Forensic Science International 288 (2018) 72–80

Contents lists available at ScienceDirect

Forensic Science International

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

Critical analysis of forensic cut-offs and legal thresholds: A coherent approach to inference and decision

A. Biedermann[a,*], F. Taroni[a], S. Bozza[b,a], M. Augsburger[c], C.G.G. Aitken[d]

[a] University of Lausanne, School of Criminal Justice, 1015 Lausanne-Dorigny, Switzerland
[b] Ca'Foscari University Venice, Department of Economics, 30121 Venice, Italy
[c] University Center of Legal Medicine Lausanne – Geneva, Forensic Toxicology and Chemistry Unit, 1000 Lausanne 25, Switzerland
[d] University of Edinburgh, School of Mathematics, EH9 3FD Edinburgh, Scotland, United Kingdom

ARTICLE INFO

Article history:
Received 13 April 2017
Received in revised form 18 February 2018
Accepted 16 April 2018
Available online 25 April 2018

Keywords:
Toxicological analyses
Forensic science
Interpretation
Cut-offs
Legal threshold

ABSTRACT

In this paper we critically discuss the definition and use of cut-off values by forensic scientists, for example in forensic toxicology, and point out when and why such values – and ensuing categorical conclusions – are inappropriate concepts for helping recipients of expert information with their questions of interest. Broadly speaking, a cut-off is a particular value of results of analyses of a target substance (e.g., a toxic substance or one of its metabolites in biological sample from a person of interest), defined in a way such as to enable scientists to suggest conclusions regarding the condition of the person of interest. The extent to which cut-offs can be reliably defined and used is not unanimously agreed within the forensic science community, though many practitioners – especially in operational laboratories – rely on cut-offs for reasons such as ease of use and simplicity. In our analysis, we challenge this practice by arguing that choices made for convenience should not be to the detriment of balance and coherence. To illustrate our discussion, we will choose the example of alcohol markers in hair, used widely by forensic toxicologists to reach conclusions regarding the drinking behaviour of individuals. Using real data from one of the co-authors' own work and recommendations of cut-offs published by relevant professional organisations, we will point out in what sense cut-offs are incompatible with current evaluative guidelines (e.g., [31]) and show how to proceed logically without cut-offs by using a standard measure for evidential value. Our conclusions run counter to much current practice, but are inevitable given the inherent definitional and conceptual shortcomings of scientific cut-offs. We will also point out the difference between scientific cut-offs and legal thresholds and argue that the latter – but not the former – are justifiable and can be dealt with in logical evaluative procedures.

1. Introduction

Many analytical branches, in particular forensic toxicology, commonly rely on what are called cut-offs. These are numerical values against which measurements – known as sets of results – made on questioned items (specimens) are compared in order for scientists to proffer, support or complement an interpretation or a conclusion in a forensic toxicological assessment regarding, for example, a person of interest.[1] Examples for sets of results are concentrations of toxic or controlled substances in blood, or of target substances (e.g., metabolites) in hair. Such analyses are of wide interest and include, for example, workplace safety contexts, child custody disputes and sports (e.g., suspected doping cases). A further area where cut-offs are used is ink dating in forensic document examination. In this context, a numerical value for an ageing parameter – referring to certain components of ink entries (e.g., solvents) – is compared against predefined values in order to reach a conclusion regarding the ink entry's age (see [7] for an example).

In forensic toxicology, the intended use of scientific cut-offs can, broadly speaking, be summarised as follows: individuals of a group with a particular behaviour (e.g., abusive drinkers, doping athletes) can be shown to exhibit detectable features that are not – or less – typically found with people who do not belong this group (e.g., non-drinkers, clean athletes). The idea is that if one analyzes a sample from a person for whom it is not known whether they fall in one or the other category of individuals, the comparison of the

* Corresponding author.
E-mail address: alex.biedermann@unil.ch (A. Biedermann).

[1] Another topic, not discussed in this paper, is the use of the term cut-off in the strict analytical sense of detection limits, such as the limit of detection (LOD) and the lower limit of quantification (LLOQ).

https://doi.org/10.1016/j.forsciint.2018.04.030
0379-0738/

measured value against the cut-off will provide discriminative information. However, a main problem in practice is to define the cut-off value in a way such that it separates, in some sense, the two populations of interest. Generally, it is not possible[2] to have cut-offs that allow one to derive categorical conclusions of the kind 'the measured value is above the threshold, hence the person of interest is a drug addict'. There are several logical problems with such statements that we will analyse and expose in this paper. We will also point out the difference with respect to legal thresholds where the definition of limiting values is justifiable and not targeted by our critique.

From a statistical point of view, the above topic is a classic problem of discrimination. Theory on this problem is well developed and widely described in literature, yet – surprisingly – in many instances of forensic science practice it is not properly taken into account. What is more, as we will point out in this paper, the notion of cut-off is used by scientists to derive conclusions in ways that cannot be logically supported, and that are contrary to current evaluative guidelines (such as [31]). Logically unsound interpretations of comparisons with cut-off values may even be found in recommendations of professional associations (e.g., [18,26]), which is a cause for concern. In this paper, we will clarify the exact sense in which common interpretations may be flawed, and contrast them with a balanced, logical and transparent approach. Specifically, we will argue – contrary to predominant perceptions – that scientific cut-offs are *not* needed when the questions of interest to recipients of expert information relate to discrimination (e.g., between abstinent individuals and chronic excessive consumers). Thus, we will justify our position that cut-offs should be eliminated from some types[3] of forensic reporting practice. This is in order to avoid the fallacious perception among both scientists and members of the judiciary that comparison of a set of results with a cut-off value is sufficient to form a conclusion about a particular proposition, especially by scientists, in the absence of both an explicit alternative proposition and consideration of task-relevant information.

An important feature of the cut-offs we discuss is that they have been elaborated within the scientific community as a concept to ease and harmonise interpretations and conclusions *without* there being any legal requirement for such a limiting value. Stated otherwise, the legal question in the first place does not focus on a threshold, but deals with discrimination, which is whether a given individual belongs to one group of individuals rather to than another. The notion of cut-off arises here only because scientists try to help with this issue by reformulating the problem of discrimination as one of comparison against a cut-off value. It is this particular practice that we consider inappropriate and on which we will focus our critique.

This does not mean, however, that threshold values are not helpful in principle. It is important to emphasise a distinction with respect to applications in which limiting values have their place. This is the case, for example, with legal and regulatory criteria formulated in terms of threshold values for substances such as ethanol and other toxic substances in blood, including doping issues in sports. Here, limiting values serve the purpose of defining situations considered to conform or not with the law, irrespective of the actual condition of the particular person of interest. For example, given a concentration of blood alcohol measured above a legal limit, two different persons may exhibit different physical capacities; one person may still show some capacity to drive a vehicle, whereas the other person may not. Yet, the legal limit applies to both persons. It serves as a standard to decide between situations that conform and situations that do not conform to the law. We will refer to such quantitative legal criteria broadly as limits, limiting values, or legal thresholds and distinguish them from the above terminology of cut-offs used primarily in scientific applications (e.g., consensus values derived within scientific communities).

The distinction between scientific cut-offs and legal thresholds is not material for the general analysis that we will present. We emphasise the generality of our approach because it allows us to deal with thresholds where the law defines them, and to avoid scientific cut-offs in applications where the law does not define limiting values. To point this out, our paper is structured around two main types of questions:

1. Is the person of interest, for example, a heavy drinker (or a doping athlete, a drug addict, etc.)?
2. Is the concentration of the target substance in the blood of person of interest above a given legal limit?

Section 2 will critically examine current scientific approaches based on cut-offs intended to help address situations exemplified by question 1. It is here that we will argue that scientific cut-off values need *not* be defined in order to help with the question of interest; what is more, cut-offs should in fact *not* be used because they are an obstacle to balanced evaluations. In Section 3, we will present an approach that allows one to avoid the limitations and drawbacks of cut-offs identified in Section 2. Section 4 will point out that the same approach is also applicable for situations in which scientists are requested to help with questions of the second type mentioned above. Discussion and conclusions are presented in Section 5.

2. Critical analysis of scientific cut-offs: the example of hair analyses

2.1. The classic inversion fallacy

The analysis of target substances in hair is a typical example of a forensic area of practice where cut-off values are widely used [17,25]. Professional organisations, such as the Society of Hair Testing (SoHT, http://www.soht.org), regularly publish consensus documents (e.g., [15,18]) with cut-offs that are intended to assist scientists in reaching conclusions regarding, for example, the drinking habits of persons of interest. These cut-offs relate to substances – also called alcohol markers – such as the minor ethanol metabolite ethyl glucuronide (EtG), which is recommended because the measurement of ethanol directly in hair is not feasible for various reasons (e.g., volatility). The suitability of hair analyses for forensic purposes is not unanimously accepted among scientists, as is demonstrated by controversies over factors such as hair washing and intra- and inter-individual differences in metabolism [16,29].

Notwithstanding, at least since 2009 [15], the interpretation of measured EtG concentrations is generally based on the cut-off of 30 pg/mg. In its most recent consensus document, the SoHT conveys the use of this threshold as follows:

> "A concentration of >30 pg/mg EtG in the proximal scalp hair up to 6 cm strongly suggests chronic excessive alcohol consumption."[4]

[2] See [23] for a recent example of claims to the contrary, discussed here later in Section 2.1.

[3] An exception are purely analytical uses as mentioned in footnote [1].

[4] 2016 Consensus for the Use of Alcohol Markers in Hair for the Assessment of both Abstinence and Chronic Excessive Alcohol Consumption (http://soht.org/images/pdf/Revision%202016_Alcoholmarkers.pdf, last accessed November 18th 2016)

Similarly, the European Workplace Drug Testing Society (EWDTS) issued the following guideline:

> "The cut-off for EtG in hair to strongly suggest chronic excessive alcohol consumption is proposed at 30 pg/mg scalp hair measured in the 0–3 cm up to 0–6 cm proximal segment." [26,at p. 1002]

Although the SoHT insists on the need to conduct careful case assessments, in particular that "[i]t is not advisable to use the results of hair testing for alcohol markers in isolation; all relevant factors surrounding a case must be considered when providing expert interpretation and opinion" [27,Section 1.3.], the instruction for evaluating measurements reduces, in essence, to the following pattern of reasoning:

> Compare the measured EtG concentration with the cut-off at 30 pg/mg, and if the measured value is greater than the threshold, (it is permissible to) conclude that the result 'strongly suggests chronic excessive alcohol consumption'.

But is this a sound way of reasoning? Is it logical to say that a measured value above a certain level 'strongly suggests chronic excessive alcohol consumption'? Presumably, the rationale for the above interpretative instruction is based on the following line of argument: 1. Chronic excessive alcohol consumers tend to have high concentrations (typically above 30 pg/mg) of EtG in their hair, 2. *If* one finds a value above 30 pg/mg, *then* this is strongly suggestive of the person of interest being a chronic excessive alcohol consumer. This, however, is a classic inversion fallacy described abundantly in forensic literature [e.g.,1].

To clarify this further, let us restate the above argument using more formal language. From the observation that many chronic excessive alcohol consumers tend to have high concentrations of EtG in hair, we may formulate the following sentence: if a person of interest is a chronic excessive alcohol consumer, then there is a high probability that the measured EtG concentration in hair of that person will be high. Formally, this is a deductive mode of reasoning that makes a statement about the results (measured concentration) given an assumption (hypothesis) about the person of interest belonging to a particular group of individuals (i.e., chronic excessive consumers). This, however, does not allow us to make a direct transformation to an inductive statement of the following kind: if we measure a high concentration of EtG in the hair of a particular person, then there is a high probability that this person is a chronic excessive alcohol consumer. More generally, the fallacy consists of interpreting a high probability for the evidence given a hypothesis as a high probability of this hypothesis given the evidence.

It might be argued that the above consensus statement of the SoHT is not a genuine inversion fallacy because it does not express a direct opinion about a hypothesis of interest (e.g., 'this person is (probably) a chronic excessive drinker'). It only contains the weaker expression 'strongly suggests' (excessive consumption by the person of interest). Notwithstanding this argument is at least an ambiguous formulation because it is at risk of being interpreted in the mind of the recipient of the statement as it being 'probable that the person of interest is a chronic excessive drinker'.

Other guidance documents contain more explicit instances of the inversion fallacy. One recent example can be found in the 'New Hair Testing Conclusions' issued by the FBI:

> "(...) we developed a new set of reporting criteria that allow FBI toxicologists to opine that an individual has *used* cocaine, if the following two criteria are met: Cocaine is identified in the hair specimen above 500 pg/mg (...) Two hydroxycocaine metabolites are identified in the hair specimen above 5 pg/mg." [23,at p. 1]

Another instance of a direct conclusion about a proposition on the sole basis of a single measurement can be found in [3,at p. 175]: "Strict abstinence is improbable or can be excluded at EtG greater than 7 pg/mg in hair."

The above two reporting schemes lead to categorical (or almost categorical) statements about hypotheses (here, drug use) if the measured concentration of particular target substances in hair is greater than particular cut-offs, and thus provide explicit examples for an inversion fallacy. Three further shortcomings of cut-off based reporting schemes are described in the next section.

2.2. Further shortcomings of cut-off based reporting schemes

2.2.1. Abrupt changes in conclusions for near cut-off findings

Reporting based on a cut-off introduces a rigid perspective in the sense that results are considered 'informative' as long as they are above the cut-off. No guidance is provided, however, on what to do if results are below, or, especially, just below, the cut-off. As an example, consider again the SoHT EtG cut-off at 30 pg/mg, associated with the qualifier 'strongly suggest chronic excessive consumption'. Such a cut-off is problematic in two ways. First, if a measurement just below the cut-off is obtained, such as 29 pg/mg, it is not clear what phrase *exactly* other than a 'strong suggestion' should be used.[5] Second, if a measurement only slightly above the cut-off is obtained, such as 31 pg/mg, the conclusion would be the same also for any other value greater than the cut-off. This bears a potential for under-evaluating and over-evaluating results close to the cutoff: results just slightly below the cut-off may be too easily dismissed as valueless, whereas values slightly above the cut-off too easily accepted as probative. Indeed, the latter will be taken as probative as any other value greater than the cut-off. To avoid over-evaluation, cut-offs may be adjusted towards higher values, but this will not resolve the issue, as the probability of under-evaluation will increase.

As an aside, let us note that a similar drawback is encountered with the reporting schemes for fingermarks – though now widely abolished – based on numerical standards (e.g., the so-called 12 point rule). According to these rules, a fingermark was considered 'identified' if a certain number of corresponding minutiae were found, and 'not identified' otherwise. Here, too, a number of corresponding points just below the numerical cut-off (without unexplainable differences) is considered categorically as 'inconclusive', while a number of minutiae just above the numerical standard is considered as 'probative' as any other number of corresponding minutiae above the numerical standard.

More generally, the sudden change from non-probative to probative as the findings increase from below a cut-off to above a cut-off is also known in literature as the 'fall of the cliff effect'.[6] Curran et al. [10,at p. 81] give an example in the context of the comparison of refractive index measurements on glass fragments.

2.2.2. Undefined verbal qualifiers

An inevitable but practically important question is what verbal qualifiers such as 'strongly suggest' mean. If the intended meaning is to express a probability for a proposition, then this is unhelpful because – as explained in Section 2.1 – there is no logically defensible way to transform a statement about a result directly into a probability statement about a single proposition of interest (e.g.,

[5] The vagueness of what to conclude for measurements below a cut-off is illustrated by the FBI reporting conventions for results of hair analyses: "If these criteria are not met, then the hair specimen will be reported as "negative", "contaminated" or "consistent with cocaine exposure", depending on the quantity of cocaine measured, as well as the specific metabolites that are identified" [23,at p. 2]. For a critical discussion of vague expressions, see also [20].

[6] The origin of this expression is commonly assigned to Ken Smalldon [e.g.,11].

chronic excessive drinking behaviour). If the intended meaning is not the expression of a probability statement about a proposition, but merely the expression of the extent to which a given proposition is strengthened, whatever the probability of this proposition was prior to the consideration of the result, then this is also not feasible. There are at least two reasons for this. First, no alternative proposition is specified. Indeed, as stipulated by the principles of forensic interpretation, the evaluation of the uncertainty of any given proposition requires the consideration at least one alternative proposition [2,13]. Second, to be useful, a verbal qualifier needs an underlying numerical scale that must be defined. This, too, is a general requirement applicable to any concept such as (physical) weight, volume or monetary value: it is the quantitative value in standard measurement units (e.g., g, mg, US$) that provides the basis for discussion. Whether, for example, a quantity *x* of US$ corresponds to 'few', 'much' or 'very much' money is a matter for convention, and strongly depends on the context (e.g., on the person uttering the statement). The same holds for the measurement and expression of probative value for forensic evidence, based on likelihood ratios [31,22]: an expression of 'strong support' for a proposition is meaningless without a definition of what this qualifier means in terms of a numerical likelihood ratio.

2.2.3. Incompatibility with current guidelines for evaluative reporting in forensic science

Current European guidelines for evaluative reporting in forensic science stipulate three main principles [2,8,31], all of which are not respected by reporting schemes based on cut-offs such as the SoHT consensus document [18]. The first principle says that forensic scientists ought to focus on the probability of their results given relevant propositions, not the contrary. Reporting whether or not a given measurement is above a cut-off, and then giving an opinion on the proposition (e.g., of chronic excessive consumption) violates this principle. The second principle is intended to ensure balance. Scientists should consider the probability of their result given at least one alternative proposition. This principle is also not observed by the SoHT consensus document because that document permits scientists to provide a conclusion such as 'the results strongly suggest chronic excessive consumption' without the requirement to consider the probability of the results if an alternative proposition (i.e., other consumption behaviour) is true. The third principle requires scientists to condition their evaluation on task-relevant case information. Reduction of the evaluation to simply the comparison of results with a predefined cut-off does not comply with these principles.

The incompatibility of reporting schemes based on cut-offs with those that are based on current guidelines for evaluative reporting in forensic science reveals a fundamental mismatch in the principles of the schemes as to how probative value may be interpreted. A scheme based on a cut-off suggests, first, that probative value is directly related to the results (i.e., if the result is greater than the cut-off, it is supportive of a given proposition, and not supportive if it is less than the cut-off), and, second, that the problem of interpretation consists only in determining this allegedly inherent character of a set of results being supportive of one proposition or another. This is different for the evaluative approach based on the three principles outlined above [31,13]. The evaluative approach does not suppose that there is a predefined conclusion associated with individual sets of results. Instead, it supposes that the probative value of a given set of results depends on the comparison of the probability of obtaining the set of results of interest, given a particular proposition and the relevant information (e.g., gender), with the probability of the set of results given at least one alternative proposition and the relevant information. This implies that a given set of results may have a different probative value depending on the propositions of interest and the conditioning information. We will make this understanding formally precise in Section 3 and discuss two illustrative examples.

3. Logical evaluation, inference and decision without scientific cut-offs

3.1. Example 1: Evaluation of a low measurement of EtG concentration in hair

Imagine a case in which a woman[7] is monitored for abstinence as required, for example, in contexts relating to child custody or driving ability (e.g., after an accident) issues. Hair of the person of interest is analysed following accepted analytical procedures. The analyses reveal an EtG concentration of 5 pg/mg in the proximal segment up to 6 cm. Assessing the value of this measurement result according to the principles of forensic interpretation (Section 2.2.3, and [31]) requires, first, definitions of the propositions of interest, based on the main issue with which the scientist is asked to help. The issue, here, is abstinence. Thus, the first proposition may be defined as 'The person of interest is a teetotaller'. The alternative proposition may be defined as 'The person of interest is a social drinker (i.e., a non-teetotaller, or low-risk drinker)'. The proposition that the person is a chronic excessive drinker is not taken into account as there is abundant other evidence, based on the person's detailed medical history and documented data, to show that such a possibility is not an issue in this case.

In order to assess what the measurement of 5 pg/mg EtG means with respect to the above propositions of interest, it is necessary to answer two questions: first, what is the probability of obtaining this result if the person of interest is a teetotaller, and second, what is the probability of obtaining this result if the person of interest is a social drinker. These two probabilities correspond to, respectively, the numerator and denominator of the likelihood ratio. To answer these questions, we refer to data published by [14]. The following are relevant considerations (see also Fig. 1):

- In [14], 29 women were monitored in the category teetotaller, defined as subjects who declared not to consume any alcohol (0 g/day) during the last 12 months. The assignment to the category teetotaller was based on a Daily Alcohol Self-Monitoring log (DASM log). For 25 women EtG was not detected at a limit of detection of 2 pg/mg. For four women, EtG concentrations of, respectively, 3.3, 6.5, 7.6 and 10.2 pg/mg, were found.
- Low-risk female drinkers are defined in [14] according to WHO recommendations as persons who consumed $\leq$20 g/day. Among the 30 low-risk women monitored by [14], EtG was not detected at a limit of detection of 2 pg/mg in 6 cases. The remaining 24 showed concentrations in the range up to 19.0 pg/mg. The mean measured concentration was 3.69 pg/mg.

From a data-analytic point of view, the above data are – although relevant for the questions of interest – rather scarce. Take, for example, the data on teetotallers: these data consist mainly of results below the limit of detection, only a few are above the limit of detection. It is not obvious, with these data, how to construct and defend a suitable continuous probability distribution function to help with the task of assigning a probability for the numerator of the likelihood ratio, that is a

[7] Specifying the sex of the person of interest is important in the context here because population studies found different levels of EtG incorporation in hair of men and women.

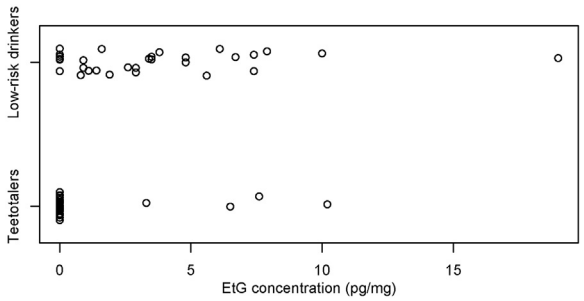

Fig. 1. EtG concentrations measured in hair of female teetotalers ($N = 29$) and low-risk drinkers ($N = 30$) (data obtained from the authors of [14]). Note that within each of the two categories displayed on the y-axis, the vertical scattering of the values has no particular meaning other than serving the purpose of visually separating values with the same or very similar value on the x-axis (i.e., EtG concentration).

probability *density* for the particular measurement 5 pg/mg. The scientist might feel that the provision of a probability density is not robust in the light of the available data.

Unless we collect more data, one way to avoid this complication consists in reformulating the question more generally as one of assigning a probability for a measurement above the limit of detection (rather than the particular result of 5 pg/mg), using the observed proportion of such non-zero values, that is 4 out 29, and interpret[8] this proportion as a probability assignment of 0.14. Similarly, the scientist could look at the data on female low-risk drinkers and consider that 24 of the 30 monitored persons had values above the limit of detection, corresponding to a proportion of 0.8. The scientist could then say that their probability for seeing a value different from zero is approximately six times greater[9] given the second proposition, the person of interest being a low-risk drinker, than given the first proposition, the person of interest being a teetotaller. This represents limited support for the proposition that the person of interest is a social drinker, rather than being a teetotaller. It is important, however, to emphasise that the evidence being evaluated is not the actual measurement of 5 pg/mg, but a more coarse and qualitative description of this result (i.e., a statement of whether the result is or is not different from zero).

An advantage of this more general approach is that the probability assignments are more robust, because they are derived from the proportions of zero and non-zero measurements among a sample of teetotallers and a sample of low-risk drinkers respectively. This robustness, however, comes at the price of being less case-tailored: indeed, the same likelihood ratio will be obtained for all non-zero measurements. This is a loss of resolution because the general understanding is that as EtG measurements increase, the probability of observing them among teetotallers decreases, which should lead to higher likelihood ratios in support of the proposition of low-risk drinking versus teetotaller. However, to make this understanding quantitatively more precise, more data from controlled studies such as [14] (Fig. 1) are needed.

To alleviate this loss of resolution, measurement results different from zero could be grouped into different ranges of values (e.g., 0–5, 5-... pg/mg etc.), as suggested for example by Bossers and Paul [6]. But this is only a partial solution because it raises the question of how to define the ranges of values. Moreover, depending on the size and the number of ranges, scarce data may again become an obstacle for robust probability assignment.

The trade-off between the level of detail at which the measurements are grouped, and the robustness of the probability assignment may be seen as a drawback, posing a limit to the usability of the likelihood ratio reporting format. This, however, is not a problem of the likelihood ratio, but of the intricacy of the real world problem under study, and the extent of relevant scientific knowledge available (here, population data). In principle, the likelihood ratio is able to deal with as much or as little background data as there may be. If, however, the scientist feels that there are not enough data to quantify the two likelihoods, even on a coarse level, it is important to stress that there is nothing better that the scientist could report as a suitable substitute for the expression of a probative value. Specifically, descriptive expressions such as 'consistent with' (e.g., this measurement result is consistent with 'low-risk drinking behaviour') do not convey any discriminative information [12,20] essentially because they do not inform about whether or not the particular result is also 'consistent with' something else. Even if such an additional statement (i.e., consistency 'with something else') would be given, it will not help the recipient of expert information discriminate between competing accounts of the case at hand. Only the likelihood ratio provides this information.

Even though the likelihood ratio in a particular case may only have a low value (because of the scarcity of information), as illustrated in the above example, the ratio's most valuable feature is that it directs evaluators to ask the relevant questions and it helps them avoid dilemmas commonly exposed in theory and practice. An example for such a dilemma is the following statement:

> "However, the most serious obstacle of using EtG in hair for a real abstinence determination is that social drinkers frequently produce negative results." [24,p. e6]

A perspective based on the likelihood ratio is particularly well suited for the evaluation of such evidence. Practical forensic measurements, of which EtG concentrations are just one example among many others, are almost never exclusively related with only one particular category of people[10] (see, e.g., Fig. 1). Arguably, the best scientists can do is to inform about the relative probability with which the item of evidence arises given each of the competing propositions of interest. The summary provides no more information but also no less information about the value of evidence.

3.2. Example 2: Evaluation of a large concentration of EtG measured in head hair

Consider the case of a man suspected of excessive alcohol consumption. The forensic medical diagnosis of excessive alcohol consumption may be necessary, for example, for reasons of safety in the workplace. As in the previous example, hair of the person of interest is analysed following accepted analytical procedures. The analysis of hair revealed a concentration of 28 pg/mg in the proximal segment up to 6 cm.

Comparison of this result with the SoHT cut-off at 30 pg/mg would lead to the conclusion that there is no 'strong suggestion' for chronic excessive alcohol consumption. The application of a cut-off based perspective is limited to such a conclusion. However, this conclusion begs the question of what probative value the particular result of 28 pg/mg has with respect to selected competing propositions of interest. Generally, a value of 28 pg/mg is something less typically found among non-heavy drinkers. In the case here, suppose that the two competing propositions of interest are: 'The

[8] Note that this amounts to interpreting a relative frequency as a probability, which may be criticised in many ways (see, e.g., [19]), especially when there are zero counts of the event of interest.

[9] The likelihood ratio is $0.8/0.14 \approx 6$.

[10] Notice that a perfect item of evidence is one that *only* arises with one proposition, and never arises with any other mutually exclusive proposition.

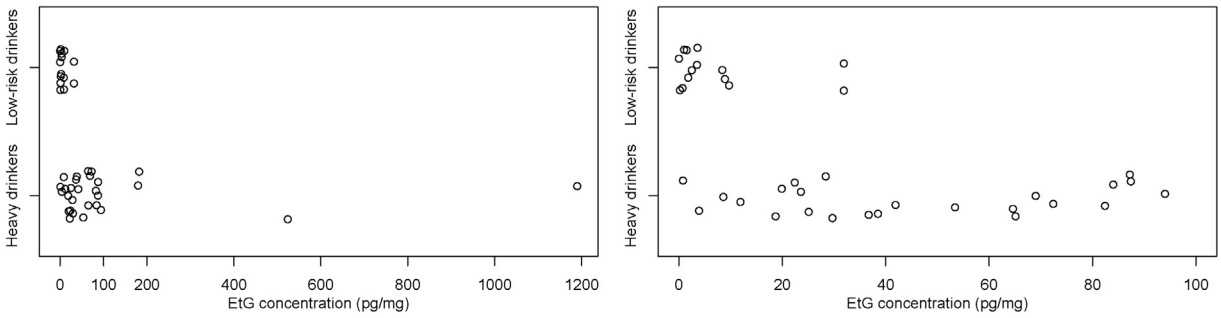

Fig. 2. Left: EtG concentrations measured in hair of male low-risk drinkers ($N = 14$) and heavy drinkers ($N = 28$) (data obtained from the authors of [14]). Right: Representation of the same data limited to concentrations below 100 pg/mg EtG. Note that within each of the two categories displayed on the *y*-axis, the vertical scattering of the values has no particular meaning other than serving the purpose of visually separating values with the same or very similar value on the *x*-axis (i.e., EtG concentration).

person of interest is a chronic excessive drinker', and 'The person of interest is a low-risk (i.e., social) drinker'. Note that 'teetotaller' is not a relevant proposition in this case because there is other uncontested information in the case, including the person's self-declared drinking habits, that indicates that the person of interest consumes alcohol, at least to some extent.

As in the previous example (Section 3.1) application of the likelihood ratio framework for evaluation of the measurement of 28 pg/mg EtG with respect to the above propositions of interest requires the answer to two questions: first, what is the probability of obtaining this result if the person of interest is a chronic excessive drinker, and second, what is the probability of obtaining this result if the person of interest is a low-risk drinker? To answer these questions, we refer again to data published by [14] (see also Fig. 2):

- 'At-risk male drinkers' in the study of [14] are men who consume more than 30 grams per day of alcohol. In this category, 28 subjects have been monitored. EtG concentrations in the hair of these men were found in the range up to 1189.9 pg/mg, though only four of them had concentrations above 100 pg/mg. The value observed in the case considered here, 28 pg/mg, is not found among these data, though there is one occurrence of the concentration 28.4 pg/mg.
- 'Low-risk male drinkers' (14 subjects) are defined in [14] as men who declare (according to the DASM log) they drink less than 30 grams per day. Their EtG concentrations in hair were found in the range up 31.9 pg/mg. There are only two subjects with values greater than 20 pg/mg in the relevant data set.

Although these are data from a carefully designed study, focusing on individuals relevant to the case studied in this example, they are scarce. While the results for low-risk drinkers are concentrated in the range of values below 10 pg/mg (i.e., 12 out of the 14 measurements), the measurements for at-risk drinkers range from 0 to 100 pg/mg. Based on these data, it may thus be argued that it is *more* probable to observe the particular result 28 pg/mg given the first proposition, the person of interest is an 'at-risk male drinker', than given the alternative proposition, the person of interest is a 'low-risk male drinker'. Thus, the result of 28 pg/mg supports the proposition 'at-risk male drinker' versus the proposition 'low-risk male drinker' *despite* the result being below the SoHT threshold of 30 pg/mg.

The crucial question is, however, the strength of the support and its order of magnitude. The answer to this question depends on the assigned probabilities. As in the previous example (Section 3.1), determination of probability densities for the particular measurement 28 pg/mg with the limited data available will be sensitive to small changes in the data. A more coarse summary of the data is required. It may thus be reported, for example, that a measurement result in the range 20–40 pg/mg is about two times more probable given the first proposition (heavy drinker) than given the second proposition (low-risk drinker). This statement is based on the observation that there are 7 out of the 28 monitored heavy drinkers who showed values in the range 20–40 pg/mg, leading to the probability assignment 0.25 for a measurement to be observed in the stated range for heavy drinkers and on the observation that there are only two observations out of 14 low-risk drinkers in that range, leading to the probability assignment of $2/14 = 0.14$ (value rounded). In combination, using the values 0.25 and 0.14 for, respectively, the numerator and the denominator, leads to a likelihood ratio of approximately 1.78.

The result of a likelihood ratio of approximately 2 is not an expression of probative value for the particular result of 28 pg/mg, but for a more coarse description of the findings in terms of an observation of a value within a particular range of values (here, 20–40 pg/mg). But even with this approach, there remains a broad scope for personal appreciation by the scientist. For example, suppose that the range of values would have been restricted to 20–30 pg/mg, rather than 20–40 pg/mg. In such a case, there would have been no observation of values for low-risk drinkers, suggesting[11] a very low probability for obtaining a value in this range (i.e., a low value for the denominator of the likelihood ratio), hence leading to a high likelihood ratio. This shows that there is not a single right way for proceeding. Conclusions about probative value depend on the particular assumptions made and on the questions that are asked. The scarcity of data is a further source for differences in the answers obtained. Notwithstanding these caveats, it is worth noting that *despite* the different levels of detail at which the above evaluation example has been framed, and the two different answers obtained, they both resulted in a likelihood ratio greater than one (of which one is theoretically infinite), indicating support for the proposition of an 'at-risk male drinker' rather than a 'low-risk male drinker'. Even with scarce data, there is an indication of qualitative robustness in the answers that can be provided.

3.3. Coherent decision criteria

The approach outlined in Sections 3.1 and 3.2 focused on the metric that indicates how scientific findings – analytical data such as EtG concentrations measured in head hair – logically help to distinguish between competing propositions (e.g., teetotaller versus low-risk drinking). It is important to underline that this metric, the likelihood ratio, does not tell how probable the two

[11] This is an instance illustrating the limitation of interpreting a relative frequency as a probability. The absence of an observation for the event of interest would lead to an assigned probability of zero (see also footnote [8]).

competing propositions are. The likelihood ratio only indicates how the probabilities of the propositions change – strengthened, weakened or left unchanged – from whatever those probabilities were prior to considering the scientific findings. The next question thus is how to decide rationally about the competing propositions, for example to decide whether a person is a teetotaller rather than a low-risk drinker. We recognise that the topic of decision is a broad one, encompassing several dimensions, including the study of how people actually make decisions and perceive the task of decision making. Yet other dimensions regard procedural constraints and features of jurisdictional frameworks. Below, we will only focus on aspects of a formal analytical approach to the question of decision, independently of particular contexts of application. This will help us point out that even on an abstract level of analysis, observations alone (i.e., measurements) – especially cut-offs – provide an incomplete basis for decision.

Formally, one way to answer this question is to use statistical decision theory (e.g., [5]), according to which two elements must be considered. First, the posterior probabilities for the competing propositions, given the scientific findings (i.e., EtG measurements), and second, the utilities or losses assigned to the various consequences of the decision (i.e., correct and incorrect 'classifications'). Leaving aside the technicalities of how probabilities and value assignments for decision consequences are to be combined, this decision-theoretic account highlights the following points:

- The decision as to what to conclude about the drinking behaviour of a person of interest depends, in part, on the probabilities for the various competing propositions given *all* the information available. Measurement of EtG concentration is only one such item of information among others. The probabilities for the competing propositions are a necessary, but not sufficient, element for a rational decision.
- A rational decision about the drinking habits of a person of interest also depends on the decision-maker's preferences among the consequences (e.g., utilities or losses assigned to correct and erroneous decisions).

It follows from these points that:

- a rational decision about the drinking behaviour of a person of interest does *not* reduce to the sole consideration of a scientific result: the result is only one item of information among others to contribute to the decision;[12]
- a cut-off defined on earlier sets of results is inappropriate in principle as a basis for a rational decision, because none of the fundamental ingredients (e.g., definition of the decision space) are covered.

4. Logical inference for numerical legal thresholds

4.1. Conceptual differences between scientific cut-offs and legal thresholds

In Section 3 we have argued that a cut-off for sets of results is a conceptually and logically improper means for scientists to help recipients of expert information make decisions about properties (e.g., drinking behaviour) of persons of interest. This does not mean, however, that numerical criteria are necessarily inappropriate in all respects. Numerical thresholds serve the purpose, for example, of defining the situations that conform to applicable laws in particular jurisdictions. A well-known example for this is the concentration of alcohol in the blood of car drivers. Similarly, limiting values exist for substances other than alcohol, such as cannabis.

We refer here to such legally defined numerical criteria as legal thresholds and distinguish them from the notion of cut-offs defined exclusively by scientists. Take, for example, the SoHT cut-off at 30 pg/mg for EtG in hair. There is no legal standard that demands concentrations be below this value. Thus, it is not illegal to have values above 30 pg/mg in hair, though it is illegal to have high alcohol concentrations in blood when driving a vehicle. In the former case, the concentration of EtG in hair refers to a scientific cut-off (not to a legal numeric criterion), whereas, in the latter case, the alcohol in blood refers to a legal threshold. In other words, the question of legal interest in the context of hair analyses is the general medical condition of the person of interest not whether that person is legally allowed to drive a car. Throughout this paper, we have argued that helping address this kind of question – which is one of diagnosis – does *not* require a scientific cut-off.

From the above, the following crucial distinction becomes clear:

- In the case of numerical legal thresholds, limiting values serve the purpose of defining the competing propositions of interest. For example, the law may prescribe the highest admissible quantity of toxic substance in the blood of a person of interest (e.g., a car driver), so that the propositions of interest become whether or not the person of interest is below or above the legal limit. The question of interest thus is, explicitly, whether the actual value of the parameter of interest, for the person at hand, is in one range of possible values rather than in another range of possible values. The point of separation between the two ranges of values is defined by the legal threshold. In summary, thus, the law defines both, the target substance (e.g., ethanol in blood) as well as a limiting value.
- In contrast to the above, a scientific cut-off does *not* define propositions of interest. It is intended as a concept for distinguishing between propositions that are defined *independently* of the results. As discussed above, the notion of cut-off is not conceptually appropriate. Moreover, there is nothing in the legal question in the first place that enquires about whether the actual value of the characteristic of interest (e.g., EtG concentration) is in a particular range of values. There is not even a question of focusing on a particular compound, such as EtG in hair.

4.2. Evaluative reporting of scientific results for legal thresholds

Consider again a general example for a legal threshold, such as the quantity of THC (short for tetrahydrocannabinol) in blood [30]. A question of legal interest may be whether the quantity of THC in blood of a person of interest is above or below a particular value, known as the legal threshold. In order for scientists to help recipients of expert information deal with this question, it is not necessary to introduce new principles. It is still appropriate for the scientist to ask, as emphasized in Section 3, two main questions:

- First, what is the probability of obtaining the result on the blood specimen if the person of interest is truly above the legal threshold?
- Second, what is the probability of obtaining the result on the blood specimen if the person of interest is truly below the legal threshold?

The combination of the answers to the above questions provides an indication of the probative strength (known as the likelihood ratio) of the result. The logic of the questions is the same

[12] Note, however, that there are applications in which legislators define single (numerical) criteria as a basis for decision, as discussed below in Section 4.

as for the examples in Section 3. There is a difference in interpretation, however. Here, the propositions relate to a legal threshold. In Section 3 the propositions relate to a behavioural pattern of the person of interest. The focus in both situations is on the probabilities of obtaining the particular result *given* competing assumptions and propositions.

More technically, determination of the answers to the above two questions concerning the legal threshold requires distinct distributional considerations for both the results and the unknown characteristic of interest (e.g., the actual concentration of the target substance in the blood of the person of interest) as described, for example, in [30]. The crucial point in this development is that the analysis clearly separates the question of the probative value of the scientific findings – the data – and the main questions of interest (e.g., whether the person's condition violates the legal threshold). Consideration of the main question of interest involves more than the scientific results alone, as similarly discussed in Section 3.3.

5. Discussion and conclusions

In many applications of forensic toxicology, measured target substances – such as EtG concentrations in hair – have limitations in the sense that they cannot provide a basis for categorical conclusions. The fundamental problem is that the evidential value of particular measurements – given the empirical knowledge currently available – should not be determined based on only one proposition of interest (e.g., the person of interest being a teetotaller). As illustrated by the two examples discussed in Sections 3.1 and 3.2, measurements on the hair of people from different categories (e.g., teetotaller, low-risk drinker, etc.) cover broad ranges of values and, most importantly, the ranges overlap. Therefore, the idea that there could be key values, so-called cut-offs, that would allow one to make categorical determinations about the drinking behaviour of a person of interest is not tenable. Yet, paradoxically, scientists and professional organisations continue to concentrate their efforts on establishing cut-offs (see, e.g., [27] and more recently the letter to the editor [9]). Opinions that dissent from the need to establish cut-offs exist [6,29][6,29, e.g.,], though they represent a minority.

There are exceptions to the criticism of the use of a cut-off. For example, it can be argued that very high values of EtG concentrations (e.g., above 100 pg/mg) are found almost exclusively in the hair of individuals who are heavy drinkers, thus apparently justifying the existence of a cut-off. However, this argument is of little value because it has only limited applicability. Many results are not greater than 100 pg/mg and such a rule would provide no guidance as to what ought to be done in these other cases. For an evaluative framework to be of value, and operationally helpful, it needs to be able to cope with any result.

Arguably, the best that scientists can do is quantify the probability of obtaining the set of results given different competing propositions and advise, if possible, whether and to what extent the results are more probable in one setting than another. This echoes an evaluative perspective, expressed elsewhere as follows: “Whether these results could be observed if one proposition for the event is true rather than another proposition is the central relevant matter on which the forensic scientist may comment.” [21,at p. 796] This reporting format is also compatible with recent evaluative guidelines, such as [31], and prevents scientists from inappropriately opining directly on propositions, rather than on an interpretation of the value of the results. This supports the point emphasized in [21], that scientists do not present evidence if their opinion relates directly to the prosecution or for the defense as if they were party to the matter.

For the above reasons, we do not see any role for scientific cut-offs in evidence evaluation. This conclusion is based on the limitations described above, reinforced by additional factors – not discussed in this paper – such as the differences in the performance of analytical procedures (e.g., extraction efficiency) and alternative modes of target substance absorption, that all lead to further variation in the results. Instead we argue, as done previously by [6], in favour of a more balanced, transparent and flexible approach that provides an indication of the value of the results in terms of their capacity for discriminating between competing propositions. Probative value, according to this perspective, is defined in terms of the ratio of the probabilities of the set of results, given the competing propositions of interest.

The two examples presented in Sections 3.1 and 3.2 broadly illustrate the rationale behind this approach and show how it allows recurrent drawbacks associated with cut-off based interpretation schemes (pointed out in Section 2) to be overcome. It is also worth noting that, unlike predefined and fixed cut-offs, the probabilistic approach is also able to cope with special circumstances – as described for example in [28] – where EtG is detected for reasons other than the drinking behaviour of the person of interest.

We concede that the application of the probabilistic approach would benefit from further relevant data, as noted clearly in our examples (and elsewhere in literature [e.g.,6]), though limitations in current data should not be considered a reason for not using the approach. It is important to emphasize that the probabilistic approach can function with as many or as few data as are available, and that the cut-off approach depends at least as much on data. We thus advocate the development and strengthening of probabilistic evaluative procedures, as emphasized also by [4]. It may be that the resulting assignments of probative values in some cases turn out to be only moderate, and hence seem less helpful. However, such assignments provide a more balanced, transparent and logical evaluation than that provided by a cut-off.

Acknowledgements

Funding: Alex Biedermann gratefully acknowledges the support of the Swiss National Science Foundation through grant No. BSSGI0_155809 and the University of Lausanne. This paper has been presented at the 6th International Conference on Evidence Law and Forensic Science, Baltimore (Maryland), August 14th–16th 2017. The authors are grateful to the participants of this conference for their helpful comments. Colin Aitken acknowledges the support of the Leverhulme Trust, grant number EM2016-027.

References

[1] C.G.G. Aitken, F. Taroni, Statistics and the Evaluation of Evidence for Forensic Scientists, 2nd ed., John Wiley & Sons, Chichester, 2004.
[2] C.G.G. Aitken, P. Roberts, G. Jackson. Fundamentals of Probability and Statistical Evidence in Criminal Proceedings (Practitioner Guide No. 1), Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses, Royal Statistical Society's Working Group on Statistics and the Law, 2010. www.rss.org.uk/Images/PDF/influencing-change/rss-fundamentals-probability-statistical-evidence.pdf.
[3] M.E. Albermann, F. Musshoff, B. Madea, Comparison of ethyl glucuronide (EtG) and fatty acid ethyl esters (FAEEs) concentrations in hair for testing abstinence, Anal. Bioanal. Chem. 400 (2011) 175–181.
[4] E. Alladio, A. Martyna, A. Salomone, V. Pirro, M. Vincenti, G. Zadora, Evaluation of direct and indirect ethanol biomarkers using a likelihood ratio approach to identify chronic alcohol abusers for forensic purposes, Forensic Sci. Int. 271 (2017) 13–22.
[5] J.M. Bernardo, A.F.M. Smith, Bayesian Theory, 2nd ed., John Wiley & Sons, Chichester, 2000.
[6] L.C.A.M. Bossers, R. Paul, Letter to the editor: Application of Bayesian theory to the reporting of results in alcohol hair testing, Forensic Sci. Int. 242 (2014) e56–e58.
[7] J.H. Bügler, H. Buchner, A. Dallmayer, Age determination of ballpoint pen ink by thermal desorption and gas chromatography-mass spectrometry, J. Forensic Sci. 53 (2008) 982–988.

[8] C. Champod, A. Biedermann, C. Vuille, S. Willis, J. De Kinder, ENFSI Guideline for evaluative reporting in forensic science: a primer for legal practitioners, Crim. Law Just. Wkly. 180 (2016) 189–193.
[9] C.L. Crunelle, M. Yegles, M. De Doncker, D. Cappelle, A. Covaci, A.L.N. van Nuijs, H. Neels, Hair ethyl glucuronide concentrations in teetotalers: should we re-evalute the lower cut-off? Forensic Sci. Int. 274 (2017) 107–108.
[10] J.M. Curran, T.N. Hicks, J.S. Buckleton, Forensic Interpretation of Glass Evidence, CRC Press, Boca Raton, FL, 2000.
[11] I.W. Evett, Interpretation: a personal odyssey, in: C.G.G. Aitken, D.A. Stoney (Eds.), The Use of Statistics in Forensic Science, Ellis Horwood, New York, 1991, pp. 9–22.
[12] I.W. Evett, The logical foundations of forensic science: towards reliable knowledge, Philos. Trans. R. Soc. Lond. B: Biol. Sci. 370 (2015) 1–10.
[13] I.W. Evett, B.S. Weir, Interpreting DNA Evidence, Sinauer Associates Inc, Sunderland, 1998.
[14] H. Kharbouche, M. Faouzi, J.B. Sanchez, N. Daeppen, M. Augsburger, P. Mangin, C. Staub, F. Sporkert, Diagnostic performance of ethyl glucuronide in hair for the investigation of alcohol drinking behavior: a comparison with traditional biomarkers, Int. J. Legal Med. 126 (2012) 243–250.
[15] P. Kintz, Consensus of the Society of Hair Testing on hair testing for chronic excessive alcohol consumption 2009, Forensic Sci. Int. (2010) 2.
[16] P. Kintz, Reply to the letter to the editor: caveats against an improper use of hair testing to support the diagnosis of chronic excessive alcohol consumption, following the "Consensus" of the Society of Hair Testing 2009 [Forensic Science International 196 (2010) 2], Forensic Sci. Int. 207 (2011) e71.
[17] P. Kintz, Drugs in hair, in: J.H. Siegel, P.J. Saukko (Eds.), Encyclopedia of Forensic Sciences, 2nd ed., Academic Press, San Diego, 2013, pp. 360–364.
[18] P. Kintz, 2014 consensus for the use of alcohol markers in hair for assessment of both abstinence and chronic excessive alcohol consumption, Forensic Sci. Int. (2015) A1–A2.
[19] D.V. Lindley, Understanding Uncertainty, John Wiley & Sons, Hoboken, 2006.
[20] T.D. Lyon, J.J. Koehler, The relevance ratio: evaluating the probative value of expert testimony in child sexual abuse cases, Cornell Law Rev. 82 (1996) 43–78.
[21] P. Margot, Commentary on 'The need for a research culture in the forensic sciences', Univ. Calif. Law Rev. 58 (2011) 795–801.
[22] R. Marquis, A. Biedermann, L. Cadola, C. Champod, L. Gueissaz, G. Massonnet, W.D. Mazzella, F. Taroni, T. Hicks, Discussion on how to implement a verbal scale in a forensic laboratory: benefits, pitfalls and suggestions to avoid misunderstandings, Sci. Just. 56 (2016) 364–370.
[23] M. Montgomery, M. LeBeau, C. Morris-Kukoski, New hair testing conclusions, J. Anal. Toxicol. (2016) (in press).
[24] F. Pragst, Interpretation problems in a forensic case of abstinence determination using alcohol markers in hair, Forensic Sci. Int. 217 (2012) e4–e7.
[25] F. Pragst, M. Yegles, Alcohol markers in hair, in: P. Kintz (Ed.), Analytical and Practical Aspects of Drug Testing in Hair, CRC Press, Boca Raton, 2007, pp. 287–323.
[26] A. Salomone, L. Tsanaclis, R. Agius, P. Kintz, M.R. Baumgartner, European guidelines for workplace drug and alcohol testing in hair, Drug Test. Anal. 8 (2016) 996–1004.
[27] Society of Hair Testing. 2016 consensus for the use of alcohol markers in hair for the assessment of both abstinence and chronic excessive alcohol consumption (accessed 12.11.17).
[28] F. Sporkert, H. Kharbouche, M. Augsburger, C. Klemm, M. Baumgartner, Positive EtG findings in hair as a result of a cosmetic treatment, Forensic Sci. Int. 218 (2012) 97–100.
[29] F. Tabliaro, F. Bortolotti, G. Viel, S.D. Ferrara, *Caveats* against an improper use of hair testing to support the diagnosis of chronic excessive alcohol consumption, following the "Consensus" of the Society of Hair Testing 2009 [Forensic Science International 196 (2010) 2], Forensic Sci. Int. 207 (2011) e69–e70.
[30] F. Taroni, A. Biedermann, S. Bozza, J. Vuille, M. Augsburger, Toxic substances in blood: an analysis of current recommendations under a Bayesian (decision) approach, Law Probab. Risk 13 (2014) 27–45.
[31] S.M. Willis, L. McKenna, S. McDermott, G. O'Donell, A. Barrett, B. Rasmusson, A. Nordgaard, C.E.H. Berger, M.J. Sjerps, J.J. Lucena-Molina, G. Zadora, C.C.G. Aitken, T. Lovelock, L. Lunt, C. Champod, A. Biedermann, T. N. Hicks, F. Taroni, ENFSI guideline for evaluative reporting in forensic science, Strengthening the Evaluation of Forensic Results Across Europe (STEOFRAE), Dublin, 2015.