New statistical framework for interlaboratory evaluation of anti-doping testing results by WADA

Meija, Juris; Possolo, Antonio; Garrido, Bruno Carius; Kisoona, Sanjana; Barroso, Osquel

Publisher's version / Version de l'éditeur:
https://doi.org/10.1007/s00769-024-01595-w
Accreditation and Quality Assurance, special issue, 2024-05-08

Accreditation and Quality Assurance
https://doi.org/10.1007/s00769-024-01595-w

RESEARCH

New statistical framework for interlaboratory evaluation of anti-doping testing results by WADA

Juris Meija[1] · Antonio Possolo[2] · Bruno Carius Garrido[3] · Sanjana Kisoona[4] · Osquel Barroso[4]

Received: 20 January 2024 / Accepted: 26 March 2024

Abstract
The World Anti-doping Agency (WADA) International Standard for Laboratories (ISL), developed as part of the World Anti-Doping Program, requires satisfactory laboratory performance in the WADA External Quality Assessment Scheme (EQAS) in order to obtain and maintain WADA accreditation. Under this mandate, WADA regularly distributes urine and blood test samples to anti-doping laboratories to continuously monitor their proficiency. Over the years, WADA has employed classical, generic statistical methods, in accordance to ISO 13528, to evaluate quantitative EQAS results. Here, we set out the rationale for a modern statistical approach that recognizes and addresses the particular features of the measurement results typically obtained in such tests and present an approach involving Bayesian measurement models and statistical data analysis that is tailored specifically to anti-doping testing.

Keywords Interlaboratory comparisons · Proficiency tests · Bayesian methods · Quality assurance · Anti-doping testing

Introduction

The main goals of the External Quality Assessment Scheme (EQAS) administered by the World Anti-doping Agency (WADA) include the following: (i) to evaluate the testing proficiency of WADA-accredited laboratories; (ii) to improve the uniformity and mutual consistency of test results between these laboratories; and (iii) to provide educational opportunities to them. There are currently 30 WADA-accredited anti-doping testing laboratories worldwide, which enables data-rich intercomparisons of their measurement results.

The WADA EQAS framework consists of three types of regular interlaboratory proficiency testing activities [3, 45]:

i) *Blind EQAS*, where laboratories are aware that the samples originate from EQAS because the samples are delivered to the laboratories by the WADA EQAS sample provider. However, the laboratories do not know the composition of the samples. Typically, there are three rounds of blind EQAS tests per year, each comprising five samples, for a total of 15 samples per year. Such comparisons are similar to those organized frequently by Consultative Committees of the International Committee for Weights and Measures (CIPM), under the *Mutual Recognition Arrangement* (CIPM MRA) — a framework through which National Metrology Institutes regularly demonstrate the equivalence of their measurement capabilities [42].

ii) *Double-blind EQAS*, where laboratories are not aware that the samples are EQAS samples since they are delivered by Anti-doping Organizations acting as testing authorities and are indistinguishable from routine anti-doping samples. Neither do the laboratories know the composition of the samples. Such a testing scheme is also adopted in some forensic laboratory testing [27]. These testing samples are

✉ Juris Meija
juris.meija@nrc-cnrc.gc.ca

Antonio Possolo
antonio.possolo@nist.gov

Bruno Carius Garrido
garridobrunoc@gmail.com

Sanjana Kisoona
sanjana.kisoona@wada-ama.org

Osquel Barroso
osquel.barroso@wada-ama.org

[1] National Research Council Canada, Ottawa, ON, Canada

[2] National Institute of Standards and Technology, Gaithersburg, MD, USA

[3] National Institute of Metrology, Quality and Technology, Duque de Caxias, Brazil

[4] World Anti-Doping Agency, Montreal, QC, Canada

Published online: 08 May 2024

delivered by the WADA EQAS sample provider to Anti-doping Organizations, which incorporate them into their regular testing missions for delivery to the laboratories. The *double-blind EQAS* typically consists of five samples per year delivered over three rounds. However, laboratories providing services during major sports events, such as the Olympic and Paralympic Games, receive additional sets of *double-blind EQAS* samples during the games [47].

iii) *Educational EQAS*, where samples may be provided as non-blinded samples, in which case both the provenance and contents of the EQAS sample are known, or as blind or double-blind samples. This approach is used for educational purposes or for data gathering and involves 2–3 samples each year.

A typical, regular EQAS round consists of several urine samples containing prohibited substances and/or their metabolites or precursors. EQAS samples are prepared either through spiking the substances of interest into a blank urine matrix or, preferentially (especially for the *double-blind EQAS*), are obtained from human excretion studies. WADA follows the ISO/IEC 17043 requirements for the production of EQAS samples. Therefore, the suitability of the EQAS samples is evaluated beforehand, which includes testing for batch homogeneity, stability assessment at different temperatures (−20 °C, +4 °C, +20 °C, and +40 °C) aimed at simulating the sample storage and transportation conditions, and determining the levels of prohibited substances in them. Blank urine samples are also routinely distributed as part of EQAS.

Testing reports for EQAS samples include qualitative results (attesting the presence or absence of prohibited substances) and quantitative results for prohibited threshold substances, markers of the urinary steroid profile, specific gravity of urine (SG), and isotopic composition of carbon as determined by gas chromatography-combustion-isotope ratio mass spectrometry (GC/C/IRMS). This study concerns only the evaluation of the quantitative results.

Rationale for change

The statistical evaluation of the quantitative results from WADA EQAS tests has traditionally relied on the statistical methods that ISO 13528:2022 [15] recommends for use in proficiency testing by interlaboratory comparison. The first edition of this ISO guide was published in 2005, 12 years after the publication of an International Harmonized Protocol for Proficiency Testing of Laboratories of Analytical Chemistry, as a Technical Report of the International Union of Pure and Applied Chemistry (IUPAC) [37], which has since been updated [39].

The main reason to deviate from the recommendations presented in the Annex C of ISO 13528 is that these involve *ad hoc* statistical methods motivated by generic considerations of "robustness," rather than being based on explicit statistical measurement models that can be tuned to be responsive to the specific features of the measurement results obtained for samples relevant to the WADA mission and satisfy criteria of transparency and scientific rigor that WADA abides by.

In addition to inadequate statistical modeling, current statistical data evaluations (often described as "data reduction") pay little attention to data structure and ignore informative parts of the data. For example, traditional data reduction methods do not take into account the measurement uncertainty associated with the reported measured values, most notably for prohibited threshold substances and carbon isotope ratio measurements, whose associated measurement uncertainties are simply ignored (left panel of Fig. 1).

The left panel of Fig. 1 also shows that the measurement results that the laboratories provide for the same measurand, which is carbon-13 isotope delta for pregnanediol in this example, can be mutually inconsistent, in the sense that the laboratory results are more dispersed than what their associated uncertainties suggest they should be. Such mutual inconsistency is a source of uncertainty that the current procedures of data reduction ignore. This is the *dark uncertainty* [36] that reveals itself only once the results from the different laboratories are intercompared.

The traditional data reductions also ignore the fact that the measurement errors affecting the determinations of most steroid profile markers and threshold substances are approximately proportional to the measured values (center panel of Fig. 1) and proceed on the unrealistic assumption that measurement errors are unrelated to the levels of the measurands of interest.

Finally, the current data reduction methods do not take into account the rounding that the Technical Document on Decision Limits (TD DL) [48] requires for some measured values. For example, the urine SG must be reported to three decimal digits. The right panel of Fig. 1 shows that this rounding can be a source of uncertainty that the traditional methods of data analysis are unable to take into account.

Besides addressing the aforementioned deficiencies, we also show that the rank-sum analysis of the steroid profile marker results, which WADA currently employs to identify persistent laboratory biases, is unnecessary.

New statistical framework

The EQAS interlaboratory comparison data are modeled using what ISO 5725-1 calls a "basic statistical model" (ISO 5725-1:1994 Art. 5.1), by considering that every measured value in a sample is the sum of three components: the true quantity value of the analyte in the sample,

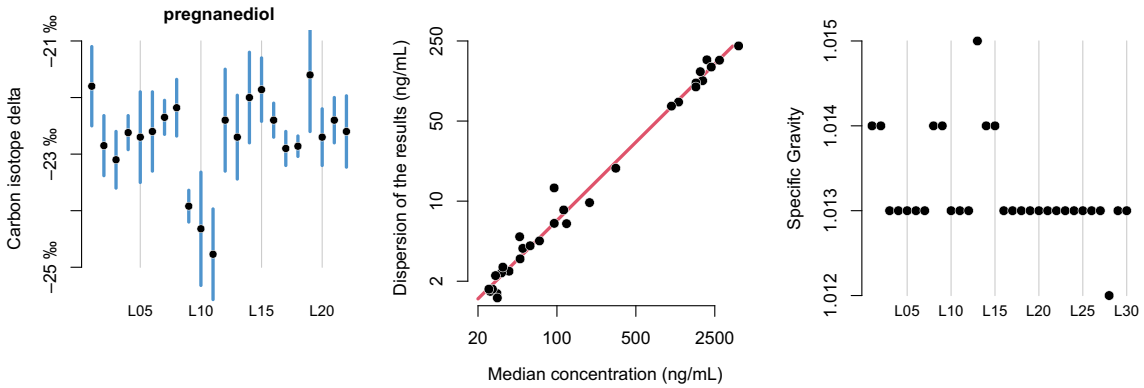

Fig. 1 Features of the WADA EQAS results that the current data reduction methods neither recognize nor address. LEFT PANEL: The standard uncertainties associated with the carbon isotope ratio measurements for pregnanediol (half the length of each vertical line) and the *dark uncertainty* [36] attributable to the "excess" dispersion of the reported results above and beyond what their standard uncertainties suggest, are both ignored. CENTER PANEL: The dispersion around the median results of steroid profile marker measurements is proportional to the median concentration of the steroid markers in the analyzed samples. RIGHT PANEL: The specific gravity (or, relative density) measurements of urine samples are reported to three decimal digits, which is an appreciable, yet disregarded source of uncertainty. The panels show anonymized results from the blind EQAS-2017-2 (left) and EQAS-2023-1 (center, right) intercomparison rounds

the laboratory bias, and the measurement error. However, neither ISO 5725-1:1994 nor ISO 13528 links the assumptions made about the components of that "basic model" to procedures of data reduction consistent with these assumptions. ISO 13528 recommends generic "robust" procedures intended to produce valid inferences across a broad spectrum of assumptions [15] without considering explicit statistical measurement models tailored to the data at hand.

Instead, we adopt explicit models that can recognize and take into account the specific features of the data obtained in WADA interlaboratory studies, which are mixed effects models [29], including laboratory random effects models [40, 41], which have a long history of use in measurement science, in the context of interlaboratory studies [23, 24] and of round-robins [8].

In each round of EQAS interlaboratory studies, the participating laboratories measure the levels of six steroids and the testosterone to epitestosterone ratio (markers of the urinary steroid profile) in several, typically five, samples that generally have different steroid profiles. Similarly, each laboratory also provides measurement results for the SG of all urine samples.

The mixed effects measurement models that we employ can leverage measurements that each laboratory makes of the same analyte in the several samples of interest, to assess the relative measurement repeatability of each laboratory.

Data reduction methods outlined in ISO 13528 give pride of place to the median [15, Annex C] in multiple roles. Indeed, the median is often employed to summarize the interlaboratory comparisons conducted by the CIPM Consultative Committee for Amount of Substance (CCQM) [25]. Modeling laboratory biases and measurement errors as random variables with Laplace distributions affords a similar outcome because the resulting estimates of consensus values are approximately weighted medians [34]. We call "consensus value" an estimate of a true value like μ in Equation (1) and in subsequent statistical measurement models.

We adopt Bayesian methods not only because they rely on explicit statistical modeling, but also because they deliver robust performance. Typically, two facets of robustness are relevant: robustness of validity and robustness of efficiency [28, p.16]. The former concerns the accuracy of confidence intervals, for example, while the latter concerns the length of the same intervals (the shorter the better, provided they are similarly accurate). Properly selected and calibrated Bayesian models can deliver both these aspects of robustness similar to conventional robust procedures, while also accomplishing so much more than these conventional procedures can accomplish.

For example, Bayesian models offer the means to express prior knowledge not only about parameters in the statistical model, while recognizing the particular constraints that they must satisfy, but also about measurement uncertainty [30, p.214]. In addition, Bayesian models excel at propagating contributions from sources of uncertainty, including the presence of dark uncertainty that expresses laboratory biases [26], as previously demonstrated in the context of interlaboratory studies that function as proficiency tests [9]. Reference [26] provides an overview of Bayesian models and methods and describes the underlying concepts, illustrated with examples of application to measurement problems.

Threshold substances

Some prohibited substances are subjected to threshold levels that represent the maximum permissible levels of the

substance in a doping control sample. This includes stimulants (such as ephedrine or methamphetamine), opioids (morphine), beta-2 agonists (such as salbutamol or formoterol), cannabinoids (carboxy-THC), and peptide hormones such as human growth hormone (hGH) and chorionic gonadotrophin (hCG) [48]. Therefore, the determinations of these substances in doping control require the application of quantitative analytical procedures.

The ISL requires that results from quantitative confirmation procedures applied to threshold substances be based on the mean of three independent determinations [45]. In this vein, averages of replicated determinations of concentrations of threshold substances reported by each laboratory (X_l) are described using mixed effects statistical models with proportional laboratory biases and measurement errors:

$$X_l = \mu \cdot (1 + B_l + E_l), \text{ for } l = 1, \ldots, L, \tag{1}$$

where L denotes the number of laboratories, μ represents the true value of the threshold substance concentration in the distributed EQAS sample, B_l is the laboratory-specific relative bias, and E_l is the laboratory-specific relative measurement error. The true value, μ, is treated as a fixed (or, persistent) effect, and B_l and E_l are both treated as random (or, volatile) effects.

We use a weakly informative, zero-truncated Gaussian prior distribution for μ, centered at the median of the reported measured values, and with coefficient of variation (relative standard deviation) of 50%, a choice recognizing that μ cannot be negative and that it is believed to lie in the general vicinity of the reported results:

$$\mu \sim \text{GAUSS}(\text{mean} = \text{median}(X),\ \text{sd} = \text{median}(X)/2) \quad (\mu \geq 0). \tag{2}$$

Relative laboratory biases, B_l, are modeled as outcomes of a truncated Laplace distribution whose mean and standard deviation (before truncation) are 0 and τ (both with the same units as X_l) :

$$B_l \sim \text{LAPLACE}(\text{mean} = 0,\ \text{sd} = \tau) \quad (B_l \geq -100\%). \tag{3}$$

The choice of Laplace distribution for the laboratory biases ensures that the consensus value, which is the estimate of μ, is approximately a weighted median of the measurement results, with data-dependent weights estimated in the process of fitting the model to the data. In addition, the relative bias cannot be smaller than −100% because the threshold substance concentrations cannot be negative.

The prior distribution for the standard deviation of the relative laboratory biases (the parameter τ that quantifies the dark uncertainty) is modeled with a half-Cauchy distribution whose median is set to half of the maximum allowed relative measurement uncertainty established by WADA in the TD DL [48]:

$$\tau \sim \text{CAUCHY}(\text{median} = u_{\text{c,Max}}/2) \quad (\tau \geq 0) \tag{4}$$

The use of the half-Cauchy distribution to model variance components and the choices for the prior distribution of the laboratory biases are as recommended by [13], widely accepted [30], and adopted and implemented in the *NIST Consensus Builder* [18].

Relative measurement errors are also modeled as outcomes of a truncated Laplace distribution with zero mean and laboratory-specific relative standard deviation, $u_{\text{c},l}$:

$$E_l \sim \text{LAPLACE}(\text{mean} = 0,\ \text{sd} = u_{\text{c},l}) \quad (E_l \geq -100\%). \tag{5}$$

The relative laboratory-specific measurement uncertainty, $u_{\text{c},l}$, is provided by the laboratories, and the input data and the parameters of this statistical measurement model are summarized in Table 1.

In some cases, a classical linear mixed effects model [29] of the form $Y_l = \mu + B_l + E_l$, where $Y_l = X_l$ or $Y_l = \log(X_l)$ delivers results similar to those corresponding to the model in Equation (1), when such linear model is fitted to the same data using the method of restricted maximum likelihood (REML), for example as implemented in the R function `lmer` of package `lme4` [2]. The Bayesian counterparts of this classical model, for example as implemented in R function `stan_lmer` defined in R package `rstanarm` [4, 14], can also deliver similar results.

Steroid profile

All urine samples are subjected to the measurement of steroidal markers, including the following six endogenous anabolic androgenic steroids (EAAS) in the Urinary Steroid

Table 1 Overview of the input data and measurement model parameters for threshold substance measurements

Type	Symbol	Description
Data	X_l	Measured value reported by laboratory l ($l = 1, \ldots, L$)
Data	$u_{\text{c},l}$	Relative measurement uncertainty reported by laboratory l
Data	$u_{\text{c,Max}}$	Maximum allowed relative measurement uncertainty set by WADA
Parameter	μ	Consensus value
Parameter	B_l	Relative laboratory bias for laboratory l
Parameter	τ	Standard deviation of the relative laboratory biases $B_1, \ldots, B_L$

Profile of the Athlete Biological Passport [46]: androsterone, etiocholanolone, 5α-androstane-3α,17β-diol, 5β-androstane-3α,17β-diol, testosterone (T), and epitestosterone (E). In addition, the T/E ratio is also reported by laboratories.

Unlike for threshold substances, a single reported value for each steroid profile marker is reported for each urine sample. These reported results are modeled as follows:

$$X_{s,l} = \mu_s \cdot (1 + B_l + E_{s,l}), \tag{6}$$

for each sample s and laboratory l, where μ_s denotes the true steroid concentration in the distributed urine sample $s = 1 \ldots S$, B_l is the laboratory-specific relative bias for laboratory $l = 1 \ldots L$, and $E_{s,l}$ is the laboratory-specific relative measurement error incurred by laboratory l when measuring sample s for a particular steroid.

This statistical model assumes that the relative laboratory biases and the relative measurement errors are the same across all samples for each particular steroid—an assumption that is corroborated by the historical data from previous EQAS rounds. In other words, the laboratory biases and measurement errors for each steroid are proportional to the concentrations of the steroid markers. The choice of proportional errors is also supported by comparing the adequacy to the data obtained with several alternative measurement models. The assumption of proportional errors can easily be extended to allow for constant errors at low analyte levels [33].

We use a weakly informative, zero-truncated Gaussian prior distribution for the consensus values μ_s centered at the median of the reported measured values, with a large relative standard deviation (50%):

$$\mu_s \sim \text{GAUSS}(\text{mean} = \text{median}(X_s),\ \text{sd} = \text{median}(X_s)/2) \quad (\mu_s \geq 0). \tag{7}$$

The same as before, this choice recognizes that the true values of the mass concentrations cannot be negative and that they lie in the general vicinity of the reported measured values.

Relative laboratory biases are modeled as outcomes of a truncated Laplace distribution with zero mean and standard deviation τ (before truncation), to prevent steroid marker concentrations from being negative:

$$B_l \sim \text{LAPLACE}(\text{mean} = 0,\ \text{sd} = \tau) \quad (B_l \geq -100\%). \tag{8}$$

This Laplace distribution for the laboratory biases ensures that the consensus values are approximately weighted medians of the measured values, and the truncation is applied to reflect that steroid marker concentrations cannot be negative.

The prior distribution for the standard deviation of relative laboratory biases is modeled with a half-Cauchy distribution whose median is set to half of the maximum allowed relative measurement uncertainty established by WADA [46]:

$$\tau \sim \text{CAUCHY}(\text{median} = u_{\text{c,Max}}/2) \quad (\tau \geq 0). \tag{9}$$

Since the the half-Cauchy distribution has a very heavy right tail, the aforementioned prior distribution amounts to an order-of-magnitude guess for the prior distribution of τ.

Since each laboratory provides a single measured value for each sample, to evaluate the dispersion of the relative biases of the laboratories one leverages the fact, illustrated in the two examples shown in Fig. 2, that laboratories that produce biased results for a particular steroid, tend to do so consistently across all urine samples analyzed in the same round. The mixed effects model can then effectively assess the laboratory biases, which are persistent across samples, separately from the volatile measurement errors.

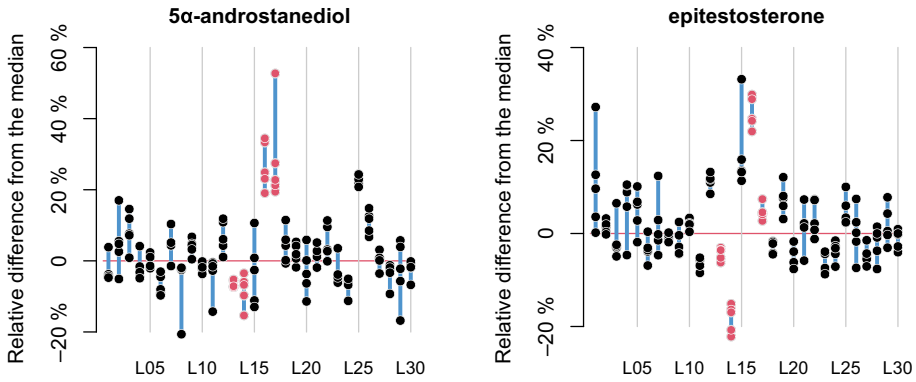

Fig. 2 Relative deviation of the reported steroid results from their median (laboratory bias) across five urine samples. Each dot represents a relative difference, $(X_{l,s} - A_s)/X_{l,s}$, between the value, $X_{l,s}$, that laboratory $l = 1, \ldots, 30$ reported for sample $s = 1, \ldots, 5$, and the median A_s of the values of all the laboratories for sample s. In these particular examples from EQAS-2023-1, laboratories 13 and 14 measured the mass concentration of both 5α-androstanediol and epitestosterone persistently low in all five samples, and laboratories 16 and 17 measured the same measurands persistently high

Relative measurement errors are modeled as outcomes of a Laplace distribution with zero mean and a laboratory-specific relative standard deviation u_l:

$$E_{s,l} \sim \text{LAPLACE}(\text{mean} = 0,\ \text{sd} = u_l) \quad (E_{s,l} \geq -100\%). \tag{10}$$

The prior distribution for the relative measurement uncertainties is a half-Cauchy distribution whose median equals half the maximum relative uncertainty established by WADA in the TD EAAS [46]:

$$u_l \sim \text{CAUCHY}(\text{median} = u_{\text{c,Max}}/2) \quad (u_l \geq 0). \tag{11}$$

These statistical measurement models for the steroid marker concentrations are somewhat different from the models for the threshold substance results because the laboratories do not report laboratory-specific measurement uncertainty evaluations. The input data and the parameters of this statistical measurement model are summarized in Table 2.

It would be very beneficial if the reporting requirements would include the specification of such uncertainties, and, in addition, would provide sufficient information to extend the model so that it could take into account correlations between laboratory biases, and between laboratory-specific measurement errors, for different steroid markers. In addition, a more elaborate error model could be adopted to recognize measurement errors near quantification limits [44].

Carbon isotope ratio measurements

Measurements of carbon isotope delta values, $\delta(^{13}\text{C})$, for steroids, made using GC/C/IRMS, are reported for urine samples which require the application of this procedure [49]. These results are expressed on the international VPDB scale as parts-per-thousand (permille, ‰) deviations [6] and, like threshold substance measurements, are accompanied by explicit measurement uncertainties.

In addition to evaluating $\delta(^{13}\text{C})$ measurements of target compounds (such as androsterone or testosterone) and endogenous reference compounds (such as pregnanediol), the difference in the $\delta(^{13}\text{C})$ values between target compounds and endogenous reference compounds is also evaluated [49].

The results for a given steroid are modeled with a linear mixed effects statistical model:

$$X_l = \mu + B_l + E_l, \tag{12}$$

where μ denotes the true value for the given steroid, B_l is the laboratory bias for measurements of that same steroid, and E_l is the laboratory-specific measurement error incurred by laboratory l when measuring that steroid. This statistical model recognizes that the laboratory biases and measurement errors are largely unaffected by the magnitude of $\delta(^{13}\text{C})$.

We use a weakly informative Gaussian prior distribution for μ that is centered at the median of the reported results for each sample, with a large standard deviation:

$$\mu \sim \text{GAUSS}(\text{mean} = \text{median}(X),\ \text{sd} = 5\ ‰). \tag{13}$$

This choice of prior distribution for μ serves simply to place the consensus value, which is the estimate of μ, in the general vicinity of the measured values, the implied assumption being that, as a group, the participating laboratories are well calibrated.

Laboratory biases are modeled as outcomes of a Laplace distribution with zero mean and standard deviation τ:

$$B_l \sim \text{LAPLACE}(\text{mean} = 0,\ \text{sd} = \tau). \tag{14}$$

As before, this Laplace distribution for the laboratory biases ensures that the consensus values are approximately weighted medians of the measured values. No truncation is applied here since isotope delta values can be negative or positive.

The prior distribution for the standard deviation of laboratory biases is modeled with a half-Cauchy distribution whose median is set to half of the maximum allowed relative measurement uncertainty established by WADA [46]:

$$\tau \sim \text{CAUCHY}(\text{median} = u_{\text{c,Max}}/2) \quad (\tau \geq 0). \tag{15}$$

Laboratory-specific measurement errors are modeled as outcomes of a Laplace distribution with zero mean and laboratory-specific standard deviation, $u_{\text{c},l}$:

$$E_l \sim \text{LAPLACE}(\text{mean} = 0,\ \text{sd} = u_{\text{c},l}). \tag{16}$$

Table 2 Overview of the input data and measurement model parameters for steroid profile measurements

Type	Symbol	Description
Data	$X_{s,l}$	Reported value measured by laboratory l for sample s
Data	$u_{\text{c,Max}}$	Maximum allowed relative measurement uncertainty set by WADA
Parameter	μ_s	True steroid marker concentration for sample s
Parameter	B_l	Relative bias of laboratory l
Parameter	τ	Standard deviation of relative laboratory biases
Parameter	u_l	Standard deviations of relative measurement errors for laboratory l

The laboratory-specific measurement uncertainty, $u_{c,l}$, is reported by the laboratories. The input data and the parameters of this statistical measurement model are summarized in Table 3. The statistical model adopted here for carbon isotope delta measurements follows closely the Laplace random effects model for interlaboratory studies [34] which has been implemented in the *NIST Consensus Builder* [17, 18].

Specific gravity (SG) of urine

SG of urine (which is the ratio of its density to that of water at 4 °C, and that IUPAC calls relative density, d) plays an important role in anti-doping analysis because the decision limits (DL) established for threshold substances, and the estimated concentrations of non-threshold substances for which WADA has established minimum reporting levels (MRLs), are adjusted when the SG of a sample, d_s, exceeds 1.018 [48]. This adjustment involves a factor,

$$k_s = \frac{1.020 - 1.000}{d_s + 2u_{c,\mathrm{Max}} - 1.000},$$

which makes the SG measurements particularly influential.

SG measurements are reported as a single value for each sample and rounded to three decimal digits. These results are modeled using a linear mixed effects statistical measurement model with an added consideration for the effect of rounding, which is imposed on all laboratories in accordance with the TD DL [48]:

$$X_{s,l} = \mu_s + B_l + E_{s,l} + R_{s,l}. \tag{17}$$

Here, μ_s is the true value of the specific gravity for urine sample s, B_l is the bias of laboratory l, $E_{s,l}$ is the laboratory-specific measurement error incurred when laboratory l measures sample s, and $R_{s,l}$ is the effect of rounding.

We use a weakly informative, zero-truncated Gaussian prior distribution for μ_s, centered at the median of the values measured for each sample, with large standard deviation that covers the SG values normally seen in the population [19]:

$$\mu_s \sim \mathrm{GAUSS}(\mathrm{mean} = \mathrm{median}(X_s),\ \mathrm{sd} = 0.020) \quad (\mu_s \geq 0). \tag{18}$$

The same as before, this choice recognizes that the consensus values cannot be negative and that they lie in the general vicinity of the reported measured values.

The laboratory biases are modeled as outcomes of a Laplace distribution with zero mean and standard deviation τ:

$$B_l \sim \mathrm{LAPLACE}(\mathrm{mean} = 0,\ \mathrm{sd} = \tau). \tag{19}$$

The choice of a Laplace distribution for the laboratory biases ensures that the consensus values are approximately weighted medians of the corresponding measured values.

The prior distribution for the standard deviation of laboratory biases is a half-Cauchy distribution whose median is set to half of the maximum allowed measurement uncertainty established by WADA (currently, $u_{c,\mathrm{Max}} = 0.001$) [48]:

$$\tau \sim \mathrm{CAUCHY}(\mathrm{median} = u_{c,\mathrm{Max}}/2) \quad (\tau \geq 0). \tag{20}$$

The joint analysis of specific gravity measurements from multiple samples, under the assumption that the standard uncertainty of such errors is largely independent of the sample, enables the assessment of measurement uncertainty. This is made possible by evaluating the dispersion of biases of each laboratory's results from the consensus values across the multiple analyzed samples whose specific gravity the laboratory measured. Measurement errors are modeled as outcomes of a Laplace distribution with zero mean and a laboratory-specific standard deviation u_l:

$$E_{s,l} \sim \mathrm{LAPLACE}(\mathrm{mean} = 0,\ \mathrm{sd} = u_l). \tag{21}$$

The same as with the laboratory biases, the prior distribution for the measurement uncertainty is a half-Cauchy distribution whose median equals half the maximum allowed value set by WADA in the TD DL [48]:

$$u_l \sim \mathrm{CAUCHY}(\mathrm{median} = u_{c,\mathrm{Max}}/2) \quad (u_l \geq 0). \tag{22}$$

The effect of rounding is modeled using a uniform (or, rectangular) distribution:

$$R_{s,l} \sim \mathrm{UNIFORM}(-0.001/2, +0.001/2). \tag{23}$$

The input data and the parameters of this statistical measurement model are summarized in Table 4.

Table 3 Overview of the input data and measurement model parameters for carbon isotope ratio measurements in steroid profile markers

Type	Symbol	Description
Data	X_l	Reported value measured by laboratory l
Data	$u_{c,l}$	Measurement uncertainty reported by laboratory l
Data	$u_{c,\mathrm{Max}}$	Maximum allowed measurement uncertainty set by WADA
Parameter	μ	Consensus value
Parameter	B_l	Bias for laboratory l
Parameter	τ	Standard deviation of the laboratory biases $B_1 \ldots B_L$

Table 4 Overview of the input data and measurement model parameters for urine-specific gravity measurements

Type	Symbol	Description
DATA	$X_{s,l}$	Reported value measured by laboratory l for sample s
DATA	$u_{\mathrm{c,Max}}$	Maximum allowed relative measurement uncertainty set by WADA
PARAMETER	μ_s	True SG value for sample s
PARAMETER	B_l	Bias of laboratory l
PARAMETER	τ	Standard deviation of laboratory biases
PARAMETER	u_l	Standard deviations of measurement errors for laboratory l

Data format

To facilitate automated uptake of the results from an interlaboratory study, quantitative EQAS results are transcribed in the form of a wide-format database (Table 5) where each row corresponds to an entry by an individual participating laboratory and each column, in turn, contains the reported outcome along with the associated descriptors (metadata). This approach ensures sufficient level of annotation and machine readability.

Model fitting

Modern data analysis cannot be conducted without modern software tools. Indeed, many requirements in modern data analysis necessitate tools that go beyond Excel spreadsheets.

Working with the aforementioned custom-tailored statistical models requires great flexibility, and we have adopted Markov chain Monte Carlo (MCMC) methods to perform the parameter fitting and to obtain random draws from the posterior distributions of the fitted model parameters [26].

These tasks can be performed using any one of the several open-source software platforms for statistical modeling, such as BUGS [20], Stan [5], or JAGS [10]. We have adopted the latter and executed the MCMC sampling using the R environment for statistical modeling, computing, and graphics [32].

The entire process of data uptake, fitting, and sampling from the posterior distribution was developed as a Shiny application in R which provides a web-based user interface for data retrieval and visualization [12]. The resulting tables and figures are generated programmatically.

Outlier handling

In the context of key comparisons carried out under the purview of the CIPM, "automatic" identification and suppression of apparent outliers is frowned upon for reasons articulated in [18] and [31]. However, in a production environment as prevails for WADA EQAS, greater freedom is warranted to set aside occasional, suspicious reported values, especially if this is done as conservatively as described next. Thus, the quantitative EQAS data reduction proceeds first by fitting the mixed effects model to the data using conventional approaches (based on the method of maximum likelihood estimation [30, p.191]) and obtain predicted values for each observation. Values that are more than 5σ apart from their model predictions are considered anomalous, and the subsequent calculation of the consensus values with the Bayesian methods will not include the anomalous values identified above.

More specifically, for each measured value, one calculates the difference between the reported and predicted values and divides this difference by the robust estimate (scaled mean absolute deviation about the median, `mad` in R) of the standard deviation of the residuals. However, z-scores are still calculated for these anomalous value(s) using the "clean" consensus value and the maximum allowed combined standard measurement uncertainty ($u_{\mathrm{c,Max}}$). While the 5σ threshold is arbitrary, it follows established practices of data rejection for proficiency test data [16].

Evaluation of laboratory bias

The evaluation of laboratory biases constitutes a major task when conducting interlaboratory studies for proficiency testing. Depending on the analyte, laboratory bias takes either an absolute or a relative form:

Table 5 An example of a wide-format database of quantitative EQAS results

EQAS	LAB	SAMPLE	ANALYTE	TYPE	VALUE	UNIT	u_{c} (%)	$u_{\mathrm{c,Max}}$ (%)
2023-1	L1	S1	Morphine	TS	1.36	µg/mL	9.9	15
2023-1	L2	S1	Morphine	TS	1.81	µg/mL	13	15
2023-1	L3	S1	Morphine	TS	1.56	µg/mL	8.3	15

$$B_l = X_l - \mu \qquad \text{(for SG, GC/C/IRMS)}, \tag{24}$$

$$B_l = (X_l - \mu)/\mu \qquad \text{(for TS, SP)}. \tag{25}$$

In all cases, we use the posterior means of laboratory biases which are the best estimates in the context of the statistical measurement models used, and from the viewpoint of mean squared error. These best estimates of laboratory bias will invariably differ from the apparent (naïve) laboratory biases, resulting from the simple calculations that we discuss in the results section.

The uncertainty associated with the laboratory bias, $u(B_l)$, takes a form of a predictive uncertainty and includes the uncertainty contributions associated with the consensus value $u(\mu)$, and the repeatability of the replicate determinations used to evaluate the bias, the dark uncertainty τ, as well as the correlation between X_l and μ. The uncertainty associated with the laboratory bias is determined as the standard deviation of the Markov Chain Monte Carlo draws from the posterior distribution of B_l.

Measurements of steroid profile markers and SG are performed for multiple samples and therefore present an opportunity to evaluate the overall laboratory performance. While various methods have been used to assess the laboratory bias in such cases, including rank-sum analysis [35] or pooling the z-scores [39, 43], the use of a mixed effects model provides a direct assessment of the bias of each laboratory and its uncertainty without the need for further analysis.

Standard deviation for proficiency assessment

An important part of proficiency testing is to determine the scaling factor σ_{PT}, formally known as the standard deviation for proficiency assessment [39, 3.5], which is used to determine the z-scores for the results reported by each laboratory:

$$z_l = \frac{B_l}{\sigma_{\text{PT}}}. \tag{26}$$

The z-scores of the participating laboratories are critical outputs of the intercomparison since z-scores $|z| \geq 3.0$ may lead to the assignment of penalty points against the corresponding laboratories, which, in turn, can lead to sanctions against them [45, Clause 7.3].

Traditionally, the value of σ_{PT} has been set by WADA to be equal to the maximum allowed combined standard measurement uncertainty, $u_{\text{c,Max}}$ [45, Clause 7.1.2.1]. In addition, the ISL:2021 notes that σ_{PT} can also be set as the robust estimate of the reproducibility of the results from all the participating laboratories (s_R) [45].

In this vein, we estimate the value of s_R using the statistical model that we propose, and use it to regularly assess the fitness-for-purpose of the $u_{\text{c,Max}}$ values adopted by WADA. For this purpose, we calculate the model-based predictive reproducibility standard deviation for each analyte — an overall posterior estimate of dispersion that combines all the random effects recognized in the statistical measurement model.

Validation of calculations

The EQAS calculator developed to support this work is built on the same basic modeling principles as the *NIST Consensus Builder* [18], and its results have been validated using several benchmark datasets and calculations as outlined in Table 6.

The steroid module results were evaluated using datasets from the Appendix 3 of the IUPAC Technical Report on Proficiency Testing [39] and datasets from the Appendix E of ISO 13528:2022 [15]. Since the methods outlined in IUPAC and ISO guidance documents rely on different statistical models, the goal of the comparison is simply to demonstrate that the results obtained are in reasonable mutual agreement. In all cases, the consensus values obtained are consistent with those from the IUPAC and ISO guides to within 2σ uncertainty despite the fact that some datasets include results with apparently bimodal and skewed distributions.

The threshold substance module results were validated using the dataset for arsenic in kudzu with triplicate measurement results from 22 laboratories [30, p.151]. Here, our results are compared against those obtained with a different probabilistic programming language, Stan [5], which is widely considered a benchmark implementation of Bayesian modeling and computation. In this particular case, both calculations rely on the same statistical model and produce identical results, differing numerically only by insignificant amounts due to the stochastic nature of Monte Carlo estimates.

Table 6 Comparison of the published consensus value estimates from benchmark interlaboratory datasets [39] with the Bayesian approach described in the foregoing

Dataset	Size	$\mu_{\text{classical}}$	μ_{Bayes}
IUPAC/A3.1	$L = 68$	53.24(8)	53.28(9)
IUPAC/A3.2	$L = 32$	85(2)	88(3)
IUPAC/A3.3	$L = 65$	101.5(16)	98.4(17)
ISO 13528/E.3	$L = 34$	0.2570(68)[a]	0.2596(90)

The numbers in parentheses denote standard uncertainties and apply to the least significant digits. For instance, 53.24(8) stands for $\mu = 53.24$ and $u(\mu) = 0.08$

a. The uncertainty was reduced by 25% to adhere to IUPAC recommendation $u(\mu) = s_R/\sqrt{L}$ instead of $u(\mu) = 1.25 \cdot s_R/\sqrt{L}$

Results

Consensus values

Here, we use the results from the blind EQAS-2023-1 round to compare the results of the currently conventional approach with the proposed Bayesian approach. Under the current WADA approach, the consensus value $\mu_{\text{classical}}$ is the robust (Huber H15) mean, and the uncertainty associated with the consensus value is taken as $u(\mu_{\text{classical}}) = k \cdot s_R/\sqrt{L}$ where s_R is the robust (Huber H15) standard deviation of the L laboratory results and $k = 1$. Note that the choice of k itself has been questioned in the literature: ISO recommends $k = 1.25$ [15] as an approximation to the $k = \sqrt{\pi/2}$ that corresponds to the median as an estimate of μ based on a large sample drawn from a Gaussian distribution [7, p.369], whereas IUPAC recommends $k = 1.00$ [39], which is the more appropriate choice for the robust mean as an estimate of μ.

In addition to the controversial nature of the factor k, obtaining the H15 estimates of location and scale can be a tedious process [1] and most users, like WADA, rely on the widely used Excel macro created by the Analytical Methods Committee (AMC) of the Royal Society of Chemistry Analytical Science Community. (An implementation in R function `huberM`, which is defined in the `robustbase` package [21], streamlines the process, yet involves choices of tuning constants.)

In most cases, we observe that the classical robust statistical method provides similar consensus values and their associated uncertainties, with both methods differing by $\pm 1u(\mu)$ in most of the cases (Table 7). In addition, as shown in the last line of Table 7, the classical robust methods can fail to evaluate the uncertainty associated with the consensus value of SG (by giving it a zero estimate). Such shortcomings of classical statistical methods are not unique to this example [26].

Laboratory bias

Traditional interlaboratory comparisons provide a simple estimate of the bias by calculating the difference (or relative difference) between the reported value and the consensus value — the *apparent* laboratory bias. In the Bayesian model framework, this is not the best estimate of the bias. Instead, the mean posterior estimate of the laboratory bias effect, B_l, is. The latter tend to be smaller than the aforementioned apparent estimates of bias as they are shifted toward their prior estimates which are set to zero (Fig. 3). Formally known as *shrinkage*, this effect was discovered in a classical (non-Bayesian) setting when Charles Stein stunned the world of statistics by proving that the sample mean is not the best estimate of the mean of a multivariate Gaussian distribution (in more than two dimensions) according to the mean squared error criterion [11].

Overall, the use of Bayesian measurement models is very appropriate in regulatory performance assessments, such as the WADA EQAS, because our Bayesian model starts out by assuming that all laboratories are performing well. Thus, the prior distribution assigned to each laboratory bias is centered at zero and acts as a mathematical device to convey the assumption that all the laboratories are performing well unless the data reveal cogent evidence to the contrary.

Measurement uncertainty

Evaluation of measurement uncertainty for each participating laboratory is an integral part of EQAS intercomparisons. Since the estimated laboratory biases are not used to correct any future results, we adopt the view that the bias should be part of the combined measurement uncertainty [22, 38]:

Table 7 Comparison of the consensus value estimates obtained by classical robust method ($k = 1.00$) and by the Bayesian method developed herein using the results from blind EQAS intercomparisons

Analyte	Quantity	EQAS	$\mu_{\text{classical}}$	μ_{Bayes}	Unit
Pregnanediol	$\delta_{\text{VPDB}}(^{13}\text{C})$	2017-2/S6	−22.59(12)	−22.69(13)	‰ vs VPDB
Androsterone	$\delta_{\text{VPDB}}(^{13}\text{C})$	2017-2/S6	−26.69(12)	−26.80(10)	‰ vs VPDB
Morphine	Concentration	2023-1/S1	1.61(2)	1.61(2)	mg/L
5α-androstanediol	Concentration	2023-1/S1	51.2(8)	50.2(7)	ng/mL
5β-androstanediol	Concentration	2023-1/S1	195.6(2.7)	194.5(2.4)	ng/mL
Androsterone	Concentration	2023-1/S1	2126(33)	2130(22)	ng/mL
Etiocholanone	Concentration	2023-1/S1	1712(21)	1706(20)	ng/mL
Testosterone (T)	Concentration	2023-1/S1	29.5(3)	29.8(3)	ng/mL
Epitestosterone (E)	Concentration	2023-1/S1	38.3(5)	37.9(5)	ng/mL
T/E ratio	Amount ratio	2023-1/S1	0.776(9)	0.786(9)	1
Urine	Specific gravity	2023-1/S1	1.016(0)	1.0162(1)	1

The numbers in parentheses denote standard uncertainties and apply to the least significant digits. For instance, 0.776(9) stands for $\mu = 0.776$ and $u(\mu) = 0.009$

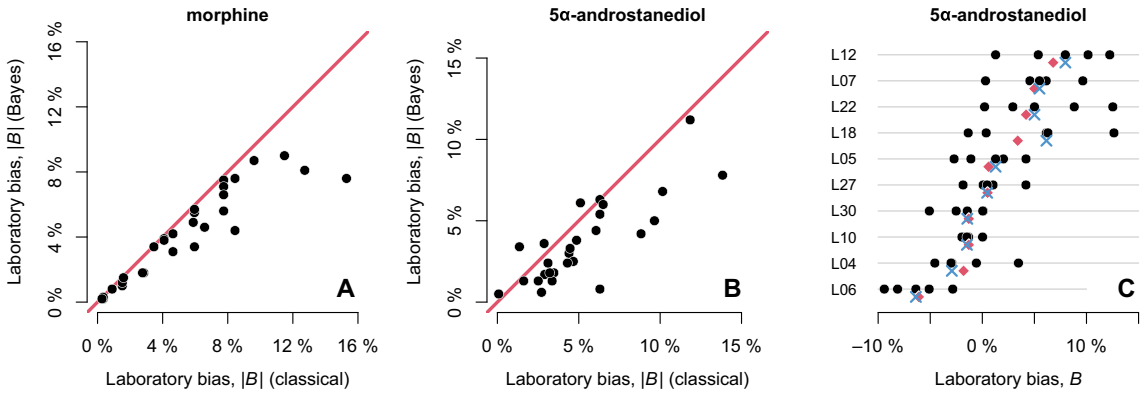

Fig. 3 Evaluating the Laboratory bias using Bayesian mixed effects measurement models. Panel A shows the estimated absolute laboratory biases for morphine measurements in urine using Bayesian and classical methods. It is evident that the Bayesian method provides lower estimates of the laboratory bias—an effect known as *shrinkage*. Panel B shows the same effect for 5α-androstanediol measurements in urine. In addition, Panel C illustrates the workings of the mixed effect model where black dots represent the individual laboratory biases observed for each of the five analyzed urine samples. Red diamonds represent the corresponding Bayesian estimates whereas the blue crosses represent the (classical) medians of the individual biases. All panels show the anonymized results from the blind EQAS-2023-1 intercomparison round

$$u_{\mathrm{c}} = \sqrt{u^2_{\mathrm{meas}} + B^2 + u^2(B)}. \tag{27}$$

To estimate the combined measurement uncertainty for each laboratory, we take the draws from the posterior distribution of measurement errors (E) and biases (B) and combine them in quadrature:

$$U^2_{\mathrm{MCMC},i} = E^2_{\mathrm{MCMC},i} + B^2_{\mathrm{MCMC},i} \quad (i = 1 \ldots N_{\mathrm{MCMC}}). \tag{28}$$

The approach of combining the individual MCMC draws incorporates the uncertainty surrounding the estimates of laboratory biases as well as the uncertainty surrounding the estimates of u_{meas} itself. The median of $\{U_{\mathrm{MCMC},i}\}$ is used as the estimate of the combined measurement uncertainty and the lower 5 % quantile of $U_{\mathrm{MCMC},i}$ is used to determine whether $u_{\mathrm{c}} \leq u_{\mathrm{c,Max}}$.

Overall performance

As noted above, fitting the statistical models to the EQAS data provides, among other results, the overall predictive reproducibility standard deviation (s_R) for each analyte. This measure of dispersion combines the standard deviation of laboratory biases (τ) and median measurement uncertainty of all laboratories:

$$B \sim \mathrm{LAPLACE}(\mathrm{mean} = 0, \mathrm{sd} = \tau), \tag{29}$$

$$E \sim \mathrm{CAUCHY}(\mathrm{median} = \mathrm{median}(u_{\mathrm{meas}})), \tag{30}$$

$$U = E + B. \tag{31}$$

The estimate of s_R is then obtained as the robust standard deviation of the combined draws (U) representing the overall laboratory bias and typical measurement errors. For SG, the calculation of s_R also includes the rounding error component. Table 8 shows the comparison for steroid profile analysis from two blind EQAS rounds. In most cases, the s_R is approximately half of the maximum allowed uncertainty, as expected.

Conclusions

We have outlined a data-driven statistical measurement model for interlaboratory comparisons conducted regularly by WADA. By embracing modern statistical methods and tools, we are able to employ finely tuned models that deliver reliable estimates of performance (z-scores) and realistic uncertainty evaluations, making the most efficacious use of the data.

Table 8 Comparison of the maximum allowed measurement uncertainty and the overall predictive reproducibility standard deviation (s_R) for steroid profile marker measurements from blind EQAS-2023-1 ($s_{R,1}$) and EQAS-2023-2 ($s_{R,2}$) intercomparisons

Analyte	$u_{\mathrm{c,Max}}$ (%)	$s_{R,1}$ (%)	$s_{R,2}$ (%)
5α-androstanediol	25	12	14
5β-androstanediol	25	12	12
Androsterone	20	10	10
Etiocholanone	20	10	11
Testosterone (T)	20	10	11
Epitestosterone (E)	20	10	11
T/E	15	10	11

Despite the nearly identical consensus values obtained by the classical robust methods and the Bayesian method, the latter is open to data-driven tuning of the underlying model assumptions and provides easy access to uncertainties associated with every model parameter and derived quantities. The method developed herein can be further explored in proficiency testing as a data-driven decision-making tool to evaluate the performance of analytical methods in other fields.

Acknowledgements This work was presented at the Eurachem's 10th Workshop on Proficiency Testing in Analytical Chemistry, Microbiology and Laboratory Medicine which was held in Windsor (UK) in September 2023.

Author contributions JM and AP contributed to the statistical design, OB and BG contributed to the project design, and SK contributed to the data gathering and result evaluation. JM, AP, and OB wrote the manuscript with input from all authors.

Funding Open access funding provided by National Research Council Canada library.

Data availability No datasets were generated during the current study.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

References

1. Analytical Methods Committee (1989) Robust statistics-how not to reject outliers Part 1. Basic concepts. Anal 114(12):1693–1697. https://doi.org/10.1039/an9891401693
2. Bates D, Mächler M, Bolker B et al (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67(1):1–48. https://doi.org/10.18637/jss.v067.i01
3. Boghosian T, Barroso O, Ivanova V et al (2012) Ensuring high quality in anti-doping laboratories. Bioanalysis 4(13):1591–1601. https://doi.org/10.4155/bio.12.136
4. Brilleman S, Crowther M, Moreno-Betancur M et al (2019) Joint longitudinal and time-to-event models for multilevel hierarchical data. Stat Methods Med Res 28:3502–3515. https://doi.org/10.1177/0962280218808821
5. Carpenter B, Gelman A, Hoffman MD et al (2017) Stan: a probabilistic programming language. J Stat Softw. https://doi.org/10.18637/jss.v076.i01
6. Coplen TB (1994) Reporting of stable hydrogen, carbon, and oxygen isotopic abundances (IUPAC Technical Report). Pure Appl Chem 66(2):273–276. https://doi.org/10.1351/pac199466020273
7. Cramér H (1922) Mathematical methods of statistics, 1st edn. Princeton University Press, Princeton
8. Crowder M (1992) Interlaboratory comparisons: Round robins with random effects. J R Stat Soc Ser C (Appl Stat) 41:409–425. https://doi.org/10.2307/2347571
9. Demeyer S, Fischer N (2017) Bayesian framework for proficiency tests using auxiliary information on laboratories. Accred Qual Assur 22(1):1–19. https://doi.org/10.1007/s00769-017-1247-y
10. Depaoli S, Clifton JP, Cobb PR (2016) Just another Gibbs sampler (JAGS): flexible software for MCMC implementation. J Educ Behav Stat 41(6):628–649. https://doi.org/10.3102/1076998616664876
11. Efron B, Morris C (1973) Stein's estimation rule and its competitors—an empirical Bayes approach. J Am Stat Assoc 68(341):117. https://doi.org/10.2307/2284155
12. Gebauer JE, Adler J (2023) Using Shiny apps for statistical analyses and laboratory workflows. J Lab Med 47(4):149–153. https://doi.org/10.1515/labmed-2023-0020
13. Gelman A (2006) Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Anal 1(3):515–533. https://doi.org/10.1214/06-BA117A
14. Goodrich B, Gabry J, Ali I, et al (2023) rstanarm: Bayesian applied regression modeling via Stan. https://mc-stan.org/rstanarm/, R package version 2.26.1
15. ISO (2022) Statistical methods for use in proficiency testing by interlaboratory comparison, 3rd edn. International Organization for Standardization (ISO), Geneva, Switzerland, ISO 13528:2022(E)
16. Jerome S, Harms A (2023) Proficiency test data interpretation and data rejection. Appl Radiat Isot 194:110678. https://doi.org/10.1016/j.apradiso.2023.110678
17. Koepke A, Lafarge T, Possolo A (2017a) NIST Consensus Builder - User's Manual. National Institute of Standards and Technology, Gaithersburg, MD https://consensus.nist.gov
18. Koepke A, Lafarge T, Possolo A et al (2017) Consensus building for interlaboratory studies, key comparisons, and meta-analysis. Metrologia 54(3):S34–S62. https://doi.org/10.1088/1681-7575/aa6c0e
19. Kuiper JR, O'Brien KM, Ferguson KK et al (2021) Urinary specific gravity measures in the US population: Implications for the adjustment of non-persistent chemical urinary biomarker data. Environ Int 156:106656. https://doi.org/10.1016/j.envint.2021.106656
20. Lunn D, Spiegelhalter D, Thomas A et al (2009) The BUGS project: evolution, critique and future directions. Stat Med 28(25):3049–3067. https://doi.org/10.1002/sim.3680
21. Maechler M, Rousseeuw P, Croux C, et al (2023) robustbase: Basic Robust Statistics. R package version 0.99-1 http://robustbase.r-forge.r-project.org/
22. Magnusson B, Ellison SLR (2007) Treatment of uncorrected measurement bias in uncertainty estimation for chemical measurements. Anal Bioanal Chem 390(1):201–213. https://doi.org/10.1007/s00216-007-1693-1
23. Mandel J, Paule R (1970) Interlaboratory evaluation of a material with unequal numbers of replicates. Anal Chem 42(11):1194–1197. https://doi.org/10.1021/ac60293a019
24. Mandel J, Paule R (1971) Correction—interlaboratory evaluation of a material with unequal numbers of replicates. Anal Chem 43(10):1287–1287. https://doi.org/10.1021/ac60304a001

25. Meija J, Possolo A (2022) Interlaboratory comparisons of chemical measurements: Quo vadis? Accred Qual Assur 28(3):89–93. https://doi.org/10.1007/s00769-022-01505-y
26. Meija J, Bodnar O, Possolo A (2023) Ode to Bayesian methods in metrology. Metrologia. https://doi.org/10.1088/1681-7575/acf66b
27. Mejia R, Cuellar M, Salyards J (2020) Implementing blind proficiency testing in forensic laboratories: motivation, obstacles, and recommendations. Foren Sci Int Synergy 2:293–298. https://doi.org/10.1016/j.fsisyn.2020.09.002
28. Mosteller F, Tukey JW (1977) Data analysis and regression. Addison-Wesley Publishing Company, Reading
29. Pinheiro JC, Bates DM (2000) Mixed-effects models in S and S-plus. Springer-Verlag, New York. https://doi.org/10.1007/b98882
30. Possolo A, Meija J (2022) Measurement Uncertainty: A Reintroduction, 2nd edn. Sistema Interamericano de Metrologia (SIM), Montevideo, Uruguay, https://doi.org/10.4224/1tqz-b038
31. Possolo A, Koepke A, Newton D et al (2021) Decision tree for key comparisons. J Res Nat Inst Stand Technol 126:126007. https://doi.org/10.6028/jres.126.007
32. R Core Team (2023) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria, https://www.R-project.org/
33. Rocke DM, Lorenzato S (1995) A two-component model for measurement error in analytical chemistry. Technometrics 37(2):176–184. https://doi.org/10.1080/00401706.1995.10484302
34. Rukhin AL, Possolo A (2011) Laplace random effects models for interlaboratory studies. Comput Stat Data Anal 55(4):1815–1827. https://doi.org/10.1016/j.csda.2010.11.016
35. Steel RGD (1961) Some rank sum multiple comparisons tests. Biometrics 17(4):539. https://doi.org/10.2307/2527854
36. Thompson M, Ellison SLR (2011) Dark uncertainty. Accred Qual Assur 16:483–487. https://doi.org/10.1007/s00769-011-0803-0
37. Thompson M, Wood R (1993) The International Harmonized Protocol for the proficiency testing of (chemical) analytical laboratories. Pure Appl Chem 65(9):2123–2144. https://doi.org/10.1351/pac199365092123
38. Thompson M, Ellison SLR, Fajgelj A et al (1999) Harmonized guidelines for the use of recovery information in analytical measurement. Pure Appl Chem 71(2):337–348. https://doi.org/10.1351/pac199971020337
39. Thompson M, Ellison SLR, Wood R (2006) The International Harmonized Protocol for the proficiency testing of analytical chemistry laboratories (IUPAC Technical Report). Pure Appl Chem 78(1):145–196. https://doi.org/10.1351/pac200678010145
40. Toman B, Possolo A (2009) Laboratory effects models for interlaboratory comparisons. Accred Qual Assur 14:553–563. https://doi.org/10.1007/s00769-009-0547-2
41. Toman B, Possolo A (2010) Erratum to: laboratory effects models for interlaboratory comparisons. Accred Qual Assur 15:653–654. https://doi.org/10.1007/s00769-010-0707-4
42. Werhahn O, Olson DA, Kuanbayev C et al (2023) The CIPM MRA—success and performance. Metrologia 60(4):042001. https://doi.org/10.1088/1681-7575/ace191
43. Wilson DJ (2019) The harmonic mean p-value for combining dependent tests. Proc Nat Acad Sci 116(4):1195–1200. https://doi.org/10.1073/pnas.1814092116
44. Wilson MD, Rocke DM, Durbin B et al (2004) Detection limits and goodness-of-fit measures for the two-component model of chemical analytical error. Anal Chim Acta 509(2):197–208. https://doi.org/10.1016/j.aca.2003.12.047
45. World Anti-Doping Agency (2021a) International Standard for Laboratories
46. World Anti-Doping Agency (2021b) Technical Document TD2021EAAS. Measurement and Reporting of Endogenous Anabolic Steroid (EAAS) Markers of the Urinary Steroid Profile
47. World Anti-Doping Agency (2022a) Report of the Independent Observers: XXIV Olympic Winter Games, Beijing 2022
48. World Anti-Doping Agency (2022b) Technical Document TD2022DL. Decision limits for the confirmatory quantification of exogenous threshold substances by chromatography-based analytical methods
49. World Anti-Doping Agency (2022c) Technical Document TD2022IRMS. Detection of Synthetic Forms of Prohibited Substances by GC/C/IRMS