Perspective

Drug Testing and Analysis

Received: 26 May 2012 Revised: 10 July 2012 Accepted: 10 July 2012 Published online in Wiley Online Library: 12 September 2012

(www.drugtestinganalysis.com) DOI 10.1002/dta.1398

Validity of stable isotope data in doping control: perspectives and proposals

Ulrich Flenker,[a]* Leo N. Geppert[b] and Katja Ickstadt[b]

$\Delta^{13}C$ and $\delta^{13}C$ values of endogenous urinary steroids represent physiological random variables. Measurement uncertainty and biological scatter likewise contribute to the variances. The statistical distributions of negative controls are well investigated, but there is little knowledge about the corresponding distributions of steroid-users. For these reasons valid discrimination of steroid users from non-users by $^{13}C/^{12}C$ analysis of endogenous steroids requires elaborate statistical treatment. Corresponding Bayesian approaches are presented following an introduction to the rationale. The use of mixture models appears appropriate. The distribution of routine data has been deconvolved and characterized accordingly. The mixture components, which presumably represent steroid users and non-users, exhibit considerable overlap. The validity of a given result depends on both the analytical uncertainty and the prior probability of doping offenses. Low analytical uncertainties but high prior probabilities facilitate valid detection of doping offenses. Two recommendations can be deduced. First, before starting an $^{13}C/^{12}C$ analysis, any initial suspicion should be well-substantiated. This precludes use of permissive criteria derived from the steroid profile. Secondly, knowledge of relevant $^{13}C/^{12}C$ distributions is required. This must cover representative numbers of authentic steroid users. Finally, it is desirable that the conditional probability for steroid administration rather than the measurement uncertainty is calculated and reported. This quantity possesses superior validity and it is largely independent of laboratory bias. The findings suggest and facilitate flexible handling of decision limits. Proposals for the evaluation of stable isotope data are presented. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: Bayesian mixture model; steroids; doping; IRMS; validity

Introduction

Many introductions to specific issues of data analysis will start with an urn model. This is a particular type of gedankenexperiment, favoured by statisticians for a number of reasons. The following one is quoted from Sivia[1] with slight modifications.

Consider an urn containing two types of balls; say, five red and seven green. Now, imagine somebody draws a ball from the urn at random. It is not replaced in the urn and you do not know whether a red or a green ball has been drawn. Then it is your turn and you find a red ball in your hand. Does this observation affect the probability for the – unknown – outcome of the first draw?

There is no physical nexus between the draws, but there is a logical one. This becomes clearer when the number of balls in the urn is changed to one green and eleven red balls. You happen to pick the green ball in the second draw. Surely, i.e. with a probability of 1, a red ball was produced during the first draw. However, before there was knowledge of the outcome of the second draw, the corresponding probability was merely 11/12.

This kind of reasoning resembles the situation in doping control to an astonishing degree. Out there, the pool of athletes contains both dopers and non-dopers. A sample is picked from the pool. Preliminarily, it is of course unknown whether it belongs to an athlete who is a doper or a non-doper. The sample is then shipped to an anti-doping laboratory, where many scientists are busy generating relevant knowledge. Because the results have measurement uncertainties,[2] this knowledge at least partly represents another random event. In turn, this knowledge, affected by random error, serves to estimate, or rather to update, the probability of having picked a doper.

Before any relevant analytical data has become available, the probability of having picked a positive sample obviously equals the proportion of dopers. This proportion is generally known as *prevalence*. In the context employed here, this will be termed the *a priori* or *prior probability*, briefly the *prior*. It can be expressed as a real number between 0 and 1.

The data acquired from the sample then serve to update the prior. The updated probability again falls into [0, 1]. Ideally, it will be close to one of the endpoints of this interval. This probability will be termed the *a posteriori* or *posterior probability*, briefly again, the *posterior*. Obviously, the prior has to be combined with empirical data in order to calculate the posterior.

Often, probabilities are conditional, for example, the probability of picking a red ball in the second draw in the urn model mentioned. In analytical sciences, the quantity's 'specificity' and 'sensitivity' pertain to conditional probabilities. Sensitivity is defined as the probability of a positive result *given* the sample is truly positive. Accordingly, specificity expresses the probability of a negative result *given* the sample is truly negative. These conditional probabilities will be important for all considerations following.

* *Correspondence to: Ulrich Flenker, Institute of Biochemistry, German Sports University, Cologne, Germany. E-mail: u.flenker@biochem.dshs-koeln.de*

a *Institute of Biochemistry, German Sports University, Cologne, Germany*

b *Faculty of Statistics, TU Dortmund University, 44221 Dortmund, Germany*

Bayesian reasoning

There is a fundamental relationship known as Bayes' theorem which relates prior and posterior probabilities. It can be formulated in different ways. For the issues discussed here, the following will be helpful:

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}. \quad (1)$$

Let us consider an easy example to illustrate Equation (1). Let the hypothesis *H* state 'The athlete has administered synthetic steroids'. Let *E* be the event that evidence against the athlete has been found by isotope-ratio mass spectrometry (IRMS) analysis. We are ultimately interested in the posterior probability $P(H|E)$, i.e. the probability that the athlete has administered synthetic steroids given there is evidence *E*.

$P(E|H)$ represents the sensitivity of the IRMS method, i.e. the chance of finding evidence given the athlete has indeed administered synthetic steroids. *P(H)* is the prior probability of *H*, because it gives the probability of *H* before IRMS tests have been carried out. *P(H)* will be discussed in detail later on. *P(E)* is the total probability of a positive result. It is composed of the total proportions of true and false positives and is calculated according to $P(E|H) \times P(H) + P(E|\bar{H}) \times P(\bar{H})$, where $\bar{H}$ reads *non-H*. The second addend may also be expressed as (1-specificity) × (1-prior). So, the sensitivity, the specificity, and the prior probability are required for an estimate.

As mentioned before, *P(H)* is the prior probability that the athlete has administered synthetic steroids. Reasonably, this will be equivalent to our prior degree of belief concerning the status of the sample. If we do not know anything about the sample beforehand, for example, when conducting a routine analysis, *P(H)* is equal to the prevalence of athletes administering synthetic steroids. However, if a pretest has been conducted which leads to conspicuous results, we will doubt the athlete's innocence and increase *P(H)*.

Reasonably, in elite sports, the prior probability of doping offenses will never exactly equal zero. However, in the special case of IRMS analysis – among other criteria – there is currently the testosterone/epitestosterone criterion (T/E).[3] The quantitation of this parameter often predates IRMS analysis. A critical T/E ratio of four has been stipulated. When exceeded, follow-up analysis by IRMS is mandatory.

To illustrate the role of the prior probability, let us consider an example. It is well-documented in the literature that the specificity of IRMS testing is better than 99.9 %.[4,5] For our example, we assume a specificity of 0.999. Let us further assume the sensitivity of the IRMS testing is 0.5, i.e. 50 % of the steroid administrations are correctly detected by the method. This might appear low. But testosterone is metabolized and excreted comparably fast. Moreover, most users will administer the substance discontinuously and possibly in only small amounts. For the time being, let us assume a T/E criterion of six. Roughly 5 % of the T/E ratios larger than six empirically result in clearly positive IRMS tests. Considering the sensitivity of 0.5, the corresponding prevalence may therefore be approximated by 10 %. With these values we get a posterior probability of

$$P(H|E) = \frac{0.5 \times 0.1}{0.5 \times 0.1 + (1 - 0.999) \times (1 - 0.1)} = 0.9823.$$

Now let us assume that the T/E criterion is reduced to four. This results in a smaller prior probability where a reasonable estimate is 0.01. *Ceteris paribus*, the value of $P(H|E)$ will drop to 0.8347. Although the analytical result is still the same, the posterior probability that it is due to a doping offense is only 83%. This invalidation of the method could still be enforced if we decided to apply IRMS to all samples regardless of their steroid profile. Reasonably assuming reduction of the prior probability by another factor of 10 gives $P(H|E) = 0.3336$. Thus, only one in three positive findings is due to a doping offense, whereas the majority is due to other factors, for example, analytical errors.

It is obvious that the prior probability exerts a big influence on the posterior probability of an athlete having administered synthetic steroids given the analytical results. This holds true even if the method has high sensitivity and specificity. For that reason high prior probabilities are advantageous to detect doping offenses successfully. There are several ways to achieve this. One could apply more specific initial measures, any kind of black-box pattern recognition, or simply an expert's rating. Probably, the latter approach represents the most efficient one. Having said this, the attitude to apply IRMS when relevant data suggest the sample will test positive anyway is perfectly rational and reasonable. In Bayesian analysis, it is also valid to update information sequentially. We will come back to this later.

Complementarily, the copious analysis of inconspicuous samples should be avoided. Due to the resulting poor prior probabilities, this will mostly produce low posterior probabilities and thus meaningless data. At the same time the chance to obtain false positive results is considerably increased.

Equation (1) can be rewritten in order to use densities or distributions instead of probabilities

$$\pi(\theta|\mathrm{x}) = \frac{f(\mathrm{x}|\theta)\pi(\theta)}{f(\mathrm{x})}. \quad (2)$$

θ is the parameter of interest, for example, the value of $\Delta^{13}C$, and $\pi(\theta)$ its prior density. If information about the density of the parameter of interest are available they can be incorporated in $\pi(\theta)$. The density of $\mathbf{x}$ given θ is the likelihood $f(\mathbf{x}|\theta)$, which indicates how likely an observation $\mathbf{x}$ is given a specific parameter value θ. $f(\mathbf{x})$ is the so-called marginal density of the data, which normalizes the term. See Gelman *et al.*[6] for details. Again, we are ultimately interested in the value on the left-hand side. $\pi(\theta|\mathbf{x})$ is the posterior density. It combines both the prior density and the likelihood.

IRMS data in doping control

It will be helpful to consider the nature and the meaning of IRMS data acquired for doping control purposes here. The method relies on different isotopic compositions of synthetic and endogenous steroids. The measurands are given by the $^{13}C/^{12}C$ ratios of various steroids. These ratios are expressed as $\delta^{13}C$ [per mil] values relative to an international standard (IS). The IS is assigned a value of 0. Typical values are −22 per mil for endogenous and −28 per mil for synthetic steroids. So both compounds are *depleted* in ^{13}C *vs.* the IS, but to different degrees. Differences between $\delta^{13}C$ values are termed $\Delta^{13}C$ values. Administration of synthetic steroids will typically change the $\Delta^{13}C$ between steroids from different metabolic pathways. Definitions, conventions, and terminology in stable isotopes have recently been compiled by Coplen.[7]

Although a decision limit for relevant $\Delta^{13}C$ values has been stipulated, this should not be confused with a true threshold value. At first sight, this distinction might appear captious. But in case of a threshold, the presence of certain amounts of a prohibited substance is tolerated. While the true concentration remains unknown, the question is whether it exceeds the threshold beyond acceptable levels of uncertainty. The presence of the compound itself is typically not questionable because it mostly represents a xenobiotic.

Qualitatively, the situation is largely different with IRMS. Here, we want to analyze carbon isotopes in order to tell whether synthetic steroids have been administered at all. This is strictly prohibited and there is no tolerable dose, for example, of synthetic testosterone. Consequently, there cannot be a threshold. By contrast, our true interest pertains to the probability of steroid administration, conditional on the analytical data. As should be clear from the previous sections, this probability depends not only on the observed $\Delta^{13}C$ value itself, but also on the prior distribution, which expresses the combination of physiological and analytical random variability.

Inference based on a standard Bayesian analysis

Implicitly, the current method to evaluate IRMS data in doping controls assumes that $\Delta^{13}C$ values lower than -3 per mil (TC - ERC) cannot be reconciled with the absence of synthetic steroids. Consequently, the probability of guilt is by definition identical to the probability that the true $\Delta^{13}C$ value of the sample falls below the decision limit.

It will be clear immediately that this probability is a function of the observed $\Delta^{13}C$ itself and of its measurement uncertainty. However, the prior distribution of this parameter needs to be taken into account as well. A model, which describes the prior distribution of relevant $\Delta^{13}C$ values, is required first. This can be found in reference distributions described in the literature.[4,5,8] Even better, it can be compiled in the respective laboratory. This prior is then combined with a given analytical result (the likelihood) in order to calculate the parameter's posterior distribution, confer Equation (2).

In contrast to many other biological parameters, we are well justified to assume Gaussianity for the reference distribution.[4,5,8] The between-subject variance of $^{13}C/^{12}C$ ratios of endogenous steroids ultimately comes from slightly different proportions of these isotopes in the diet. These deviations add up and result in a Gaussian to be an appropriate model. From internal quality controls it is also well known that analytical errors in IRMS follow Gaussian distributions.

The Gaussianity of both, biological variance, and analytical error, nicely facilitates employment of standard methods.[6,9–11] So, the prior is chosen to be a Gaussian with mean μ_0 and standard deviation τ_0, conventionally denoted by $\mathcal{N}(\mu_0, \tau_0)$.

The posterior in this so-called conjugate model is a Gaussian again which allows for easy interpretation. The posterior mean μ_1 is calculated according to

$$\mu_1 = \mu_0 \frac{1/\tau_0^2}{1/\tau_0^2 + n/\sigma^2} + \bar{x} \frac{n/\sigma^2}{1/\tau_0^2 + n/\sigma^2}, \qquad (3)$$

and the posterior standard deviation τ_1 is calculated according to

$$\tau_1 = \left(1/\tau_0^2 + n/\sigma^2\right)^{-1/2}. \qquad (4)$$

The standard deviation σ of the analytical method must be known beforehand unless there is a sufficiently large number of replicates n. In practice, n will be very small. Therefore σ should be inferred from quality control procedures, which are routinely conducted in laboratories. For a number of reasons, it is recommendable to employ the method of Thompson and Howarth here.[12,13] The procedure exploits the information present in routine samples and thus gives very valid estimates. Briefly, the required precision – the measurement uncertainty – should be defined in advance. Subsequently, routine duplicates serve to check whether the criterion is met. $\bar{x}$ is the empirical mean of the results. It may represent also a single value ($n = 1$).

Our prior belief about the true value of the sample is updated using the IRMS results. This yields the posterior distribution of the value, which is expressed by the Gaussian $\mathcal{N}(\mu_1, \tau_1)$. It is now easy to evaluate the result by the corresponding density.

Figure 1 illustrates the rationale: A routine analysis has yielded a $\Delta\delta$ value of -4 per mil. The measurement uncertainty σ is assumed to be ± 0.6 per mil. The results and the – known – measurement uncertainty serve to calculate the likelihood function of the result. The likelihood is combined with the prior in order to calculate the posterior density of the true $\Delta^{13}C$ value θ. The prior parameters μ_0 and τ_0 have been set to 0 per mil and to ± 1 per mil, respectively. These are arbitrary but reasonable values. The decision limit D is compared to the posterior density. The shaded area expresses the probability that the true value θ falls below D.

Note that this posterior density expresses the probability for a hypothesis *conditional on given analytical data*. By contrast, the uncertainty merely represents an attribute of the data and does not allow for immediate inference. Unless the prior is sufficiently characterized, the uncertainty of a result will not yield any significant information for the purpose intended here. Rather, knowledge about the measurement uncertainty must be considered a prerequisite to allow for reasonable inference.

As can be seen from the example in Figure 1, it is difficult to demonstrate validly that the true value θ falls beyond D, which is assumed to be -3 per mil. The corresponding probability is merely ca. 45% for a single analysis yielding a value of -4.0 per mil. Figure 2 illustrates the effects of varying results ($\Delta^{13}C$) and measurement uncertainties (σ) on the conditional probability for θ to fall below -3 per mil. All other parameters are as

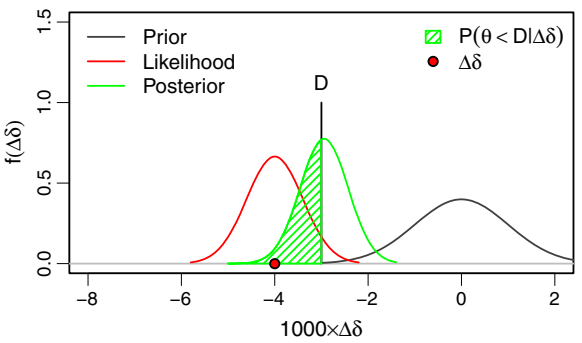

Figure 1. Rationale of standard Bayesian inference based on a single reference prior. $\Delta\delta$: Result (-4.0 per mil); θ: true value of the sample; D: decision limit (-3 per mil); f: probability density; $P(\theta < D|\Delta\delta)$: Probability for θ to fall below D conditional on the result $\Delta\delta$. Mean and standard deviation of the Gaussian prior are assumed to be 0 and 1, respectively. Measurement uncertainty σ assumed to be ± 0.6 per mil (standard deviation).

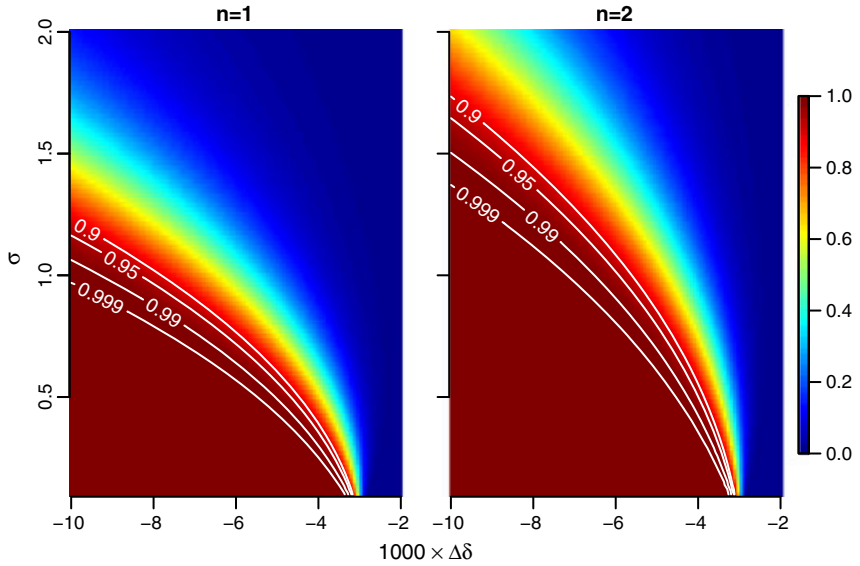

Figure 2. Probabilities for the true $\Delta^{13}C$ value (θ) of a sample to fall below a decision limit (D) of -3 per mil conditional on the measured $\Delta^{13}C$ value and on the measurement uncertainty (σ). Calculations based on a prior mean (μ_0) of 0 per mil and a prior standard deviation (τ_0) of ± 1 per mil. Results are shown for single ($n = 1$) and duplicate analyses ($n = 2$).

described above. The calculations have been performed for single and duplicate analyses. At a measurement uncertainty of ± 0.6 per mil, a single observed value of -6 per mil suffices to meet the conventional criterion with 99.7% probability. However, significant progress will be achieved when sample preparation and analysis are replicated.

Remember that the significance of any evidence also depends on the nature of the prior. Because it has been derived from negative reference samples, the prior does not consider the existence of positive samples at all. As long as real doping controls are merely compared to negative references, it is difficult to validly detect doping offenses. When dealing with probabilities as in Equation (1), this results in a prior probability *P(H)* of 0. The posterior probability for any corresponding criterion to be met likewise must be 0. But when dealing with densities as in Equation (2), the prior probability exceeds 0 for any decision limit. Still the probability for the true value to exceed given limits is low if only reference distributions are taken into account.

Obviously, the knowledge of reference distributions is necessary. But it is not sufficient. Instead, additional knowledge about the distribution of data from true positive samples should be taken into account. We suggest incorporating knowledge of true positive samples via a mixture model.

Inference based on mixture models

By definition, the distribution of steroid isotope ratios reflects distinct sources, namely endogenous and synthetic steroid hormones. Consequently, an appropriate prior distribution can be regarded as a mixture of at least two components. Mathematically, this mixture is simply the sum of its components where each component is given a weight. These weights represent the proportions of the different classes, steroid users and non-users here. Obviously, the sum of the weights must be 1. A comprehensive introduction into mixture models has been presented by Frühwirth-Schnatter.[14]

Assuming Gaussianity holds for each of the components, the calculations for the posterior means and variances can be performed according to Equations (3) and (4). The calculations for a given analytical result are simply performed separately for each class. Consequently, we will end up with one posterior mean and one posterior variance for each class.

However, we are also in need of the posterior weights. These are equivalent to the probabilities of falling into the respective groups and are calculated according to Equation (5). The calculation is immediately derived from Bayes' Theorem (Equation (1)):[14]

$$P(S_i = k|\mathbf{x_i}, \vartheta) = \frac{p(\mathbf{x}_i|\theta_k, S_i = k)\cdot\eta_k}{\sum_{j=1}^{K} p(\mathbf{x}_i|\theta_j, S_i = j)\cdot\eta_j}. \quad (5)$$

The left-hand side denotes the probability P that the sample S with index i belongs to class k conditional on the empirical data set $\mathbf{x}_i$ and on the parameter vector ϑ of the prior distribution. The term $p(\mathbf{x}_i|\theta_k, S_i = k)$ expresses the likelihood of the data $\mathbf{x}_i$, given S_i falls into class k. The likelihood is the product of the probability densities of the data points in $\mathbf{x}_i$. These densities are calculated according to the distribution parameters represented by θ_k, i.e. the mean and standard deviation of the Gaussian corresponding to class k. η_k reflects the *a priori* weight of class k. The denominator contains the sum of these expressions over all K classes. Finally, the posterior probability $P(S_i = k|\mathbf{x_i}, \vartheta)$ is calculated for each class.

Remember that the prior in the standard model is based on reference distributions only and does not take into account the presence of positive samples. Therefore, the algorithm reasonably gives very low posteriors for the extreme values associated with steroid abuse. The results presented in Figure 3 are in principal based on the same data and parameters as employed in Figure 1. However, in addition, the presence of 10% positives has been assumed which are characterized by a prior mean of -6 per mil and by a prior standard deviation of ± 1.25 per mil. Simply compare the resulting posterior distributions.

Obviously, the introduction of mixture models is suited for our purpose. But it should be noted that starting with a high prior probability of picking a positive sample still makes convicting

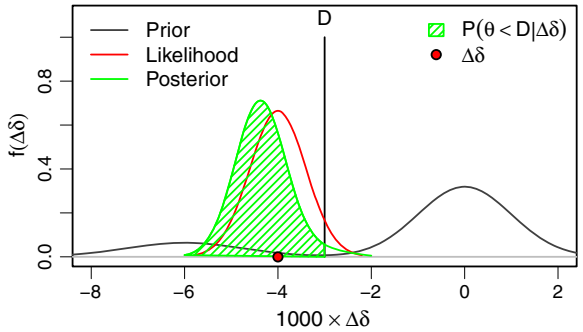

Figure 3. Bayesian inference based on a Gaussian mixture prior. The prior represents 90% negative samples and 10% positive samples; mean and standard deviation of the positives −6 and ±1.25 per mil, respectively. All other data and parameters identical to those in Figure 1 (mean and standard deviation of the negatives 0 and 1, respectively; measurement uncertainty σ ±0.6 per mil).

dopers easier. In case of mixture models, this means that the prior weight for the class, which represents users of synthetic steroids, should not be close to 0. In Figure 4, the effect of different prior proportions of positives is demonstrated. The measured $\Delta^{13}C$ value has been set to −4 per mil while measurement uncertainty σ is still ±0.6 per mil. A single observation has been assumed. With exception of the proportions, the parameters of the prior are unchanged. At a prior proportion of 0 in respect to the positives, the mixture model collapses to the standard model. Assuming a prevalence of 5%, the posterior still gives a considerable chance that the result is not due to a doping offense. Note the comparably high posterior densities for $\Delta^{13}C$ values between −2 and −3 per mil. The vertical line corresponding to the conventional decision limit of −3 per mil may be considered for orientation only. Nonetheless, even an *a priori* proportion of 25% positives still leaves some chance of having picked a negative sample which just exhibits some unusual values.

It has been suggested to employ likelihood ratios in order to assess the strength of given evidence.[9] However, this procedure implicitly assumes identical prior weights for both distributions, i.e. a prior ratio of 1. In a Bayesian context this will be appropriate if no relevant information is available beforehand. This, however, is not the case. By contrast, it has been demonstrated that, for example, the T/E criterion results in only poor prior probabilities for steroid application.[3]

Note that it is possible to calculate the sensitivity and specificity of the testing procedure using the mixture components. Simply speaking, they give the probability of erroneously classifying a sample as belonging to the wrong component. Both sensitivity and specificity depend on the posterior variances of the components, which in turn depend on the biological scatter and the analytical uncertainty, as well as on the posterior weights. This is another reason why it is advantageous not to have a low prior weight for the class of steroid users.

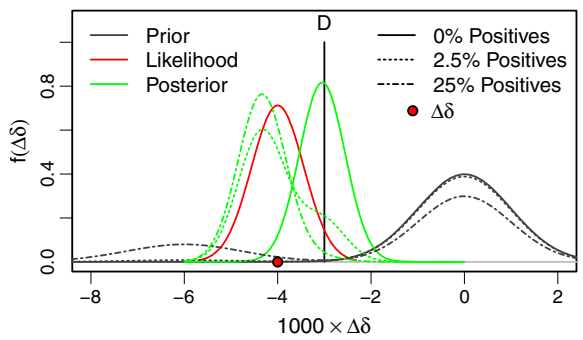

Figure 4. Effects of different prior proportions (0%, 2.5%, and 25%) of positives on the posterior distribution of the true value θ. Single analysis with a result of a $\Delta^{13}C$ value of −4 per mil at a measurement uncertainty of ±0.6 per mil is assumed. Other parameters identical to those in Figure 3 (mean and standard deviation of the negatives 0 and 1 per mil, respectively; mean and standard deviation of the positives −6 and ±1.25 per mil, respectively; measurement uncertainty σ ±0.6 per mil).

Application to empirical data

Application to empirical data requires some assumptions concerning the nature of the respective components, but it is now possible to deconvolve mixture distributions. This falls beyond the scope of this paper, and the reader must be referred to the description of the EM algorithm (expectation maximization) as presented, for example, by the Analytical Methods Committee of the Royal Society of Chemistry (AMC).[15] This method is employed by the `mclust` library[16–18] (model-based clustering) for the `R` statistical computing system[19,20]. The `Mclust` function was applied to the $\Delta^{13}C$ values of androsterone *vs.* 11OH-androsterone from routine samples analyzed in 2009 and 2010. Often, initialization of the IRMS procedure was formally due to T/E ratios >4. But many samples exhibited much higher T/E ratios or were highly conspicuous from other patterns in the steroid profile.

The Bayesian information criterion (BIC[21]) was employed to estimate the number of groups present in the data set. The population is in fact best described by a mixture of two groups with unequal variances. This is suggested by the highest BIC value of −1398 for the two group solution compared to BIC values of −2086 and −1408 for the one and three group solution, respectively.

Table 1 shows the corresponding parameter estimates. Table 1 also shows the parameters of the corresponding reference distribution taken from the literature.[5] These reference parameters and the parameters of group 1 are virtually identical. Group 2 is characterized by $\Delta^{13}C$ values that are depleted *vs.* group 1 and by a large scatter.

Figure 5 shows the resulting probability density functions separately for groups 1 and 2. The segments perpendicular to the *x*-axis correspond to the analytical results. The density function corresponding to the reference population is also shown. The slight shift towards higher densities is due to the fact that this group has a proportion of 100% by definition. If a proportion of 91% of this group in the routine data is assumed, the reference curve will be indistinguishable from that for group 1.

Table 1. Parameters of the distributions of $\Delta^{13}C$ values of androsterone *vs.* 11-OH-androsterone in the reference population[5] (RP) and following deconvolution of routine data into two groups by the EM algorithm

	Mean	Standard Deviation	Proportion
RP	−0.20	±0.48	–
group 1	−0.19	±0.46	91%
group 2	−2.47	±2.56	9%

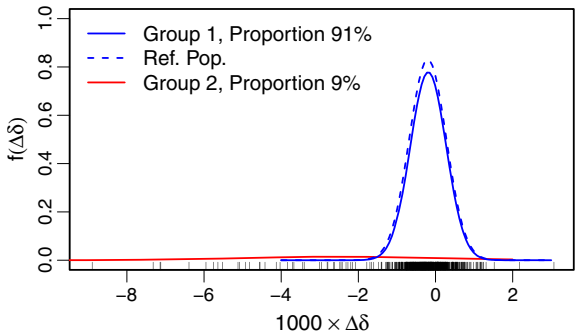

Figure 5. Deconvolution of the density of $\Delta^{13}C$ values (represented by ticks on the *x*-axis) obtained from routine controls in 2009 and 2010 (androsterone – 11OH-androsterone). Parameters of the normal distributions of the components presented in Table 1. The dashed line is the density function of the reference distribution.[5]

As mentioned before, the data are taken from routine analysis. For that reason they reflect a mixture of steroid users and non-users. As the reference population and group 1 are very similar (Table 1), it is very likely that group 1 consists of steroid non-users and group 2 expresses the presence of steroid users in the set of analyzed samples. It can also be seen from Table 1 that the steroid users make up 9% of all samples.

Assuming two distributions with these proportions and parameters it is possible to perform a classification into one of the groups. Additional to the parameters of the mixture distribution the probability to fall in one of the groups is conditional on the analyzed values and on the measurement uncertainty. Assuming the parameters from Table 1, Figure 6 shows the corresponding probabilities to fall in group 2 (the 'positives') in the left panel. The calculations have been performed for single analyses only. Note that for replicates (multiple analyses), Bayesian inference is more complicated. However, for any given real sample the classification based on multiple analyses can be performed with ease.

Instead, the calculations have been repeated for different proportions of group 2. As can be expected, an increase of the prior proportion of steroid users significantly facilitates detection of steroid administration. Note the strong right-shift of the posterior probabilities when the prior proportion is increased to 50%. Because the parameters of the mixture model have been estimated from real samples, these estimates are likely to be realistic. Note that under the current sampling regime, a single $\Delta^{13}C$ value of -3 per mil at a measurement uncertainty of ± 0.6 per mil reflects a probability of 96.5% for steroid abuse. By contrast, this value rises to 99.6% when the prior is increased to 50% dopers. Roughly, this reduces the probability for a wrong decision by a factor of 10. Conversely, this strength of evidence drops to ca. 74% when the prevalence of steroid abuse is assumed to be merely 1% (not shown in Figure 6). This will probably resemble the situation if samples are analyzed by IRMS indiscriminately.

Let us get back to multiple analyses. As an example, let us assume the sample has been prepared and analyzed twice. The results are, say, -2.8 and -3.2 per mil. *Ceteris paribus* and assuming a proportion of 9% dopers, the posterior probability to fall into the group of steroid users increases to roughly 99.99% (as compared to 96.90%). It must be stressed, that this procedure will require replication of the complete method, i.e. sample preparation and analysis.

Discussion and conclusions

In diagnostic testing, 'validity' evidently pertains to the degree to which samples are classified correctly.[22] This concept obviously is also applicable to doping control. However, there is a fundamental difference between clinical and forensic diagnostics. In clinical testing, unambiguous *ex post* classification is typical. In forensics this is not the case and therefore the validity of testing procedures is difficult to assess. The appropriate choice and management of priors is all the more important.

In particular, this applies to the evaluation of physiological parameters, such as stable isotope ratios, blood profiles, etc. The data evaluation must be performed in fundamentally different manner as compared to analytics of xenobiotic

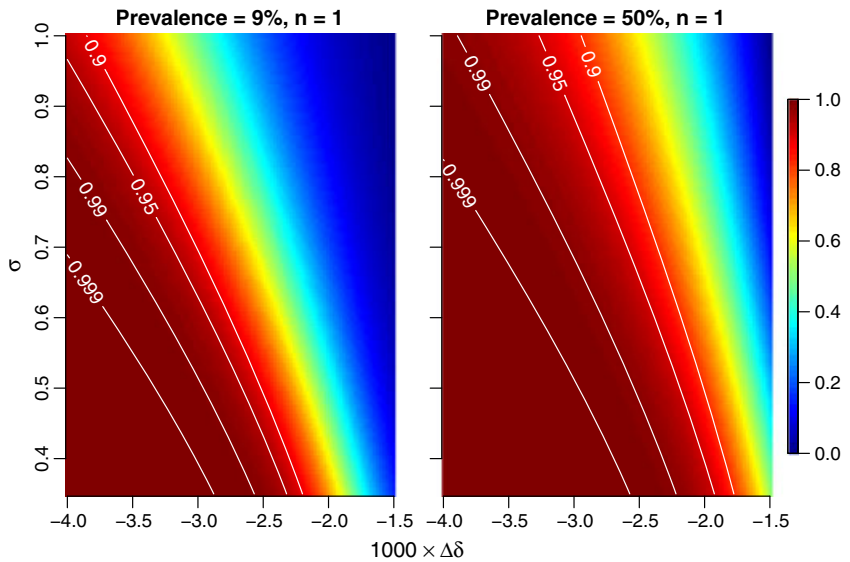

Figure 6. Presumed effects of different proportions of positives (9% and 50%, respectively) on the posterior probability for steroid administration conditional on the result ($\Delta\delta$) and on the measurement uncertainty (σ). Calculations performed for the $\Delta\delta$-values of androsterone *vs*. 11OH-androsterone. All parameters were estimated from routine samples as presented in Table 1.

compounds. Physiological data are statistical by nature. Their variability partly depends on biological factors and partly on analytical bias and measurement uncertainty. Hibbert *et al.*[23] hold a similar view. These authors considered the physiological distribution of total blood CO_2 in horses. They likewise suggested employing a Bayesian approach to validly detect doping offenses in horse racing.

So, elaborate statistical treatment is required here. Actually, the evaluation of IRMS data in doping control represents a classification problem. Although likewise based on quantitative measures, it cannot be overemphasized that the rationale is different from quantitation. In principle, quantitation of xenobiotics may be performed unconditional on further assumptions because the identity of the corresponding signal typically is not questionable. The fact, that, for example, ephedrine is present in a given urine sample will generally be undoubted. However, for the problem discussed here classification is preferably done conditional on analytical data and conditional on prior assumptions and information. Quantitation by contrast, is used to infer a quantity conditional on analytical data but unconditional on further assumptions.

Mere definition of threshold values for physiological measures is therefore over-simplistic and inappropriate. It neglects the fact that the parameters of the statistical distributions corresponding to truly positive or negative samples are required for correct evaluation, i.e. classification. For the mixture of Gaussians considered here, the corresponding means, the variances, and the proportions, i.e. the prior probabilities, are assumed to be known.

The sources of the respective variances are rather insignificant. One could spend considerable efforts deconvoluting the sources without noteworthy gain of information. It is not very important to which degree the scatter originates from analytical error or from biological processes. Therefore the calculation of confidence intervals associated with the estimation of uncertainties is of only minor interest. This will hardly contribute to the creation of significant knowledge. The combined variances must however be modelled realistically.

By contrast, the likelihood of the measurement results conditional on the nature of the sample is really important. The likelihood is conditional on the measurement uncertainty (Equation (5)). However, a rough estimate such as, for example, 'better than ± 0.7 per mil' does suffice.

The fact that the Bayesian approach immediately expresses the probability of a given hypothesis or event is a great advantage. The approach favourably expresses the probability of the doping offense itself. By contrast and as indicated, the expressiveness of uncertainties is rather restricted, unless they are taken into account by the model. Otherwise, these measures apply to nothing else but to a given number resulting from an analytical procedure. Unless an appropriate prior has been defined, there is no immediate nexus between an uncertainty and the probability of a hypothesis in question. Ultimately, analytical results have to be converted into a conditional probability for a doping offense anyway. This always implies a kind of Bayesian reasoning because the result is to be interpreted conditional on numerous factors. Measurement uncertainty represents only one of them.

The evaluation of stable isotope data in doping control according to fixed decision limits should be reconsidered. As has been demonstrated here, the proportion of steroid users present in the tested population plays a pivotal role for the validity of the results. This proportion may vary according to the tested population, but it is preferable that it be as large as possible. Any decision limit 'fit-for-purpose' will have to be more conservative at smaller proportions of steroid-users. Therefore, for testing purposes applying IRMS to inconspicuous samples should be avoided. It does not matter how a suspicion is generated as long as this is done rationally, i.e. based on relevant facts.

Many readers will probably feel uncomfortable with the idea of flexible decision limits and with the suggested drop of uncertainties. This seemingly violates the demand for reproducible scientific results. However, this will not be the case if the probability of a given hypothesis is considered rather than the mere analytical data. One can simply stipulate that the conditional probabilities for steroid administration must exceed critical values. This probability immediately depends on the relevant prior distribution. Assuming different prior distributions additionally provides us with the advantage that possible bias between laboratories can be dealt with.

Another advantage is the possibility of updating priors. This is innate to Bayesian statistics. Posterior information can be used as prior information for subsequent Bayesian analyses. In the situation at hand, one could start, for example, by adopting the distribution parameters presented here. Possibly, the variances initially could be augmented. After several analyses have been performed, these results can be used to update the distribution. Thus, the specific situation in the respective laboratory will be reflected more and more precisely.

In principle, the same rationale can also be applied to only one specific sample. Following administration of synthetic steroids, typically more than one metabolite will exhibit a conspicuous isotope signature. Therefore, each corresponding result out of a sequence of measurements changes the prior probability to obtain another conspicuous one. Thus, each measurement sequentially increases the validity of the overall result. In principle, the calculation of the conditional probabilities is straightforward. But this is beyond the scope of the present study.

For practical purposes, the following conclusions can be derived.

- Indiscriminate testing by IRMS should be avoided.
- By contrast, it will be helpful to analyze many authentic samples in order to describe and to define relevant prior distributions.
- Because these priors are likely to be different in different populations and laboratories, the $\Delta^{13}C$ values that lead to a positive result should be handled flexibly. Note, however, that we suggest fixed limits on the conditional probability of steroid application.
- Rather than the analytical result and its uncertainty, the conditional probability of steroid application is of fundamental interest. Consequently, this entity should be indicated in the reports.
- The validity of positive results improves considerably when conspicuous samples are prepared and analyzed at least in duplicate. It therefore should be considered to stipulate repeated analyses.
- Replication of critical routine samples will also facilitate most valid estimation of measurement uncertainties.

Acknowledgements

The authors wish to thank Prof. Dr Wilhelm Schänzer (Institute of Biochemistry, German Sport University Cologne) for scientific support and two reviewers for valuable comments.

References

[1] D.S. Sivia, J. Skilling. *Data Analysis – A Bayesian Tutorial*, 2nd Edn, Oxford University Press: Oxford, New York, **2006**.
[2] Joint Committee for Guides in Metrology (JCGM). Evaluation of measurement data – Guide for the expression of uncertainty in measurement. Available at: http://www.bipm.org/en/publications/guides/gum.html [1 August 2012].
[3] U. Mareck, H. Geyer, G. Fusshöller, A. Schwenke, N. Haenelt, T. Piper, *et al*. Reporting and managing elevated testosterone/epitestosterone ratios - novel aspects after five years' experience. *Drug Test. Anal.* **2010**, *2*, 637.
[4] U. Flenker, U. Güntner, W. Schänzer. δ^{13}C-values of endogenous urinary steroids. *Steroids* **2008**, *73*, 408.
[5] T. Piper, U. Mareck, H. Geyer, U. Flenker, M. Thevis, P. Platen, *et al*. Determination of C-13/C-12 ratios of endogenous urinary steroids: method validation, reference population and application to doping control purposes. *Rapid Commun. Mass Spectrom.* **2008**, *22*, 2161.
[6] A. Gelman, J. Carlin, H. Stern, D. Rubin. *Bayesian Data Analysis*, 2nd Edn, Chapman & Hall/CRC: Boca Raton, London, New York, Washington DC, **2004**.
[7] T. Coplen. Guidelines and recommended terms for expression of stable-isotope-ratio and gas-ratio measurement results. *Rapid Commun. Mass Spectrom.* **2011**, *25*, 2538.
[8] T. Piper, U. Flenker, U. Mareck, W. Schänzer. $^{13}C/^{12}C$ ratios of endogenous urinary steroids investigated for doping control purposes. *Drug Test. Anal.* **2008**, *1*, 65.
[9] C. Aitken, F. Taroni. *Statistics and the Evaluation of Evidence for Forensic Scientists*, 2nd Edn, Wiley: Chichester, **2004**.
[10] Analytical Methods Committee. A glimpse into Bayesian statistics. AMC Technical Brief No. 14. Royal Society of Chemistry: London, **2003**. Available at: http://www.rsc.org/images/bayesian-statistics-technical-brief-14_tcm18-214861.pdf [1 August 2012].
[11] P.D. Hoff. *A First Course in Bayesian Statistical Methods*. Springer: New York, **2009**.
[12] M. Thompson, R. Howarth. New approach to estimation of analytical precision. *J. Geochem. Explor.* **1978**, *9*, 23.
[13] Analytical Methods Committee. A simple fitness-for-purpose control chart based on duplicate results obtained from routine test materials. AMC Technical Brief No. 9. Royal Society of Chemistry: London, **2002**. Available at: http://www.rsc.org/images/duplicate-results-technical-brief-9_tcm18-214876.pdf [1 August 2012].
[14] S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer: New York, **2006**.
[15] Analytical Methods Committee. Mixture models for describing multimodal data. AMC Technical Brief No. 23. Royal Society of Chemistry: London, **2006**. Available at: http://www.rsc.org/images/models-describing-multimodal-data-technical-brief-23_tcm18-214842.pdf [1 August 2012].
[16] C. Fraley, A.E. Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. *J. Class.* **2007**, *24*, 155.
[17] C. Fraley, A.E. Raftery. MCLUST Version 3 for R: Normal Mixture Modeling and Model-based Clustering **2006**. (Revised in 2009).
[18] C. Fraley, A.E. Raftery. Model-based clustering, discriminant analysis and density estimation. *J. Am. Stat. Ass.* **2002**, *97*, 611.
[19] R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria **2012**. Available at: http://www.R-project.org/ [1 August 2012].
[20] R Development Core Team. An Introduction to R **2008**. Available at: http://cran.r-project.org/ [1 August 2012].
[21] G. Schwarz. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461.
[22] J. Buttner. Diagnostic validity as a theoretical concept and as a measurable quantity. *Clin. Chim. Acta* **1997**, *260*, 131.
[23] D. Hibbert, N. Armstrong, J. Vine. Total CO_2 measurements in horses: where to draw the line. *Accr. Qual. Assur.* **2011**, *16*, 339.