Combined statistical decision limits based on two GH-2000 scores for the detection of growth hormone misuse

W. Liu[a], F. Bretz[b], D. Böhning[a], R.I.G. Holt[c], Y. Han[d]
W. Böhning[c], N. Guha[e], D. A. Cowan[f]

[a]S3RI and School of Mathematics
University of Southampton, SO17 1BJ, UK
[b]Novartis Pharma AG
Basel, 4002, Switzerland
[c]Human Development and Health Academic Unit
Faculty of Medicine, Southampton General Hospital
University of Southampton, SO16 6YD, UK
[d] Department of Mathematics,
University of Manchester, Manchester M13 9PL, UK
[e]Chemical Pathology and Metabolic Medicine Department of
Clinical Biochemistry, John Radcliffe Hospital
Oxford, OX3 9DU, UK
[f] Department of Analytical and Environmental Sciences,
King's College London, London SE1 9NH, UK

Abstract

The GH-2000 biomarker method, based on the measurements of insulin-like growth factor-I (IGF-I) and the amino-terminal pro-peptide of type III collagen (P-III-NP), has been developed as a powerful technique for the detection of growth hormone (GH) misuse by athletes. IGF-I and P-III-NP are combined in gender specific formulas to create the GH-2000 score, which is used to determine whether GH has been administered. To comply with World Anti-Doping Agency regulations, each analyte must be measured by two methods. IGF-I and P-III-NP can be measured by a number of approved methods, each leading to its own GH-2000 score. Single decision limits for each GH-2000 score have been originally developed by Bassett and co-workers (Erotokritou-Mulligan *et al.* 2012) and further developed in Holt *et al.* (2015) and Böhning *et al.* (2019). These have been incorporated into the guidelines of the World Anti-Doping Agency. Erotokritou-Mulligan *et al.* (2012) and Holt *et al.* (2015) constructed a joint decision limit based on the sample correlation between the two GH-2000 scores generated from an available sample in order to increase the sensitivity of the biomarker method. This paper takes this idea further into a fully developed statistical approach. It constructs combined decision limits when two GH-2000 scores from different assay combinations are used to decide whether an athlete has been misusing GH. The combined decision limits are directly related to tolerance regions and constructed using a Bayesian approach. It is also shown to have highly satisfactory frequentist properties. The new approach meets the required false-positive rate with a pre-specified level of certainty.

Keywords: Bayesian tolerance regions; Growth hormone misuse detection; GH-2000 scores; Decision Limits; Tolerance limits; Tolerance regions.

1 Introduction

As a powerful anabolic agent of considerable therapeutic value, growth hormone (GH) is misused in sport to enhance performance (cf. Holt, 2009). In order to preserve the fairness of competition, its use is prohibited by the World Anti-Doping Agency (WADA) (cf. WADA, 2014, 2016). Two methods for the detection of GH misuse are currently available and approved by WADA: the isoform test developed by Bidlingmaier *et al.* (2000) (see also WADA, 2014, 2016) and the GH-2000 biomarker test developed by the GH-2000 and GH-2004 projects (cf. Holt *et al.*, 2015). The latter method depends on the measurements of two GH sensitive biomarkers, the insulin-like growth factor-I (IGF-I) and the amino-terminal pro-peptide of type III collagen (P-III-NP), both of which rise in response to exogenous GH administration (cf. Longobardi *et al.*, 2000, and Dall *et al.*, 2000). The measured concentrations of the two biomarkers are combined in sex-specific and age-adjusted discriminant functions (cf. Powrie *et al.*, 2007, Erotokritou-Mulligan *et al.*, 2012, Holt *et al.*, 2015, and Böhning *et al.*, 2016) to allow the calculation of a score, the GH-2000 score. It is possible that the score may take a negative value.

The measurements of IGF-I and P-III-NP are carried out by choosing two specific assays. As the measured results differ slightly from one assay to another, each assay pair generates an assay-specific GH-2000 score, which differs from other GH-2000 scores generated by different assays. These are the basis of the data generating process. Currently, there are three IGF-I assays and two P-III-NP assays approved by WADA. The IGF-I assays are: a mass spectrometry (MS) based approach, Immunotech A15729 IGF-I IRMA (Immunotech SAS, Marseille, France), and Immunodiagnostic Systems iSYS IGF-I (Immunodiagnostics Systems Limited, Boldon, UK). The P-III-NP assays are: UniQ$^{\text{TM}}$ P-III-NP RIA (Orion Diagnostica, Espoo, Finland), and Siemens ADVIA Centaur P-III-NP (Siemens Healthcare Laboratory

Diagnostics, Camberley, UK). For more details and background on these assays, see Holt *et al.* (2015). As any GH-2000 score requires a pair of IGF-I assay and P-III-NP assay, there are six possible GH-2000 scores. Depending on the available technology, laboratories choose the appropriate GH-2000 scores for evaluating their samples, and a decision limit based on one single GH-2000 score has been developed in Holt *et al.* (2015) and Böhning *et al.* (2019) by assuming that a GH-2000 score from an athlete without GH misuse has a normal distribution.

As stated in Holt *et al.* (2015), a result will be declared as an adverse analytical finding (i.e. indicative of doping) only if the confirmation procedure results in GH-2000 scores greater than the decision limits for two pairs of analytes. These decision limits are constructed on the basis of the univariate normal distributions of the associated GH-2000 scores. Erotokritou-Mulligan *et al.* (2012), with further details in Holt *et al.* (2015), construct combined decision limits on the basis of a bivariate normal distribution. The idea is motivated by the intuition that a reduced correlation between the two GH-2000 scores could lead to reduced decision limits, and thus increase the sensitivity of GH misuse detection. We describe details on this method in Subsection 3.3.

The purpose of this paper is to provide a valid construction method of combined decision limits when two GH-2000 scores, based on two different pairs of IGF-I and P-III-NP assays, are used jointly in assessing the compliance of a sample. It is shown here how the combined decision limits are directly related to a particular tolerance region, and can be constructed so that the false positive rate (FPR) is controlled at a pre-specified level $1-\beta$, say 1 in 10,000, with a pre-specified confidence or belief $1-\alpha$, e.g. 95%, about the possible value of $(\boldsymbol{\mu}, \Sigma)$, assuming the two GH-2000 scores have a bivariate normal distribution $N_2(\boldsymbol{\mu}, \Sigma)$. This is in contrast to the previously mentioned method discussed in Erotokritou-Mulligan *et al.* (2012) where such a property is assumed to hold *bona fide*.

A Bayesian approach is adopted in this paper since a frequentist solution is much harder to construct and not available thus far (see Subsection 3.2 for more details). The frequentist property of the Bayesian combined decision limits is also assessed by simulation, which shows that the Bayesian combined decision limits can also be interpreted as frequentist combined decision limits.

The paper is organized as follows. Section 2 collects some known distributional results that will be used in Section 3. Section 3 considers the construction of decision limits. A very brief review of the construction of a single decision limit for one GH-2000 score is given in Subsection 3.1. Subsection 3.2 constructs Bayesian combined decision limits for two GH-2000 scores. The method is then illustrated with a real data set in Subsection 3.3. A simulation study is presented in Subsection 3.4 to assess the frequentist property of the Bayesian combined decision limits given in Subsection 3.3. Finally the paper closes with a brief discussion in Section 4.

2 Preliminary distributional results

In this section, we collect some known distributional results which are used in Section 3 for the construction of decision limits. More details about these results can be found in Guttman (1970), Box and Tiao (1992) and Anderson (2003).

Following Holt *et al.* (2015) and Böhning *et al.* (2019), we assume that the GH-2000 scores $\mathbf{x} = (x_1, \ldots, x_k)'$ from an athlete without GH misuse have a k-variate normal distribution $N_k(\boldsymbol{\mu}, \Sigma)$, with both $\boldsymbol{\mu}$ and Σ unknown. For the problem considered in this paper, we are only interested in the case of $k = 2$ since only two GH-2000 scores are involved. We assume further that we have observed a random sample from the population $N_k(\boldsymbol{\mu}, \Sigma)$:

$$\mathbf{x}_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{1k} \end{pmatrix}, \ldots, \mathbf{x}_n = \begin{pmatrix} x_{n1} \\ \vdots \\ x_{nk} \end{pmatrix} \overset{i.i.d.}{\sim} N_k(\boldsymbol{\mu}, \Sigma)$$

Denote $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$, $\bar{\mathbf{x}} = \sum_{i=1}^n \mathbf{x}_i/n$, and $V = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'/(n-1)$.

In this paper, the non-informative reference prior distribution of $(\boldsymbol{\mu}, \Sigma^{-1})$, given by

$$p(\boldsymbol{\mu}, \Sigma^{-1}) \propto p(\boldsymbol{\mu})P(\Sigma^{-1}) \propto |\Sigma^{-1}|^{-\frac{(k+1)}{2}}, \tag{1}$$

is used since the sample $\mathbf{X}$ is all we have. An additional incentive for using the non-informative reference prior is that the Bayesian decision limit for the special case of $k = 1$ is also the frequentist decision limit; see Section 3.1 below.

The posterior distribution of $(\boldsymbol{\mu}, \Sigma^{-1})$ based on the observed data $\mathbf{X}$ is then given by

$$p(\boldsymbol{\mu}, \Sigma^{-1}|\mathbf{X}) \propto |\Sigma^{-1}|^{\frac{(n-k-1)}{2}} \exp\left\{-\frac{1}{2} tr \Sigma^{-1} \left[(n-1)V + n(\boldsymbol{\mu} - \bar{\mathbf{x}})(\boldsymbol{\mu} - \bar{\mathbf{x}})'\right]\right\},$$

where trA denotes the trace of matrix A. Integrating out $\boldsymbol{\mu}$ gives the posterior distribution of Σ^{-1}

$$\Sigma^{-1}|\mathbf{X} \sim W_k([(n-1)V]^{-1}, n-k) \tag{2}$$

in the notation of Box and Tiao (1992), and the posterior conditional (on Σ^{-1}) distribution of $\boldsymbol{\mu}$ is

$$p(\boldsymbol{\mu}|\Sigma^{-1}, \mathbf{X}) = \frac{p(\boldsymbol{\mu}, \Sigma^{-1}|\mathbf{X})}{p(\Sigma^{-1}|\mathbf{X})} \sim N_k(\bar{\mathbf{x}}, \Sigma/n). \tag{3}$$

3 Decision limits

In this section, we consider the construction of decision limits. In Subsection 3.1, we provide a very brief review of the construction of decision limit for one single GH-2000 score, that is, for the case of $k = 1$, which helps the understanding of Subsection 3.2. Subsection 3.2 studies the construction of combined decision limits based on two GH-2000 scores, that is for the case of $k = 2$. Subsection 3.3 illustrate the computation of the decision limits by using the dataset on GH-2000 scores given in Holt *et al.* (2015). Subsection 3.4 presents the results of simulation studies.

3.1 Decision limit for one GH2000 score

Let $a = a(\mathbf{X})$ denote the decision limit for one GH-2000 score. Hence a future sample observation y is declared to be positive if and only if y is larger than $a(\mathbf{X})$. In order to control the FPR at the pre-specifed level $1 - \beta$, it is desirable to have

$$P_{y|\mu,\sigma^2}\{y > a(\mathbf{X})\} \leq 1 - \beta$$

which is equivalent to

$$P_{y|\mu,\sigma^2}\{y \leq a(\mathbf{X})\} \geq \beta \tag{4}$$

under the assumption that y is from the population distribution $N(\mu, \sigma^2)$. Here the probabilities are calculated with respect to the distribution of y conditional on (μ, σ^2). Note that the probability in (4) depends on the value of (μ, σ^2), for which we have only the posterior distribution $p(\mu, \sigma^2|\mathbf{X})$ after observing the data $\mathbf{X}$. Hence this probability cannot be guaranteed to be at least β for every possible value of $(\mu, \sigma^2) \sim p(\mu, \sigma^2|\mathbf{X})$ and, instead, we guarantee with a pre-specified $1-\alpha$ (close to one) belief (or confidence) with respect to possible values of (μ, σ^2) that the probability in (4) is at least β, that is,

$$P_{\mu,\sigma^2|\mathbf{X}} \left\{ P_{y|\mu,\sigma^2}\{y \le a(\mathbf{X})\} \ge \beta \right\} = 1-\alpha. \tag{5}$$

One recognizes this is the defining equation of that $(-\infty, a(\mathbf{X})]$ is a Bayesian $1-\alpha$ confidence and β content upper tolerance interval for the population $N(\mu, \sigma^2)$. Bayesian tolerance intervals were first introduced in Aitchison (1964), and Guttman (1970, 2006) are excellent references on the topic.

Following Guttman (1970, pp.140-141), the $a(\mathbf{X})$ that solves equation (5) under the non-informative prior for $(\boldsymbol{\mu}, \Sigma)$ given in (1) with $k = 1$ is given by

$$a(\mathbf{X}) = \bar{\mathbf{x}} + \frac{1}{\sqrt{n}}\sqrt{V}\, t_{n-1,\sqrt{n}z_\beta,1-\alpha}\,, \tag{6}$$

where z_β denotes the β quantile of the standard normal distribution $N(0,1)$, and $t_{n-1,\sqrt{n}z_\beta,1-\alpha}$ denotes the $1-\alpha$ quantile of the non-central t distribution with df $n-1$ and non-centrality parameter $\sqrt{n}z_\beta$. This Bayesian $1-\alpha$ confidence and β content upper tolerance interval $(-\infty, a(\mathbf{X})]$ is also the frequentist $1-\alpha$ confidence and β content upper tolerance interval (cf. Guttman, 1970, pp.141). This is the additional incentive for using the non-informative reference prior in this paper.

3.2 Combined decision limits

In this subsection we have $k = 2$. Hence a future sample observation $\mathbf{y} = (y_1, y_2)'$ is declared to be positive if and only if both $y_1 > a_1(\mathbf{X})$ and $y_2 > a_2(\mathbf{X})$ as stated in Holt *et al.* (2015), that is,

$$\mathbf{y} \in S(\mathbf{X}) \quad \text{with} \quad S(\mathbf{X}) = \{\mathbf{y} : \, y_1 > a_1(\mathbf{X}) \text{ and } y_2 > a_2(\mathbf{X})\} \,. \tag{7}$$

In order to control the FPR at the pre-specified level $1-\beta$, it is desirable that

$$P_{\mathbf{y}|\boldsymbol{\mu},\Sigma^{-1}}\{\mathbf{y} \in S(\mathbf{X})\} \leq 1-\beta$$

which is equivalent to

$$P_{\mathbf{y}|\boldsymbol{\mu},\Sigma^{-1}}\{\mathbf{y} \in \bar{S}(\mathbf{X})\} \geq \beta \tag{8}$$

under the assumption that $\mathbf{y}$ is from the population distribution $N_2(\boldsymbol{\mu}, \Sigma)$, where the probabilities are calculated with respect to the distribution of $\mathbf{y}$ conditional on $(\boldsymbol{\mu}, \Sigma^{-1})$, and $\bar{S}(\mathbf{X})$ denotes the complement of $S(\mathbf{X})$. As in the case of $k=1$ in Subsection 3.1, the probability in (8) depends on the value of $(\boldsymbol{\mu}, \Sigma^{-1})$, for which we have only the posterior distribution $p(\boldsymbol{\mu}, \Sigma^{-1}|\mathbf{X})$ after observing the data $\mathbf{X}$. Hence this probability cannot be guaranteed to be at least β for every possible value of $(\boldsymbol{\mu}, \Sigma^{-1})$ from the posterior distribution $p(\boldsymbol{\mu}, \Sigma^{-1}|\mathbf{X})$, and we guarantee with a pre-specified $1-\alpha$ belief (or confidence) about the possible values of $(\boldsymbol{\mu}, \Sigma^{-1})$ that the probability in (8) is at least β, that is,

$$P_{\boldsymbol{\mu},\Sigma^{-1}|\mathbf{X}}\left\{P_{\mathbf{y}|\boldsymbol{\mu},\Sigma^{-1}}\{\mathbf{y} \in \bar{S}(\mathbf{X})\} \geq \beta\right\} = 1-\alpha. \tag{9}$$

One recognizes immediately that $\bar{S}(\mathbf{X})$ is a Bayesian $1-\alpha$ confidence and β content tolerance region for the population $N_2(\boldsymbol{\mu}, \Sigma)$. But this particular tolerance region has not been considered before to the best of our knowledge.

Under frequentist framework, tolerance intervals/regions were introduced first by Wilks (1941). Guttman (1970, 2006), Hahn and Meeker (1991), Krishnamoorthy and Mathew (2009) and Meeker *et al.* (2017) are excellent references on tolerance intervals/regions.

The R package `tolerance` (Young, 2010) allows the computation of many tolerance intervals/regions. Until very recently, the only available frequentist β content and $1-\alpha$ confidence tolerance region specifically for multivariate normal distribution $N_k(\boldsymbol{\mu}, \Sigma)$ is of the ellipsoidal form

$$R(\mathbf{X}) = \{\mathbf{y} : \, (\mathbf{y} - \bar{\mathbf{x}})' V^{-1} (\mathbf{y} - \bar{\mathbf{x}}) \leq c\}$$

where c is the critical constant that needs to be determined so that

$$P_{\bar{\mathbf{x}},V} \left\{ P_{\mathbf{y}|\bar{\mathbf{x}},V} \{\mathbf{y} \in R(\mathbf{X})\} \geq \beta \right\} \, = \, 1 - \alpha \,, \tag{10}$$

where the probability $P_{\mathbf{y}|\bar{\mathbf{x}},V}\{\cdot\}$ is calculated with respect to the random variable $\mathbf{y}$ conditional on $(\bar{\mathbf{x}}, V)$, and $P_{\bar{\mathbf{x}},V}\{\cdot\}$ is calculated with respect to $(\bar{\mathbf{x}}, V)$. The central ellipsoidal tolerance region of Dong and Mathew (2015), also of the form $R(\mathbf{X})$ above but with a larger c, is conservative, i.e. the probability on the left side of the equation in (10) is strictly larger than $1 - \alpha$.

One key factor in the choice of the $R(\mathbf{X})$ above is that the probability $P_{\bar{\mathbf{x}},V}\{\cdot\}$ in (10) does not depend on the unknown parameters $(\boldsymbol{\mu}, \Sigma)$. But even in this case the computation of c is very challenging and only approximation methods are available; see, for example, Krishnamoorthy and Mathew (1999), Krishnamoorthy and Mondal (2006), and Mbodj and Mathew (2015). If $R(\mathbf{X})$ is replaced by $\bar{S}(\mathbf{X})$ then the corresponding probability expression depends on the unknown Σ in a complicated manner and so the computation of tolerance regions of forms different from $R(\mathbf{X})$ is much harder.

But most recently, rectangular (including one-sided or mixed-sided) tolerance regions of β content and $1 - \alpha$ confidence, specifically for multivariate normal distribution, have been constructed in Lucagbo (2021, Section 4.7) by using parametric bootstrap. These tolerance regions are of different forms from the tolerance region $\bar{S}(\mathbf{X})$ in (9) considered in this paper.

For nonparametric rectangular (including one-sided or mixed-sided) tolerance regions, the reader is referred to Young and Mathew (2020) and Lucagbo (2021, Sections 5.6 and 5.7) for the latest development.

In the Bayesian framework, there is no published work on $1-\alpha$ confidence and β content tolerance region of the form $R(\mathbf{X})$ for $N_k(\boldsymbol{\mu}, \Sigma)$ even with $k = 2$. The reader is referred to Chen (2021, Chapter 3) for the latest development on the construction of nonparametric Bayesian tolerance regions.

To determine $(a_1(\mathbf{X}), a_2(\mathbf{X}))$ of $S(\mathbf{X})$ in (7) from the only constraint in (9), we set

$$a_1(\mathbf{X}) = a_1(\lambda, \mathbf{X}) = \bar{x}_1 + \lambda\sqrt{V_{11}}, \quad a_2(\mathbf{X}) = a_2(\lambda, \mathbf{X}) = \bar{x}_2 + \lambda\sqrt{V_{22}} \tag{11}$$

where $\bar{x}_i$ is the i-th element of $\bar{\mathbf{x}}$, V_{ii} is the i-th diagonal element of V, $i = 1, 2$, and λ is the critical constant that needs to be determined from (9). These two expressions of $a_i(\mathbf{X})$ are sensible if one compares them with the decision limit $a(\mathbf{X})$ in (6) for the case of one GH-2000 score. Hence $S(\mathbf{X}) = S(\lambda, \mathbf{X})$, and λ is solved from

$$P_{\boldsymbol{\mu},\Sigma^{-1}|\mathbf{X}}\left\{P_{\mathbf{y}|\boldsymbol{\mu},\Sigma^{-1}}\{\mathbf{y} \in S(\mathbf{X})\} \leq 1-\beta\right\} = 1-\alpha \tag{12}$$

which is equivalent to (9).

Algorithm 3.2 for computing λ by simulation for given $\mathbf{X}$

- *Step 1*: simulate one $(\boldsymbol{\mu}, \Sigma^{-1})$ from the posterior distribution $p(\boldsymbol{\mu}, \Sigma^{-1}|\mathbf{X})$.

- *Step 2*: given the simulated $(\boldsymbol{\mu}, \Sigma^{-1})$ in *Step 1*, solve λ from $P_{\mathbf{y}|\boldsymbol{\mu},\Sigma^{-1}}\{\mathbf{y} \in S(\mathbf{X})\} = 1-\beta$.

- *Step 3*: repeat *Steps 1* and *2* for a large number of L times, $L = 100,000$ say, to get the corresponding $\lambda_1, \cdots, \lambda_L$; order these values as $\lambda_{[1]} \le \cdots \le \lambda_{[L]}$ and use $\lambda_{[\langle(1-\alpha)L\rangle]}$ as the λ we want. Here $\langle(1-\alpha)L\rangle$ denotes the integer part of $(1-\alpha)L$.

We use a simulation method, given by Algorithm 3.2, to compute the λ from (12). It is well known that the $(1-\alpha)$ sample quantile $\lambda_{[\langle(1-\alpha)L\rangle]}$ in Algorithm 3.2 converges almost surely to the $(1-\alpha)$ population quantile λ that solves (9) as $L \to \infty$. Hence $\lambda_{[\langle(1-\alpha)L\rangle]}$ can be regarded as accurate so long as the number of simulations L is large enough. The computation results given in (the penultimate paragraph of) Section 3.3 below show that $L = 100,000$ is sufficiently large for the problem considered in this paper.

Now *Step 1* can be implemented by using the distributional results in (2) and (3) in the following way. We first simulate one Σ^{-1} from $W_2([(n-1)V]^{-1}, n-2)$ and then one $\boldsymbol{\mu}$ from $N_2(\bar{\mathbf{x}}, \Sigma/n)$ to generate one $(\boldsymbol{\mu}, \Sigma^{-1})$. To simulate one Σ^{-1} from $W_2([(n-1)V]^{-1}, n-2)$, we use the Bartlett decomposition (cf. Smith and Hocking, 1972, and the references therein) in the following way. Step (a): generate independent random variables $u_{11} \sim \chi^2_{n-1}$, $u_{22} \sim \chi^2_{n-2}$ and $u_{12} \sim N(0,1)$ to form matrix $U = \begin{pmatrix} \sqrt{u_{11}}, u_{12} \\ 0, \sqrt{u_{22}} \end{pmatrix}$. Step (b): set $\Sigma^{-1} = [(n-1)V]^{-1/2}U'U[(n-1)V]^{-1/2}$ which has the required Wishart distribution. This simulation method for Σ^{-1} works for a general $k \ge 2$. Alternatively one can directly use, for example, the R package `rWishart`.

For *Step 2*, we have from (7) and (11) that

$$\begin{aligned} & P_{\mathbf{y}|\boldsymbol{\mu},\Sigma^{-1}}\{\mathbf{y} \in S(\mathbf{X})\} \\ = \;& P_{\mathbf{y}|\boldsymbol{\mu},\Sigma^{-1}}\left\{y_1 \ge \bar{x}_1 + \lambda\sqrt{V_{11}},\; y_2 \ge \bar{x}_2 + \lambda\sqrt{V_{22}}\right\} \\ = \;& P_{\mathbf{y}|\boldsymbol{\mu},\Sigma^{-1}}\left\{Z_1 \le -\frac{\bar{x}_1 - \mu_1 + \lambda\sqrt{V_{11}}}{\sqrt{\sigma_{11}}},\; Z_2 \le -\frac{\bar{x}_2 - \mu_2 + \lambda\sqrt{V_{22}}}{\sqrt{\sigma_{22}}}\right\} \end{aligned} \tag{13}$$

where $Z_1 = -(y_1 - \mu_1)/\sqrt{\sigma_{11}}$ and $Z_2 = -(y_2 - \mu_2)/\sqrt{\sigma_{22}}$ have distribution $N_2(\mathbf{0}, \begin{pmatrix} 1, \rho \\ \rho, 1 \end{pmatrix})$, with $\Sigma = (\sigma_{ij})$ and $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$. The probability in (13) can be computed directly by using the function `pmvnorm` of the R package `mvtnorm`; See Genz and Bretz (2009) and Genz *et al.* (2020) for more details. Furthermore, note this probability is monotone decreasing in λ. Hence the unique solution λ of $P_{\mathbf{y}|\boldsymbol{\mu},\Sigma^{-1}}\{\mathbf{y} \in S(\mathbf{X})\} = 1 - \beta$ can be easily computed by using a numerical searching algorithm, for example, the bisection method is used in our coding. From our experience, the computation of one λ in *Step 2* takes only a small fraction of a second on an ordinary PC; see more details in the next subsection.

If one uses, in *Step 2*, an inner loop of simulation to compute an approximation to λ, similar to an idea used in, for example, Krishnamoorthy and Mathew (1999) to construct the frequentist ellipsoidal tolerance region $R(\mathbf{X})$, then the computation is much more time-consuming and the resultant λ is much less accurate. Hence this is not recommended for computing λ.

3.3 Applications to the dataset on GH-2000 scores

In this subsection, we compute the decision limits given in the last two subsections using the available sample observations on GH-2000 from Holt *et al.* (2015). For the purpose of illustration, we focus on the following two GH-2000 scores: (1) Siemens IDS generated by using the P-III-NP assay Siemens ADVIA Centaur and IGF-I assay Immunodiagnostic Systems iSYS IGF-I, and (2) Orion LC-MS/MS generated by P-III-NP assay UniQ$^{\text{TM}}$ P-III-NP RIA and IGF-I assay Liquid chromatography-tandem mass spectrometry. These two GH-2000 scores are available for a sample of $n = 917$ female athletes, and plotted by the 917 dots in Figure 1. There were 932 female athletes in the sample originally. But some had Siemens IDS readings missing, and some had Orion LC-MS/MS readings missing. Hence only the $n = 917$ female athletes having both readings available are used in the analysis below.

We set FPR $1-\beta = 1/10000$ and confidence level $1-\alpha = 95\%$ which are currently adopted by WADA. If one wants to use the GH-2000 score Siemens IDS (y_1) only to decide whether a future female athlete with reading $\mathbf{y} = (y_1, y_2)'$ is positive on GH misuse, then the single decision limit is computed from the formula in (6) and given by $a(\mathbf{X}) = 9.3445$. It is depicted by the vertical dotted line in Figure 1. Hence, a future female athlete is judged to be positive if and only if $y_1 > 9.3445$. If one wants to use the GH-2000 score Orion LC-MS/MS (y_2) only to decide whether a future female athlete is positive, then the single decision limit is computed again from the formula in (6) and given by $a(\mathbf{X}) = 8.5703$. It is depicted by the horizontal dotted line in Figure 1. Hence, a future female athlete is judged to be positive if and only if $y_2 > 8.5703$.

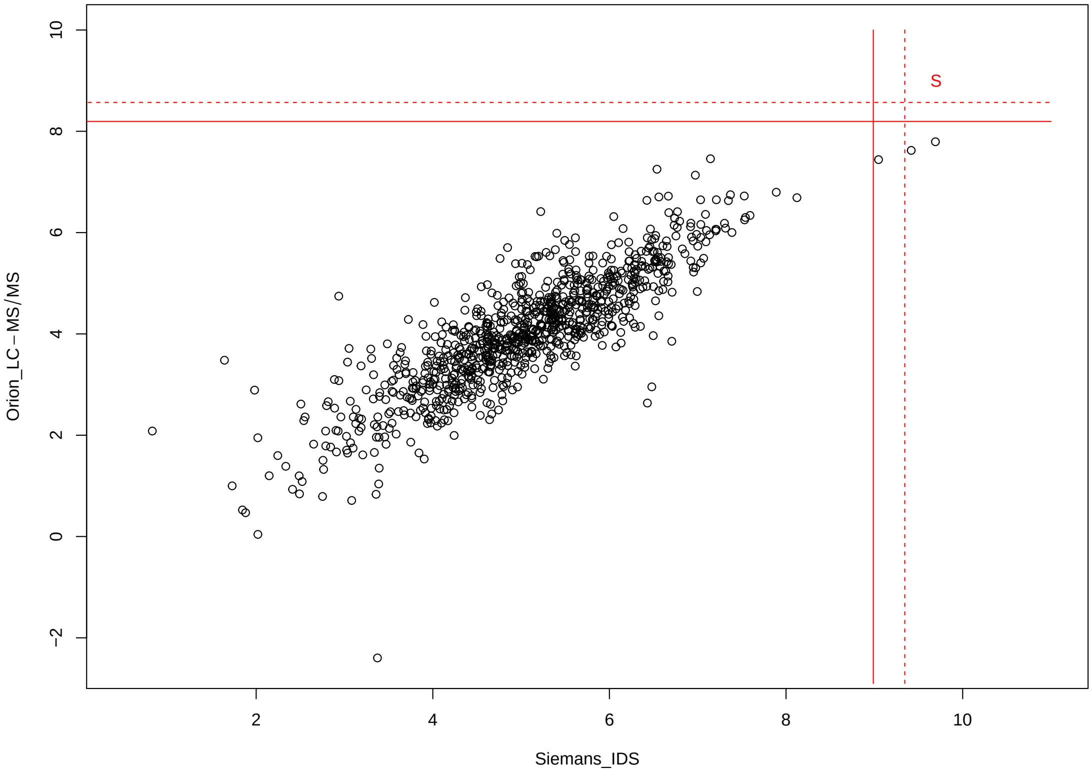

Figure 1: *Plots of the single and combined decision limits based on the observed data. Single decision limits are given by the dotted lines. Combined decision limits are given by the upper-right quadrant S bounded by two solid lines*

On the other hand if one wants to use both the GH-2000 scores Siemens IDS and Orion LC-MS/MS to decide whether a future female athlete with reading $\mathbf{y} = (y_1, y_2)'$ is positive, then the combined decision limits in (11) are used and computed by using Algorithm 3.2. By using $L = 100,000$ simulations, λ is calculated to be 3.5578 which gives $a_1(\mathbf{X}) = 8.9881$ and $a_2(\mathbf{X}) = 8.1952$. These two decision limits are depicted by the two solid lines respectively in Figure 1, and the set $S(\mathbf{X})$ is given by the upper-right quadrant formed by these two solid lines and indicated by the letter S in the figure. Hence, a future female athlete is judged to be positive if and only if both $y_1 > 8.9881$ and $y_2 > 8.1952$.

The decision limits constructed according to the property in (9) or (12) have the following interpretation. With $1 - \alpha$ belief or confidence about the possible value of $(\boldsymbol{\mu}, \Sigma)$ that, the FPR is no more than 1 in 10,000 that a future athlete, whose GH-2000 reading $\mathbf{y}$ follows the distribution $N_2(\boldsymbol{\mu}, \Sigma)$, is wrongly judged to be positive.

The computation of λ based on $L = 100,000$ simulations takes about 210 seconds on an ordinary Window's PC (Intel(R) Core(TM) i5-6600 CPU 3.30GHz, RAM 8.0 GB). We have tried five different random seeds for the random number generator which give the corresponding λ-values: 3.5572, 3.5574, 3.5567, 3.5594 3.5579. This indicates that λ value computed using $L = 100,000$ is likely to be accurate to the second decimal place at least. Indeed, one computation we have done using $L = 1,000,000$ simulations produces $\lambda = 3.5572$ and takes about 2210 seconds (37 minutes). Hence the computation method proposed is fast and accurate enough for practical purpose with $L = 100,000$.

It is valuable to compare the proposed combined decision limits with the combined decision limits of Erotokritou-Mulligan *et al.* (2012) and Holt *et al.* (2015) which is mentioned in Section 1. They are given by $\tilde{a}_1(\mathbf{X}) = \bar{x}_1 + \tilde{\lambda}\sqrt{V_{11}}$ and $\tilde{a}_2(\mathbf{X}) = \bar{x}_2 + \tilde{\lambda}\sqrt{V_{22}}$, and so of similar form as the new combined decision limits given in (11). However, the critical constant is given by $\tilde{\lambda} = \tilde{k} + z_{1-\alpha}\sqrt{(1+\tilde{k}^2/2)/n}$ with $\tilde{k}$ being solved from $P\left\{W_1 > \tilde{k},\ W_2 > \tilde{k}\right\} = 1-\beta$, where $(W_1, W_2)'$ has distribution $N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1, \tilde{\rho} \\ \tilde{\rho}, 1 \end{pmatrix}\right)$ with $\tilde{\rho}$ being the usual sample correlation coefficient between the two GH-2000 scores x_1 and x_2 based on the sample $\mathbf{X}$. For the case considered in this subsection, it is computed that $\tilde{\rho} = V_{12}/\sqrt{V_{11}V_{22}} = 0.852$, $\tilde{k} = 3.4049$, $\tilde{\lambda} = 3.5465$, $\tilde{a}_1(\mathbf{X}) = 8.9755$ and $\tilde{a}_2(\mathbf{X}) = 8.1820$.

While $\tilde{a}_1(\mathbf{X}) = 8.9755$ and $\tilde{a}_2(\mathbf{X}) = 8.1820$ are quite close to $a_1(\mathbf{X}) = 8.9881$ and $a_2(\mathbf{X}) = 8.1952$ respectively for the specific sample observed, the construction of $\tilde{a}_1(\mathbf{X})$ and $\tilde{a}_2(\mathbf{X})$ does not guarantee that $\{\mathbf{y} : y_1 \le \tilde{a}_1(\mathbf{X}) \text{ or } y_2 \le \tilde{a}_2(\mathbf{X})\}$ forms a $1-\alpha$ confidence and β content tolerance region for the population $N_2(\boldsymbol{\mu}, \Sigma)$. Indeed a simulation study in the next subsection shows that the true probability of $\{\mathbf{y} : y_1 \le \tilde{a}_1(\mathbf{X}) \text{ or } y_2 \le \tilde{a}_2(\mathbf{X})\}$ covering β content of the population $N_2(\boldsymbol{\mu}, \Sigma)$ could deviate from the nominal level $1-\alpha$ in both directions.

3.4 Simulation studies

In this subsection, a simulation study is carried out to assess whether the Bayesian $1-\alpha$ confidence and β content tolerance region $\bar{S}(\mathbf{X})$ in Section 3.3 is also a frequentist $1-\alpha$ confidence and β content tolerance region for $N_2(\boldsymbol{\mu}, \Sigma)$. Specifically we assess whether

$$P_{\bar{\mathbf{x}},V}\left\{P_{\mathbf{y}|\bar{\mathbf{x}},V}\{\mathbf{y} \in \bar{S}(\mathbf{X})\} \ge \beta\right\} \ge 1-\alpha \tag{14}$$

holds for all possible values of $\boldsymbol{\mu}$ and Σ; here the probability $P_{\mathbf{y}|\bar{\mathbf{x}},V}\{\cdot\}$ is calculated with respect to $\mathbf{y} \sim N_2(\boldsymbol{\mu}, \Sigma)$ conditional on the sample mean and covariance matrix $(\bar{\mathbf{x}}, V)$, and $P_{\bar{\mathbf{x}},V}\{\cdot\}$ is calculated with respect to $(\bar{\mathbf{x}}, V)$ which depends on the random sample $\mathbf{X}$ from $N_2(\boldsymbol{\mu}, \Sigma)$.

It can be shown that $P_{\bar{\mathbf{x}},V}\left\{P_{\mathbf{y}|\bar{\mathbf{x}},V}\{\mathbf{y}\in\bar{S}(\mathbf{X})\}\geq\beta\right\}$ does not depend on $\boldsymbol{\mu}$. Hence it is only necessary to assess whether (14) holds for all possible values of $\Sigma=\begin{pmatrix}\sigma_{11},\rho\sqrt{\sigma_{11}\sigma_{22}}\\ \rho\sqrt{\sigma_{11}\sigma_{22}},\sigma_{22}\end{pmatrix}$ with $\boldsymbol{\mu}=0$. From the observed sample on GH-2000 scores in Section 3.3, the 99% confidence intervals for ρ, σ_{11} and σ_{22} are given respectively by $(0.827, 0.874)$, $(1.097, 1.396)$ and $(1.215, 1.546)$. So the following three values $(0.83, 0.85, 0.87)$ are used for ρ, $(1.10, 1.25, 1.40)$ for σ_{11}, and $(1.22, 1.38, 1.55)$ for σ_{22} in the simulation study, with a total of 27 combinations of Σ. The range of these 27 combinations cover the likely true value of Σ. Furthermore, FPR $1-\beta=1/10,000$, confidence level $1-\alpha=95\%$ and sample size $n=917$ are used as in Section 3.3.

Algorithm 3.4 for computing the (outer) probability in (14) by simulation

- *Step 1*: simulate one sample $\mathbf{X}=(\mathbf{x}_1,\cdots,\mathbf{x}_n)$ from $N_2(\mathbf{0},\Sigma)$.
- *Step 2*: use the sample $\mathbf{X}$ to compute the region $\bar{S}(\mathbf{X})$ by using Algorithm 3.2; the number of simulations used to compute λ is $L=100,000$.
- *Step 3*: compute $P_{\mathbf{y}|\mathbf{X},V}\{\mathbf{y}\in\bar{S}(\mathbf{X})\}=1-P_{\mathbf{y}|\mathbf{X},V}\{\mathbf{y}\in S(\mathbf{X})\}$ by using an expression similar to (13) with $\boldsymbol{\mu}=0$ for $P_{\mathbf{y}|\mathbf{X},V}\{\mathbf{y}\in S(\mathbf{X})\}$, and the R function `pmvnorm`.
- *Step 4*: repeat *Steps 1-3* for a large number, say $M=1,000$, times; the proportion of times that $P_{\mathbf{y}|\mathbf{X},V}\{\mathbf{y}\in\bar{S}(\mathbf{X})\}\geq\beta$ is used as the required probability.

The (outer) probability $P_{\bar{\mathbf{x}},V}\{\cdot\}$ in (14) is approximated by a proportion using Algorithm 3.4. It takes about 60 hours to compute each probability in Table 1 on the same computer as mentioned in Section 3.3, with most computation time spent on computing the λ-values in the $M=1,000$ repetitions.

Table 1 presents the simulation results on the probability in (14). It is clear from Table 1 that the probabilities are very close to $1-\alpha=0.95$ for all the 27 configurations of ρ,σ_{11}

and σ_{22}. Since the range of these 27 configurations most likely covers the true value of Σ, it follows therefore that it is most likely the inequality in (14) holds for the unknown true value of Σ. Hence the Bayesian combined decision limits can also be interpreted as the frequentist combined decision limits of approximate $1-\alpha$ confidence.

Table 1: the probability in (14) for given Σ

		$\sigma_{11} = 1.10$	$\sigma_{11} = 1.25$	$\sigma_{11} = 1.40$
$\rho = 0.83$	$\sigma_{22} = 1.22$	0.951	0.951	0.951
	$\sigma_{22} = 1.38$	0.951	0.951	0.951
	$\sigma_{22} = 1.55$	0.951	0.951	0.951
$\rho = 0.85$	$\sigma_{22} = 1.22$	0.951	0.951	0.951
	$\sigma_{22} = 1.38$	0.951	0.951	0.951
	$\sigma_{22} = 1.55$	0.951	0.951	0.951
$\rho = 0.87$	$\sigma_{22} = 1.22$	0.953	0.953	0.953
	$\sigma_{22} = 1.38$	0.953	0.953	0.953
	$\sigma_{22} = 1.55$	0.953	0.953	0.953

The results in Table 1 seem to indicate that the probability in (14) depends on Σ, i.e. σ_{11}, σ_{22} and ρ, only through ρ. But this is difficult to prove analytically since the critical constant λ of $S(\mathbf{X})$ depends on Σ in a complicated manner in *Step 1* of Algorithm 3.2. We have done further simulation study on the probability in (14) with $\sigma_{11} = \sigma_{22} = 1$ and various values of ρ in the wider range $[-0.9, 0.9]$. The results are given by Prob (new) in Table 2.

The results on Prob (new) in Table 2 indicate that, if the probability in (14) does depend on Σ only through ρ, then this probability seems to be quite close to $1-\alpha = 0.95$ across the wide range $[-0.9, 0.9]$ of ρ-values.

Finally, we have carried out a simulation study to assess the probability corresponding to the probability in (14) but for the combined decision limits $\tilde{a}_1(\mathbf{X})$ and $\tilde{a}_2(\mathbf{X})$ of Erotokritou-Mulligan *et al.* (2012) and Holt *et al.* (2015). It is clear that this probability depends on

Σ only through ρ, and the simulation results on this probability are given by Prob (old) in Table 2. As pointed out in the last subsection, its construction does not guarantee that $\{\mathbf{y} : y_1 \leq \tilde{a}_1(\mathbf{X}) \text{ or } y_2 \leq \tilde{a}_2(\mathbf{X})\}$ is a $1-\alpha$ confidence and β content tolerance region for the population $N_2(\boldsymbol{\mu}, \Sigma)$. From the results in Table 2, it can be seen that the probability of $\{\mathbf{y} : y_1 \leq \tilde{a}_1(\mathbf{X}) \text{ or } y_2 \leq \tilde{a}_2(\mathbf{X})\}$ covering $\beta = 0.9999$ content of the population $N_2(\boldsymbol{\mu}, \Sigma)$ tends to be a bit smaller than the nominal level $1-\alpha = 0.95$ when the true value of ρ is around 0.7. Strong deviations from the nominal level occur for ρ smaller than -0.5. In practice, negative correlations between two GH-2000 scores are unlikely to occur. But in other applications where the two scores have a large negative correlation, the method of Erotokritou-Mulligan *et al.* (2012) and Holt *et al.* (2015) will produce $\tilde{a}_1(\mathbf{X})$ and $\tilde{a}_2(\mathbf{X})$ that are larger than necessary. In contrast, with the new combined decision limits, the probabilities Prob(new) are consistently close to the nominal level across the range of ρ values.

Table 2: the probability in (14) for given Σ with $\sigma_{11} = \sigma_{22} = 1$

$\rho =$	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1	0
Prob(new) =	0.949	0.950	0.950	0.950	0.949	0.950	0.950	0.951	0.950	0.950
Prob(old) =	0.998	0.995	0.986	0.975	0.963	0.956	0.953	0.952	0.949	0.948
$\rho =$	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	
Prob(new) =	0.955	0.950	0.950	0.953	0.958	0.961	0.954	0.950	0.951	
Prob(old) =	0.945	0.940	0.939	0.941	0.948	0.950	0.950	0.943	0.946	

4 Discussion

Decision limits based on the GH-2000 scores produced by the various pairs of analytical assays employed have been published. These scores are used individually but scores for two pairs of assays must be exceeded before an athlete has to answer a case for the misuse of GH. In other words, WADA mandated the measurement of each analyte by two methods which meant that each sample had two GH-2000 scores. Hence it is natural to use the correlation structure involved in the two scores to develop naturally decreased decision limits which would increase the sensitivity of the biomarker method. The biomarker test then would be more sensitive the lower the correlation between the two GH-2000 scores under consideration would be.

While combined decision limits have their benefits there are some drawbacks. First, depending on which of the other pair of assays was used, the decision limit for one assay pair could change and that could lead to confusion. Ideally, for a given GH-2000 score one would like to have a unique decision limit and not one that depends on which other GH-2000 score is used in the pair. Secondly, it became possible to measure IGF-I by mass spectrometry as the preferred choice to measure IGF-I. WADA does not mandate measurement by a second assay when an analyte is measured by mass spectrometry because of the greater reliability and traceability of the method compared with immunoassays. Hence, using the same assay for IGF-I in two GH-2000 scores leads to an increase in the correlation and the potential for an increased sensitivity of the biomarker test diminishes. On the other hand, it might be that in the near future two mass spectometric methods for IGF-I (intact and digest) will be available with a potential of a decrease in the correlation of two GH-2000 scores involved in the pair.

Having said that, it is valuable nevertheless to have a statistical theory for constructing combined decision limits. These combined decision limits should have the same pre-specified $1 - \alpha$ confidence and $1 - \beta$ FPR as the single decision limit. A Bayesian approach is used in this paper to construct the combined decision limits. Our simulation study in Section 3.4 shows that the Bayesian combined decision limits also have satisfactory frequentist property and so can be regarded as frequentist combined decision limits too. The R code available from the authors allow the method ready to be used.

Combined decision limits of other forms are worth investigation too in future. For example, it seems also sensible to use combined decision limits of the form $T(\mathbf{X}) = \{\mathbf{y} : y_1 > a_1(\mathbf{X})$ or $y_2 > a_2(\mathbf{X})\}$ with $a_1(\mathbf{X})$ and $a_2(\mathbf{X})$ of the forms in (11). That is, a future athlete is judged to be positive if either of the two readings is too high. The corresponding $\bar{T}(\mathbf{X})$ becomes a one-sided rectangular tolerance region for a bivariate normal distribution considered recently in Lucagbo (2021, Section 4.7). It would be interesting to compare the tolerance region $\bar{T}(\mathbf{X})$ constructed using Bayesian method as in this paper with the tolerance region $\bar{T}(\mathbf{X})$ constructed using parametric bootstrap of Lucagbo (2021).

The computation method of Subsection 3.2 can potentially be explored in the construction of a frequentist tolerance region of ellipsoidal form $R(\mathbf{X})$ in (10) for $k = 2$ at least, which is probably the most useful case in applications of tolerance regions. Furthermore, construction of Bayesian tolerance region of ellipsoidal form $R(\mathbf{X})$ can also be investigated, even though it is not of direct interest to GH misuse detection.

While the GH misuse detection motivates this work, one can envisage other potential applications of the methodology developed in this paper. For example, suitable decision limits can be constructed to trigger alert on whether a child of a given age is over/under weight or over/under height.

For nonparametric tolerance regions, the reader is referred to Young and Mathew (2020), Lucagbo (2021, Chapter 5) and Chen (2021, Chapter 3) for the latest development.

Acknowledgments We thank the referees for constructive comments.

5 References

Aitchison, J. (1964). Bayesian tolerance regions. *J. Roy. Statist. Soc. Ser. B*, 26, 161-175.

Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis, 3rd ed.*. Wiley: New York.

Bidlingmaier, M., Wu, Z., Strasburger, C.J. (2000). Test method: GH. *Baillieres Best Pract Res Clin Endocrinol Metab.*, 14(1), 99-109.

Böhning, D., Böhning, W., Guha, N., Cowan, D.A., Sönksen, P.H. and Holt, R.I.G. (2016). Statistical methodology for age-adjustment of the GH-2000 score detecting growth hormone misuse. *BMC Medical Research Methodology*, 16: 147 (DOI: 10.1186/s12874-016-0246-8)

Böhning, D., Liu, W., Holt, R.I., Böhning, W., Guha, N., Sönksen, P.H., Cowan, D.A. and Liang, T. (2019). Exact statistical calculation of the uncertainty term in the decision limits based on the GH2000 score for growth hormone misuse detection (doping). *Statistical Methods in Medical Research*, Vol. 28(3), 928936. (doi: 10.1177/0962280217739452).

Box, G.E.P. and Tiao, G.C. (1992). *Bayesian Inference in Statistical Analysis.* Wiley: New York.

Chen, X. (2021). Prediction Sets via Parametric and Nonparametric Bayes: With Applications in Pharmaceutical Industry. Unpublished PhD dissertation, Leiden University.

Dall, R., Longobardi, S., Ehrnborg, C., Keay, N., Rosen, T., Jorgensen, J.O., *et al.* (2000). The effect of four weeks of supraphysiological growth hormone administration on the insulin-like growth factor axis in women and men. GH-2000 Study Group. *J. Clin Endocrinol Metab.*, 85(11):4193-200.

Dong, X. and Mathew, T. (2015). Central tolerance regions and reference regions for multivariate normal population. *Journal of Multivariate Analysis*, 134, 50-60.

Erotokritou-Mulligan, I., Guha, N., Stow, M., Bassett, E.E., Bartlett, C., Cowan, D.A., Sönksen P.A., Holt, R.I.G. (2012). The development of decision limits for the implementation of the GH-2000 detection methodology using current commercial insulin-like growth factor-I and amino-terminal pro-peptide of type III collagen assays . *Growth Hormone & IGF Research*, 22(2): 53-58. (doi:10.1016/j.ghir.2011.12.005)

Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities.* Springer Lecture Notes in Statistics.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F. and Hothorn, T. (2020). `mvtnorm`: Multivariate Normal and t Distributions. R package version 1.1-1, https://CRAN.R-project.org/package=mvtnorm.

Guttman, I. (1970). *Statistical Tolerance Regions: Classical and Bayesian.* Griffin: London.

Guttman, I. (2006). Tolerance Regions, in *Encyclopedia of Statistical Sciences, 2nd edition*, edited by Kotz S *et al.*, 8644-8659, Wiley: New York.

Hahn, G. and Meeker, W.Q. (1991). *Statistical Intervals: A Guide t o Practitioners.* New York: Wiley.

Holt, R.I. (2009). Is human growth hormone an ergogenic aid? *Drug Testing and Analysis*, 1: 412-418. (doi: 10.1002/dta.58)

Holt, R.I., Böhning, W., Guha, N., Bartlett, C., Cowan, D.A., Giraud, S., Bassett, E.E., Sönsken, P.H., Böhning, D. (2015). The development of decision limits for the GH-2000 detection methodology using additional insulin-like growth factor-I and amino-terminal propeptide of type III collagen assays. *Drug Testing and Analysis*, 7(9):745-755. (doi: 10.1002/dta.1772)

Krishnamoorthy, K. and Mathew, T. (1999). Comparison of approximation methods for computing tolerance factors for a multivariate normal population. *Technometrics*, 41, 234-249.

Krishnamoorthy, K. and Mathew, T. (2009). *Statistical Tolerance Regions – Theory, Applications, and Computation.* Wiley: New York.

Krishnamoorthy, K. and Mondal, S. (2006). Improved tolerance factors for multivariate normal distributions. *Comm. Statist. Simulation Comput.*, 35, 461-478.

Longobardi, S., Keay, N., Ehrnborg, C., Cittadini, A., Rosen, T., Dall, R., *et al.* (2000). Growth hormone (GH) effects on bone and collagen turnover in healthy adults and its potential as a marker of GH abuse in sports: a double blind, placebo-controlled study. The GH-2000 Study Group. *J. Clin Endocrinol Metab.*, 85(4):1505-12.

Lucagbo, M.D. (2021). Rectangular Statistical Regions with Applications in Lab oratory Medicine and Calibration. Unpublished PhD dissertation, University of Maryland, Baltimore County, USA.

Mbodj, M. and Mathew, T. (2015). Approximate ellipsoidal tolerance regions for multivariate normal populations. *Statistics and Probability Letters*, 97, 41-45.

Meeker, W.Q., Hahn, G.J. and Escobar, L.A. (2017). *Statistical Intervals: A Guide For Practitioners And Researchers, 2nd ed..* New York: Wiley.

Powrie, J.K., Bassett, E.E., Rosen, T., Jorgensen, J.O., Napoli, R., Sacca, L., Christiansen, J.S., Bengtsson, B.A., Sonksen, P.H. (2007). Detection of growth hormone abuse in sport. *Growth Horm IGF Res.*, 17: 220-226.

Smith, W.B. and Hocking, R.R. (1972). Algorithm AS 53: Wishart Variate Generator. *J. Roy. Stats. Soci. Series C*, 21(3), 341-345.

Wilks, S.S. (1941). Determination of sample sizes for setting tolerance limits. *Annals of Mathematical Statistics*, 12, 91-96.

World Antidoping Agency. (2021). Laboratory Guidelines - Human Growth Hormone (hGH) Biomarkers Test. https://www.wada-ama.org/sites/default/files/resources/files/wada_guidelines_hgh_-biomarkers_test_v3_jan_2021_eng.pdf

World Antidoping Agency. (2019). Human Growth Hormone (hGH) Isoform Differential Immunoassays for Doping Control Analyses. https://www.wada-ama.org/sites/default/files/-resources/files/-td2019gh_v1_final_eng.pdf

World Antidoping Agency. (2016). The World Anti-Doping Code International Standard: Prohibited List 2016, https://wada-main-prod.s3.amazonaws.com/resources/files/wada-2016-prohibited-list-en.pdf

Young, D.S. (2010). *tolerance*: An R Package for Estimating Tolerance Intervals. *Journal of Statistical Software*, 36, 1-39.

Young, D.S. and Mathew, T. (2020). Nonparametric hyperrectangular tolerance and prediction regions for setting multivariate reference regions in laboratory medicine. *Statistical Methods in Medical Research*, Vol. 29(12) 35693585