Forensic Science International 276 (2017) 142–153

Contents lists available at ScienceDirect

Forensic Science International

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

A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation

Didier Meuwly [a,b,*], Daniel Ramos [c], Rudolf Haraksim [d]

[a] Netherlands Forensic Institute, Laan van Ypenburg 6, 2497GB The Hague, The Netherlands
[b] University of Twente, Drienerlolaan 5, 7522NB Enschede, The Netherlands
[c] ATVS – Biometric Recognition Group, Escuela Politecnica Superior, Universidad Autonoma de Madrid, C/Francisco Tomas y Valiente 11, 28049 Madrid, Spain
[d] LTS5 – Signal Processing Laboratory, École Polytechnique Fédérale de Lausanne, Faculty of Electrical Engineering, Station 11, CH-1015 Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 4 June 2015
Received in revised form 23 March 2016
Accepted 24 March 2016
Available online 26 April 2016

Keywords:
Method validation
Automatic interpretation method
Strength of evidence
Accreditation
Validation report

ABSTRACT

This Guideline proposes a protocol for the validation of forensic evaluation methods at the source level, using the Likelihood Ratio framework as defined within the Bayes' inference model. In the context of the inference of identity of source, the Likelihood Ratio is used to evaluate the strength of the evidence for a trace specimen, e.g. a fingermark, and a reference specimen, e.g. a fingerprint, to originate from common or different sources.

Some theoretical aspects of probabilities necessary for this Guideline were discussed prior to its elaboration, which started after a workshop of forensic researchers and practitioners involved in this topic. In the workshop, the following questions were addressed: *"which aspects of a forensic evaluation scenario need to be validated?"*, *"what is the role of the LR as part of a decision process?"* and *"how to deal with uncertainty in the LR calculation?"*. The questions: *"what to validate?"* focuses on the validation methods and criteria and *"how to validate?"* deals with the implementation of the validation protocol.

Answers to these questions were deemed necessary with several objectives. First, concepts typical for validation standards [1], such as *performance characteristics*, *performance metrics* and *validation criteria*, will be adapted or applied by analogy to the LR framework. Second, a validation strategy will be defined. Third, validation methods will be described. Finally, a validation protocol and an example of validation report will be proposed, which can be applied to the forensic fields developing and validating LR methods for the evaluation of the strength of evidence at source level under the following propositions:
H_1/Hss: The trace and reference originate from the same source.
H_2/Hds: The trace and reference originate from different sources.

1. Introduction

1.1. Preliminary considerations

This Guideline aims at providing assistance to the forensic practitioners in determining the scope of validity[1] and applicability of the LR methods developed and to validate the LR's produced as forensic evidence in practice. Even though the empirical examples given (taken over from forensic fingerprints) are shaped around the LRs computed from scores of a biometric system (namely *score-based LRs*), the Guideline proposed is general and can be applied to any forensic method producing LR values, whether it is biometric or not, and whether it is score-based or feature-based.

It is worth noting, that there is an on-going discussion in the forensic community regarding issues related to the concepts of probability and of the Likelihood ratio (LR). Especially concerning is the concept of uncertainty of computed LRs, which leads to different methods for the measurement of performance of LRs methods, which may not necessarily be compatible. This has direct consequences on the definition of the criteria for the validation of computer-assisted LR methods developed for forensic evaluation. Therefore, the points of view regarding the concepts of probability and of the LR will be discussed prior to the introduction of the performance characteristics and criteria.

* Corresponding author at: Netherlands Forensic Institute, Laan van Ypenburg 6, 2497GB The Hague, The Netherlands. Tel.: +31 708886344.
E-mail addresses: d.meuwly@nfi.minvenj.nl, d.meuwly@utwente.nl (D. Meuwly).

[1] The scope of validity is to be understood as the range of conditions for which the method has been tested.

http://dx.doi.org/10.1016/j.forsciint.2016.03.048
0379-0738/

1.2. Definitions

In the context of the interpretation of the evidence by LR values we understand validation as the process followed in order to determine the scope of validity of a method used to compute LR values. The latter means that we allow the method to be used in forensic casework in the future.

Here, we define important concepts that are typical in validation strategies in other contexts. Later our definitions adapt to the LR methodology.

- A *performance characteristic* is a characteristic of a LR method that is thought to have an influence in the validation of a given method. For instance, LR values should be discriminating in order to be valid, provide clear distinction between comparisons under different hypotheses. In this case, discriminating power is a performance characteristic.
- A *performance metric* is a variable whose numerical or categorical value measures a performance characteristic. For instance, the minimum log-likelihood ratio cost ($minC_{llr}$) can be interpreted as a measure of discriminating power, and therefore it can be used as a performance metric of the discriminating power.
- A *validation criterion* presents a condition related to the performance characteristics that has to be met as a necessary condition for the LR method to be deemed as valid. For instance, a validation criterion can be formulated as follows: *only methods producing rates of misleading evidence smaller than 1% can be considered as valid*. Note that a single validation criterion is not sufficient in general, and therefore several validation criteria might be necessary in order to determine the validity of the method.

2. Computation of likelihood ratios for forensic evaluation

Many different methods have been described in the literature to compute LR values [2–7], feature-based [3,4] and score-based [5,7–11]. This Guideline considers both classes of LR methods, score and feature-based, and an example of comparison of these methods can be found in [5].

In a score-based method illustrated in Fig. 1, the LR values are calculated from the comparison scores [7,10], which are typically the result of a comparison performed by pattern-recognition algorithms. These extract and compare the features of trace (T) and reference (R) specimens. The score (E) resulting from this comparison is used to compute a likelihood ratio with the LR method (Bayes' inference model), using a dataset of trace specimens (DB Traces) and a dataset of reference specimens (DB References). Score-based approaches are traditionally used in forensic biometrics and a typical example can be found in [2].

Feature-based LR methods illustrated in Fig. 2 exploit directly the features of the specimens in comparison and produce a LR value without the previous computation of a comparison score. Several examples of feature-based LR methods are described in [3–5]. These methods involve statistical modeling at the level of the features, using for example probability density functions for either of the propositions to produce the LR values. Feature-based approaches are traditionally used in forensic chemistry and examples can be found in [5,12,13].

2.1. The LR as part of the forensic evaluation process

Forensic research makes progress in the field of evaluation of forensic evidence. Currently, a uniform and logical inference model is used for evaluating and reporting forensic evidence [14]. It uses a likelihood ratio (LR) approach based on the Bayes inference model (Theorem of conditional probabilities). Standards and Guidelines have been proposed for the formulation of evaluative forensic science expert opinion first in UK by the Association of Forensic Science Providers (AFSP) [15] and then in Europe, within the European Network of Forensic Science Institutes (ENFSI) [16].

The LR methods are extensively used, for example, for the interpretation of DNA profiles. Some recommendations on the interpretation of the DNA mixtures have been issued in 2006 [17]:

R1: "LR is the preferred approach to (DNA) mixture interpretation".
R2: "Even if the legal system does not implicitly appear to support the use of the likelihood ratio, it is recommended that the scientist is trained in the methodology and routinely uses it in case notes".

Even though this Guideline does not use examples from the DNA, we endorse and follow these recommendations, because the logic of the inference model remains, independently of the type of traces considered [18].

Computer-assisted methods have been developed to compute LRs, assisting forensic practitioners in their role of forensic evaluators to perform inferences at source level [19]. Very early principles for using the LR approach in forensic evaluation can be found in the analysis of glass microtraces [3]. It has also been used in forensic fields focusing on human individualization, such as fingermark [20,21], earmark [22], speaker recognition [7,23] and hair [24]; or object individualization such as toolmarks [25], fibre [26] and glass microtraces [3,6,12,13,27] (which represents a very early practical example of the use of the LR approach). But the LR approach has been firstly implemented in a casework process as a standard for the evaluation of DNA profiles [14] and several computer-assisted methods are being developed and validated to assess the value of DNA mixture profiles [28–32].

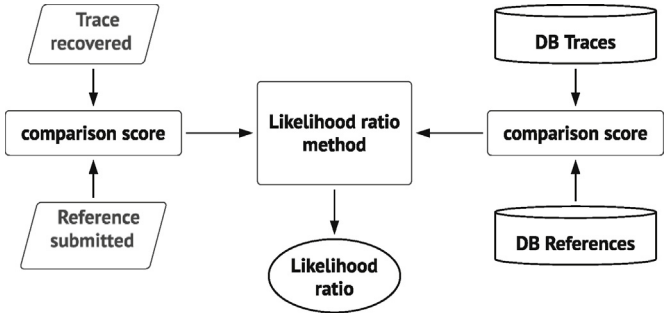

Fig. 1. Score-based LR computation.

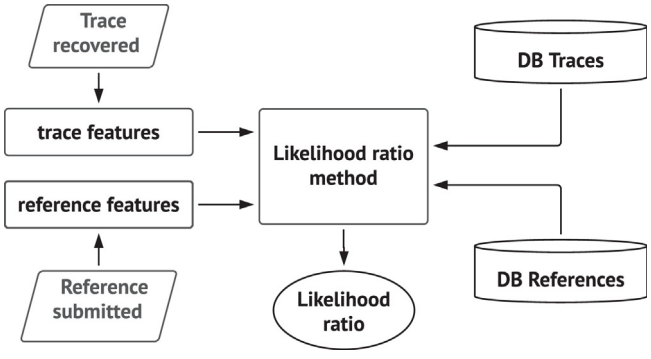

Fig. 2. Feature-based LR computation.

2.2. The LR as a part of the decision process

The development of likelihood ratio methods implies combination of profiles and roles of the personnel involved. The first role focuses on forensic methodology; it is for forensic scientists to conceive new methods and solutions to specific forensic open questions.

The second role focuses on the development and validation of these new methods and solutions; forensic scientists, statisticians and engineers create new technologies or adapt existing technologies for some specific forensic purpose, like for example the development of computer-assisted methods for forensic evidence evaluation. The validation of these methods consists of finding their scope of validity when used in forensic environment. In order to determine this scope the LR is considered as part of a decision process. The LR method is tested in a full Bayes' inference model, for an extensive range of prior and posterior probabilities related to the propositions, thresholds and decision costs/utilities. Such tests simulate the functionality of the method for the whole range of conditions and the results allow establishing the limits in which the method provides a reliable strength of evidence to the trier of fact.

The third role focuses on the evaluation of forensic evidence in practice. Forensic practitioners introduce new methods, for example using computer-assisted LR methods, and use them in casework to asses the strength of the forensic evidence regarding alternative propositions provided by the trier of fact (at least one proposition from the prosecution and one from the defence). In this evaluator role the forensic practitioner plays a role of neutral facilitator. Therefore as an evaluator, the responsibility for the forensic practitioner is to obtain the most relevant alternative propositions to be considered in the case, using the most suitable method to provide the most correct strength of the evidence in form of a LR [19]. The forensic practitioner has also the responsibility to understand the scope and limitations of the method used, which are described in the validation report.

3. Validation of LR methods

The EU Council Framework Decision 2009/905/JHA [33] on the *"Accreditation of forensic service providers carrying out laboratory activities"* regulates issues related to the quality standards in two forensic areas: DNA-profile and fingerprint/fingermark data. This decision framework seeks to ensure that the results of laboratory activities carried out by accredited forensic service providers in one member state are recognized by the authorities responsible for the prevention, detection and investigation of criminal offences within any other member state. Equally reliable laboratory activities carried out by forensic service providers are sought to be achieved by the EN ISO/IEC 17025 accreditation of these activities [34]. For this reason, this framework focuses on the *General requirements for the competence of testing and calibration laboratories* as described in the EN ISO/IEC 17025:2005 standard, and particularly on the requirements for the validation of non-standard methods described in the Section 5.4.4 of this ISO standard, as we consider the LR methods used for forensic evaluation as non-standard methods. In essence, alike the analytical methods, the computer-based LR methods need to be validated as well.

The computer-assisted methods for forensic evaluation are still very recent and the EN ISO/IEC17025:2005 [34] does not address the question of their validation yet. In 2010, the Dutch accreditation body has issued an explanatory document [35] that provided some Guidelines for the validation of the opinions and interpretations of forensic practitioners. In short, the criteria proposed for the validation of instrumental analytical methods are based on performance and the approach for the validation of the human-based methods used for interpretation is based on competence assessment. As the existing criteria used for interpretation only focus on human-based methods, they are not suitable for the validation of computer-assisted methods developed for forensic evaluation.

Even the newer ILAC-G19:2014 Guideline for forensic laboratories [36] only addresses the question of the validation of instrumental methods used for analytical purpose. But the Guideline pinpoints that *"interpretations of results and findings shall be based on robust studies and documented procedures"*. It also suggests that, *"when developing their processes, forensic units shall show with objective evidence that they have assessed the factors that can influence the results and have recorded these"*. Finally, it advises, *"where software is used it shall be demonstrated as being fit for purpose. This may be a verification check of the software functionality. [...] 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"*.

3.1. Validation strategy

A theoretical or an empirical approach can be used to validate a LR method. The theoretical validation of a LR method rests upon the mathematical proof or falsification, and it is not the focus of this Guideline. On the other hand, the empirical validation rests upon the acceptance or rejection of validation criteria.

3.1.1. Theoretical validation

Where applicable, the theoretical validation is handled using the falsifiability approach [37], focusing on proving/disproving mathematical formulae, propositions, lemmas and theorems, in general assuming that there is a ground truth (trueness) of a given statement that can be falsified (disproved or nullified). This part of the validation is deductive (deductive reasoning), since it relies on mathematical properties and does not imply assumptions.

Although theoretical validation is necessary, the choice of any (LR) method also needs to be validated empirically using appropriate measures of performance, even if it seems *theoretically so well grounded* (for example the Bayes formula) that it is considered as mathematically correct. The term *theoretically so well grounded* should be approached with moderation; it refers to situations where the choices within a method are solidly grounded, for example based on deductive reasoning, justifying its use by proofs and mathematical rigor.

3.1.2. Empirical validation

The empirical validation focuses on the acceptance or rejection of chosen validation criteria. This part of the validation is inductive, as it implies assumptions regarding the inference model(s) used for the evidence evaluation. The empirical validation implies the definition of a validation protocol and experiments, in order to demonstrate the acceptance/rejection of the chosen validation criteria. Where a validation process leads to quantitative results, a range of variables in which the LR method passes each validation criterion should be presented.

The elements involved in the structure of the validation protocol have been defined before: performance characteristics, performance metrics and validation criteria. In order to define such a protocol, the performance characteristics and the related performance metrics need to be identified. The validation criteria need to be established, such as the numerical threshold expressed in terms of the performance metrics chosen. An experiment (or series of experiments) needs to be designed for the LR method under evaluation and appropriate sets of data have to be chosen. Each performance metric produced on this basis is confronted with

an appropriate validation criterion, in order to achieve a validation decision, which would ideally take a binary form – either accept or reject the LR method as validated.

But the example of validation report to be published in Data in Brief [56] demonstrates, that the concept of a "validated method" is not absolute, it is relative and limited to the scope of validation. This scope depends on the type and quantity of data used for the validation and the validation decision is intrinsically linked to the relevance of the validation protocol, justifying the initiative of proposing such a Guideline. Finally, this example spotlights that the determination of the validation criteria cannot be entirely justified scientifically. In the example given the determination of the validation criteria for a completely new method depends on the performance of a baseline method. In case of the validation of a method improving an existing one, the determination of the validation criteria depends on the performance of the existing one. In order to achieve transparency in the validation process, the validation report need to document explicitly all these aspects, and the EN ISO/IEC 17025 format fulfils this requirement.

The scope of validation should be defined prior to the empirical validation of a LR method, and documented in the validation report. There is not a universal rule to select the validation criteria during the definition of the validation protocol. Some recommendations are provided here. Validation criteria can be obtained by a comparison with the *"state-of-the-art"*, if possible. In absence of existing validation criteria due to the novelty of the LR method, the validation criteria can be specified based on the functionality of a *"baseline method"*. The selection of performance characteristics and metrics is important in order to adequately measure the desired characteristics of the LR method. In this Guideline, we give recommendations about performance characteristics in the next sections.

According to the EN/IEC 17025:2005 standard, the validation report of a method provides confirmation by examination and the provision of objective evidence that the particular requirements for a specific intended use are fulfilled. This is the instrument that interfaces science and quality in the accreditation procedure of forensic services, demonstrating the capabilities and limits of the method under evaluation, and therefore determining the extent of the scope when used under accreditation.

The following main scientific questions will be answered: *"What to measure?"* is addressed in the Section 3.2 entitled *"performance characteristics"*, *"How to measure?"* is addressed in the Section 3.3 entitled *"performance metrics"* and the question *"What values should be observed or deemed satisfactory?"* is addressed in the Section 3.5 entitled *"validation criteria"*.

3.2. Performance characteristics

As an outcome of the validation workshop held at the Netherlands Forensic Institute in 2011 (see acknowledgements), several performance characteristics have been identified for the validation of computer-assisted LR methods developed for forensic evaluation. All of these were already defined, though the workshop helped to organise them and to clarify their role. They are now structured in primary and secondary performance characteristics. The primary characteristics of the LR method under evaluation are related directly to performance metrics and focus on desirable properties of the LR methods (e.g. goodness of a set of LR values, in which we are assessing whether a set of LR values is good or bad, adequate or non-adequate, whether it has desirable properties or not). The secondary characteristics describe how the primary metrics behave in different situations, in some cases simulating the typical forensic casework conditions (e.g., specimens of degraded quality, varying quality conditions between the training data and the crime scene specimens, etc.). The secondary characteristics are related to a single primary metric. For instance, generalization may refer to the variation in accuracy (primary characteristic) when modifying the type of data.[2]

Originally, several performance characteristics have been defined in the context of validation of analytical methods for the measurement of physical and chemical quantities (metrology). The definitions of these performance characteristics can be found in the International Vocabulary of Metrology (VIM) [38]. The performance characteristics proposed in this Guideline for the forensic evaluation methods (shown below in Table 2) have been chosen based on their similarity with the original performance characteristics defined for the validation of analytical methods, but have a different meaning. To prevent confusion between the original and newly defined performance characteristics, we present both definitions in parallel in the following subsections. Where the VIM does not provide an exact definition, analogous definitions are extracted from sources cited in the ENFSI 2013 Guidelines for the single laboratory Validation of Instrumental and Human Based Methods in Forensic Science [39], keeping in mind that the fact that the two documents do not have the same status.

3.2.1. Proposed primary performance characteristics

For forensic evaluation methods, three primary performance characteristics have been identified (presented below in Table 1):

3.2.2. Proposed secondary performance characteristics

The following secondary characteristics are recommended in this Guideline, and presented below in Table 2:

3.3. Performance metrics and corresponding graphical representations

For each performance characteristic, the performance metrics and the associated graphical representations recommended in this Guideline will be presented in this section. A summary of all of them is shown in Table 3, and a more detailed description is given below.

3.3.1. Detection error trade-off (DET) plot and equal error rate (EER)

The DET plots and the EER both measure the discriminating power of a LR method. In the DET plot, we threshold a log(LR) in order to exploit the ability of an inference model to make decisions based on the Detection errors – the False Acceptance Rate (FAR) and the False Rejection Rate (FRR). A DET plot then represents a trade-off between these Detection errors.

The DET plot as defined in [46] is a 2-dimensional graphical representation in which the FAR is plotted as a function of the FRR. Fig. 3 shows an example of DET plots. The error rates are plotted in 2 dimensions with a Gaussian-warped scale. Thus, linearity of the DET curves happens when de distribution of the log(LR) values is normal. The closer the curves to the coordinate origin, the better are the discriminating capabilities of the method. The intersection of a DET curve with the main diagonal of the DET plot marks the Equal Error Rate (*EER*) which is used as a performance measure to show the discriminating power. Even if the DET plot is meant to characterize a discrimination system (implying a decision), the information provided indirectly informs about the discriminating power of the LR method. Worth noting, that although a threshold is used to draw DET plots, a particular value of that threshold is not selected, instead all the thresholds are represented. As a consequence, this shows the discriminating power of the set of

[2] Definitions of generalization and accuracy are in more detail provided in Sections 3.2.1 and 3.2.2.

Table 1
Definitions of the primary performance characteristics for LR methods, and contrast with respect to the definitions in VIM [38].

Performance characteristics	VIM definition or other authoritative definition	New definitions for LR-based forensic evaluation methods
Accuracy[a]	"Closeness of agreement between a measured quantity value and a true quantity value of a measure" Closely linked to the accuracy is the precision, in VIM defined as follows: "Closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions"[b]	Closeness of agreement between a LR computed by a given method and the ground truth status of the proposition in a decision-theoretical inference model. The LR is accurate if it helps to lead to a decision that is correct according to the ground truth of the propositions In case of source level inference, the ground truth typically relates to the following pair of propositions: • H_1: the pair of specimens tested originate from the same source (SS) • H_2: the pair of specimens tested originate from two different sources (DS) If the performance metric of a set of LR values is to be computed, and the corresponding ground-truth labels of each of the LR values are known, then a given LR value is evaluated as more accurate if it supports the true (known) proposition to a higher degree, and vice-versa
Discriminating power	"*Discriminating power of a series of k attributes is defined as probability that the two distinct samples selected at random from the parent population would be discriminated in at least one attribute if the series of attributes were determined. The distribution of each attribute over the population is assumed to be known from a study of a large number of samples*" [40]	Performance property representing the capability of a given method to distinguish amongst forensic comparisons where different propositions are true
Calibration (Calibration loss)	*"Operation that, under specified conditions, in a first step, establishes a relation between the quantity values with measurement uncertainties provided by measurement standards and corresponding indications with associated measurement uncertainties and, in a second step, uses this information to establish a relation for obtaining a measurement result from an indication."* The concept of calibration used in the context of analytical methods has nothing to do with the definition of calibration used in statistics	A property of a set of LRs. Perfect calibrations imply that the LR is exactly as big or small as is warranted by the data, or that *the LR of the LR is the LR* [42]. Perfect calibration of a set of LR's means that those LR's can probabilistically be interpreted as the strength of evidence of the comparison result for either proposition. The strength of evidence of well-calibrated LRs tends to increase with the discrimination power for a given method [42,43]

[a] In analytical methods accuracy and precision imply the existence of a true magnitude of certain physical phenomena that is to be measured. One can for instance measure the short side of a standard credit card, and performing 100,000 measurements to arrive to a certain distribution of data. There is a "true" (exact) value in this case – the exact value of the short side of a credit card is in reality 53.98 mm. By performing additional measurement (obtaining a size of 63.98 mm) the accuracy then represents distance (10 mm in this case) between the reference value and the "true" value.
On the other hand we understand, that due to the definition of the LR as being the result of a *probabilistic inference* and not a *measurement*, no quantitative ground truth exists for the LR because of the "*Bayesian interpretation of probabilities as a degree of belief*" [14]. Therefore, it is not possible to establish univocal relation between a pair of specimens and a numerical likelihood ratio value.

[b] In [41] the accuracy is deemed equal to validity and precision is deemed equal to reliability. In this work we do not follow those definitions, since the validity is regarded as the outcome of a whole validation process, rather than a single measurable entity.

Table 2
Definitions of the secondary performance characteristics and contrast with respect to the definitions on VIM [38].

Performance characteristics	VIM definition or other authoritative definition	New definitions for LR-based forensic evaluation methods
Robustness	The robustness/ruggedness of an analytical procedure is a measure of its capacity to remain unaffected by small, but deliberate variations in method parameters and provides an indication of its reliability during normal use" definition given in [38]	The ability of the method to maintain a performance metric when a measurable property in the data changes. For instance, method A is more robust to the lack of data than method B if, as the data gets sparser, a performance metric of method A degrades relatively less than the same performance metric of method B Note: Robustness in the LR context usually refers to the stability of the LR methods to varying conditions (e.g. quality/quantity of the data), which prevent reliable measurement of the information or of the features carrying this information
Coherence	Not defined in the VIM [38] From Oxford Dictionary: • The quality of being logical or consistent • The quality of forming a unified whole	The ability of the method to yield LR values with better performance with the increase of intrinsic quantity/quality of the information present in the data. For example the quantity of minutiae in the fingerprint field or the signal to noise ratio in the speaker recognition field
Generalization	Not defined in the VIM [38]. From Collins English Dictionary (Logic): • Any statement ascribing a property to every member of a class (universal generalization) or to one or more members (existential generalization). Example: every function is a relation but not every relation is a function	Property of a given method to maintain its performance under dataset shift. A simple illustration of dataset shift [44] can be the amount of the difference in descriptive statistics of two datasets For instance, the bigger the difference in means of two datasets, the bigger their dataset shift. The dataset shift can be also measured with metrics of distance between their probability distributions, for example using the Kullback–Leibler divergence [53]

Table 3
Performance characteristics and metrics recommended in this Guideline, with their corresponding graphical representations.

Validation aspects	Performance characteristic	Performance metric	Graphical representation
Primary performance characteristics	Accuracy = discriminating power + calibration	Cllr [45]	ECE plot [43,51]
	Discriminating power	EER, $Cllr^{min}$	ECE^{min} plot DET plot
	Calibration	$Cllr^{cal}$	Tippett plot ECE plot
Secondary performance characteristics	Robustness	Cllr, EER, range of LR values	ECE plot DET plot Tippett plot
	Coherence	Cllr, EER	ECE plot DET plot Tippett plot
	Generalization	Cllr, EER	ECE plot DET plot Tippett plot

LR values for all possible values of the prior odds and decision costs involved in a decision.

It is important to highlight here that the forensic scientist uses these kinds of plots in the development stage of validating a method, where simulation of prior odds or decision costs can be used in order to check the performance of methods. We do not suggest by any means that decisions are going to be made by the practitioner in the evidence evaluation stage. The practitioner does not set the prior odds all on his own.

3.3.2. Tippett plot

The Tippett plot [2,47,48] is a representation of the complement of the empirical cumulative distributions of the LR values. This representation has been named after the initial comparative experiments of pairs of paints flakes originating from the same or two different sources realised by Tippett et al. in the late 1960's [49]. Each of the curves represents the empirical cumulative distribution of the proportion of the LRs in a given set of LR values, assuming that each of competing propositions is true. In the Tippett plot (Fig. 4), the rates of misleading evidence can be observed when either of the propositions about the common/different origin is true. These rates are visible at the intersection of each of the inverse cumulative density lines for either the LRs resulting from same source (at a source level) comparisons or the LRs resulting from different sources (at a source level) comparisons and the imaginary vertical line going through value zero on the *X*-axis. The log(LR) value zero on the *X*-axis on the log scale corresponds to the neutral LR value of 1.

In the Tippett plot shown in Fig. 4 we can with relative ease distinguish the quantity of the evidential information within the LR values captured by the LR method presented with datasets in different conditions. However, this may not always be possible [12,51]. A Tippett plot of a LR method evaluating the strength of evidence in fingermarks with five minutiae (smaller area encapsulated within the two dashed lines) and 10 minutiae (greater area encapsulated within the two solid lines) configurations are presented in Fig. 4.

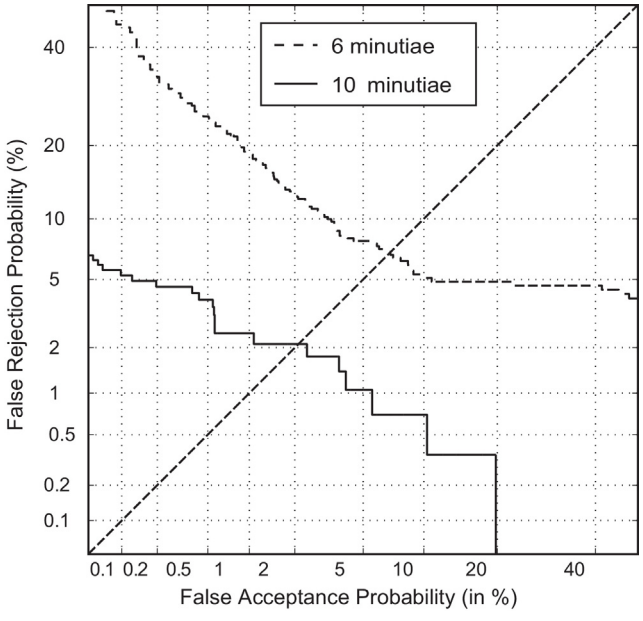

Fig. 3. Two DET plots representing the discriminating power of the same LR method with different quantity of information. Dashed curve represents a method showing less discriminating power in the LR of fingermark to fingerprint comparison for six minutiae, while the solid line shows more discriminating power contained within the LRs of fingermark to fingerprint comparison for 10 minutiae configuration. The dashed straight diagonal indicates that the EER should be measured at the intersection of the DET curves with the diagonal (in this example, EERs are ca. 2.5% for 10 minutiae and ca. 7% for six minutiae).

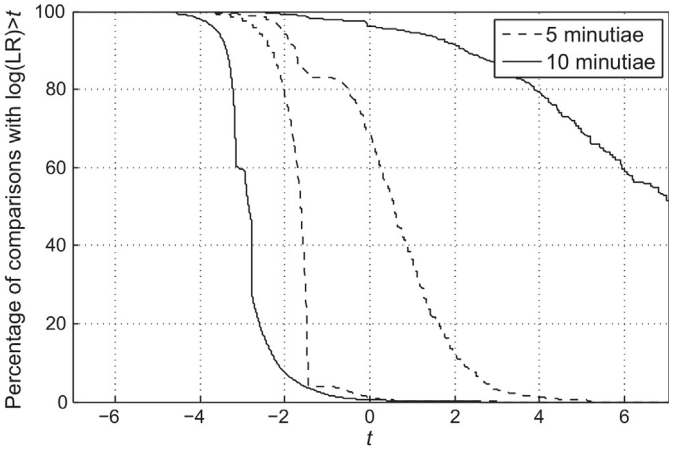

Fig. 4. In this graph, the Tippett plot presents the complement of the empirical cumulative distributions of the same LR method with different quantity of information. The dashed lines represent a method showing less evidential information captured in the LR of fingermark to fingerprint comparison for five minutiae, while the solid lines show more evidential information captured in the LRs of fingermark to fingerprint comparison for 10 minutiae configuration.

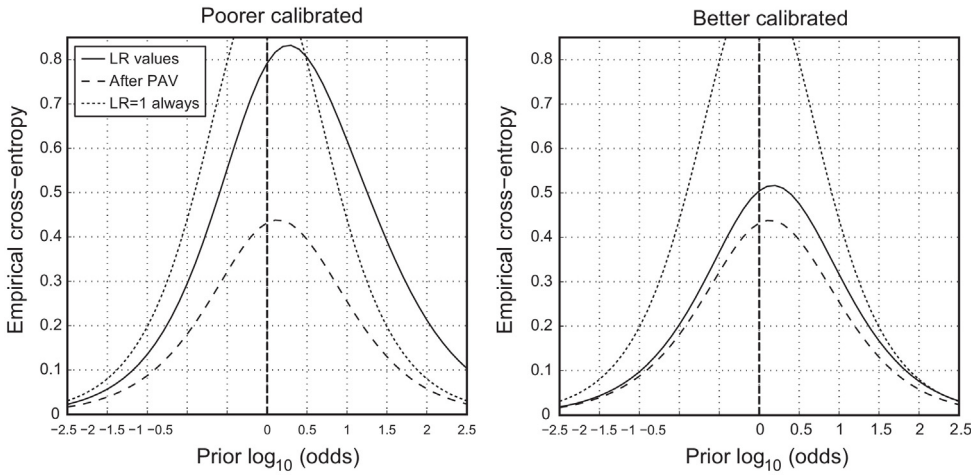

Fig. 5. The ECE plots of the two LR methods using the same data. The LR method on the right presents better calibration than the one on the left. The main drawback of the poorer-calibrated LR method is the fact that around Prior log10(odds) = 0.5 the poorer-calibrated LR method crosses the reference method (outputting always LR = 1). Loosely translated for Prior log10(odds) > 0.5 the poorer-calibrated method performs poorer than not evaluating the evidence at all (the latter represented by a method constantly returning LR = 1).

3.3.3. Empirical cross-entropy (ECE) plot and the log-likelihood-ratio cost (Cllr)[3]

Here, we introduce two performance measures that have been proposed in forensic science: the so-called *log-likelihood-ratio cost* (Cllr) [45,50], and the Empirical Cross-Entropy (ECE) [43,51]. ECE and Cllr are measures of performance based on *strictly proper scoring rules*, widely studied in Bayes' statistics [43,45]. Both measures tend to indicate better performance when the likelihood ratio leads to the correct decision. As both of them measure *cost* or *loss*, its numerical value will be lower as the performance increases. Their difference relies on the interpretation of both measures. Cllr is interpreted as the cost of decisions, averaged for all prior probabilities and costs involved in the decision process. Therefore, it does not provide the performance by fixing any prior probability and cost, but for all possible priors and costs. On the other hand, the ECE has an information-theoretical interpretation as the information needed to support the correct value of the proposition, on average in a given set of LR values. The ECE is represented as an ECE-plot, showing its value for a certain range of priors (Fig. 5). In fact, both measures, Cllr and ECE, are related [43,51], as the Cllr can be seen as a summary of the ECE plot. In this sense, the ECE and Cllr are a general and interpretable performance metric in a forensic context, in which no decision is to be made by the forensic evaluator and in which the value of the prior is expected to be different from one case to another. The ECE plot also appears to be more suitable for the forensic practice, in which the aim is to show the range of possible prior probabilities in which the LR method is deemed to be validated, and the prior probabilities are in general unknown to the forensic evaluator. On the other hand, the Cllr is a single scalar measure, useful for ranking and comparison, and it in fact summarizes ECE. In this sense of scalar/graphical performance measures, the pair Cllr/ECE plot is comparable to the pair EER/DET plot. An example of ECE plots can be seen in Fig. 5.

The measures of *accuracy* ECE and Cllr can be decomposed into discriminating power and calibration [43,45]. In can be shown that both Cllr and ECE can be decomposed into its additive components: discrimination (namely $Cllr^{min}$ and ECE^{min}) and calibration (namely $Cllr^{cal}$ and ECE^{cal}) [45,50,51].

Moreover, the relationship between ECE and Cllr is as follows. The *Cllr* can be found on the intersection of the solid curve in the ECE plot with the $Prior_{log}odds = 0$ (the lower the *Cllr* the better performance of the system); the $Cllr^{min}$ can be found on the intersection of the dashed curve with the $Prior_{log}odds = 0$ (the lower the $Cllr^{min}$ the better the discrimination of the LR method – see [43,46] for details); while the difference between these two lines on the intersection with the $Prior_{log}odds = 0$ represents the $Cllr^{cal}$ (the smaller the distance, the better the calibration of the LR method).

Besides the information-theoretical aspect, the ECE provides another interesting insight – that is the "range of application" of the LR method under evaluation, understood as the range of prior odds where the method is deemed to be valid. We can safely assume that one of the most desirable properties of a LR method should be to obtain useful[4] performance for the whole range of prior odds.

Fig. 5 presents the ECE plots of the LR method using fingermarks with five minutiae configuration with different calibration performances – poorer-calibrated and better-calibrated. A better-calibrated method also extends the range of application of this method. While the range of application of the poorer-calibrated LR method in terms of $Prior\ log_{10}(odds)$ is [−2.5,0.5] (intersection of the solid line and the black dotted line in the ECE plot in Fig. 4 left), the range of application of the better-calibrated LR method is [−2.5,2.5] (Fig. 4 right). As the prior odds and decision costs are not known by forensic examiners in general, it is necessary to aim to obtain useful methods for the biggest possible ranges of application. Worth noting, the range [−2.5,2.5] is arbitrary, and aims to reflect a wide range of prior log-odds.

3.4. Validation experiment

Before entering validation experiments, validation criteria should be formulated. In order to get more insight on the expected LR method performance, either a comparison with the current state of the art or with a baseline LR method can be performed, which will provide the initial set of validation criteria.

The validation experiment itself should be divided into two stages – the method development stage and the method validation stage.

[3] The formulae for calculating the Cllr and ECE are beyond the scope of this Guideline and the reader is advised to consult the corresponding reference material.

[4] The term "useful LR method" used here refers to "a LR method with better accuracy than a method producing LR=1 always", as the two methods are compared.

- In the method development stage we propose to deal with processes related to the method/model selection, method/model training and method/model testing, and measure basically the primary performance characteristics with a given development dataset. The aim is to provide the best performance with the most representative dataset for the widest possible range of applications.
- In the validation stage we evaluate the LR method performance using the validation dataset (with a known ground truth) and measure the method performance with previously unseen data in forensic conditions. Here, we will measure both primary and secondary performance characteristics [52]. The aim here is to test the LR method in the most similar conditions as the real forensic casework, and to arrive to a validation decision as an outcome.

The proposed flowchart of the validation procedure is shown in Fig. 6.

3.4.1. Method development stage

As previously mentioned, the main objective of the LR method development stage is to establish inference models with the most relevant data in order to provide the best-performing LRs in the widest scope of applications possible. Although secondary performance characteristics can also be measured in the developments stage and provide additional insight into the functionality of the LR method, we use mainly the primary performance characteristics with the proposed performance metrics and graphical representations. Later, in the validation stage, the LR method will be tested using realistic forensic data.

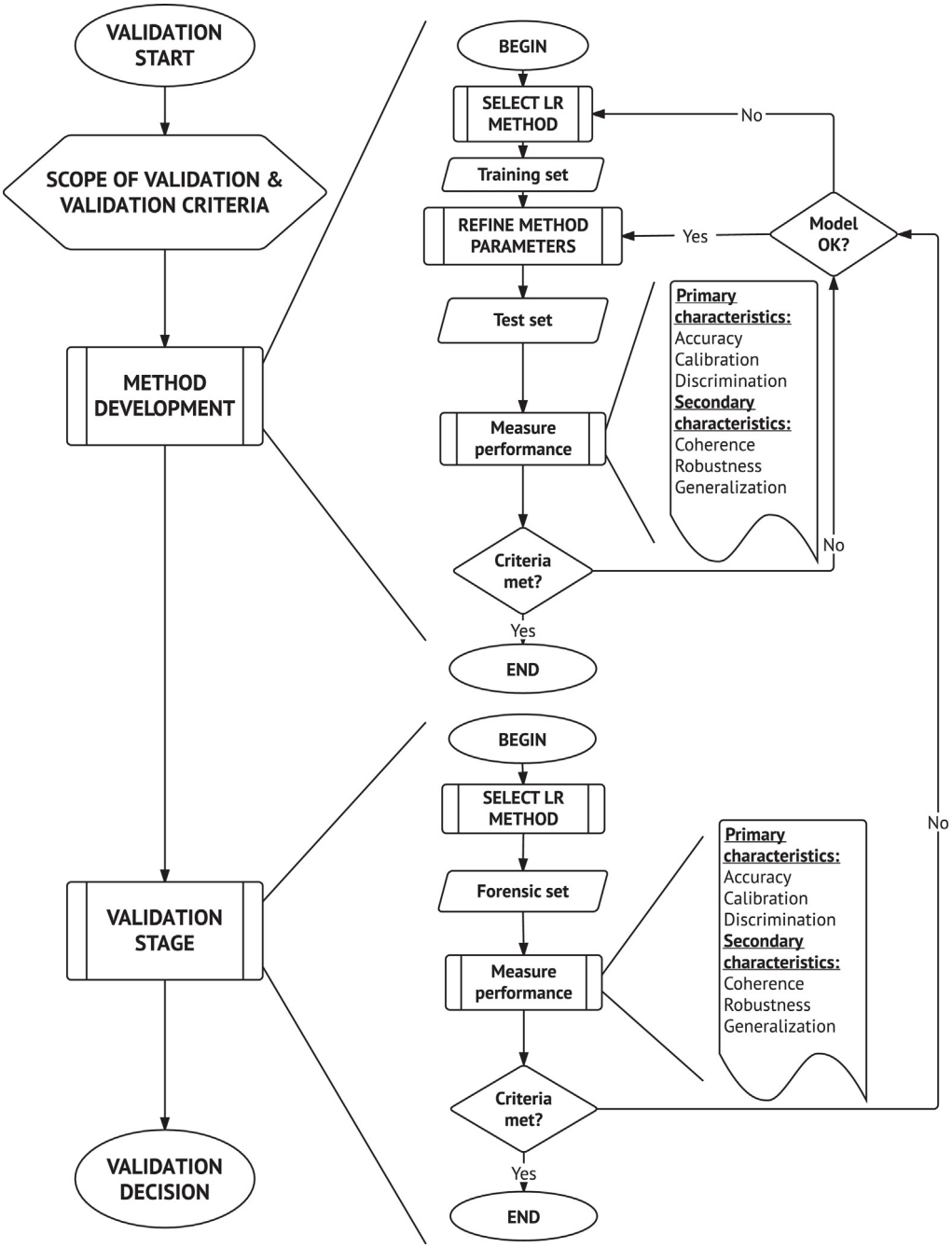

Fig. 6. Flowchart of a validation procedure.

At the development stage, there are two typical datasets involved, which receive different names in different disciplines. Typically, the whole dataset used in the development stage is dubbed *development* dataset, and can be divided into the training and test dataset. The training dataset is used to calculate the parameters of the LR method, while the test dataset is used to establish the performance of the LR method to the previously unseen data in the development stage. The LR method development stage of the validation Guideline uses independent datasets for the training and the test phase, in order to confront the method to previously unseen data of comparable quality. The training dataset is used to define the parameters of the given method, while the testing dataset is used to establish the primary and secondary measures of the method on previously unseen data. The real difficulty is to determine a priori, whether the properties of the test data (previously unseen) are really comparable to the training data. This is easier to accomplish when splitting one dataset, but it can pose a significant challenge when two datasets acquired in different conditions are used (e.g. simulated and real data). A single dataset can be split into a training and test subsets, which should be non-overlapping, independent (previously unseen) and representative. More detailed aspects about the data selection for validation depend on the discipline and are out with the scope of this Guideline.

An inadequate split of one single dataset can cause issues known as under- or overfitting. In underfitting, the LR method will provide a bad representation of the dataset, while in overfitting the LR method will fit too closely to the training dataset and will perform badly on the previously unseen data. Splitting the dataset into the training and test sets, the performance of the LR method will be measured in fair conditions, where the testing data is not known in advance.

3.4.1.1. Training dataset. In a simple case where we aim for example to fit a normal distribution to the set of scores or features, our objective in the training phase is to use the training dataset to obtain the parameters of the normal probability density function (mean vector and covariance matrix). Usually more complex methods are used instead, such as linear logistic regression, kernel density functions, beta or gamma distributions to name a few. However the principle is the same – use the training dataset to learn the parameters for the LR method.

3.4.1.2. Test dataset. The test dataset is intended as the sanity check regarding the basic functionality of the LR method. It is used in the method development stage to evaluate the performance of the LR method to the previously unseen data. In a simple case, we take the LR method developed on the training dataset and measure its performance in the test dataset. Since the test dataset appears to the LR method as previously unseen, lower performance of the LR method is expected.

3.4.2. Validation stage

The dataset used for the validation stage should represent realistic forensic conditions found in casework. Also, it should be independent and representative with respect to the dataset used in the method development stage. Ideally the real forensic data should be kept for the validation stage, ensuring the functionality of the method developed in real forensic conditions. It is common and expected that well-performing methods in the development stage lose some of their performance when subjected to real forensic data.

If the performance of the LR method in the development stage meets the validation criteria (established previously to the validation experiment), one can proceed further to the validation stage. In the validation stage the LR method developed in the method development stage uses the forensic dataset to measure the primary and secondary performance characteristics.

As mentioned earlier, the LR method developed using the development dataset may show lower performance on the previously unseen dataset, mainly due to the dataset shift between the datasets used in the method development stage and the validation stage.

Should the validation criteria not be met by the LR method on the forensic dataset at the validation stage, a logical step is to move back to the method development stage, and then proceed with any of the following possible steps:

- Refine the training parameters of the LR models.
- Use alternative models to compute the LR.
- Relax the validation criteria.

The order in which the steps are to be applied should be critically assessed – based on the time/budget constraints. One can spend months trying to refine parameters of a completely ill-performing LR method, whereas an alternative LR method might give comparable (if not better) results. It might be therefore a good idea to measure and compare the performance of different LR methods in the development stage prior to moving to the validation stage.

3.5. Validation criteria

Biometric technologies used as black boxes are sometimes subject to empirical evaluation using standardized datasets. In fingerprints, good examples are the fingermark/print databases NIST SD04 or NIST SD27 of the National Institute of Standards and Technology. If the result of these evaluations, expressed as measures of performance, are to be used to establish validation criteria, we shall refer to this approach as "*a comparison with the state-of-the-art*", since the validation criteria can be deduced based on the performance of state-of-the-art algorithms. It should be noted here, that establishing the validation criteria as strictly equal to the performance of the state-of-the-art only makes sense in the case of either using the state-of-the-art algorithm or being sure that the LR method proposed will be able to directly compete against and/or outperform the state-of-the-art, which might be rather challenging.

If such a specific database and/or previous experiments do not exist and the comparison with state-of-the-art methods is not an option; a baseline method can be developed. An example could be a LR model based on the score distributions of a training dataset computed by a biometric system (under same-source and different-source propositions). Then, we can measure performance of the baseline method with a given dataset, and establish validation criteria according to this performance. We shall refer to this approach "*a comparison with the baseline*". The criteria should not be set according to the selection of the worse possible baseline. In that sense, the state-of-the-art methods should be considered indirectly (as a reference), as a sanity check in order to establish the proposed baseline to be competitive/comparable to the state-of-the-art methods.

Alternatively, multiple LR methods can be developed at the same time on the development stage, of which one might play the role of the baseline method from which the validation criteria will be defined. The LR methods proposed, including the baseline, should be fit for purpose. For example, a gamma probability distribution function (pdf) will not be a good representation of features in a training dataset showing a distribution very similar to a normal pdf. Thus, LR methods obviously not fit for purpose should be eliminated.

Table 4
Forensic process.

Forensic process			
Crime-scene	Analysis	Interpretation	Reporting
Initial issues What, where, when, how, who	Case assessment -Reporting -Interpretation -Analysis	Intelligence	Cases links (intelligence)
Scene investigation - Refine scenarios and hypotheses - Trace recovery	Analytical methods - Indicative methods - Human-based - Computer-assisted	Or investigation	Rank list (investigation)
Crime scene assessment - Description - Requests	Comparative methods - Human-based - Computer-based	Or evaluation - Human-based - Computer-assisted	Strength of evidence (evaluation)

3.6. Validation decision

A validation procedure[5] should yield a *validation decision*, defined as a binary expression (e.g. pass/fail) regarding the LR method being fit/not fit for forensic evaluation casework. This validation decision may be obtained from a set of comparisons of performance metrics with validation criteria, as proposed in this Guideline.

A set of recommendations can be issued alongside the validation decision, addressing mainly the shortcomings and limitations of the LR method under evaluation. These may contain applicability range of a LR method, clarity/distortion limits, description of sampling procedures, comparison algorithms used etc.

3.7. Validation report

The validation of non-standard methods is described in the EN ISO/IEC 17025:2005 standard (General requirements for the competence of testing and calibration laboratories) in section 5.4.4. *"When it is necessary to use methods not covered by standard methods, these shall be subject to agreement with the customer and shall include a clear specification of the customer's requirements and the purpose of the test and/or calibration. The method developed shall have been validated appropriately before use."* In the section 5.4.4 the ISO standard also lists the information recommended:

a) appropriate identification
b) scope
c) description of the type of item to be tested or calibrated
d) parameters or quantities and ranges to be determined
e) apparatus and equipment, including technical performance requirements
f) reference standards and reference materials required
g) environmental conditions required and any stabilization period needed
h) description of the procedure, including
   - affixing of identification marks, handling, transporting, storing and preparation of items
   - checks to be made before the work is started
   - checks that the equipment is working properly and, where required, calibration and adjustment of the equipment before each use
   - the method of recording the observations and results
   - any safety measures to be observed;
- criteria and/or requirements for approval/rejection
- data to be recorded and method of analysis and presentation
- the uncertainty or the procedure for estimating uncertainty.

Prior to starting the validation of a LR method, a validation plan should be drawn by a forensic scientist. It is mandatory for the reader to keep in mind, that the EN ISO/IEC 17025:2005 standard was predominantly developed for the validation of analytical methods, therefore not all of the recommended information is applicable to the validation of LR methods. Especially the points e), f), g), h), j) and k) will be rather challenging to defend in the interpretation of forensic evidence. In compliance with the remaining recommendations from the EN ISO/IEC 17025:2005 standard the validation plan should contain (but is not limited to) the following:

- Identification of the LR method – point a)
- The intended use – point b)
- The performance characteristics – point d)
- The performance metrics – point d)
- The validation criteria – point i)
- The scope of the validation (range of application of the LR method) – point b)
- Validation time span (applicable in cases in which the datasets used in the LR method development/validation stage are envisaged to get obsolete).

An example of the validation report is presented in an independent support document published in the Elsevier journal "Data in Brief" [56].

4. Conclusions

The forensic process is composed of four types of activities: crime-scene, analysis, interpretation and reporting (Table 4). The forensic crime-scene activities are accredited under the EN ISO/IEC 17020:2012 standard [54], the forensic laboratory activities and the human-based interpretation methods for forensic evaluation are accredited under the EN ISO/IEC 17025:2005 standard [34]. On the other hand there is currently no standard for the accreditation of computer-based methods for forensic evaluation and to our knowledge there was even no Guideline for the validation of likelihood ratio methods used for forensic evidence evaluation.

This Guideline is a first attempt to foster an open discussion within the scientific community and to offer an opportunity for comments and suggestions. Our aim is evolve towards an agreement regarding the performance characteristics, the performance metrics and the procedure to be followed to validate these methods, in order to integrate them into a ISO/IEC standard, with a preference for the EN/IEC 19795-2:2007 standard [55].

[5] Understood as the procedure that uses the validation protocol.

Acknowledgements

It should be noted here, that this validation Guideline reflects the authors point of view on "*how to validate*" LR methods used for forensic evidence evaluation, based on the inputs gained on the workshop, described below. Even though a global consensus by all the participants to the validation workshop was not reached regarding some aspects, the main objective was to foster further discussions and to provide a decent starting point for validation of computer-based LR methods. Special credit belongs to all the participants in the validation workshop meeting. Without those inputs it would be impossible to draft this document.

The validation workshop, held in The Hague on 19th and 20th October 2011 was organised by the authors of this article to define Guidelines for the validation of computer-assisted LR methods developed for forensic evaluation. Topics discussed, namely "*the LR as part of a decision process*" and "*dealing with uncertainty in the LR calculation*" were deemed necessary to define a validation strategy based on validation criteria.

Names of the participants and their affiliation (at the time of the workshop):

Christophe Champod – Université de Lausanne, Switzerland
Niko Brümmer – AGNITIO, Madrid, Spain
R.N.J. Veldhuis – University of Twente, The Netherlands
D. Ramos, J. Gonzalez-Rodriguez – Universidad Autonoma de Madrid Spain
D. Meuwly, M. Sjerps, A. Bolk, A. Ruifork, Ch. Berger, D. van Leeuwen, I. Alberink, H. Hanned, A. de Jongh, P. Vergeer, K. Slooten, L. Peschier, R. Haraksim, J, Leegwater, A. Lubach, J. Vermeulen, K. Herlaar – Netherlands Forensic Institute, The Netherlands

The research was conducted in scope of the BBfor2 – European Commission Marie Curie Initial Training Network (FP7-PEOPLE-ITN-2008 under Grant Agreement 238803) in cooperation with the Netherlands Forensic Institute and the ATVS Biometric Recognition Group at the Universidad Autonoma de Madrid.

References

[1] S.L. Ellison, Uncertainties in qualitative testing and analysis, Accredit. Qual. Assur. 5 (8) (2000) 346–348.
[2] D. Meuwly, Forensic individualization from biometric data, Sci. Just. 46 (2006) 205–213.
[3] D.V. Lindley, A problem in forensic science, Biometrika 64 (1977) 207–213.
[4] C.G.G. Aitken, F. Taroni, Statistics and the Evaluation of Evidence for Forensic Scientists, John Wiley & Sons, Chichester, 2004.
[5] A. Bolck, C. Weyermann, L. Dujourdy, P. Esseiva, J. van den Berg, Different likelihood ratio approaches to evaluate the strength of evidence of MDMA tablet comparisons, Forensic Sci. Int. 191 (1) (2009) 42–51.
[6] F. Taroni, D. Bozza, A. Biedermann, P. Garbolino, C. Aitken, Data Analysis in Forensic Science: A Baxyesian Decision Perspective, 88, John Wiley & Sons, 2010.
[7] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D.T. Toledano, J. Ortega-Garcia, Emulating DNA: rigorous quantification of evidential weight in transparent and testable forensic speaker recognition, IEEE Trans. Audio Speech Lang. Process. 15 (September (7)) (2007) 2104–2115.
[8] T.L. Ali, J. Spreeuwers, R.N.J. Veldhuis, A review of calibration methods for biometric systems in forensic applications, in: 33rd WIC Symposium on Information Theory in Benelux, Boekelo, Netherlands, (2012), pp. 126–133, WIC. ISBN 978-90-365-3383-6, May.
[9] I. Alberink, A. de Jongh, C. Rodriguez, Fingermark evidence evaluation based on automated fingerprint identification system matching scores: the effect of different types of conditioning on likelihood ratios, J. Forensic Sci. (2013), http://dx.doi.org/10.1111/1556-4029.12105.
[10] A.B. Hepler, C.P. Saunders, L.J. Davis, J. Buscaglia, Score-based likelihood ratios for handwriting evidence, Forensic Sci. Int. 219 (1–3) (2012) 129–140.
[11] C. Neumann, C. Champod, M. Yoo, T. Genessay, G. Langenburg, Quantifying the weight of fingerprint evidence through the spatial relationship, directions and types of minutiae observed on fingermarks, Forensic Sci. Int. 248 (2015) 154–171.
[12] A. Martyna, K. Sjastad, G. Zadora, D. Ramos, Analysis of lead isotopic ratios of glass objects with the aim of comparing them for forensic purposes, Talanta 105 (2013) 158–166.
[13] D. Ramos, G. Zadora, Information-theoretical feature selection using data obtained by Scanning Electron Microscopy coupled with and Energy Dispersive X-ray spectrometer for the classification of glass traces, Anal. Chim. Acta 705 (1–2) (2011) 207–217.
[14] I. Evett, Toward a uniform framework for reporting opinions in forensic science casework, Sci. Just. 38 (3) (1998) 198–202.
[15] Association of the Forensic Science Providers (AFSP), Standards for the formulation of evaluative forensic science expert opinion, Sci. Just. 49 (2009) 161–164.
[16] S.E. Willis, ENFSI Guideline for Evaluative Reporting in Forensic Science - Strengthening the Evaluation of Forensic Results across Europe, European Network of Forensic Science Institutes, 2015 (STEOFRAE).
[17] P. Gill, C.H. Brenner, S.J. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayr, N. Morling, M. Prinz, P.H. Schneider, B.S. Weir, DNA commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures, Forensic Sci. Int. 160 (2006) 90–101.
[18] C.E. Berger, J. Buckleton, C. Champod, I.W. Evett, G. Jackson, Evidence evaluation: a response to the court of appeal judgment in R v T, Sci. Just. 51 (2) (2011) 43–49.
[19] R. Cook, I.W. Evett, G. Jackson, P.J. Jones, J.A. Lambert, A method for case assessment and interpretation, Sci. Just. 38 (3) (1998) 151–156.
[20] C. Neumann, I.W. Evett, J.E. Skerrett, I. Mateos-Garcia, Quantitative assessment of evidential weight for fingerprint comparison i. generalization to the comparison of a mark with a set of ten prints from a suspect, Forensic Sci. Int. 207 (1:3) (2011) 101–105.
[21] C. Champod, I.W. Evett, A probabilistic approach to fingerprint evidence, J. Forensic Identif. 51 (2001) 101–122.
[22] C. Champod, I. Evett, B. Kuchler, Earmarks as evidence: a critical review, J. Forensic Sci. 46 (2001) 1275–1284.
[23] C. Champod, D. Meuwly, The inference of identity in forensic speaker recognition, Speech Commun. 31 (2000) 193–203.
[24] K. Hoffman, Statistical evaluation of the evidential value of human hairs possibly coming from multiple sources, J. Forensic Sci. 36 (1991) 1053–1058.
[25] C. Champod, D. Baldwin, F. Taroni, S.J. Buckleton, Firearms and tool marks identification: the Bayesian approach, AFTE J. 35 (2003) 307–316.
[26] C. Champod, F. Taroni, Interpretation of Evidence: the Bayesian Approach, Taylor and Francis, London, 1999, pp. 379–398.
[27] G. Zadora, A. Martyna, D. Ramos, C. Aitken, Statistical Analysis in Forensic Science: Evidential Values of Multivariate Physicochemical Data, John Wiley and Sons, 2014 January.
[28] C.D. Steele, J.D. Balding, Statistical evaluation of forensic DNA profile evidence, Annu. Rev. Stat. Appl. 1 (2014) 361–384.
[29] P. Gill, H. Haned, A new methodological framework to interpret complex DNA profiles using likelihood ratios, Forensic Sci. Int.: Genet. 7 (2013) 251–263.
[30] D.J. Balding, Evaluation of mixed-source, low-template DNA profiles in forensic science, PNASs 110 (30) (2013) 12241–12246.
[31] D. Taylor, J.A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int.: Genet. 7 (2013) 516–528.
[32] M. Perlin, M. Legler, C. Spencer, J. Smith, W. Allan, J. Belrose, B. Duceman, Validating true allele DNA mixture interpretation, J. Forensic Sci. 56 (6) (2011) 1430–1447.
[33] COUNCIL FRAMEWORK DECISION 2009/905/JHA of 30 November 2009 on Accreditation of forensic service providers carrying out laboratory activities, Official Journal of the European Union, 9.12.2009, online http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L: 2009:322:0014:0016:EN: PD.F.
[34] International Organization for Standardization EN ISO/IEC 17025, General Requirements for the Competence of Testing and Calibration Laboratories, ICS: 03.120.20, stage 90/93, 15/12/2010.
[35] Dutch Accreditation Council (RvA), Explanation of ISO/IEC 17025:2005, 30 November 2010.
[36] International Laboratory Accreditation Cooperation, ILAC-G19:2014 Guidelines for Forensic Science Laboratories, 2014.
[37] K. Popper, The Logic of Scientific Discovery (Taylor & Francis e-Library ed.), Routledge/Taylor & Francis e-Library, London–New York, 2005.
[38] Bureau International des Poids et Measures, International vocabulary of metrology – basic and general concepts and associated terms, JCGM 200 (2012), pp. 91.
[39] European Network of Forensic Science Institutes, Guidelines for the Single Laboratory Validation of Instrumental and Human Based Methods in Forensic Science, 2013 working version 04/11/.
[40] K.J. Smalldon, A.C. Moffat, The calculation of discrimination power for a series of correlated attributes, J. Fortensic Sci. Soc. 13 (Oct (4)) (1973) 291–295.
[41] G.S. Morrison, Measuring the validity and reliability of forensic likelihood-ratio systems, Sci. Justice 51 (2011) 91–98.
[42] D. van Leeuwen, N. Brummer, The distribution of calibrated likelihood-ratios in speaker recognition, in: Proceedings of Interspeech, Lyon, France, 2013.
[43] D. Ramos, J. Gonzalez-Rodriguez, Reliable support: measuring calibration of likelihood ratios, Forensic Sci. Int. (10) 230 (1–3) (2013) 156–169.
[44] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, N.D. Lawrence, Dataset Shift in Machine Learning, The MIT Press, 2009.
[45] N. Brümmer, J. du Preez, Application independent evaluation of speaker detection, Comput. Speech Lang. 20 (2–3) (2006) 230–275.
[46] A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, The DET Curve in Assessment of Detection Task Performance, National Institute of Standards and Technology (NIST) Gaithersburg, 1997 MD 20899 8940.
[47] J. Lucena-Molina, D. Ramos, J. Gonzalez-Rodriguez, Performance of likelihood ratios considering bounds on the probability of observing misleading evidence, Law Probability Risk (2015) mgu022.
[48] D. Meuwly, Reconnaissance de Locuteurs en Sciences Forensiques: L'apport d'une Approche Automatique, 2001 PhD thesis.

[49] C. Tippett, V. Emerson, M. Fereday, F. Lawton, A. Richardson, L. Jones, S. Lampert, The evidential value of the comparison of paint flakes from sources other than vehicles, J. Forensic Sci. Soc. 8 (2) (1968) 61–65.
[50] D. van Leeuwen, N. Brümmer, An introduction to application-independent evaluation of speaker recognition systems, in: Christian Müller (Ed.), Speaker Classification I: Fundamentals, Features, Methods, Springer, 2007.
[51] D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, C. Aitken, Information-theoretical assessment of the performance of likelihood ratio computation methods, J. Forensic Sci. 58 (6) (2013) 1503–1518.
[52] R. Haraksim, D. Ramos, D. Meuwly, C.E.H. Berger, Measuring coherence of computer-assisted likelihood ratio methods, Forensic Sci. Int. 249 (2015) 123–132.
[53] R. Haraksim, Validation of Likelihood Ratio Methods Used for Forensic Evidence Evaluation: Application in Forensic Fingerprints, Enschede, The Netherlands, 2014, PhD thesis.
[54] International Organization for Standardization EN ISO/IEC 17020:2012(E), Conformity Assessment – Requirements for the Operation of Various Types of Bodies Performing Inspection, 2012.
[55] International Organization for Standardization EN ISO/IEC 19795-2:2007, Information Technology – Biometric Performance Testing and Reporting – Part 2: Testing Methodologies for Technology and Scenario Evaluation, 2007.
[56] D. Meuwly, D. Ramos, R. Haraksim, Validation of Likelihood Ratio Methods: an Example of Validation Report in Fingerprints, Data In Brief, under revision, 2016.