Law, Probability and Risk, 2025, **24**, 1–20
https://doi.org/10.1093/lpr/mgaf012
Research Article

OXFORD

Distinguishing exogenously and endogenously defined reliability from individual report accuracy in expert and machine evidence

Alex Biedermann [1]*

[1]Faculty of Law, Criminal Justice and Public Administration, School of Criminal Justice, University of Lausanne, 1015 Lausanne–Dorigny, Switzerland

*Corresponding author. Faculty of Law, Criminal Justice and Public Administration, School of Criminal Justice, University of Lausanne, 1015 Lausanne–Dorigny, Switzerland. E-mail: alex.biedermann@unil.ch

Abstract

Concerns about the accuracy and reliability of forensic methods and techniques, both in general and when applied in specific cases, permeate current debates about how to think about forensic science and its use in legal proceedings, regardless of whether applications involve humans, machines or a combination of the two. While accuracy and reliability are relevant topics of inquiry when considering the admissibility of specialized ('expert') witness opinion testimony, especially when it is machine-based, they raise the intricate question of what exactly one is entitled to infer from testimony that has been deemed reliable enough to be admissible. This article systematically examines the conceptual aspects of this problem using an analytical approach and formal methods from systems engineering and probabilistic epistemology. Based on the notions of *exogenously* and *endogenously* defined reliability, as well as the concepts of feature selectivity and examiner diagnosticity, the analyses presented here explain that what is generally referred to as the reliability of an information source, i.e. accuracy in the aggregate case, is conceptually different from the accuracy of an individual report produced in the instant case. In the account given here, the former can be seen as an empirical matter, whereas the latter is defined deductively and depends on more than data from empirical testing. This finding is relevant in that it shows that the widespread calls for more empirical testing (e.g. through black-box studies), while necessary, are not sufficient to resolve assessments of the accuracy of individual reports. The hope that (more) empirical testing *alone* will solve the reliability/accuracy problem in forensic science is therefore unwarranted in the light of the account given here. This article concludes that the subtle properties of individual report accuracy limit our practical ability to know, let alone control, the extent to which accuracy can be taken for granted in actual cases.

Keywords: accuracy; accuracy fallacy; black box studies; empirical testing; feature selectivity; probabilistic epistemology; scientific evidence; reliability; validity.

1. Introduction

Not least since the publication of the report of the President's Council of Advisors on Science and Technology (PCAST 2016), the concepts of reliability, validity, and accuracy have come to the forefront of the attention of evidence commentators, academics, legal practitioners, and forensic scientists. Broadly speaking, these terms revolve around the question of whether forensic examiners are capable of doing what they claim to be able to do. The inconvenient fact is that over the past few decades a steady stream of publications has called into question the trustworthiness of many forensic science disciplines (e.g. Risinger et al., 1989; Saks and Koehler 2005; Fabricant 2022), leaving virtually no forensic discipline free from serious doubt. The baseline

Received: 15 February 2025. **Revised:** 29 June 2025. **Accepted:** 12 September 2025

demand in these critiques is that forensic scientists should be able to provide information about the rate at which they make errors (Koehler 2008). In response to these challenges, researchers in several forensic disciplines, particularly those involving feature comparison, have begun to conduct empirical studies of examiner performance. Hicklin et al. (2022a,b) reported examples in the fields of handwriting and footwear mark examination. In these so-called black-box studies, 'many examiners render decisions about many independent tests (typically, involving 'questioned' samples and one or more 'known' samples) and the error rates are determined' (PCAST 2016: 5, 6). Despite these initiatives to strengthen forensic science, critical observers continue to expose what they perceive as notoriously flawed disciplines, such as firearms examination (Faigman et al., 2022; Cuellar et al., 2024).

The above concerns about the performance of forensic science methods and the examiners who use them stem from the role that reliability plays in the legal rules of certain jurisdictions governing the admissibility of *specialized witness opinion evidence*.[1] A prime example in US law is Federal Rule of Evidence (FRE) 702, widely known for its in-depth interpretation by the US Supreme Court in the *Daubert* decision.[2] The Rule mentions the notion of reliability twice. Specifically, FRE 702(c) requires that the testimony of a witness qualified as an expert be 'the product of reliable principles and methods' and FRE 702(d) requires that 'the expert's opinion reflects a reliable application of the principles and methods to the facts of the case'. The importance of FRE 702 is emphasized by the fact that in June 2022, the Judicial Conference Committee on Rules of Practice and Procedure (Standing Committee) approved amendments, effective from 1 December 2023, intended to mitigate the misapplication of the rule.[3]

English law, as a contrasting example, is more pragmatic, as there is no primary legislation directly comparable to FRE 702. Instead, England and Wales relies on what Roberts (2018: 53) calls 'hardworking soft law', such as the Criminal Procedure Rules (CPR) and Practice Directions.[4] These rules also invoke and rely on the notion of reliability. For example, CPR 19.4 states that the expert's report must 'include such information as the court may need to decide whether the expert's opinion is sufficiently reliable to be admissible as evidence'. In turn, Section 7.1.1 of The Criminal Practice Directions 2023 states that '[e]xpert opinion evidence is admissible in criminal proceedings if (…) the expert opinion is sufficiently reliable', while Section 7.1.2 lists '[f]actors which the court may take into account in determining the reliability of expert opinion, and especially of expert scientific opinion'. The purpose here is not to provide a comprehensive review of regulatory and procedural frameworks. However, the examples are sufficient to illustrate the common underlying purpose of ensuring that a party seeking to introduce particular expert testimony can demonstrate that it is, in some sense, *reliable*.[5]

While these concerns rightly dominate much of the current debate about expert evidence and its admissibility, the notion of reliability as conveyed raises the deeper question of what exactly one is logically entitled to infer from testimony that has been deemed reliable enough to be admissible. After all, concerns about reliability are rooted in concerns about the accuracy of fact-finding. The key question, then, is how assessments of reliability, and in particular degrees of reliability, (should) relate to or affect conclusions about accuracy. One might ask, for example, whether it is logically justified to conclude that the more reliable an information source is, the higher the probability that a given report from that source is accurate. On the one hand, this may seem a plausible intuition. On the other hand, the well-known weaknesses of human reasoning under uncertainty (Kahneman et al., 1982), especially when it comes to expressing uncertainty in terms of probability, give us ample reason to distrust our intuitions.[6] In addition, there is limited guidance from a legal point of view because, at least in English law, there are no prescriptions regarding the credit to be given to information, such as particular categories of scientific evidence (e.g. Dennis 2017: 4–008). Overall, therefore, the logical relationship between

[1] For simplicity, the term expert evidence will be used in the remainder of this article.
[2] *Daubert v. Merrell Dow*, 125 L Ed 2d 469; 113 S Ct 2786 (1993).
[3] See also https://www.govinfo.gov/content/pkg/CDOC-118hdoc33/pdf/CDOC-118hdoc33.pdf (at p. 18, accessed 22 December 2024).
[4] https://www.legislation.gov.uk/uksi/2020/759/part/19 (accessed 22 December 2024).
[5] See e.g. *R v Dlugosz* [2013] EWCA Crim 2.
[6] As will be shown later in this paper, this intuition is indeed not warranted.

reliability and accuracy does not seem to have an obvious answer, which makes it a relevant topic of investigation.

The purpose of this article is to address the relationship between reliability and accuracy using formal methods from systems engineering and probabilistic epistemology. Particular attention will be given to the notions of *endogenous* and *exogenous* source reliability (Bovens and Hartmann 2003), and their formal representation in terms of graphical probabilistic models, i.e. Bayesian networks (Cowell et al., 1999). This methodological choice will facilitate the logical extension of reliability considerations to the notion of accuracy.

This article is structured as follows. Section 2 reviews and compares common definitions of reliability and validity, which are often used synonymously. These definitions will serve to highlight areas where this article will develop alternative views. Sections 3–5 present and discuss the notions of endogenous and exogenous source reliability according to Bovens and Hartmann (2003), as applied here in the context of forensic science and law, and existing work in these fields.[7] These notions are extended in Section 6 to the notion of individual report accuracy and its structural relationship to reliability. Bayesian networks are used to illustrate the structural relationship between reliability and accuracy and to examine the inferential properties of the proposed formal model. Section 7 concludes this article by discussing the results of the analyses presented in the previous sections in the light of ongoing debates about the sound and safe use of scientific evidence in the legal process.

2. Common accounts of reliability and validity

Commentators in the forensic science literature have noted that the term reliability is used in a variety of ways, creating confusion where there should be clarity. For example, Robertson et al. (2016) criticize that the term *reliable* 'appears to have no fixed meaning' (at p. 6), especially in ordinary language where it can be used to refer to several concepts, including but not limited to validity, accuracy, precision, sensitivity, and specificity (at 62). For this reason, the same authors suggest that the term reliability should be avoided (Robertson et al., 2016: 62). However, this advice seems of little help as the term continues to be widely used in legal literature and terminological clarity is key to effective communication between lawyers and scientists.

Fortunately, in its 2016 report, the PCAST Committee[8] has been clear about its understanding of reliability and validity, and the relationship between these two concepts. The Committee has defined the reliability of a metrological method as the combination of the three properties of repeatability, reproducibility, and accuracy (PCAST 2016: 47). On this basis, the Committee then went on to define *scientific validity* as being established when a 'method has shown, based on empirical studies, to be reliable with levels of repeatability, reproducibility, and accuracy that are appropriate to the intended application' (PCAST 2016: 48). The Committee further distinguished between foundational validity and validity as applied. The former refers to a method that is considered reliable in principle.[9] The latter, validity as applied, refers to the question of whether the examiner in the instant case properly applied the method. It is worth noting that the Committee did not intend these directions to be understood as opinions on legal standards (of admissibility), but only as 'guidance concerning *scientific* standards for scientific validity' (PCAST 2016: 4). More recently, Swofford et al. (2024) introduced the terms method performance and method conformance, which appear to be analogous to PCAST's foundational validity and validity as applied.

While these understandings are largely uncontroversial, a few comments are in order to highlight peculiarities and blind spots that will be explored in the remainder of this article. First, the above account of validity involves a subtle distinction between performance at the level of the method and performance at the level of the examiner. The investigation of these two dimensions of performance requires different means, to the extent that methods and examiners can reasonably be investigated separately. This goes a long way towards dispelling the hope that

[7] For a discussion of endogenous and exogenous reliability in the context of the study of rational approaches to argumentation, see Hahn et al. (2013).

[8] Hereinafter referred to as the 'Committee'.

[9] "Foundational validity means that a method can, *in principle*, be reliable" (PCAST, 2016, at p. 56, emphasis as in original).

concluding validity in actual cases can be reduced to a simple one-off assessment. Secondly, it is important to understand that the above account of validity is deeply rooted in an empirical perspective. Indeed, the Committee insists on establishing validity through 'empirical studies' (PCAST 2016: 48), a view that has a long history in the legal literature, at least since the mid-1990s. More specifically, the argument goes, the validity of a method is to be examined through validity studies, whereas the demonstration of an examiner's ability to apply a given method is based on proficiency studies or tests (e.g. Imwinkelried 1995; Koehler 2008).

So far, so good, one might think, but the focus of this empirical view is exclusively on performance measurement in the *aggregate case*. As noted by Cole (2006), a validation study yields an 'accuracy *rate*: latent print identification is accurate approximately such-and-such percent of the time' (at p. 11, emphasis added). This alludes to the long-run arguments commonly associated in statistics with frequentism and its known limitations for case-specific inferential reasoning. The blind spot here is that an accuracy *rate* of this kind does not tell us, at least not directly, what credibility we should assign to testimony given in the case at hand. At the most fundamental level, a rate is just a descriptive summary of data, i.e. the proportion of times a particular outcome has been observed in a series of events of interest. Without further argument, a summary statistic is far from an inferential conclusion. Sure, for what it's worth, empiricism might suggest that conclusions about case-specific validity should be based on the trust that past performance under controlled conditions will safely extend to the present case. But such trust is fragile, as Imwinkelried (2020) notes, because it depends—among other constraints—on whether the instant case falls within the so-called *range* of validation of the method being offered. This opens up a host of further challenges, such as determining a range of validation and deciding whether or not a given case is covered by such a range.

All of the above observations are interrelated manifestations of the broader 'Group to Individual (G2i) Challenge' (Faigman et al., 2014), i.e. the question of what, if anything, general scientific knowledge (i.e. on the group-level) provided by experts allows courts to conclude in individual cases. The fundamental nature of this challenge is underlined by the fact that it is not limited to the output of human experts, assisted or not by scientific methods and techniques. The problem also affects the AI output of many current machine learning approaches, which rely on performance assessment through test data and the calculation of standard metrics such as false positive and false negative rates (Lau and Biedermann 2020). Recently, a large group of artificial intelligence researchers have called for more attention to be paid to the limitations of such aggregate metrics for performance assessment in future research and policy development (Burnell et al., 2023).

This article argues that a missing piece in bridging the gap between validation data and case-based assessments is a conceptual framework. This contention is based on the observation that what the above accounts—particularly the PCAST Report—provided is a definition of the kinds of data that are necessary but not sufficient for case-based assessments, i.e. endorsing empiricism. However, these accounts are silent on other venerable traditions in epistemology, particularly rationalism. It is probably not surprising that the PCAST Report bypassed the latter, since questions about how to legitimate beliefs (e.g. about reliability) on the basis of reason are reputedly more controversial than mere empirical questions with which the PCAST Report has already encountered considerable resistance.

In light of these convoluted matters, the remainder of this article seeks to address the epistemological gap surrounding the transition from aggregate case data to conclusions about the accuracy of reports provided by specialized witnesses in instant cases. The next two sections address this topic by presenting and discussing relevant notions from probabilistic epistemology, namely endogenous and exogenous (source) reliability, and analogous developments in the forensic science literature.

3. Exogenously defined reliability

3.1 Preliminaries: generic statement of the problem

Consider a situation where a report is available from a human source, a machine (e.g. an AI-based system), or a combination of the two. The report consists of a statement about an

unknown fact (state of the world). For example, in the context of fingermark examination there may be a report stating that a certain number of corresponding features have been observed between a fingermark found on a surface of interest and a print obtained under controlled conditions from the thumb of a person of interest (Champod et al., 2016). Here, the unknown state of the world is the actual existence of corresponding features between the items being compared. Note that the state of the world is considered unknown here because it is not directly observable by the receiver of the report. Note also that the real state of the world can be considered as a 'testable' consequence of some higher-level proposition, such as whether or not the two compared items come from the same source. Another example, in the area of digital evidence, could be the report of charging data records (in terms of geographical location and time) as provided by mobile network operators, used in cell site analysis (Tart 2020). Here, the unknown state of the world is the actual past position of the mobile device at a given time, while the higher-level proposition could refer to the whereabouts of the owner of the mobile device.

Reports received in practice are potentially misleading in the sense that they may occur even though the respective ground truth state is not true. For example, fingermark examiners may report that they have observed corresponding features when in fact the two items being compared do not have corresponding features. A source of information that provides such less-than-perfect reports is called *partially reliable*. The next section presents and discusses a first model for accounting for source reliability in situations such as those described above. The terms 'reliable' and 'reliability' are used in the sense in which they are understood in the academic literature and not with reference to legal standards as mentioned in Section 1. In addition, unless otherwise stated, the term 'source' is used to refer to a source of information that issues a report, not to a source of physical or digital traces or materials.

3.2 A minimal model for exogenously defined reliability

In probabilistic epistemology, the most basic formal account of the reliability of a source of information involves two variables, here called RP and F. The variable RP is binary and represents a report about a disputed fact or event F that is not directly observable by the person receiving the report. The variables RP and F are sometimes referred to as the report variable and the fact variable, respectively. Note that the focus of this article is on the processing of so-called positive reports, i.e. reports that affirm the occurrence of some fact or event F. Thus, the negation $\overline{RP}$ will not be examined in detail here. It is used generically to denote the absence of a report of type RP, i.e. either no report or a report that makes some other statement.

More broadly, the event F can itself be seen as a potential consequence of some higher-level proposition, commonly denoted H. This extends the basic model to 'hypothesis testing' (Bovens and Hartmann 2003), i.e. inference about hypotheses that cannot be directly investigated, but studied through measurable consequences. Thompson et al. (2003) discussed this view in the context of inference about the source of biological traces in forensic science. In their account, the report variable refers to a scientist's report of an observed correspondence between the characteristics of materials of unknown and known origin. As anticipated in Section 3.1, the unknown and unobserved event is whether or not the compared materials *actually* have corresponding features. This variable, in turn, represents the testable consequence of the hypothesis that the compared items come from the same source. See also Koehler et al. (1995) for a similar development, and Thompson (2016) for an example on scientific assessment in national intelligence investigations.

Figure 1 shows a graphical representation of the relevance relationships between the three variables RP, F, and H (Taroni et al., 2004, 2006). Interpreted probabilistically, the model

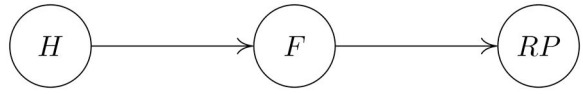

Figure 1. Graphical representation of the relevance relationships between the three binary variables *RP*, short for 'reported correspondence', *F*, short for 'the compared items have corresponding features', and *H*, short for 'the compared items come from the same source'.

represents a Bayesian network (Jensen 1996; Cowell et al., 1999). In such a model, an edge from one node, called a parent node, to another node, called a child node, means that knowledge of or information about the proposition represented by the former node is relevant to the assessment of the truth or falsity of the proposition represented by the latter node. The probabilistic relationship between pairs of adjacent nodes is modelled in terms of probability tables, one for each node. For example, a child node with exactly one parent node has a table that specifies a probability for each combination of a state of the child node and a state of the parent node. Throughout this article, node probability tables will play a central role in defining and discussing the concepts of reliability and accuracy.

The structure shown in Fig. 1 is valuable because it allows one to distinguish the reliability of an information source from other aspects, such as *feature selectivity* (or, *diagnosticity*). The diagnostic capacity of analytical features is modelled in terms of the conditional probabilities assigned to the node F, the event that the compared items have corresponding features. That is, the occurrence of corresponding features (event F) speaks to the proposition that the compared items come from the same source (H), rather than from two different sources ($\overline{H}$), whenever $\Pr(F\,|\,H,I) > \Pr(F\,|\,\overline{H},I)$.[10] The ratio of $\Pr(F\,|\,H,I)$ and $\Pr(F\,|\,\overline{H},I)$, denoted $\mathrm{V}(F;H\,|\,I)$, is well known in the forensic statistics literature as a measure of the value of knowledge of the event F for discriminating between the competing propositions H and $\overline{H}$ (e.g. Aitken et al., 2020), and similarly in legal literature (e.g. Friedman 2017). In medical literature, the event F can be interpreted as a symptom that is, to some extent, indicative of a particular disease in a patient (event H).

It could be argued that the probabilistic structure $H \rightarrow F$ models one form of reliability, i.e. feature reliability, interpreted in the following sense. A given corresponding feature combination, event F, is an indicator of proposition H whenever $\mathrm{V}(F;H\,|\,I) > 1$. More generally, for real-world feature (or, symptom) types, $\Pr(F\,|\,\overline{H}) > 0$, so that F is only a *partially reliable* (i.e. imperfect) indicator of H. This makes intuitive sense, since for a feature configuration to be a perfect indicator of H it would have to *never* occur when H is false, i.e. $\Pr(F\,|\,\overline{H}) = 0$. Of course, these considerations are basic results in probabilistic reasoning, described in a wide range of literature. Nevertheless, the results are stated here because they provide a way of interpreting reliability in formal terms, which serves as a basis for constructing various extensions in the rest of this article. Note that possible synonyms for feature reliability are feature *selectivity* and feature *diagnosticity*.

The above logic for feature reliability extends analogously to *source* or *examiner reliability*, represented here in terms of the probabilistic structure $F \rightarrow RP$. That is, the report RP given by an information source, which may be a measuring instrument with or without human involvement, is of value whenever $\Pr(RP\,|\,F) > \Pr(RP\,|\,\overline{F})$, i.e. the likelihood ratio $\mathrm{V}(RP;F\,|\,I)$ is greater than one.[11] In forensic science literature, the probability $\Pr(RP\,|\,\overline{F})$ has been referred to as the false positive probability *fpp* (Thompson et al., 2003). The term $\Pr(RP\,|\,F)$ is the probability of a hit (e.g. Schum 1994) and is equal to one minus the probability of a 'miss', $\Pr(\overline{RP}\,|\,F)$. Again, there is nothing spectacular or controversial here: the concepts and notation are widely used by advocates of the logical approach to evaluating scientific evidence.

However, challenges arise when trying to relate this formal framework to practical applications. In particular, a crucial question that arises here is how to assign values to the conditional probabilities $\Pr(RP\,|\,F)$ and $\Pr(RP\,|\,\overline{F})$. The answer to this question depends on how these terms are interpreted in an applied context. Two perspectives will be discussed here. Both invoke terminology that is on a slippery slope towards the notion of accuracy. That is, $\Pr(RP\,|\,F)$ and $\Pr(RP\,|\,\overline{F})$ are interpreted as expressions of the *accuracy* of the statement made by the source of information. The idea is to say that the greater (smaller) the probability that a source of information gives a report RP when the target event F holds (does not hold), the more accurate the

[10] The variable I represents other background and circumstantial information relevant to the assignment of probabilities, but is not explicitly represented as a node in the model. For simplicity, the variable I will be omitted from the notation in the remainder of this article.

[11] We leave aside the case of awkward sources of information that work in the opposite way, i.e. when their report is more probable to occur when $\overline{F}$ is true than when F is true. An example of such a source of information would be a weather forecaster who is more inclined to announce rain when it will *not* rain than when it will. The report from such a source of information would still have some value, but simply not in the usual sense.

source of information and hence the greater its validity. But how exactly is this account of accuracy to be understood and evaluated?

One way is to take an empirical view. The obvious way to get quantitative assignments for $\Pr(RP|F)$ and $\Pr(RP|\overline{F})$ would be to look at *how often* the information source provides *RP* when in fact *F* or $\overline{F}$ holds, respectively. As noted in Section 2, relevant data for this purpose can be found in so-called *validity* or *proficiency studies*,[12] which provide accuracy rates (Cole 2006). However, this approach is not as straightforward as it may seem. Most importantly, accuracy rates are only summary statistics that characterize the aggregate case. Strictly speaking, they cannot be used *directly* as assignments for $\Pr(RP|\cdot)$, because *RP* stands for a singular event, although these assignments can be used as proxies. The deeper point here is that a rate (or relative frequency), i.e. a summary of data, is not a probability (Lindley 1985), although shortcuts to the contrary are widely found in practice and even in the scientific literature. At best, (relative frequency) data can inform, but not fully determine, the assignment of probability, unless one adheres to the frequentist interpretation of probability. The latter is not recommended, however, given the numerous conceptual and operational limitations of frequentism (Biedermann 2015; Biedermann and Vuille 2018; Taroni et al., 2018). But even if one accepts the simplistic definition of probability as relative frequency, the assignment of probability may not be immediate, automatic or obvious. It is still necessary to justify that the case at hand can reasonably be regarded as an example of the types of cases included in the particular validity study, i.e. whether the present case falls within the range of experimental conditions that characterize the validity study. This requires an examination of what Imwinkelried (2020) has called the range of validation. And however that examination is conducted, it ultimately rests on a judgment that amounts to what Stoney has called, albeit for a slightly different purpose, a 'leap of faith' (Stoney 1991). In both theory and practice, this has proved to be the controversial bottleneck in the use of frequency data for assigning probabilities in individual cases. At one extreme, the disputes over coverage by the range of validation, i.e. the suitability of candidate data for the case at hand, can be seen as an instance of the contentious reference class problem (Roberts 2007). In one of its interpretations, this problem boils down to the view that the most appropriate data (i.e. reference class) for a given case is 'the very event under discussion' (Allen and Pardo 2007: 109), leading to analysis paralysis.

Another limitation to bear in mind is that, strictly speaking, common proficiency studies do not provide information about $\Pr(RP|F)$, but rather about $\Pr(RP|H)$. Because proficiency or black-box studies collect examiners' conclusions for comparison pairs known to come from either the same source or from different sources, the conditioning is on a pair of source-level propositions *H*, not on the event of corresponding features *F*. However, it could be argued that experiments are designed so that same and different source pairs have, respectively, corresponding and non-corresponding combinations of features, so that data of the type $RP|H$ can be assimilated to data of the type $RP|F$. But a definitional problem remains: mainstream black-box studies do not, strictly speaking, record reported correspondences, but rather the examiners' direct opinions on source-level propositions, so-called identification conclusions, which makes it difficult to adapt the resulting data to the needs of specifying the formal model discussed here.

A second way of assigning values to $\Pr(RP|F)$ and $\Pr(RP|\overline{F})$ is to adopt a pragmatic position. It does not ignore the difficulties mentioned above, but argues that summary statistics derived from validation or proficiency studies—with design characteristics that may not exactly correspond to those of the case at hand—could at least provide an anchor for assessments in individual cases.[13] Thus, in order to avoid a dead end at the above hurdles, pragmatism in probability assessment seems essential if the epistemological account of reliability is to get off the ground in practical applications. That is, either one is *able* to articulate assessments for the key quantities of interest, despite the conceptual and practical obstacles, or one must accept that the reliability of the source of information remains undefined and that the information output cannot be used. This would amount to a case of uninterpretability (Biedermann and Kotsoglou 2022).

[12] "A validity study is designed to measure the accuracy of a scientific technique" (Imwinkelried, 1995, at p. 1254).

[13] For an example of a proponent of this view, see e.g. Koehler (2008): "the industrywide error-rate estimates provide anchors or proxies for judgments about the risks of error in individual cases" (at pp. 1088–89).

But beyond the practical challenge of pinpointing specific numbers or orders of magnitude, the structural properties of the exogenous reliability model are important at a more fundamental level. In particular, the model makes it clear that reliability is not a one-off assessment, but is decomposed into at least two[14] levels of assessment. One of these levels accounts for the diagnosticity (or, alternatively, the reliability or selectivity) of the features, and the other captures the performance of the operator or examiner.[15]

Another key aspect of the model is that it clarifies how the component assessments of reliability interact and together affect the probative value $\mathrm{V}(RP;H\,|\,I)$. To reduce this expression to its essentials, consider the following simplifying assumptions. Given H, the probability of corresponding features F is close to 1, and hence $\overline{F}$ given H is close to zero. Given corresponding features, event F, the probability of a 'miss' ($\overline{RP}$) is small, so the probability of RP given F, the sensitivity of the information source, tends to 1. Also, denote by γ the rarity of the corresponding features, $\Pr(F\,|\,\overline{H},I)$, and by fpp the false positive probability $\Pr(RP\,|\,\overline{F},I)$. The likelihood ratio $\mathrm{V}(RP;H,I)$ is thus simplified as follows (Thompson et al., 2003):

$$\mathrm{V}(RP;H,I)=\frac{\overbrace{\Pr(RP\,|\,F,I)}^{\to 1}\overbrace{\Pr(F\,|\,H,I)}^{\approx 1}+\overbrace{\Pr(RP\,|\,\overline{F},I)}^{fpp}\overbrace{\Pr(\overline{F}\,|\,H,I)}^{\approx 0}}{\underbrace{\Pr(RP\,|\,F,I)}_{\to 1}\underbrace{\Pr(F\,|\,\overline{H},I)}_{\gamma}+\underbrace{\Pr(RP\,|\,\overline{F},I)}_{fpp}\underbrace{\Pr(\overline{F}\,|\,\overline{H},I)}_{(1-\gamma)}}\approx\frac{1}{\gamma+fpp(1-\gamma)} \tag{1}$$

This result shows that, contrary to what is widely suggested, the probative value of a report is not reduced to the inverse of the rarity metric of the corresponding features, but to a smaller value depending on the false positive probability. In particular, for types of evidence such as DNA, where values for γ in the order of one in a billion or much smaller are claimed, $\mathrm{V}(RP;H\,|\,I)$ is dominated by 1 over fpp.

4. Endogenously defined reliability

4.1 Preliminaries: the practical limitations of exogenously defined reliability

A key assumption in relation to the model discussed in Section 3 is that the performance characteristics of the information source, be it human, machine or a combination of the two, are known and fixed, or at least assignable based on available aggregate accuracy rates (e.g. from validation studies) for the purpose of the case at hand. Broadly speaking, these performance characteristics can be loosely assimilated to component metrics such as sensitivity and specificity used in the context of diagnostic tests. However, in many practical situations, detailed performance characteristics at this level of resolution are not available due to a lack of data from appropriate validation studies. This problem is not specific to forensic science, but is well known in the context of epistemology and philosophy of science. Researchers in these fields have proposed an alternative way of dealing with the question of reliability, based on so-called *endogenously defined reliability* (Bovens and Hartmann 2003). This perspective treats reliability in a more general way. An earlier presentation of this modelling approach can also be found in Bovens and Hartmann (2002), focusing on how to draw conclusions from the results of what they call 'less than fully reliable (LTFR) instruments' (p. 29). This section extends the notion of endogenously defined reliability and illustrates it in the context of forensic inference. The presentation here differs slightly from Bovens and Hartmann (2002) in that the focus is on the likelihood ratio of the output of the information source. Bovens and Hartmann (2002) focused on the degree of confirmation given by the difference between the posterior and the prior probability of the target proposition of interest.

[14] Other authors, in particular Schum (1994), with his account of multiple attributes of credibility, argue for more fine-grained decompositions of more than two levels of assessment.

[15] More generally, an information source can be a human, a machine or a combination of the two.

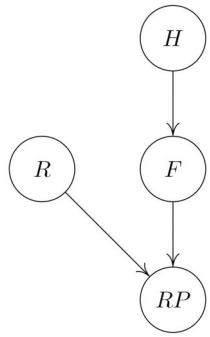

Figure 2. Graphical representation of the relevance relationships between the binary variables *RP*, short for 'reported correspondence', *F*, short for 'the compared items have corresponding features', *H*, short for 'the compared items come from the same source', and *R*, representing the reliability of the information source.

4.2 A model for endogenously defined reliability

The main idea of endogenously defined reliability is to think about reliability in a broader sense. That is, instead of expressing performance characteristics with detailed assignments, such as the false positive probability in the model for exogenously defined reliability (Section 3), one makes an overall assessment of the reliability of the information source. Technically speaking, exogenously defined reliability amounts to fixed conditional probability assignments in the node probability table of the variable *RP* ('reported correspondence'), whereas the endogenous definition of reliability amounts to the introduction of a separate binary variable, here denoted *R* (short for 'reliability').[16] As shown in Fig. 2, the variable *R* is introduced as a parent for the node *RP*. The rest of the model, in particular the network fragment $H \rightarrow F$, has the same definition as the model shown in Fig. 1, discussed in Section 3.

One way of thinking about the definition of the node *R* is to consider it in terms of the proposition that the information source is a truth-teller (Bovens and Hartmann 2003). Table 1 illustrates what this means. Clearly, if the information source is fully reliable, then *RP* is true if *F* is true: $\Pr(RP \mid R, F) = 1$. Similarly, a fully reliable information source would *not* report corresponding features if the compared items do not have corresponding features: $\Pr(\overline{RP} \mid R, \overline{F}) = 1$. Conversely, if the information source is not reliable, then it is not informative about whether or not the items have corresponding features. That is, if $\overline{R}$ is true, then the probability of obtaining a report of corresponding features would be the same whether *F* or $\overline{F}$ is true: $\Pr(RP \mid R, F) = \Pr(RP \mid R, \overline{F})$. Let this probability be denoted by α. The value to be assigned to α is not important at this stage. What is important at this point is that assigning a given value to α in the case of no reliability ($\overline{R}$) implies that the likelihood ratio $\mathrm{V}(RP; \{F, H\} \mid \overline{R})$ is 1. As an aside, note that Bovens and Hartmann (2002) refer to the information source under $\overline{R}$ as a 'randomizer' (p. 33) and to α as the 'randomization parameter' (p. 33).

To complete the numerical specification of the model shown in Fig. 2, it is necessary to provide probabilities for the root node *R*. Let $\Pr(R)$ be denoted *r* for short. Furthermore, $\Pr(\overline{R})$ is given by $1 - r$. Broadly speaking, the probability *r* represents a person's degree of belief in the proposition that the information source is reliable. This and all other probabilities used in this article are considered to be operationally defined (Lad 1996), following the work of de Finetti (e.g. 1937, 1939, 1974).[17]

Given the above definitions, it is possible to examine the likelihood ratio for a report *RP* with respect to different propositions of interest, taking into account uncertainty about the reliability of the information source. For example, consider the probative value of a report *RP* with respect to the proposition that the items being compared actually have corresponding features (*F*).

[16] Such model structures have also been discussed in the context of forensic evidence. An example is given by Fenton et al. (2013). However, this example will not be pursued here because it uses different terminology, such as "accuracy" instead of "reliability", which interferes with the deductive definition of individual report accuracy used later in this article (Section 6).

[17] See also the discussion in e.g. Lindley (1982, 1985) on the primacy of probability theory over alternative concepts such as fuzzy logic.

Table 1. Conditional probabilities assigned to the binary variable *RP*, representing the report of an information source, depending on the variables *R*, the reliability of the information source, and *F*, the proposition that the compared items have corresponding features.

	R		$\overline{R}$	
	F	$\overline{F}$	F	$\overline{F}$
RP	1	0	α	α
$\overline{RP}$	0	1	$1-\alpha$	$1-\alpha$

Based on the relevance relationships encoded by the model shown in Fig. 2, the likelihood ratio can be written and simplified as follows:

$$\begin{aligned} \mathrm{V}(RP;F) &= \frac{\overbrace{\Pr(RP \mid F,R)}^{1}\overbrace{\Pr(R)}^{r} + \overbrace{\Pr(RP \mid F,\overline{R})}^{\alpha}\overbrace{\Pr(\overline{R})}^{1-r}}{\underbrace{\Pr(RP \mid \overline{F},R)}_{0}\Pr(R) + \underbrace{\Pr(RP \mid \overline{F},\overline{R})}_{\alpha}\Pr(\overline{R})} \\ &= \frac{r+\alpha(1-r)}{\alpha(1-r)}\,. \end{aligned} \tag{2}$$

At first sight, this result may seem cryptic. However, it can readily be seen that this likelihood ratio has intuitively reasonable properties. For example, suppose that the reliability is maximal: $r = 1$. In this case, the likelihood ratio is infinite. This result is a direct consequence of the 'truth-teller'-properties specified in Table 1. That is, a reliable information source (i.e. R is assumed to be true) is one that is assumed to *never* report RP when $\overline{F}$ is the case: $\Pr(RP \mid R, \overline{F}) = 0$. In such a situation, the likelihood ratio is entirely determined by the first two rows of probabilities in Table 1. In other words, the case where F is true is the only situation in which a reliable information source (proposition R) will report the observation of corresponding features (RP). Conversely, if the information source is assumed to be completely unreliable, i.e. $r = 0$, then Equation (2) reduces to $\alpha/\alpha = 1$. As mentioned above, this can also be read directly from Table 1. If $\overline{R}$ is true, then the likelihood ratio is entirely determined by the probabilities given in the last two columns. Thus, in the two extreme cases of R being true or false, there is no need to inquire about what value to assign to α.

However, the likelihood ratio $\mathrm{V}(RP;F)$ is of rather limited interest. When we receive an expert's report that corresponding features have been observed (proposition RP), we are less interested in what such a report can tell us about whether or not the compared items actually have corresponding features (proposition F) than in what such a report can tell us about whether the compared items come from the same source (proposition H). The latter question requires the likelihood ratio $\mathrm{V}(RP;H)$. Nevertheless, clarity about the analytic form and properties of $\mathrm{V}(RP;F)$ is helpful in extending the considerations to $\mathrm{V}(RP;H)$.

The likelihood ratio for a report RP with respect to the source-level propositions H and $\overline{H}$ is somewhat more complex, because it takes into account uncertainty about both propositions R and F:

$$\mathrm{V}(RP;H) = \frac{\sum_{R,F}\Pr(RP \mid R,F)\Pr(R)\Pr(F \mid H)}{\sum_{R,F}\Pr(RP \mid R,F)\Pr(R)\Pr(F \mid \overline{H})}\,.$$

Using the notation and probability assignments introduced so far in this article, this likelihood ratio can be simplified to:

$$\mathrm{V}(RP;H) = \frac{r+\alpha(1-r)}{r\gamma+\alpha\gamma(1-r)+\alpha(1-r)(1-\gamma)}\,. \tag{3}$$

Similar to what was considered above in relation to Equation (2), it is possible to examine selected situations. For example, suppose that the information source is completely reliable ($r = 1$). In such a situation, $V(RP; H)$ reduces to $1/\gamma$, a well-known form of the likelihood ratio for forensic inference of source problems (e.g. Evett and Weir 1998; Robertson et al., 2016). This makes sense, because if the source of information is fully reliable, then, as seen when examining the properties of (2), obtaining the report RP that the compared items have corresponding features implies that proposition F is true, i.e. that the compared items do indeed have corresponding features. In turn, the value of the information that F is true with respect to the proposition H that the compared items come from the same source is given by the source-level likelihood ratio of $1/\gamma$. In other words, for a fully reliable information source, the likelihood ratio depends only on the diagnosticity of the features.

In summary, the reliability node R 'regulates', so to speak, the effect of the report variable RP with respect to the nodes F ('the compared items have corresponding features') and H ('the compared items come from the same source'). A report RP will at best provide a likelihood ratio that is the inverse of the rarity of the corresponding features if the report comes from a fully reliable source.

5. Comparison between the models for exogenously and endogenously defined reliability

5.1 Structural and definitional properties

The models of exogenously and endogenously defined reliability presented in Sections 3 and 4 differ in the way they deal with the inferential step of going from an information source's report that a pair of compared items have corresponding features (RP) to the proposition that the compared items actually have corresponding features (F). As shown in Fig. 3, this inference step is taken care of by the lower part of the two models. This part is concerned with the ability of an examiner (or instrument) to correctly recognize some aspect of the real world. In other words, the focus here is on the reliability (or: diagnosticity) of the *examiner* (or instrument). This aspect is to be distinguished from the capacity—or: selectivity (diagnosticity)[18]—of the *features*. It is modelled by the network fragment $H \rightarrow F$, which is the same in both models. Feature selectivity characterizes the potential of the features to discriminate between the target propositions (here: whether the compared items come from the same source).

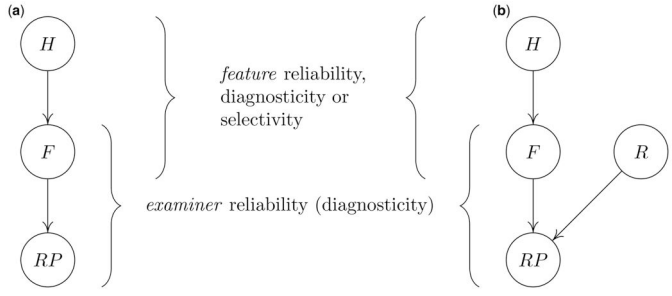

Figure 3. Comparison between the graphical models for (a) exogenously defined reliability (Section 3) and (b) endogenously defined reliability (Section 4). The nodes RP represent the proposition 'the information source reports that the compared items have corresponding features', the nodes F represent the proposition 'the compared items have corresponding features', and the nodes H represent the proposition 'the compared items come from the same source'. The node R represents the reliability of the information source. All nodes are binary.

[18] Broadly speaking, feature selectivity or diagnosticity can be viewed as a function of the rarity of the features in the population of interest: i.e. the less common a feature or combination of features is, the more valuable it is in helping to discriminate between competing propositions regarding the source(s) of a pair of items being compared.

5.2 A side note on black-box studies

The models discussed in this article are, in a sense, minimal models, because the chain of reasoning between the nodes RP and H can be further broken down into a variety of intermediate considerations. Schum (1994) discusses how this can be done by considering different attributes of the credibility of an information source. Nevertheless, the models discussed here are still more sophisticated than the level of resolution underlying the currently widely promoted idea of forensic black-box studies. As briefly mentioned in Section 1, in such studies 'many examiners render decisions about many independent tests (typically, involving 'questioned' samples and one or more 'known' samples) and the error rates are determined' (PCAST 2016: 5, 6). Structurally, this amounts to a basic two-node network fragment of the type $H \rightarrow RP'$, which represents a simplified instance of the model for exogenously defined reliability. Here, while H has the same definition as in the models shown in Fig. 3, the node RP' is defined as the examiner's reported *identification* conclusion, i.e. a statement about whether (the examiner thinks) the items being compared came from the same source. Note that this differs from the definition of RP used so far in this article, which is limited to an examiner's report of the mere *observation* of corresponding features between the compared items. In particular, note that a graphical model of the type $H \rightarrow RP'$ for black box study data represents a shortcut approach: there are no intermediate steps in going from an expert's report (RP') to a source conclusion (H). In particular, the selectivity of the features of the examined evidential material is completely ignored. The result is a fundamental shift in emphasis. The focus is no longer on the selectivity, and hence the probative value, of the features of the examined *physical* items of evidence, coupled with considerations of examiner/instrument performance. Instead, the focus is boiled down to the expert's utterance alone, and the diagnosticity of that utterance with respect to the source-level propositions, *regardless* of how selective the features of the examined items are. This amounts to feature agnosticism, because one would assign the *same* probative value to all expert utterances of identification. The result may be an over- or under-evaluation with respect to the informative value derived from the selectivity of the observed features, as well as a misrepresentation of the performance of those examiners who deviate from the average.

5.3 Analytical properties

From what has been discussed so far in this article, the two models for exogenously and endogenously defined reliability reflect two different perspectives. The model for exogenously defined reliability (Section 3) allows one to assign probabilities for well-defined events, such as the probability of a false positive (Thompson et al. 2003), especially when data are available from studies of expert performance. The model for endogenously defined reliability (Section 4), on the other hand, provides a way of characterizing reliability more broadly, i.e. without going into detailed assessments such as the probability of a false positive. Instead, it treats reliability as the general proposition that the source of information is a truth-teller (Bovens and Hartmann 2003).

While these definitional and structural differences lead to different expressions of the likelihood ratio (see Equations (1) and (3)), used to quantify the probative value of an expert's report, the two models do not necessarily behave differently in all cases. For example, it has already been noted in previous sections that, in the case of a completely unreliable source of information, both models lead to the reasonable result of a likelihood ratio of 1. In the optimal case, both models lead to a likelihood ratio bounded by the inverse of the rarity of the corresponding features. This raises the question of how the two models compare more generally.

One way of doing this, in addition to examining extreme cases as in the previous sections, is to study model behaviour through sensitivity analysis. This is particularly important for the model for endogenously defined relevance because it requires a probability that seems difficult to assign: the so-called randomization parameter α. Assigning a value to α amounts to asking what our probability is that an information source known *not* to be a truth-teller will issue a report RP. This is a difficult task because it is not clear how one could even think of an experiment to generate data that could help assign such a probability, not least because the event to which the assignment refers is unique and therefore unrepeatable.

For the sensitivity analysis of the model for endogenously defined relevance, suppose that the rarity term of the corresponding features is $\gamma = 0.05$. The probability of reliability r is varied

between 0 (the information source is completely *un*reliable) and 1 (the information source is completely reliable). For the randomization parameter α, the following assignments are chosen: 0.001, 0.01, 0.1, 0.5, 1. Figure 4a shows the likelihood ratio as a function of $\Pr(R)$ for different values of α. Clearly, for a perfectly reliable information source ($\Pr(R) = 1$), the likelihood ratio $V(RP;H)$ is $1/\gamma = 20$, regardless of the value of α. However, for less than perfectly reliable information sources ($\Pr(R) < 1$), the likelihood ratio is *lower* the *higher* the value of the randomization parameter α. Again, this property is plausible because the higher the tendency of an unreliable source to report RP, the less diagnostic a report RP becomes. In particular, for very low values of α, such as 0.001 shown in Fig. 4a, the likelihood ratio may still be close to the upper bound of $1/\gamma$ even when the probability of reliability is low.

Compare these results with a sensitivity analysis for the model where reliability is defined exogenously. Again assume that the rarity term of the corresponding features is $\gamma = 0.05$. Figure 4b shows the likelihood ratio $V(RP;H)$ as a function of the false positive probability $\Pr(RP \mid \overline{F})$, i.e. the probability that the information source reports RP (the observation of corresponding features) even though the items been compared do not have corresponding features ($\overline{F}$). A likelihood ratio of 1 is obtained in a case where the false positive probability is maximal (i.e. $\Pr(RP \mid \overline{F}) = 1$). Conversely, the likelihood ratio reaches the upper bound of $1/\gamma = 20$ when the probability of the information source issuing a false positive report is zero ($\Pr(RP \mid \overline{F}) = 0$).

Interestingly, the likelihood ratio curve for $\alpha = 1$ in Fig. 4a is a perfect mirror image of the curve in Fig. 4b. That is, the effect of the probability of reliability in the model for endogenously defined reliability (when $\alpha = 1$) is the inverse of that of the false positive probability in the model for exogenously defined reliability: i.e. a high (low) probability of reliability affects the likelihood ratio in exactly the same way as a low (high) false positive probability.

This parallel between the models for exogenously and endogenously defined reliability can be demonstrated formally. One can take Equation (3), the likelihood ratio derived from the model for endogenously defined reliability, set $\alpha = 1$ and simplify to obtain:

$$V(RP;H) = \frac{1}{1 - r(1-\gamma)}.$$

Next, one can replace r with $1 - fpp$ and rearrange the terms in the denominator to obtain Equation (1), the likelihood ratio derived from the model for exogenously defined reliability. In summary, therefore, Equations (1) and (3) are equivalent when $r = (1 - fpp)$ and $\alpha = 1$.

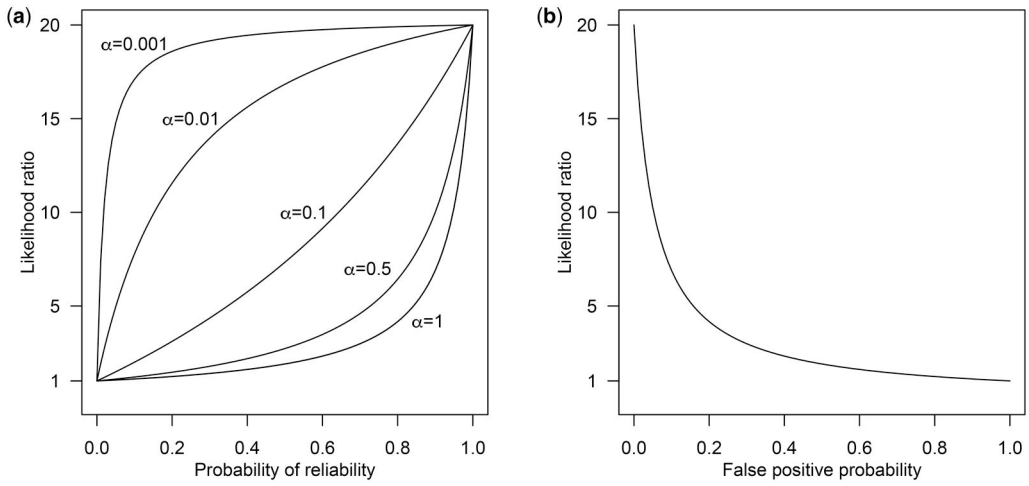

Figure 4. Sensitivity analyses for the likelihood ratio $V(RP;H)$ using the model for (a) endogenously defined reliability and (b) exogenously defined reliability. The rarity term of the corresponding features γ was set to 0.05. (a) shows the likelihood ratio as a function of the probability of reliability and for different values α of the randomization parameter. (b) shows the likelihood ratio as a function of the false positive probability.

Table 2. Conditional probabilities assigned to the binary variable A, the accuracy of a report RP regarding the occurrence of a statement regarding a fact variable F.

	RP		$\overline{RP}$	
	F	$\overline{F}$	F	$\overline{F}$
A	1	0	0	1
$\overline{A}$	0	1	1	0

6. Accuracy in the instant case and the accuracy fallacy

The two models of reliability discussed in Sections 3 and 4 can serve as a basis for approaching the notion of the *accuracy of an individual report* provided in a particular case, and for clarifying the functional relationship between this notion and reliability. Note that the notion of individual report accuracy developed here differs from the notion of *accuracy in the aggregate case*, which the PCAST report considers to be a component of reliability. Indeed, the PCAST report (PCAST 2016) defines 'accurate' to mean that, 'with known probabilities, an examiner obtains correct results both (1) for samples from the same source (true positives) and (2) for samples from different sources (true negatives)' (p. 47). In turn, the PCAST defines 'reliability' as 'repeatability, reproducibility, and accuracy' (p. 47). These definitions characterize a method or an examiner in general terms, which may be helpful during admissibility proceedings. However, these definitions do not address the question of what exactly one should conclude once a method or examiner has been deemed reliable enough to be admitted and a report has been provided. More specifically, when one receives a report from an information source for which accuracy *rates* are available, a crucial question of interest may be how sure one can be that the report aligns with the respective ground truth, i.e. that it is accurate *in the case in question*. Unfortunately, aggregate case accuracy does not answer this question. To equate the two would be to fall into what might be called the 'accuracy fallacy'.

Accuracy in the instant case, as understood here, refers to the congruence between the proposition RP, the report of a correspondence, and the unknown state of nature F, the fact variable (i.e. the ground truth state), to which the report refers. This understanding can be illustrated by adding an extension to the graphical models developed in the previous sections. Specifically, let A denote the accuracy of a report RP with respect to F. A report RP is certainly accurate if F is true. Therefore, $\Pr(A \mid RP, F) = 1$. This reflects the understanding that RP and F together logically entail accuracy. Conversely, a report RP is *not* accurate, $\overline{A}$, if $\overline{F}$ holds: $\Pr(\overline{A} \mid RP, \overline{F}) = 1$. This also implies that $\Pr(A \mid RP, \overline{F}) = 0$. Table 2 gives a summary of all conditional probability assignments for the variable A.

It is immediately apparent that the notion of *individual report accuracy*, denoted A here, differs from the classical (diagnostic) performance metric of *sensitivity*. The latter is defined as the proportion of test cases where F holds and the information source produced the statement RP. It is also sometimes used as a proxy for the so-called hit probability $\Pr(RP \mid F)$. The two should not be confused. More formally, the inequality can be written as $\Pr(A \mid RP) \neq \Pr(RP \mid F)$.[19] But if the probability of an individual report being accurate is not the same as the hit probability, what is it?

To clarify this issue, some further comments on the above distinction are in order. First, note that the probability of individual report accuracy $\Pr(A \mid RP)$ is conditioned on knowing RP, while *not* knowing F. The opposite is true for the hit probability $\Pr(RP \mid F)$, which *is* conditioned on F. Secondly, given that F is generally not known, the probability of report accuracy can be further developed and simplified, using values defined in Table 2, as follows (omitting information I for simplicity):

[19] Note that the two terms are *formally* different, but *numerically* they can take the same value under certain conditions (i.e. probability assignments).

$$\begin{aligned}\Pr(A \mid RP) &= \underbrace{\Pr(A \mid RP, F)}_{1}\ \Pr(F \mid RP) + \underbrace{\Pr(A \mid RP, \overline{F})}_{0}\ \Pr(\overline{F} \mid RP) \\ &= \Pr(F \mid RP).\end{aligned} \tag{4}$$

This result states that the accuracy of a report is equal to the posterior probability that the fact variable is true, given knowledge of the report RP. This probability is given by Bayes' theorem:

$$\Pr(F \mid RP, I) = \frac{\Pr(RP \mid F, I)\Pr(F \mid I)}{\Pr(RP \mid F, I)\Pr(F \mid I) + \Pr(RP \mid \overline{F}, I)\Pr(\overline{F} \mid I)} \tag{5}$$

Equation (5) shows that the probability of F given RP, and hence the accuracy of RP, depends not only on what is known about F based on other information I (prior to learning RP), but also on the performance metrics of the source of the report RP, in particular the hit probability $\Pr(RP \mid F, I)$ and the false positive probability $\Pr(RP \mid \overline{F}, I)$. This shows that the probability of individual report accuracy $\Pr(A \mid RP)$ is not the same as the hit probability (or, sensitivity), $\Pr(RP \mid F)$, but is a function of it.

Note that $\Pr(A \mid RP) = \Pr(F \mid RP)$ is not an initial assumption, but an insight derived from other, more basic assumptions about the structural dependencies among the key variables A, RP, and F. Note also that although assessing individual report accuracy (Equation (4)) may seem to be merely a matter of using Bayes' theorem to find the posterior probability of the ground truth variable F (Equation (5)), this does not lead back to the use of aggregate case data (i.e. relative frequencies). Equation (5) does not primarily ask for aggregate case data, but rather for *probabilities* for single, non-repeatable events, such as the hit probability $\Pr(RP \mid F)$. Relative frequencies merely *inform* probability assessment, but do not define them (Lindley 2006).

The above distinction between $\Pr(A \mid RP)$ and $\Pr(RP \mid F)$ is of value in avoiding a presumably common logical fallacy, which may be called the *accuracy fallacy*. This fallacy consists in incorrectly concluding that a case-specific report RP from an information source with a high hit probability (i.e. sensitivity) is necessarily highly accurate, because it is (incorrectly) assumed that $\Pr(RP \mid F)$ is equal to $\Pr(A \mid RP)$.[20] Informed readers will recognize this fallacy as closely related to the false positive fallacy described by Thompson et al. (2003). The false positive fallacy also leads to a false conclusion of high report accuracy, but based on a low false positive probability. That is, it is assumed that $1 - \Pr(RP \mid \overline{F}, I)$ equals $\Pr(F \mid RP, I)$, and therefore, in the context of this article (Equation (4)), $\Pr(A \mid RP, I)$.

To illustrate the above elements, it is helpful to present a numerical example based on an extension of the graphical model described in Fig. 1. The extension consists in the addition of a node A representing accuracy. Recall that the variable A monitors the congruence between the variables PR and F, as defined in Table 2. This interpretation of the notion of accuracy implies that the node A has arcs coming from PR and F respectively, as shown in Fig. 5a.[21] For the purpose of illustration, consider a hit probability $\Pr(RP \mid F, I)$ of 0.9 and a false positive probability $\Pr(RP \mid \overline{F}, I)$ of 0.05. For the fact variable F, make the reasonable assumption that it is certain that the items being compared have corresponding features if they come from the same source: $\Pr(F \mid H, I) = 1$. In the case where the items being compared come from different sources, $\overline{H}$, the probability of occurrence of corresponding features, $\Pr(F \mid \overline{H}, I)$, is assigned as $\gamma = 0.01$. Finally, again for illustrative purposes, consider prior probabilities of $\Pr(H \mid I) = 0.1$ and $\Pr(\overline{H} \mid I) = 0.9$ for the proposition that the compared items come (do not come) from the same source.

Using these assignments in Equation (5), one obtains the posterior probability of the fact variable F, and hence the probability of accuracy A, *given* the report RP, as $(0.9 \times 0.109)/(0.9 \times 0.109 + 0.05 \times 0.891) = 0.6877$. Here 0.109, the value for $\Pr(F \mid I)$, is obtained as $\sum_H \Pr(F \mid H, I)\Pr(H \mid I)$. This example thus shows that although the hit probability, $\Pr(RP \mid F, I)$, is qualitatively high (here: 0.9), the accuracy of the report RP is considerably

[20] Note that, according to Equation (4), $\Pr(A \mid RP)$ is *not* equal to $\Pr(RP \mid F)$, but to $\Pr(F \mid RP)$.
[21] Note that such a node A could also be added to the alternative model for reliability presented in Section 4, which treats the notions of reliability and accuracy *endogenously*.

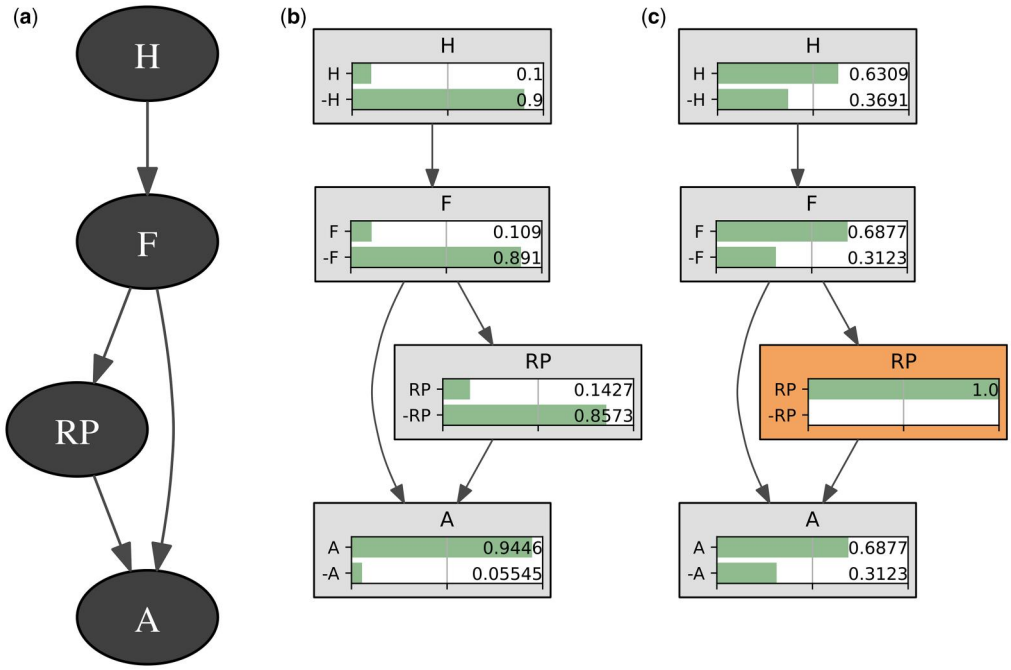

Figure 5. Extension of the Bayesian network shown in Fig. 1, constructed in Python using aGrUM/pyAgrum (version 0.20.2, https://agrum.gitlab.io, see also Ducamp et al. (2020)), by adding a node *A*, representing accuracy (as defined in Table 2): (a) network structure, (b) initial state of the Bayesian network using probability assignments as given in the main text, (c) state of the Bayesian network after instantiation of the node *RP* to *RP*.

lower, i.e. less than 0.7. The reason for this is that the initial probability that the fact variable F is true, $\Pr(F\,|\,I)$, is rather low (i.e. about 0.1). As an aside, note also that the probability of accuracy, $\Pr(A\,|\,RP, I)$, cannot be derived directly from the false positive probability, $\Pr(RP\,|\,\overline{F}, I) = 0.05$, e.g. as $1 - 0.5 = 0.95$, since this would amount to the false positive fallacy (Thompson et al., 2003).

In summary, then, there are two sides to the notion of accuracy that must not be confused. The first is the *probability of obtaining* a report *RP if* what one is trying to prove is true, i.e. the fact variable *F*. This notion of accuracy is also known as the hit probability $\Pr(RP\,|\,F, I)$. In other words, here *F* is fixed and one is thinking about the occurrence of *RP*. The higher the probability of obtaining a report *RP* when *F* is true, the more accurate the information source is said to be. Note that $\Pr(RP\,|\,F, I)$ is a *fixed*, case-specific and *exogenously defined* assignment. The second side of the notion of accuracy in the model discussed here is the probability of accuracy *A* of a report *RP* when it is *not* known whether the fact variable *F* is true. Here *RP* is known but *F* is not. The accuracy of *RP* is *not* a fixed assignment, but is defined deductively, depending on all information relevant to the fact variable *F other* than *RP*.

The bottom line of the above distinction is thus that the probability that an information source *accurately* reports *RP* when *F* is true is not the same as the probability that a report *RP*, once obtained, is accurate when one does *not* know whether *F* is true. In practice, the former type of accuracy is a necessary assignment to be made by anyone who wishes to use the output *RP* of an information source. The latter type of accuracy is *not* directly assigned, but is—in part—a function of the former.

7. Discussion and conclusions

One of the oldest and most persistent concerns about the use of external sources of information in legal contexts, particularly the use of specialized ('expert') witness opinion testimony, is the alignment of the content of expert testimony with ground truth. This concern naturally arises

from the aspiration of trials to ensure factual accuracy (Allen 2013). In recent years, concerns about the accuracy of expert testimony appear to have increased with the emergence of novel algorithms, methods and systems, some of which are commonly and collectively referred to as artificial intelligence (AI).

The interaction between humans and machines in the production of expert testimony raises several interrelated questions. From a purely procedural and descriptive perspective, an immediate question is how contemporary legal systems accommodate expert witness opinion testimony, especially when it is machine-based (or AI-informed). While helpful in addressing technical aspects of actual proceedings, such as the admissibility of proffered expert opinion, such inquiries are of limited help in addressing the *inferential* questions that factfinders face once admissibility has been granted. Mechanisms for reaching *decisions* on the admissibility of experts, such as those found in systems in the common law traditions, thus lead to a void in the sense that they provide no guidance as to how to conceptually process—i.e. *weigh*—an opinion that has been deemed admissible. However, questions of inference are important from both a theoretical and a practical point of view, and include the following: How *should* the trustworthiness of expert witnesses be understood—conceptually? Does such an understanding vary according to the category and nature of expert witness testimony, e.g. whether it is AI-based or not? What if it is integrally reduced to AI output? How, if at all, can an assessment of the trustworthiness of an expert inform an assessment of the accuracy of an individual report in the case at hand? The developments in this article shed new light on these questions.

With regard to probabilistic approaches to the notion of information source reliability, this article reviewed the models for exogenously and endogenously defined reliability (Bovens and Hartmann 2003) from a novel perspective, uncovering hitherto undiscussed analytical relationships between them. While the traditional literature in probabilistic epistemology has analysed and compared models of exogenously and endogenously defined reliability in terms of formulae for posterior probabilities of propositions of interest, the account given here has focused on the likelihood ratio. The analyses carried out here have shown that, under certain well-defined assumptions, the two accounts lead to identical results despite structurally different starting points.

A further main result of the analyses proposed in this article is the conceptual clarification of the latent difference between two distinct assessments: on the one hand, the *general* trustworthiness of an information source (here: an expert) and, on the other hand, the accuracy of a *case-specific* output produced by an information source (here called 'expert output'). The two are not the same, but in both legal scholarship and practice they often appear to be intertwined or even equated. The mainstream heuristic, simply put, seems to be that *if* a source of information is considered trustworthy in the aggregate, so should its individual output (i.e. be considered truthful). However, it is well known—as the inversion fallacy (Diaconis and Freedman 1981)—that this intuition does not translate into probabilistic terms in a straightforward way: if there is a high probability that an information source reports 'A' when 'A' is true, this does not imply that 'A' is (probably) true when the information source reports 'A'.

This article contributes to the clarification of this inferential problem by separating the accuracy of an individual report from what has so far been treated exclusively in terms of the (posterior) probability of the target proposition of interest. The graphical model described here defines report accuracy as a separate Boolean variable, defined deductively as a function of both the report and the actual ground truth. Although the result is computationally unsurprising in the sense that the probability of accuracy (i.e. agreement with the ground truth) is equal to the probability of the target proposition of interest, the value of the model is to show why the latter probabilities cannot be derived directly from empirically defined performance measures, such as the widely used standard statistics of false positives and false negatives. Instead, the accuracy of individual reports is closely related to (i.e. depends on) the beliefs held by the recipient of expert information based on other elements of the case. In this respect, legal provisions are rightly pragmatic in the sense that they recognize the accuracy of individual reports as an empirically inaccessible concept. More specifically, it can be seen that legal provisions dealing with the concept of reliability in terms of aggregate measures of source performance rely on an *empirical proxy* for the accuracy of individual reports.

Overall, the above findings contribute to ongoing debates about reliability by showing that the conventional paradigm of empirical testing, widely used in forensic science and AI disciplines, while valuable, is not sufficient to address the judiciary's primary concern about the accuracy of individual reports—whether produced by a human, a machine, or a combination of the two. These subtle conceptual properties of individual report accuracy thus limit our practical ability to know, let alone control, the extent to which accuracy can be taken for granted in actual cases.

Conflict of interest statement. The author has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article. The author is not an editor or a board member of *Law, Probability and Risk*.

Funding

This work was supported by the Swiss Benevolent Society of New York. Open access publishing was supported by the Consortium of Swiss Academic Libraries (CSAL).

References

Aitken, C. G. G., Taroni, F., and Bozza, S. (2020) *Statistics and the Evaluation of Evidence for Forensic Scientists*, 3rd edn. Chichester: John Wiley & Sons.

Allen, R. J. (2013) 'The Conceptual Challenge of Expert Evidence', *Discusiones Filosóficas*, **14**: 99–113.

Allen, R. J., and Pardo, M. S. (2007) 'The Problematic Value of Mathematical Models of Evidence', *The Journal of Legal Studies*, **36**: 107–40.

Biedermann, A. (2015) 'The Role of the Subjectivist Position in the Probabilization of Forensic Science', *Journal of Forensic Science and Medicine*, **1**: 140–48.

Biedermann, A., and Kotsoglou, K. (2022) '(Un-)interpretability in Expert Evidence: An Inquiry into the Frontiers of Evidential Assessment', *Quaestio facti (Revista Internacional sobre Razonamiento Probatorio Quaestio facti. International Journal on Evidential Legal Reasoning)*, **3**: 481–515.

Biedermann, A., and Vuille, J. (2018) 'The Decisional Nature of Probability and Plausibility Assessments in Juridical Evidence and Proof', *International Commentary on Evidence*, **16**: 1–30.

Bovens, L., and Hartmann, S. (2002) 'Bayesian Networks and the Problem of Unreliable Instruments', *Philosophy of Science*, **69**: 29–72.

Bovens, L., and Hartmann, S. (2003) *Bayesian Epistemology*. Oxford: Clarendon Press.

Burnell, R., and others (2023) 'Rethink Reporting of Evaluation Results in AI', *Science*, **380**: 136–38.

Champod, C., and others (2016) *Fingerprints and Other Ridge Skin Impressions, 2nd edn*. Boca Raton: CRC Press.

Cole, S. A. (2006) 'Is Fingerprint Identification Valid? Rhetorics of Reliability in Fingerprint Proponents' Discourse', *Law & Policy*, **28**: 109–35.

Cowell, R. G., and others (1999) *Probabilistic Networks and Expert Systems*. New York: Springer.

Cuellar, M., and others (2024) 'Methodological Problems in Every Black-box Study of Forensic Firearm Comparisons', *Law, Probability and Risk*, **23**: mgae015.

de Finetti, B. (1937) 'La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré* 7', 1–68 (English translation) in Kyburg, H. E., and Smokler, H. E. (eds) *Studies in Subjective Probability* (1980), 2nd edn. pp. 93–158. New York: Dover.

de Finetti, B. (1939) 'The Theory of Probability in Its Relationships with the Analysis' (Translation of an essay published in "Relazioni della XXVIII Riunione della Società Italiana per il Progresso delle Scienze", Pisa, 11–15, October 1939, vol. 3, Sec. A, Roma, 1940, pp. 27–35). in Monari, P., and Cocchi, D. (eds) *Bruno de Finetti, Probabilità e induzione*, pp. 365–74. Bologna: Bibliotheca di STATISTICA, 1993.

de Finetti, B. (1974) *Theory of Probability, A Critical Introductory Treatment*, Volume **1**. London: John Wiley & Sons.

Dennis, I. (2017) *The Law of Evidence*, 6th edn. London: Sweet & Maxwell.

Diaconis, P., and Freedman, D. (1981). 'The Persistence of Cognitive Illusions', *Behavioural and Brain Sciences*, **4**: 333–34.

Ducamp, G., Gonzales, C., and Wuillemin, P.-H. (2020) 'aGrUM/pyAgrum: A Toolbox to Build Models and Algorithms for Probabilistic Graphical Models in Python', In *10th International Conference on Probabilistic Graphical Models*, pp. 609–612. Denmark: Skørping.

Evett, I. W., and Weir, B. S. (1998) *Interpreting DNA Evidence*. Sunderland: Sinauer Associates Inc.

Fabricant, C. M. (2022). *Junk Science and the American Criminal Justice System*. Brooklyn: Akashic Books.

Faigman, D. L., Monahan, J., and Slobogin, C. (2014) 'Group to Individual (G2i) Inference in Scientific Testimony', *The University of Chicago Law Review*, **81**: 417–80.
Faigman, D. L., Scurich, N., and Albright, T. D. (2022) The field of firearms forensics is flawed. *Scientific American (May 25)*.
Fenton, N., Neil, M., and Lagnado, D. A. (2013) 'A General Structure for Legal Arguments about Evidence Using Bayesian Networks', *Cognitive Science*, **37**: 61–102.
Friedman, R. D. (2017). *The Elements of Evidence*, 4th edn. Saint Paul, MN: West Academic Publishing.
Hahn, U., Oaksford, M., and Harris, A. J. L. (2013). 'Testimony and Argument: A Bayesian Perspective', in Zenker, F. (ed.) *Bayesian Argumentation*, Synthese Library 362, pp. 15–38. Dordrecht: Springer.
Hicklin, R. A., and others (2022a) 'Accuracy and Reliability of Forensic Handwriting Comparisons', *Proceedings of the National Academy of Sciences*, **119**: e2119944119.
Hicklin, R. A., and others (2022b) 'Accuracy, Reproducibility, and Repeatability of Forensic Footwear Examiner Decisions', *Forensic Science International*, **339**: 111418.
Imwinkelried, E. J. (1995) 'Coming to Grips with Scientific Research in *Daubert*'s "Brave New World": The Courts' Need to Appreciate the Evidentiary Differences between Validity and Proficiency Studies', *Brooklyn Law Review*, **61**: 1247–84.
Imwinkelried, E. J. (2020) 'The Admissibility of Scientific Evidence: Exploring the Significance of the Distinction between Foundational Validity and Validity as Applied', *Syracuse Law Review*, **70**: 817–49.
Jensen, F. V. (1996) *An Introduction to Bayesian Networks*. London: University College London Press.
Kahneman, D., Slovic, P., and Tversky, A., eds (1982) *Judgement under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.
Koehler, J. J. (2008) 'Fingerprint Error Rates and Proficiency Tests: What They Are and Why They Matter', *Hastings Law Journal*, **59**: 1077–100.
Koehler, J. J., Chia, A., and Lindsey, S. (1995) 'The Random Match Probability in DNA Evidence: Irrelevant and Prejudicial?', *Jurimetrics Journal*, **35**: 201–19.
Lad, F. (1996) *Operational Subjective Statistical Methods: a Mathematical, Philosophical, and Historical Introduction*. New York: John Wiley & Sons.
Lau, T., and Biedermann, A. (2020) 'Assessing AI Output in Legal Decision-making with Nearest Neighbors', *Penn State Law Review*, **124**: 609–55.
Lindley, D. V. (1982) 'Scoring Rules and the Inevitability of Probability', *International Statistical Review / Revue Internationale de Statistique*, **50**: 1–11.
Lindley, D. V. (1985). *Making Decisions*, 2nd edn. Chichester: John Wiley & Sons.
Lindley, D. V. (2006). *Understanding Uncertainty*. London: John Wiley & Sons.
PCAST (2016) *President's Council of Advisors on Science and Technology, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. Washington, D.C: Executive Office of the President.
Risinger, D. M., Denbeaux, M. P., and Saks, M. J. (1989) 'Exorcism of Ignorance as a Proxy for Rational Knowledge: The Lessons of Handwriting Identification "expertise"', *University of Pennsylvania Law Review*, **137**: 731–92.
Roberts, P. (2007) 'From Theory into Practice: Introducing the Reference Class Problem', *International Journal of Evidence and Proof*, **11**: 243–54.
Roberts, P. (2018) 'Making Sense of Forensic Science Evidence', in Roberts, P. and Stockdale, M. (eds) *Forensic Science Evidence and Expert Witness Testimony: Reliability through Reform*?, chap. 1, pp. 27–70. Cheltenham: Edward Elgar Publishing.
Robertson, B., Vignaux, G. A., and Berger, C. E. H. (2016) *Interpreting Evidence. Evaluating Forensic Science in the Courtroom, 2nd edn*. Chichester: John Wiley & Sons.
Saks, M. J. and Koehler, J. J. (2005) 'The Coming Paradigm Shift in Forensic Identification Science', *Science*, **309**: 892–95.
Schum, D. A. (1994). *Evidential Foundations of Probabilistic Reasoning*. New York: John Wiley & Sons, Inc.
Stoney, D. A. (1991). 'What Made Us Ever Think We Could Individualize Using Statistics?', *Journal of the Forensic Science Society*, **31**: 197–99.
Swofford, H. J., and others (2024) 'Inconclusive Decisions and Error Rates in Forensic Science', *Forensic Science International: Synergy*, **8**: 100472.
Taroni, F., and others (2006) *Bayesian Networks and Probabilistic Inference in Forensic Science*. Statistics in Practice. Chichester: John Wiley & Sons.
Taroni, F., and others (2004) 'A General Approach to Bayesian Networks for the Interpretation of Evidence', *Forensic Science International*, **139**: 5–16.
Taroni, F., and others (2018) 'Reconciliation of Subjective Probabilities and Frequencies in Forensic Science', *Law, Probability and Risk*, **17**: 243–62.
Tart, M. (2020). 'Opinion Evidence in Cell Site Analysis', *Science & Justice*, **60**: 363–74.

Thompson, W. C. (2016). 'Determining the Proper Evidentiary Basis for an Expert Opinion: What Do Experts Need to Know and When Do They Know too Much?', in Robertson, C., and Kesselheim, A. (eds) *Blinding as a Solution to Bias: Strengthening Biomedical Science, Forensic Science, and Law*, pp. 133–50. Amsterdam: Academic Press.

Thompson, W. C., Taroni, F., and Aitken, C. G. G. (2003) 'How the Probability of a False Positive Affects the Value of DNA Evidence', *Journal of Forensic Sciences*, **48**: 47–54.

Law, Probability and Risk, 2025, 24, 1–20
https://doi.org/10.1093/lpr/mgaf012
Research Article