Growth Hormone & IGF Research 24 (2014) 205–215

Contents lists available at ScienceDirect

Growth Hormone & IGF Research

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

hGH isoform differential immunoassays applied to blood samples from athletes: Decision limits for anti-doping testing

James A. Hanley [a,b,*], Olli Saarela [a], David A. Stephens [b], Jean-Christophe Thalabard [c,d]

[a] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada
[b] Department of Mathematics and Statistics, McGill University, Montreal, Canada
[c] Paris Descartes University, MAP5, UMR CNRS 8145, Paris, France
[d] Endocrine Gynaecology Unit, Hôpital Cochin, Paris, France

ARTICLE INFO

Article history:
Received 17 May 2014
Accepted 2 June 2014
Available online 11 June 2014

Keywords:
Quantile
Regression
Decision limits
Isoforms
Human Growth Hormone
Doping

ABSTRACT

Objective: To detect hGH doping in sport, the World Anti-Doping Agency (WADA)-accredited laboratories use the ratio of the concentrations of recombinant hGH ('rec') versus other 'natural' pituitary-derived isoforms of hGH ('pit'), measured with two different kits developed specifically to detect the administration of exogenous hGH. The current joint compliance decision limits (DLs) for ratios derived from these kits, designed so that they would both be exceeded in fewer than 1 in 10,000 samples from non-doping athletes, are based on data accrued in anti-doping labs up to March 2010, and later confirmed with data up to February–March 2011. In April 2013, WADA asked the authors to analyze the now much larger set of ratios collected in routine hGH testing of athletes, and to document in the peer-reviewed literature a statistical procedure for establishing DLs, so that it be re-applied as more data become available.

Design: We examined the variation in the rec/pit ratios obtained for 21,943 screened blood (serum) samples submitted to the WADA accredited laboratories over the period 2009–2013. To fit the relevant sex- and kit-specific centiles of the logs of the ratios, we classified 'rec/pit' ratios based on low 'rec' and 'pit' values as 'negative' and fitted statistical distributions to the remaining log-ratios. The flexible data-driven quantile regression approach allowed us to deal with the fact that the location, scale and shape of the distribution of the modeled 'rec/pit' ratios varied with the concentrations of the 'rec' and 'pit' values. The between-kit correlation of the ratios was included in the fitting of the DLs, and bootstrap samples were used to quantify the estimation error in these limits. We examined the performance of these limits by applying them to the data obtained from investigator-initiated hGH administration studies, and in athletes in a simulated cycling stage race.

Results: The mean and spread of the distribution of the modeled log-ratios depended in different ways on the magnitude of the rec and pit concentrations. Ultimately, however, the estimated limits were almost invariant to the concentrations, and similar to those obtained by fitting simpler (marginal) log-normal and Box–Cox transformed distributions. The estimated limits were similar to the (currently-used) limits fitted to the smaller datasets analyzed previously. In investigator-initiated instances, the limits distinguished recent use of rec-hGH from non-use.

Conclusions: The distributions of the rec/pit ratios varied as a function of the rec and pit concentrations, but the patterns in their medians and spreads largely canceled each other. Thus, ultimately, the kit- and sex-specific ratio DL obtained from the simpler model was very close to the 'curve of DLs' obtained from the more complex one. Both were close to previously established limits.

1. Introduction and background

Human Growth Hormone (hGH) is a naturally occurring peptide hormone synthesized in and secreted by the pituitary (p) gland. phGH can also be medically supplemented or replaced by recombinant ["r"] hGH in the case of children's growth disorders and adult deficiencies. rhGH has been listed as a prohibited substance in sport initially by the International Olympic Committee, and then by WADA. Large amounts of rhGH were uncovered during the 1998 Tour de France, at a time when its misuse was considered undetectable by laboratory methods.

In 1999, Strasburger and colleagues [1] described how changes in serum hGH isoform composition [2] could be used to detect the presence of exogenous [externally produced] GH, and they began to develop and validate selective immunoassays intended specifically to screen for

* Corresponding author at: Department of Epidemiology, Biostatistics and Occupational Health, 1020 Pine Ave. West, Montreal H3A 1A2, Canada. Tel.: +1 514 398 6720; fax: +1 514 398 4503.
E-mail address: james.hanley@mcgill.ca (J.A. Hanley).

http://dx.doi.org/10.1016/j.ghir.2014.06.001
1096-6374/© 2014 Elsevier Ltd. All rights reserved.

and confirm hGH doping [3]. To paraphrase one of them [4], these assays are based on the principle that whereas endogenous [produced by the body] hGH consists of several isoforms (22-kDa being most abundant, followed by 20-kDa, …) – with relative abundances that are presumed to be largely invariant to the overall phGH level, and to be largely unaffected by normal activities – exogenous hGH consists of only one of these isoforms, the major, monomeric 22-kDa.

After rhGH administration, phGH release is down-regulated, and 22-kDa hGH becomes predominant. By subjecting a serum sample to 2 assays, one preferentially recognizing the monomeric 22-kDa isoform, the other a broader one recognizing a variety of isoforms, one can determine two concentrations. Instead of "22-kDa" and "general," the two assays have been named "rec" and "pit" because of their preferential binding to either rhGH or phGH. The "22-kDa"/"general" (or "rec"/"pit") ratio is taken as a measure of the relative abundance of monomeric 22-kDa hGH in the sample. An abnormally high ratio may indicate recent rhGH administration.

The WADA International Standard for Laboratories [5] establishes a requirement for a second, independent test to confirm any adverse analytical finding (AAF) that is based on the use of immunoassays. Therefore, two different kits (hereafter called "kit 1" and "kit 2"), using capture antibodies that recognize different epitopes of the target hGH molecule, were developed. Each kit (supplied by CMZ-Assay GmbH, Germany) produces its own pair of "rec" and "pit" values, and, thus, its own "rec/pit" ratio.

WADA-supervised testing of athletes to detect hGH doping began in 2004, when the first, research grade isoform differential immunoassays developed by Strasburger, Bidlingmaier and Wu [6] were applied for testing of athletes during the Athens Olympic Games. A detailed review of the history of the development and implementation of tests for detection of doping with hGH in sport can be found here [7]. The current WADA guidelines can be found at this URL [8].

The 'A' sample is used for screening, and the 'B' sample is only analyzed later if need be, if requested by the athlete. Typically only one of the two kits is used for the initial testing procedure (screening) using the 'A' sample. Currently, if, with a specific kit, the value of rec is below 0.1 ng/mL, the sample is considered negative with respect to that kit, irrespective of the value of pit or the resulting rec/pit ratio. If rec is at or above 0.1 ng/mL, but the value of pit is below the assay's limit of quantification ("loq_pit"), the ratio is calculated as rec/loq_pit rather than rec/pit. If the value of rec is at or above 0.1 ng/mL, and the rec/pit (or, if applicable, the rec/loq_pit) ratio exceeds a kit-specific decision limit (DL), the sample is considered 'positive' with respect to that kit, and the finding of the screening procedure constitutes a Presumptive Analytical Finding (PAF) which would have to be confirmed in the "A" sample and, if necessary (i.e. if requested by the athlete), in the corresponding "B" sample. During confirmation, the sample is analyzed with both kits in triplicate aliquots, i.e. it is reanalyzed with the same kit used during the screening procedure (e.g. kit 1) and also measured with the other, complementary kit (e.g. kit 2). Only when the results of the confirmation analysis are positive for both kits simultaneously (i.e. the rec/pit value obtained with each kit is higher than the gender- and kit-specific DLs) is the finding for the "A" sample reported as an AAF. If requested by the athlete, the confirmation analysis is repeated anew on the "B" sample, and it shall confirm the "A" sample findings for the AAF to hold true. In many cases, however, doping athletes opt to accept the original "A" sample finding to avoid the further embarrassment of the "B" sample confirmation.

The DLs (males 1.81; females 1.46 on kit 1; males 1.68; females 1.55 on kit 2) were established and promulgated by WADA in 2010. They were based on values for 1428 males and 691 females for kit 1 and 263 and 121 for kit 2, derived from samples of elite track-and-field athletes collected during the IAAF World Championships in Athletics in Berlin in 2009, athletes included in the German NADA anti-doping program, and data collected from 9 WADA-accredited laboratories from Jan 2009 to March 2010 following the adoption of the current kits 1 and 2 in routine anti-doping analysis.

The WADA/USADA hGH Working Group and the WADA Laboratory Expert Group decided to proceed with the publication and application of these DL values at that stage of test implementation. It was anticipated that as more data were collected by WADA-accredited laboratories, a re-evaluation of the DLs would occur. The first such re-evaluation (based on data on 2244 males and 772 females (kit 1) and 551 and 167 (kit 2) from 21 laboratories) suggested no need to revise the DLs upwards. However, the statistical procedures used to set the DLs were challenged in an appeal to the Court of Arbitration of Sport in 2011 [9].

In April 2013, WADA asked {JAH, OS, DAS} and J-CT to prepare two independent reports describing a detailed statistical procedure to establish a decision limit DL_1 for the ratio from kit 1 and a DL_2 for that from kit 2. It provided them with updated datasets containing considerably more observations than had been analyzed previously. The only stipulation was that, as had been for the previous analyses, *the DLs be such that, of 10,000 samples, from sports persons whose hGH is entirely endogenous, tested with one or other kit, and the other kit if indicated, fewer than 1 would have ratios that exceed both DLs*.

In setting reference centiles for clinical medicine and anthropometry, the focus is often on the 3rd and 97th, or 1st and 99th; moreover, abundant data are usually available, and samples always exclude those with a condition known to influence the entity in question. Here, in contrast, the focus is on the much more extreme 99.99th centile, where even our comparatively large sample sizes preclude using direct sample centiles; moreover, as we will document, the location and shape of the distribution of the log-ratio are functions of the rec and pit concentrations. For these reasons, the fitting of extreme centiles must rely on statistical models. Since the bulk of our data are from routine anti-doping tests, we are unable to identify and exclude samples influenced by the 'condition' (exogenous hGH) being screened for. Thus, depending on the extent of exogenous hGH, the fits are likely to overestimate the centile of interest — that for athletes who have not recently used hGH.

This report describes how the sex- and kit-specific distributions of the log-ratio depend on the values of rec and pit, and documents the statistical modeling used to arrive at decision limits. It proceeds by first answering the question of whether, for each sex and kit separately, and after setting aside (treating as 'negative') the ratios based on low serum hGH concentrations, a transformation could be applied to the remaining ratios that would result in a single (homogeneous) distribution – free of any systematic patterns – that could then be used to establish a single decision limit for each sex and kit. In light of the (negative) answer, it then describes how, again for each sex and kit separately, we modeled the systematic pattern using a flexible semi-parametric regression model for the log-ratios, using a function of the rec and pit concentrations as the 'regressor,' and how we used this fitted regression model to establish concentration-specific decision limits. We examine how much these limits differ from a single (independent of concentration) decision limit for the sex and kit in question. In addition to graphs, we present the DL 'functions' as Tables. We look for any evidence of systematic distributions across sports. We report how well the fitted model and resultant decision limits detect known hGH doping.

2. Materials and methods

2.1. Datasets

Table 1 describes the 3 datasets provided to us. For the analyses of the doping-control data set (the primary focus), it was not possible to identify different samples from the same athlete, but we believe the proportion is small enough that error-band corrections for this 'clustering' would be small. We included Atypical Findings. These were highly suspicious values obtained either before the DLs had been officially

Table 1
The datasets provided by WADA: provenance, content and use.

Provenance/nature	Size and content	Use/exclusions
1. Athletes/samples analyzed by both kit 1 and kit 2	816 samples (438 men, 378 women) Lab code Sample code 'rec' & 'pit' value for each kit 'Ethnicity' (88 African, 175 Caucasian, 466 Japanese, 1 Chinese, 1 Indian, 185 unspecified) Sport participated in (45 categories)	To calculate between-kit correlations in rec/pit ratios
2. All blood samples screened in WADA accredited laboratories over period 2009–March 2013.	Kit 1: 4546 females; 10,155 males; Kit 2: 2150 females; 5092 males. Lab code; date; sample code 'rec' & 'pit' value from the kit used 'Ethnicity' (89% unspecified) Sport participated in (61 categories)	To establish DLs/ 1 male athlete medically treated with hGH[a] doped athletes (8 M, 1 F)
3. One-time samples from blood donors Serial samples in subjects who participated in (i) controlled studies of investigator-administered exogenous rec-hGH [Jing et al.] (ii) elite athletes in simulated 15-day cycling tour. [Voss et al.]	'rec' & 'pit' measured by both kits Lab code; date; sample code 19 male Chinese university students 'rec' & 'pit' value from each kit Timing of each sample 21 males 'rec' & 'pit' value from each kit Timing of each sample	To test DLs

[a] Under an approved Therapeutic Use Exemption for hGH.

approved and implemented by WADA, and which triggered further target testing of the athlete, or samples for which the values of rec/pit were higher than the DLs just for one kit, but not for the other kit. We also included those Adverse Analytical Findings (3 males) that have been appealed by the athletes before arbitration courts, irrespective of how extreme these values may look with respect to the rest of the data (see below). We excluded, but show, data from 9 doped athletes i.e. those values corresponding to reported Adverse Analytical Findings for hGH from athletes who have either admitted to using recombinant hGH or have accepted the anti-doping sanctions without challenging the analytical result and thus have been sanctioned.

2.2. Statistical analysis

2.2.1. Preliminary remarks on screening data

We examined and used the distributions of 21,943 observations. Of these, most involved numerical values for both rec and pit, even if some of these values were below the laboratory limits of quantification ('loq' or 'lq'). Across the 4 kit × sex combinations, some 48, 18 and 51 (117 in total) of the records involved text (rather than purely numerical) entries (e.g., "< lq") or a mix of text and a number indicating that the rec or pit, or both values were below the respective loq's. In the mixed text-and-number cases, we did not try to extract the numbers from the text, and so classed them with the other "< lq" ones. All of the numerical and non-numerical values were used, either for the 'non-modeled', or for the 'modeled' portion.

The kit- and sex-specific distributions of rec, pit and their ratio have longer right tails (with several orders of magnitude variations), and so we present the distributions of their logs ($\log_2$ scale) in the left portion of each panel in Fig. 1. The proportion of low (below the loq) hGH values was much higher for males than females. For kit 1, the median rec and pit values are approximately 0.60 and 1.25 ng/ml in females and 0.15 and 0.35 ng/ml in males, but the medians of the sex-specific ratios are much closer to each other: 0.51 in females and 0.47 in males. For kit 2, the rec and pit values are again higher in females, but again the medians of the sex-specific ratios are closer to each other: 0.59 in females and 0.53 in males. Although one cannot readily use boxplots to visually judge log-normality, the rough log-symmetry of all 12 distributions, and in particular those of the 4 ratios, is of note. It provided a natural starting point for the statistical modeling necessary to establish sex- and kit-specific decision limits: throughout we took the log-ratio as the 'raw' dependent variate.

2.2.2. Dealing with low hGH concentrations

Laboratories are reluctant to consider limits for ratios of small (and thus less reliably quantified) quantities. Likewise, we did not wish to model the excessive variation caused by these low concentrations. Thus in order to focus on genuine inter-individual variation, we followed the current two-part approach of treating samples with low rec or pit values as automatically 'negative' and fitting a statistical model to the remaining ratios for the sex and kit in question. For the latter, we used a cutoff that allowed us to get good fitted values for ratios based on concentrations that are considered to have been reliably measured.

2.2.3. The influence of hGH concentration on the distribution of the remaining log-ratios

After setting aside the log-ratios based on low serum hGH concentrations, we looked for systematic patterns in the distributions of the remaining ones, by constructing concentration-specific boxplots. We used four measures of hGH concentration: rec alone, pit alone, the geometric mean of rec and pit, and the minimum of rec and pit. We found the same systematic patterns (see the rightmost portion of each panel in Fig. 1) with all measures, and so adopted a concentration-based 'centile regression' approach, using as the regressor the geometric mean (GM) of rec and pit. For coherence, we also used the GM to divide each sex- and kit-specific dataset into the 'concentration too low' and 'used in modeling' portions. The reasons for the choice of the geometric mean of rec and pit, rather than one or the other or some other function of both, and for the GM boundary of 0.075, are given in Appendix A.

2.2.4. Form of centile regressions

As is evident from Fig. 1, each centile regression needed to accommodate the fact that the distributions of the log-ratios at different hGH concentrations had means, medians, standard deviations and shape (possible skewness) that varied with these concentrations. When plotted against (the ranks of) concentration, the medians of the log-ratios tended to have (in women) a mostly monotonic or (in men) a more quadratic relationship, while the interquartile ranges tended to be smaller ('taper') at higher concentrations. As is commonly done when establishing reference limits for growth charts, we sought a Box–Cox transform of the log-ratios that would make the residual variation at each hGH concentration close to Normal. We accommodated these features by using the LMS model for centile regression [10], which allows the mean, standard deviation and skewness to be modeled as separate functions of the

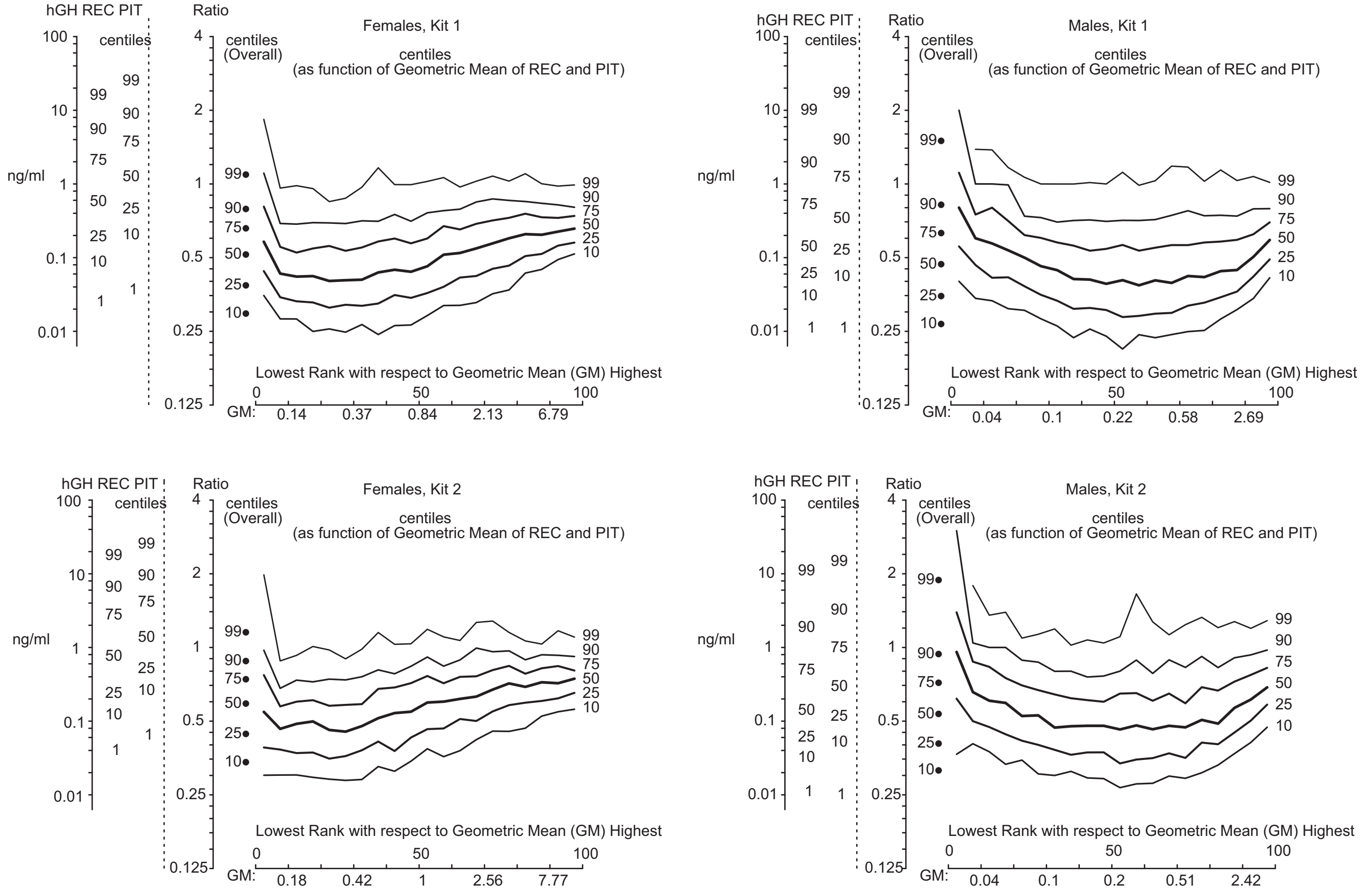

Fig. 1. Selected centiles (empirical) of the kit- and sex-specific distributions of rec and pit (shown to left of vertical dotted line) together with the centiles, both overall (shown as dots) and as a function of concentration (shown as lines, with labels to their right), of their ratio. The geometric mean (GM) of rec and pit was used as a measure of concentration. The concentration-specific centiles were calculated by dividing the kit and sex-specific dataset into 20 equal-size bins based on the GM, thereby ranking the samples (along the horizontal axis) by the 'magnitude' of the rec and pit concentrations. All 'numerical' observations whose values exceed zero were used.

hGH concentration. A more technical description, along with the fitting criteria we followed, can be found in Appendix A. A further advantage of the LMS model was the possibility to reduce it to the simplest nested case so that it yielded a single (not-concentration-specific) limit derived after a conventional (single) Box–Cox transform. The series of transformations induces a scale where fitting of parameters is less affected by extremes, and on which it is easier to fit statistical models.

2.2.5. Adjustment of kit-specific DL's for between-kit correlations in the log-ratios

The fitted regressions allowed a separate DL to be calculated for each concentration for each sex and the kit, but using a modification that reflects the correlation between the ratios on the two kits. If the decision were based on a single kit, we would have used a 99.99% DL that was 3.72 standard deviations above the mean log-ratio. Since a decision involves both kits, and the correlation between the log-ratios on kit 1 and kit 2 is less than 1, it reduces the chance that both test results would exceed 3.72 standard deviations. In samples with concentrations above the GM cutoff, the 'paired' data on athletes yielded a correlation of 0.84 in the log-ratios in males and 0.85 in females, so we used 0.85 for both sexes. Thus we used a deviate derived from a bivariate Normal distribution, where an expected proportion 0.0001 of the modeled ratios would exceed 3.40 (rather than 3.72) standard deviations on each of the two kits. After carrying out all of the limit-fitting on the Box–Cox scale, we transformed the fitted values and limits back to the original ratio scales. There are limited data on A and B samples; we did *not* build in a correction for imperfect correlation between the ratios in the A and B samples, or for the requirement, before a sanction, that the results of both kits applied to sample B would *also* have to have exceeded the kit-specific DLs.

2.2.6. Assessment of model fit

In order to assess whether the LMS model provides a reasonable fit, we used a number of diagnostic checks. These are described in more detail in Appendix A. We did not use p-values from traditional test-of-fit statistics: the sample sizes are large enough that even small deviations from the assumed models, or larger deviations involving low ratios (of lesser interest), or even a few extreme high ones, will produce small p-values. Moreover, the situation is not directly analogous to that in medicine, epidemiology and clinical chemistry, where only those subjects known to be free of the condition/behavior of interest are used to fit reference percentiles. We had not such assurance that this was the case: some of the values below (and some above) the fitted limits may well be from athletes who had recently doped with hGH. In view of this, we did not pursue models that would result in such high and particularistic (peculiar to this dataset, 'overfitted') limits that fewer than 1 value in 10,000 in the dataset would exceed them.

Moreover, even if we could have been reassured that none of the samples was taken following recent doping, it would have been difficult – without having paired values from both kits on a very large number of samples – to check the overall 1/10,000 exceedance frequency. In practice, exceeding the DL on one kit (as shown in Fig. 2) would not necessarily lead to a sanction. First, such values would lead to the use of the second kit, and if need be to a confirmation analysis, and then, if desired, to the analysis of the B sample. Each of these steps would reduce the overall false positive rate.

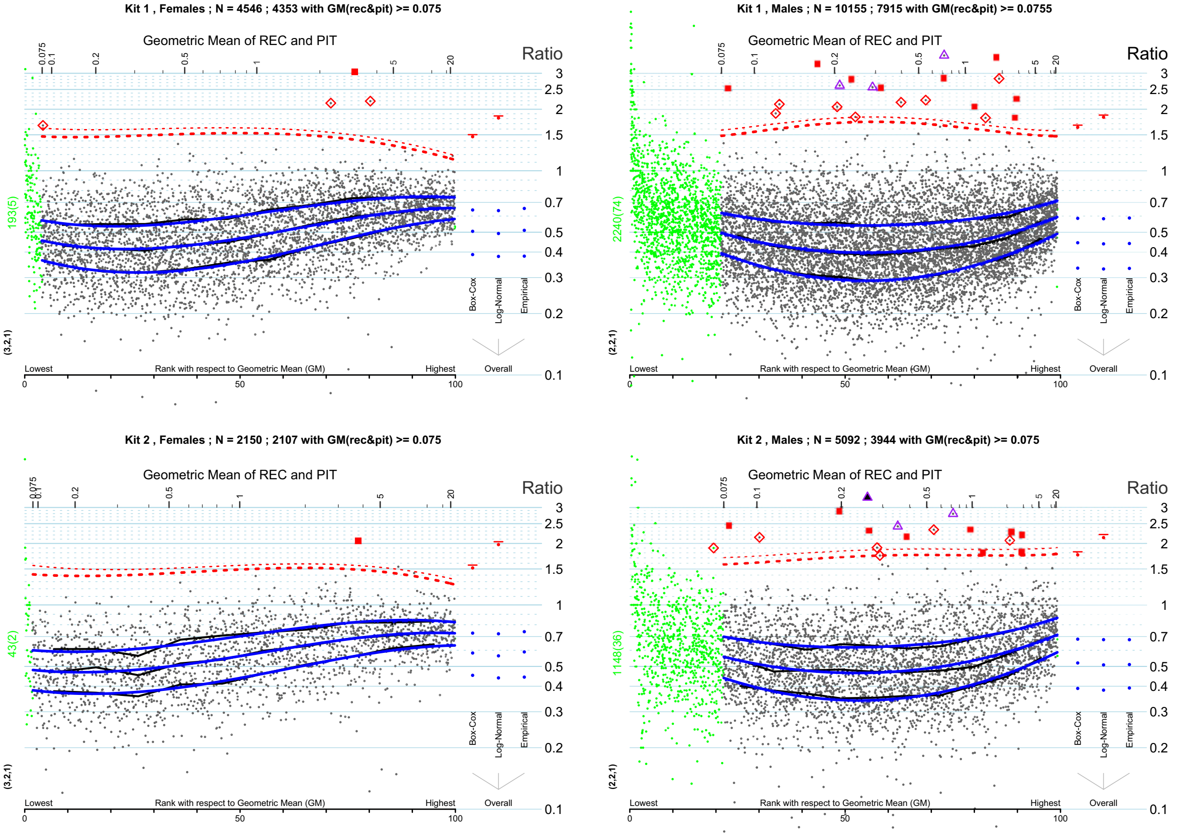

Fig. 2. Kit- and sex-specific distributions of the rec/pit ratio: raw data, empirical quantiles, and fitted quantiles and decision limits, both overall and concentration-specific. Samples with low concentrations, not used in fitting, are shown in green. Samples with higher concentrations are shown as gray dots. The concentration-specific empirical 25th, 50th and 75th centiles are shown as solid black lines. The smooth curves (the 25th, 50th and 75th centiles are in blue; the thicker red dotted lines are the DL point estimates, and the thinner ones are the 95% upper limits) were fitted using LMS models. The three numbers shown vertically at the bottom left of each sub-panel indicate the complexity of each of the 'best' fitted L, M and S curve. All ratios shown as gray dots were used in the fitting. Values corresponding to atypical findings are enclosed by hollow red diamonds; those corresponding to Adverse Analytical Findings that have been appealed by the athletes are enclosed by hollow purple triangles (all dots have been used in the fitting). Values *not used in fitting*, but measured in 1 athlete who had been medically treated with hGH (under an approved TUE for hGH) and in 8 males and 1 female with reported Adverse Analytical Findings for hGH who have either admitted to using recombinant hGH (and thus have been sanctioned) or have accepted the anti-doping sanctions without challenging the analytical result, are shown as solid red squares.

2.2.7. Confidence intervals for DLs

To reflect the fact that the fitted limit at any hGH concentration has its own statistical estimation error, we derived a standard error for the fitted DL. These standard errors are complex functions of the 'n' used in the curve fitting, the variability of the log ratios, the spread of the regressor values, and their distances from the center of the regressor-axis — and not expressible as a closed expression or formula. Instead, we computed them using 250 bootstrap samples for each panel. In the bootstrap procedure, several 'copies' of the data, each one somewhat different from the next, are made by sampling with replacement, and different estimates of the curve are obtained, and the variation between them at each concentration is used to calculate a 'standard error'. We used it to construct a 95% confidence interval at each regressor value, as shown in Fig. 2. We constructed it as a one sided interval in order to be conservative.

2.2.8. Ethnicity-specific limits

Given the complexity of the concept of 'ethnicity', the large percentage of cases where it was not reported, and the large and not easily defined, number of subgroups there would be even if it were, we did not pursue ethnicity-specific limits.

2.2.9. Variations across different sports

We collapsed the over 300 sports categories (and different spellings) to a short list. We used box-plots of the residuals for the categories with at least 50 ratios. We looked for any patterns that were consistent across genders and kits.

2.3. Performance of fitted DLs in serial blood samples from subjects in controlled studies of (a) investigator-administered exogenous rec-hGH, and (b) a simulated training and competition regime

We used the data generated by Jing et al. [11] and Voss et al. [12] to learn how often the fitted DLs would be exceeded in serial blood samples from subjects who had been administered rec-hGH at various intervals before testing, and from athletes who performed a simulated nine day cycling stage. hGH isoforms were analyzed by the official WADA immunoassays. Although we could have (as Voss et al. did) limited the data further to just those ratios based on rec and pit values above the loq, we took a worse case scenario in which we show the *full* range of variation in these ratios, regardless of concentration. We do however use symbols to indicate which ones would in practice be automatically considered 'negative'.

3. Results

3.1. Orientation to graphic display

The raw data, as well as the centiles fitted by various models, are shown in Fig. 2. So as to orient the reader to the format used, we first consider in more detail the test results for the 4546 females tested with kit 1 (upper left panel). The vertical location of each result is the rec/pit ratio (on a log scale), and the horizontal location is where the result ranked (on a 0–100 scale) with respect to the geometric mean (GM) of rec and pit. We treated the 193 with concentrations with a GM < 0.075 (shown as green dots) as 'automatically negative' and used the log-ratios in the remaining 4353 (shown as gray dots) to fit the centiles.

The 25th, 50th and 75th centiles obtained under three methods that *ignored* the horizontal location of each gray dot (i.e., the magnitudes of the two concentrations used to form the ratio) are shown on the right, as blue dots, above the word "*overall*". The three methods are labeled as Empirical: log ratios, no model assumed; Log-Normal: log ratios assumed to follow a Gaussian distribution, and Box–Cox: log ratios, after a shift to make them positive, and then subjected to a Box–Cox transformation, assumed to follow a Gaussian distribution. As one can see, the median ratios by the three approaches are all very close to 0.5. The DLs (with accompanying upper 95% confidence limits) given by the two statistical models are shown directly above them as red dots with error bars: the DL of 1.8 based only on a log-Normal distribution (the model used to establish the *current* limits) differs considerably from the 1.5 obtained by the Box–Cox approach. There is no corresponding distribution-free limit, since it is not possible to establish an extreme centile without assuming *some* distributional form.

3.2. Centile-regression based DLs

The empirical 25th, 50th and 75th centiles as a function of concentration are shown as solid black lines. The 25th, 50th and 75th centiles and the DL values fitted by centile regressions are shown respectively as solid blue and red dotted lines; the thicker red dotted lines are the point estimates, and the thinner ones are the 95% upper limits.

As would be expected from Fig. 1, the fitted median (50th centile) is a mostly-increasing function of the concentration, but when coupled with tapering SDs, and transformations induced by the Box–Cox transforms, the fitted 75th centile curve is less steep. In women, with very few ratios above 1.5, the fitted DL 'function' is almost flat, i.e., almost independent of concentration. This may reflect better-measured input values to the ratio at the upper end of the {rec,pit} scales, as well as other unknown factors. In men, with a number of ratios above 1.5, the fitted DLs are slightly higher than in women; again, however, despite the shape of the curve of medians, the DL curve is largely constant, and could, for practical reasons, be readily replaced by the 'independent-of-concentration' single value, such as that given by the 'overall' DL based on the Box–Cox transformation.

That the error bands accompanying the point estimates of the DLs are slightly wider at the extremes is an expected feature of any regression technique.

The diagnostic plots (Fig. 3) show that the 'raw' (untransformed) log ratios are not as Normal as those derived from even a single-number Box–Cox transformation, and that these in turn are generally improved by fitting the LMS model. We caution against over-interpreting some of the seeming deviations from Normality, since (because of the finite sample sizes) some seeming 'imperfections' are also seen in the data drawn from a known-to-be-Normal distribution (the histogram at the bottom right of each panel).

The 'residuals' from the fitted LMS models for male ratios do include some extreme cases, but we hesitate to try to fit (what would have to be much more particularistic and 'tailored') distributions in which virtually all observations would be below the fitted limits. Moreover, as has been addressed above, the practice of using the other kit to test a screening sample that is elevated on the screening kit, and of further

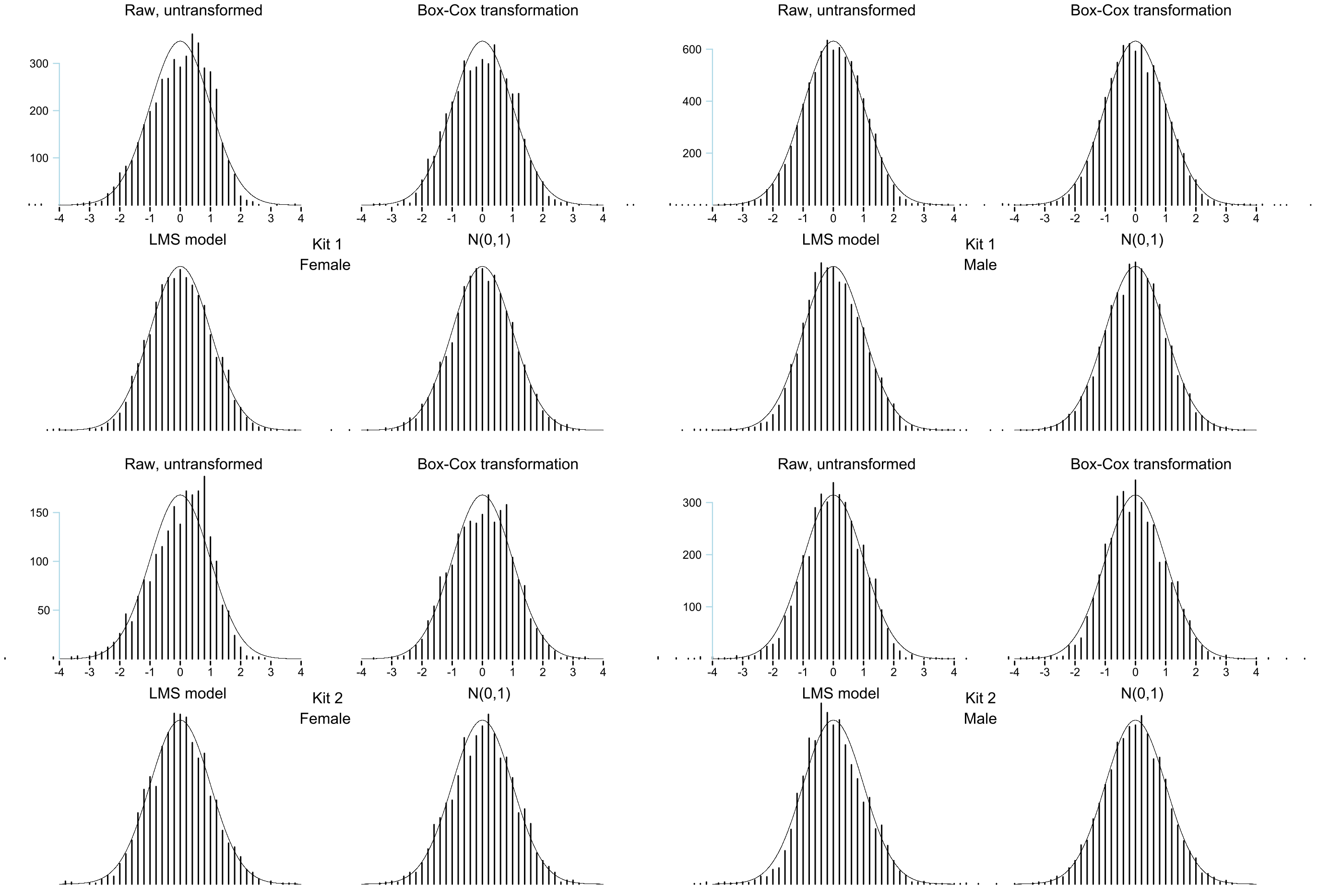

Fig. 3. Histograms (with bins 0.2 units wide, and actual frequencies shown on the y-axes) of the Z-score residuals from the various fits, with the Normal curve superimposed on them.

testing if desired, offers clean athletes additional protection against false accusations.

Since we do not have the sample sizes (or the assurance on the homogeneity of the sample) to directly check the accuracy of the DLs, we instead checked the percentage of the actual observations below the fitted 97.5% limit. We found that across the 4 panels, the percentages ranged from 97.3 to 97.8%. *Within* each panel, the variation in the percentage across the quintiles of concentration levels was also satisfactory — the accuracy did not fluctuate by more than would be expected (under binomial variation) for 5 equal subgroups of the numbers involved.

Had we excluded from the curve fitting the data points corresponding to Atypical Findings or Adverse Analytical Findings that have been appealed by the athletes (marked with a diamond or triangle — see legend, as well as above text), the limits for the ratios would have been lower than those in Fig. 2. Since the ultimate decision on where to set the DLs is for WADA to take, we present a separate Fig. 4 based on the same statistical models, but with these data points removed.

3.3. Tabulated limits

Table 2 shows the same limits as those in Fig. 2, but in tabular form. The "overall" limits shown at the bottom are derived from 'univariate' modeling that uses the geometric mean of 0.075 as a divider, but does not use the ranks of the GM values above 0.075 as a regressor. Table 3 provides the corresponding versions, but with values indicated by a diamond or triangle excluded from the fitting. In all instances, the entries represent the upper 95% limit of the fitted DL, rounded up to the 2nd decimal place.

3.4. Variations across sports

Fig. 5 shows boxplots of regression residuals for categories with 50 or more athletes. Mostly, there do not seem to be any obvious or consistent patterns. A sport that 'seems' to have higher/lower ratios in one kit in one sex doesn't necessarily have the same pattern with another kit or sex. The one notable exception is baseball, where the median ratio is consistently high. Since this could be a chance finding, we hesitate to calculate DLs with this category removed, and instead await replication in an independent dataset.

3.5. Ratios derived from samples from athletes who admitted or were sanctioned for doping

We were also provided with values measured in a male athlete who had been medically treated with hGH (under an approved TUE for hGH) and in those with reported Adverse Analytical Findings for hGH who have either admitted to using recombinant hGH or have accepted the anti-doping sanctions without challenging the analytical result. Naturally, we did not use these to fit the curve, but for interest, we have merely superimposed them on Fig. 2 and denoted them by solid squares.

One reviewer suggested that Figs. 2 and 4, by virtue of their logarithmic ordinates, may give a distorted visual impression of the data cloud and the separation between 'clean' and 'doped' samples, with the top area compressed, and that the discriminating power of the hGH detection test would be more obvious if a linear ordinate were employed. To this end, we have added in the Supplementary Material a version of Fig. 2 with a linear y-axis.

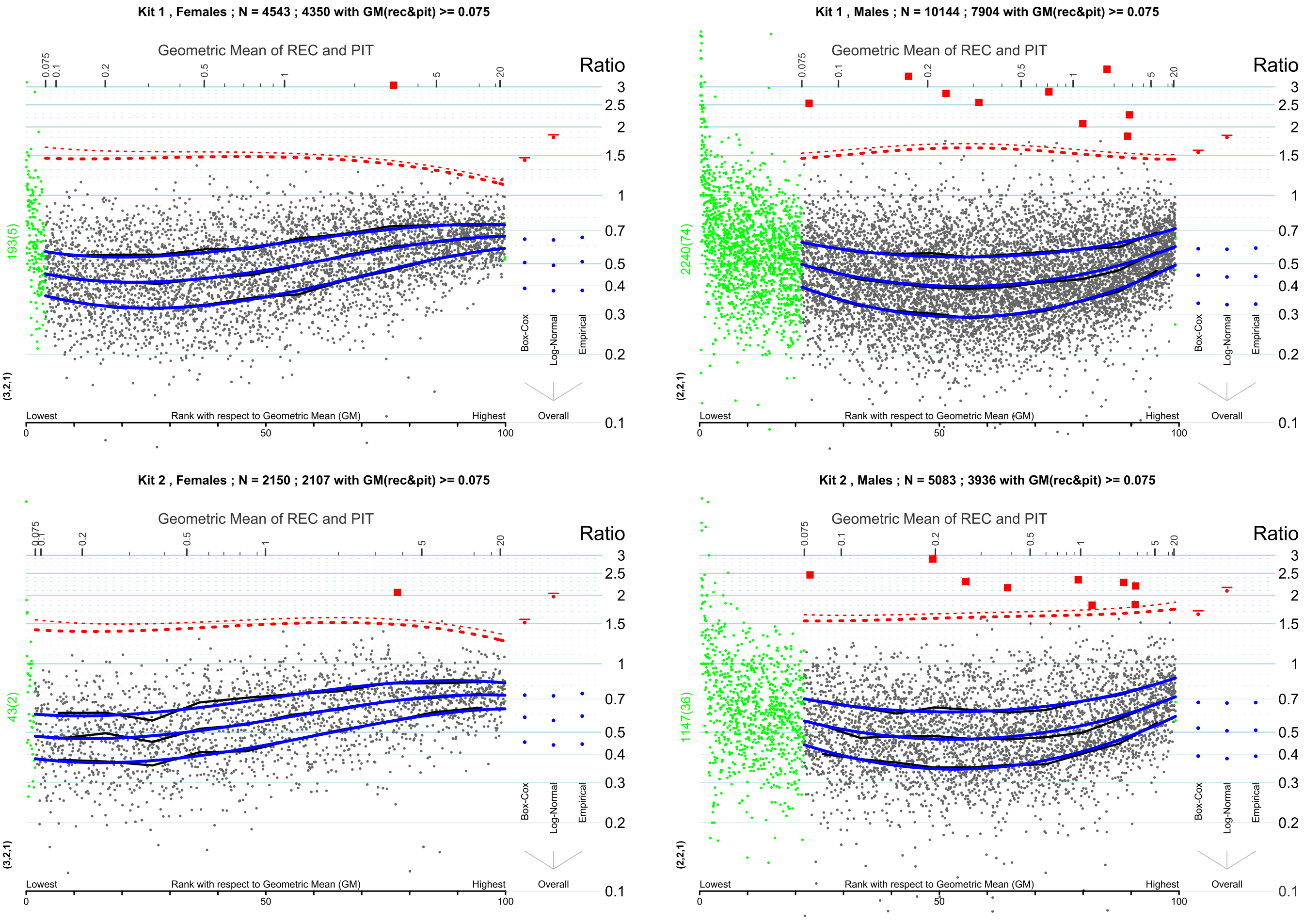

Fig. 4. As in Fig. 2, but with atypical and adverse analytical findings excluded from the fitting.

Table 2
The same fitted 'concentration-specific' decision limits as shown as the upper curves in Fig. 2 (*all* ratios based on a geometric mean (GM) concentration of 0.075 ng/ml or higher), but given in tabular form. Each entry is the upper 95% limit of the confidence interval that accompanies the point estimate. The "Overall" limits shown in the bottom row are derived from 'univariate' modeling that uses the geometric mean of 0.075 as a divider, uses a Box–Cox transformation, but does not use the ranks of the GM values above 0.075 as a regressor.

GM of rec & pit (ng/ml)	Kit 1		Kit 2	
	Females	Males	Females	Males
0.075	1.62	1.59	1.57	1.71
0.1	1.61	1.65	1.56	1.74
0.2	1.59	1.81	1.51	1.84
0.3	1.60	1.84	1.50	1.87
0.4	1.61	1.82	1.50	1.88
0.5	1.62	1.80	1.52	1.88
0.6	1.63	1.78	1.53	1.88
0.7	1.63	1.77	1.54	1.88
0.8	1.63	1.75	1.55	1.88
0.9	1.63	1.73	1.56	1.88
1	1.62	1.72	1.57	1.88
2	1.56	1.64	1.59	1.87
3	1.50	1.61	1.58	1.88
4	1.45	1.60	1.56	1.88
5	1.40	1.59	1.53	1.89
6	1.37	1.58	1.51	1.89
7	1.34	1.58	1.48	1.90
8	1.31	1.58	1.46	1.90
9	1.29	1.58	1.45	1.90
10	1.28	1.58	1.43	1.90
12	1.25	1.57	1.40	1.91
14	1.23	1.57	1.39	1.91
16	1.22	1.57	1.37	1.91
18	1.21	1.57	1.36	1.91
20	1.21	1.57	1.36	1.91
Overall	1.51	1.68	1.57	1.83

Table 3
The same fitted 'concentration-specific' decision limits as shown as the upper curves in Fig. 4 (i.e., atypical samples excluded from the fitting), but given in tabular form. Each entry is the upper 95% limit of the confidence interval that accompanies the point estimate. The "Overall" limits shown in the bottom row are derived from 'univariate' modeling that uses the geometric mean of 0.075 as a divider, uses a Box–Cox transformation, but does not use the ranks of the GM values above 0.075 ng/ml as a regressor.

GM of rec & pit (ng/ml)	Kit 1		Kit 2	
	Females	Males	Females	Males
0.075	1.63	1.54	1.57	1.65
0.1	1.62	1.58	1.56	1.64
0.2	1.57	1.68	1.51	1.66
0.3	1.56	1.69	1.5	1.68
0.4	1.56	1.68	1.51	1.69
0.5	1.56	1.66	1.52	1.7
0.6	1.56	1.65	1.54	1.71
0.7	1.55	1.63	1.55	1.71
0.8	1.55	1.62	1.56	1.72
0.9	1.54	1.61	1.57	1.72
1	1.54	1.6	1.57	1.73
2	1.48	1.55	1.6	1.76
3	1.42	1.53	1.59	1.79
4	1.38	1.52	1.57	1.8
5	1.34	1.52	1.54	1.82
6	1.31	1.52	1.51	1.83
7	1.29	1.52	1.49	1.84
8	1.26	1.52	1.47	1.84
9	1.25	1.52	1.45	1.85
10	1.23	1.52	1.43	1.85
12	1.21	1.52	1.41	1.86
14	1.19	1.52	1.4	1.86
16	1.19	1.52	1.38	1.86
18	1.18	1.52	1.37	1.86
20	1.18	1.52	1.37	1.86
Overall	1.47	1.58	1.57	1.72

3.6. Performance of fitted DLs in serial blood samples from subjects in controlled studies

The data from serial measurements in studies of (a) investigator-administered exogenous rec-hGH [11] (Jing et al.) and (b) a simulated intense training and competition regime in cyclists [12] (Voss et al.) can be used to assess how well the DL's discriminate persons who have recently doped from those who have not. Fig. 6 shows the raw data. The displayed ratios from (b) include some where the absolute concentrations were low. Despite this, they show the quite limited range of rec/pit ratios, well below the DLs (and, as can be assessed from Fig. 1 in the original publication, with no obvious 'real' between-person variation). Despite this added noise, there is a clear separation between the ratios of these cyclists and the soon-after-administration ratios of those in (a). In approximately 50% of instances of recent (within 12 h) hGH administration, ratios exceeded the fitted DLs. Moreover, these ratios are very similar to those values in athletes who have admitted to doping. Detection rates would have been lower in more distant-in-time samples, indicating the narrow time-window of detection (for more details, see the original publication [11]).

4. Discussion

One major difference between this statistical analysis and the ones that led to the current limits is our pursuit of limits that are specific to the magnitudes of the rec and pit values used to calculate the rec/pit ratio. We continued to use log-normality, but did so in the context of concentration-specific (rather than all) ratios, and after data-suggested Box–Cox transformations. The transformation is not modulated by one (constant) global parameter, but by one that changes with concentration in a smooth curve. The regression model also allowed for the fact that log-ratios were considerably more variable when the rec and pit values were small (but above the dividing boundary) and less variable when the rec and pit values were larger. The limits reflect not just this tapering, but also the fact that the overall locations (middles) of the specific distributions of log-ratios themselves vary with concentration. By allowing flexibility in how the means, and the patterns in which the observations lie on each side of them, are described, the adopted approach yields limits that are very much chosen by the patterns present in the data.

In our approach, a much larger proportion of values in males are below our GM cutoff of 0.075. As is described more fully in Appendix A, this dividing line was chosen as a compromise between the loq of 0.05 and the 0.10 threshold often used for rec, and to allow more data to improve the stability of the fitted centile regressions. Since the central tendencies in the non-modeled datapoints (in green) seem to be a smooth continuation of the patterns seen at higher concentrations, one might be inclined to set the cutpoint even lower to further increase the stability of the fitted curves, especially around the GM value of 0.075 itself. However, adopting such an approach actually led to a decrease in stability.

Interestingly, even though the concentration-specific distributions were quite different from each other, the mean and standard deviation patterns largely canceled each other: ultimately, the *single* (not-concentration-specific) DL obtained from the simpler model was quite close to the '*curve of DLs*' obtained from the more complex one.

The limits implied by these models are generally in close agreement with current ones. Possible explanations for the differences include sampling variation, our use of the Box–Cox and other modeling features, and the addition of error bars to reflect the estimation error.

Simplifications of the 4 curves for use as a basis for decision limits will of course need to be based on what would be practical and feasible in the field.

Given our imperfect knowledge as to the behavior of the athletes before the samples were taken, we did not expect to find a model that fitted all of the quantiles perfectly. The empirical performance at 97.5%, where we have sufficient data, was however very close to what would be expected with samples of these sizes. We were unwilling to

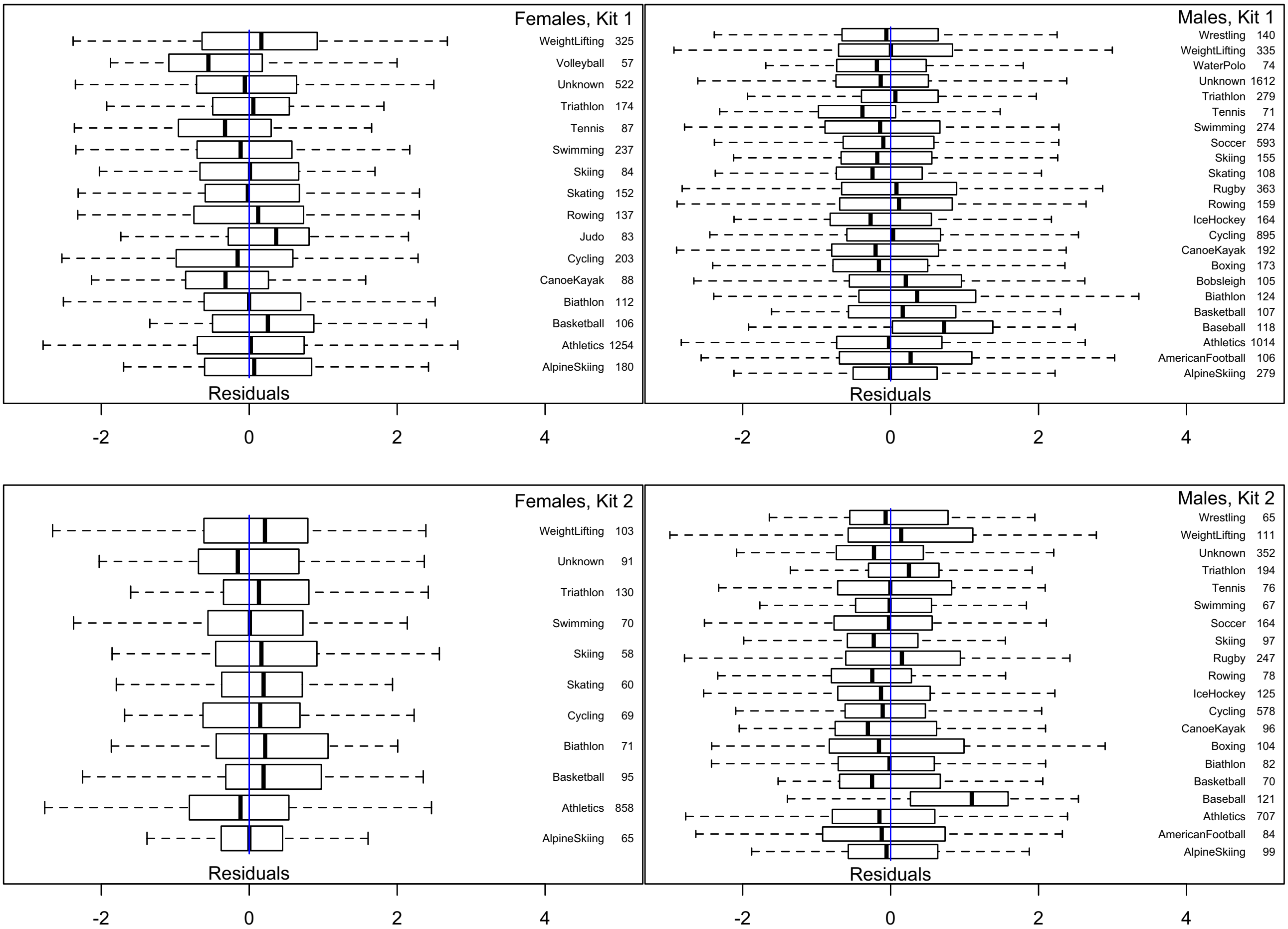

Fig. 5. Boxplots showing between-sports categories variations in the residuals from the centile regression fits in Fig. 2. Only those sports categories with samples sizes (at right) of 50 or more are shown. The left and right boundaries of each box are the 25th and 75th centiles, and the band inside the box is the median. The whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box.

tune the model further (either by increasing the flexibility of the L, M and S functions, or by modeling the residuals as a mixture) so that the fitted DL had no ratios above it. Doing so would not protect the clean athletes. Rather, it could encourage the very behavior that hGH testing is designed to minimize, and push the DLs in the next revision even higher. The only a priori reason to fit a mixture is the very reason one should not fit a mixture.

5. Summary/conclusion

The distributions of the rec/pit ratios varied as a function of the rec and pit concentrations, but the patterns in their medians and spreads largely canceled each other. Thus, ultimately, the kit- and sex-specific ratio DL obtained from the simpler model was very close to the 'curve of DLs' obtained from the more complex one. Both were close to previously established limits.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ghir.2014.06.001.

Appendix A. Additional statistical details

A.1. Choice of 'hGH concentration' metric as regressor in centile regression

We selected the geometric mean of rec and pit, rather than one or the other or some other function of both. We did so in order to be more 'neutral' and to avoid inducing the types of correlations seen when the difference of two quantities (the log of the ratio is a difference of the logs of rec and pit) is regressed against one or other of them. The difference of two random variables is often (negatively) correlated with each one of the two, whereas it is less correlated with their sum or their mean. The difference of log(rec) and log(pit) is the log of the rec/pit ratio, and their mean is the log of the geometric mean. By using the mean, we avoid having to model any correlation that would be induced by using one or the other. We note in passing that this approach is similar to that used by Bland and Altman [13], who plot the difference between two scales of measurement against their mean.

In order to avoid artifacts in the fits at the extremes of the regressor range, we did not use the GM values themselves as the regressor. Instead, (just as in Fig. 1) we used their *ranks*, so that there would be equal amounts of data at all locations on the regressor scale. When displaying the fitted curves, we use the ranks as the primary x-axis scale, but show selected GM values as well.

A dividing line of GM < 0.075 versus >=0.075 was chosen as a compromise between the loq of 0.05 and the 0.10 threshold often used for rec, and to allow for more of the data to contribute to the modeling and to improve stability. Putting the dividing line any lower (especially in men) would have involved modeling of log-ratios based on concentrations that fall below the limit of quantification, and with more complex polynomials. Moreover, as is explained in the text, setting it lower would have allowed a few extreme and not very reliable ratios to reduce the stability of the fitted regressions.

A.2. The LMS model

To quote the authors who first described it, "the distribution at each covariate value is summarized by three parameters, the Box–Cox power

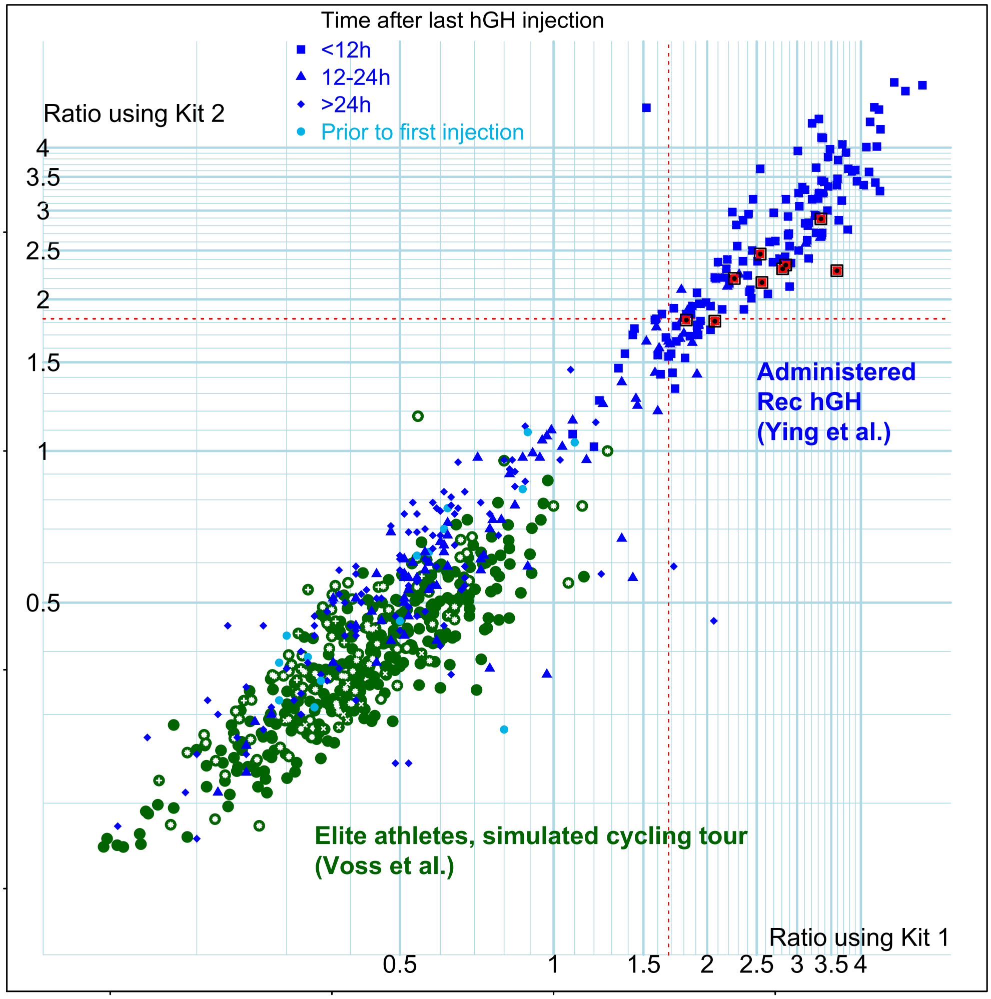

Fig. 6. rec/pit ratios from kits 1 (horizontal axis) and 2 (vertical axis) in male athletes who were administered REC hGH (athletes in Jing et al. study, blue symbols) and in male athletes who are known not to have taken REC hGH (athletes in Voss et al. study [12], green dots). Values from the Jing et al. study [11] are shown using symbols indicating how long after the last administration the sample was taken. All ratios from the Voss et al. study were included, including those shown with a special symbol that were based on low values (+: ratio1 based on value(s) < 0.05, x: ratio2 based on value(s) < 0.05) that would automatically result in a 'negative' WADA result. Also shown are corresponding values in athletes who have admitted to doping (red squares). The dotted red lines indicate the decision limits ('overall', 1.68 and 1.83 for kits 1 and 2 respectively) from the last row of Table 2.

(*l*ambda) the mean (*mu*) and the coefficient of variation (*sigma*), and the initials of the parameters give the name to the LMS method." But, rather than first fit and use a single lambda [L] in a Box–Cox transformation, and then fit a standard regression model for the mean [M] log-ratio at each concentration, with the same standard deviation [S] at each concentration, one instead fits 3 smooth 'parameter' curves, one for L, one for M, and one for S, over the range of concentrations, *i.e. each 'parameter' is a function of the concentration*. The variation described by S is assumed to have a Normal distribution.

We took two approaches to the fitting of these LMS curves. In the first (implemented in the `lms` function of the `gamlss` package for the R statistical language [14]) the method of penalized Likelihood was used to ensure that the different parameter curves (fitted as *splines*) are not too esoteric (i.e. not over-fitted); the extent of smoothing required can be expressed in terms of smoothing parameters or equivalent degrees of freedom. The number of additional degrees of freedom (d.f., how flexible the L, M and S portions of the `lms` model were allowed to be, beyond the default M spline, and linear S and L functions) was the {df.L, df.M, df.S} combination that yielded the smallest value of the Bayes Information Criterion (BIC, or the `sbc` value returned by the `lms` function). The search was over the set of models with 0 < = df.L < = df.S < = df.M < = 1. Since we have more information about the center than the spread or the shape, the M function was allowed to be more flexible, and S and L were allowed to be up to quadratic in shape. We did not wish to have the models 'chase' or be unduly influenced by unusual observations, or to be unstable at the extremes of the ordinate.

In the second, slightly simpler, approach, we were guided by the regularity of the curves seen in Fig. 1, and by the shapes of the fitted splines. Thus, we used the `gamlss` function directly to fit the M S and L curves as *polynomials*, thereby maintaining more direct control over the shapes of the fitted functions. Again, greater flexibility was allowed for the M than S and the S than L curves, with powers of 5, 3, and 1 as the upper limits, and with the final choice (displayed as three numbers on the left of each panel in Fig. 2) determined by the combination that yielded the minimum BIC. Centiles, shown in blue and red in Fig. 2, were calculated using the `centiles.pred` function applied to the polynomial fits.

A.3. Double transformations

Typically, Box–Cox transformations are applied directly to untransformed values: a 'lambda' of zero yields the log of the value, and a non-zero lambda a power transform. In the interest of greater flexibility, we first applied a log transform to all ratios (thereby treating the log ratios as the 'raw' data), and then analyzed the log-ratios with the LMS software, thereby allowing for a second transform. (We found that the direct application of the LMS transforms to the rec/pit ratios themselves 'chased' the most extreme values, and was overly sensitive).

A.4. Dealing with negative log-ratio values

Box–Cox transformations are designed for positive values, but our derived $\log_2$ ratios ranged from approximately -4 (ratio: 1/16) to 1 (ratio:2). In order to fit the LMS models, we shifted all values upwards so the minimum is at least 1, and later shifted the fitted values back. This is like multiplying all ratios by a constant so that they start at 2, and later dividing by this same constant. We tested if this mattered by shifting all log-ratios by 5, 7.5 and 10 and found that it made virtually no difference to the fitted DLs. The ones shown are based on an offset of 6.

A.5. Checks

In order to assess whether the LMS model provides a reasonable fit, we used a number of diagnostic checks. The first was the shape of distributions of the residuals, shown as a histogram in the bottom left corner of each panel in Fig. 3 (the top row shows histograms of the log-ratios before and after a simple Box–Cox transformation, but foregoing the LMS approach). As a comparison we also generated and displayed (bottom right) histograms of values from the same-sized sample but from a *known* Normal N(0,1) distribution.

In addition to visually judging how well the fitted and observed quantiles (Q_{25}, Q_{50}, Q_{75}) agree, we numerically assessed how well the fitted limit curves were calibrated at the 97.5 centile by calculating what percentage of the residuals were less than 1.96 fitted SDs above the fitted mean (there are not sufficient data to directly assess the fit of the 99th or higher percentiles). We did so both at an overall level, but also for each 20% vertical (concentration-based) slice of the data, since it is possible to have an overall pattern of residuals that looks satisfactory, without it being satisfactory all along the regressor axis. To do so, we calculated the root mean square error (RMSE) of the deviations of the 5 empirical percentages from the target of 97.5%, and compared it with the RMSE expected if the model applied to each data slice.

Conflict of interest

JH, OS, DS and JCT were asked by WADA to analyze the data and prepare independent reports. They have no commercial interest in the testing kits used to produce the data analyzed. At WADA's request, JH attended a CAS hearing in Lausanne in August 2013, but the 'McGill Report' was not addressed in the hearings.

References

[1] Z. Wu, M. Bidlingmaier, R. Dall, C.J. Strasburger, Detection of doping with human growth hormone, Lancet 353 (1999) 895.

[2] G.P. Baumann, Growth hormone isoforms, Growth Hormon. IGF Res. 19 (2009) 333–340.

[3] M. Bidlingmaier, Z. Wu, C.J. Strasburger, Test method: GH, Bailliere Clin. Endocrinol. Metab. 14 (I) (2000) 99–109.

[4] M. Bidlingmaier, J. Suhr, A. Ernst, Z. Wu, A. Keller, C.J. Strasburger, A. Bergmann, High-sensitivity chemiluminescence immunoassays for detection of growth hormone doping in sports, Clin. Chem. 55 (3) (2009) 445–453, http://dx.doi.org/10.1373/clinchem.2008.112458 (Epub 2009 Jan 23).

[5] http://www.wada-ama.org/en/World-Anti-Doping-Program/Sports-and-Anti-Doping-Organizations/International-Standards/Laboratories/ (accessed 2014.01.17).

[6] M. Bidlingmaier, Z. Wu, C.J. Strasburger, Problems with GH doping in sports, J. Endocrinol. Invest. 26 (2003) 924–931.

[7] O. Barroso, P. Schamasch, O. Rabin, Detection of GH abuse in sport: past, present and future, Growth Hormon. IGF Res. 19 (4) (2009) 369–374, http://dx.doi.org/10.1016/j.ghir.2009.04.021 (Epub 2009 May 30).

[8] World Anti-Doping Program, Guidelines: hGH isoform differential immunoassays for anti-doping analyses, June 2010 Version 1.0 http://www.wada-ama.org/Documents/Resources/Guidelines/WADA_Guidelines_hGH%20Differential%20Immunoassays_EN_June10.pdf .

[9] http://www.tas-cas.org/d2wfiles/document/6633/5048/0/256620FINAL20Award20_internet_.pdf (accessed 2014.01.17).

[10] T.J. Cole, P.J. Green, Smoothing reference centile curves: the LMS method and penalized likelihood, Stat. Med. 11 (1992) 1305–1319.

[11] J. Jing, S. Yang, X. Zhou, C. He, L. Zhang, Y. Xu, M. Xie, Y. Yan, H. Su, M. Wu, Detection of doping with rhGH: excretion study with WADA-approved kits, Drug Test. Anal. 3 (2011) 784–790.

[12] S.C. Voss, S. Giraud, M. Alsayrafi, P.C. Bourdon, Y.O. Schumacher, M. Saugy, N. Robinson, The effect of a period of intensive exercise on the isoform test to detect growth hormone doping in sports, Growth Hormon. IGF Res. 23 (2013) 105–108.

[13] D.G. Altman, J.M. Bland, Measurement in medicine: the analysis of method comparison studies, Statistician 32 (1983) 307–317.

[14] M. Stasinopoulos, B. Rigby, C. Akantziliotou, V. Voudouris, R language: version 2.15. the R foundation for statistical computing, 2012, and gamlss: the library for fitting Generalized Additive Models for Location Scale and Shape (GAMLSS models), http://www.gamlss.org/ (Repository: CRAN Built: R 2.15.1; 2012-09-30).