A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation’”. Forensic Science International 276, p. 154, 2017. DOI: https://doi.org/10.1016/j.forsciint.2017.03.011.

[Berger12] C.E.H. Berger, D. Ramos. “Objective paper structure comparison: Assessing comparison algorithms”. Forensic Science International 222(1-3), pp. 360-367, 2012. DOI: https://doi.org/10.1016/j.forsciint.2012.07.018.

[Brier50] G. W. Brier. “Verification of forecasts expressed in terms of probabilities”. Monthly Weather Review 78(1), pp. 1-3, 1950. DOI: https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[Brummer06] N. Brümmer and J. du Preez. “Application-independent evaluation of speaker detection”. Computer Speech & Language, 20 (2–3), pp. 230-275, 2006. DOI: https://doi.org/10.1016/j.csl.2005.08.001.

[Cohen04] I. Cohen and M. Goldszmidt. “Properties and benefits of calibrated classifiers”. PKDD 2004: Knowledge Discovery in Databases: PKDD 2004 (LNCS, volume 3202), Springer, pp 125-136, 2004. DOI: https://doi.org/10.1007/978-3-540-30116-5_14.

[Cook98] R. Cook, I.W. Evett, G. Jackson, P.J. Jones, J.A. Lambert. A hierarchy of propositions: deciding which level to address in casework. Science & Justice. 1998 Oct 1;38(4):231-9.

[Cover06] T. M. Cover and J. A. Thomas. “Elements of Information Theory”. John Wiley and Sons, 2006.

[Dawid82] A.P. Dawid. “The well-calibrated Bayesian”. Journal of the American Statistical Association 77 (379), pp. 605-610, 1982. DOI: http://doi.org/10.2307/2287720.

[deGroot82] M.H. DeGroot and S.E. Fienberg. “The Comparison and Evaluation of Forecasters”. Journal of the Royal Statistical Society. Series D (The Statistician) 32 (1-2), 1982. DOI: http://doi.org/10.2307/2987588.

[ENFSI14] T. De Baere, W. Dmitruk, B. Magnusson, D. Meuwly and G. O’Donnel, Guideline for the single laboratory – Validation of Instrumental and Human Based Methods in Forensic Science, ENFSI, 2014.

[ENFSI15] ENFSI, “ENFSI Guideline for Evaluative Reporting in Forensic Science”. 2015. Available at http://enfsi.eu/news/enfsi-guideline-evaluative-reporting-forensic-science/ (last access: February 2018).

[Fawcett07] T. Fawcett and A. Niculescu-Mizil. “PAV and the ROC convex hull”. Machine Learning 68(1), pp. 97-106, 2007. DOI: https://doi.org/10.1007/s10994-007-5011-0.

[ForSciReg18] Forensic Science Regulator annual report 2017. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/674761/FSRAnnual_Report_2017_v1_01.pdf

[Gneiting07a] T. Gneiting and A.E. Raftery. “Strictly proper scoring rules, prediction, and estimation”. Journal of the American Statistical Association 102 (477), pp. 359-378, 2007. DOI: https://doi.org/10.1198/016214506000001437.

[Gneiting07b] T. Gneiting, F. Balabdaoui, A.E. Raftery “Probabilistic forecasts, calibration and sharpness”. Journal of the Royal Statistical Society: Series B 69 (2), pp. 243-268, 2007. DOI: http://doi.org/10.1111/j.1467-9868.2007.00587.x.

[Haraksim15] R. Haraksim, D. Ramos, D. Meuwly, C.E.H. Berger. “Measuring coherence of computer-assisted likelihood ratio methods”. Forensic Science International 249, pp. 123-132, 2015. DOI: https://doi.org/10.1016/j.forsciint.2015.01.033.

[ISO17025] ISO/IEC standard 17025:2017, General requirements for the competence of testing and calibration laboratories, Third edition 2017-11.

[ISO21043] ISO standard 21043, Forensic sciences, draft.

[ISO22842] ISO/IEC AWI 22842, Validation of automatic biometric methods for forensic purposes, Draft.

[ILACG19] International Laboratory Accreditation Cooperation G19:08/2014, Modules in a Forensic Science Process.

[Lindley06] D. Lindley. “Understanding Uncertainty”. John Wiley and Sons, 2013. DOI: http://doi.org/10.1002/0470055480.

[Lucena-Molina15] J-J. Lucena-Molina, D. Ramos and J. Gonzalez-Rodriguez. “Performance of likelihood ratios considering bounds on the probability of observing misleading evidence”. Law, Probability and Risk 14(3), pp. 175-192, 2015. DOI: http://dx.doi.org/10.1093/lpr/mgu022.

[Martin97] A. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocki. “The DET curve in assessment of detection task performance”. In proceedings of Eurospeech, pp. 1895-1898, 1997.

[Meuwly00] D. Meuwly, Reconnaissance de Locuteurs en Sciences Forensiques: L’apport d’une Approche Automatique, (PhD thesis), 2000.

[Meuwly06] D. Meuwly “Forensic individualisation from biometric data. Science and Justice”, 2006 Oct 1;46:205-13.

[Meuwly17] D. Meuwly, D. Ramos and R. Haraksim. “A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation”, Forensic Science International 276, pp. 142-153, 2017. DOI: https://doi.org/10.1016/j.forsciint.2016.03.048.

[Murphy87] A.H. Murphy and R.L. Winkler. “A General Framework for Forecast Verification”. Monthly Weather Review 115, pp. 1330-1338, 1987. https://doi.org/10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2.

[PCAST16] President’s Council of Advisors on Science and Technology, Report to the president Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, Washington DC, 2016. https://obamawhite-house.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_foren-sic_science_report_final.pdf.

[Perfevtoolbox] https://sites.google.com/site/perfevtoolbox/

[Ramos13a] Daniel Ramos, Joaquin Gonzalez-Rodriguez, Grzegorz Zadora and Colin G. G. Aitken. “Information-Theoretical Assessment of the Performance of Likelihood Ratio Models”. Journal of Forensic Sciences 58 (6), pp. 1503-1518, 2013. DOI: http://dx.doi.org/10.1111/1556-4029.12233.

[Ramos13b] D. Ramos and J. Gonzalez-Rodriguez. “Reliable Support: Measuring Calibration of Likelihood Ratios”. Forensic Science International 230, pp. 156-169, 2013. DOI: http://dx.doi.org/10.1016/j.forsciint.2013.04.014.

[Ramos17] D. Ramos, R. Haraksim and D. Meuwly. “Likelihood ratio data to report the validation of a forensic fingerprint evaluation method”. Data in Brief 10, pp. 75-92, 2017.

[Ramos18] D. Ramos, J. Franco-Pedroso, A. Lozano-Diez and J. Gonzalez-Rodriguez. “Deconstructing Cross-Entropy for Probabilistic Binary Classifiers”. Entropy 20(3), pp. 208, 2018.

[Robertson16] B. Robertson, G.A. Vignaux and C.E.H. Berger. “Interpreting Evidence: Evaluating Forensic Science in the Courtroom”, 2nd edition. John Wiley and Sons, 2016. DOI: http://doi.org/10.1002/9781118492475.

[ST-AU12] Australian Standard AS 5388 (2012) Forensic analysis.

[SWGFACT05] Standards and Guidelines, Validation Guidelines for Laboratories Performing Forensic Analysis of Chemical Terrorism. FBI Law enforcement bulletin, 2005. 7(2).

[vanLeeuwen13] D.A. van Leeuwen and N. Brümmer. “The distribution of calibrated likelihood-ratios in speaker recognition”. In Proceedings of INTERSPEECH-2013, pp. 1619-1623, 2013.

[Wash15] S.S. Hsu, FBI admits flaws in hair analysis over decades, Washington Post, April 18, 2015.

[Wilson18] L. Wilson-Wilde, The International Development of Forensic Science Standards – A Review, Forensic Science International, In Press.

[Zadrozny02] B. Zadrozny and C. Elkan. “Transforming classifier scores into accurate multiclass probability estimates”. In Proceedings of KDD’02, the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 649-699, 2002.