Growth Hormone & IGF Research 19 (2009) 361–365

Contents lists available at ScienceDirect

Growth Hormone & IGF Research

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

Statistical issues in implementing the marker method

E. Eryl Bassett [a,*], Ioulietta Erotokritou-Mulligan [b]

[a] Institute of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury CT2 7NF, UK
[b] The Institute of Developmental Sciences, University of Southampton, UK

ARTICLE INFO

Article history:
Accepted 3 April 2009
Available online 9 June 2009

Keywords:
Growth hormone
Statistical analysis
Harmonisation
Measurement uncertainty

ABSTRACT

The detection of growth hormone (GH) abuse by athletes raises statistical problems as well as biochemical ones. We outline the statistical approaches to the various issues which have arisen during the work of the GH-2000 and GH-2004 teams; in particular, it considers the need to develop a test which detects GH abuse in any elite athlete 'beyond reasonable doubt'. The test needs to be robust enough to withstand legal challenge, while minimising the risk of false accusation. The paper identifies various issues which arise in the development of such a test, and describes how these were resolved.

Since GH is a naturally occurring hormone whose concentration varies substantially, its abuse cannot be detected by direct measurement. The methodology considered here made use of markers whose levels are more stable but are influenced by GH. The statistical methods employed aimed to make the best use of these markers, taking account of all factors contributing to errors in measurement.

There were two key steps in the statistical investigation undertaken to develop the GH detection algorithm. The first was the requirement to identify GH-dependent biomarkers which would identify GH doping reliably and robustly for a significant length of time. The second was to calibrate the GH detection method in the elite athlete population, so that the method would be applicable to all athletes, regardless of age, sex and ethnicity, and regardless of whether they had recently sustained an injury.

In practice, further work was needed to ensure that the methodology met the WADA testing protocol rules, but also that the proposed method can be used by any WADA accredited lab without placing any athlete at an unfair disadvantage and ensuring a high level of confidence in any result produced.

1. Introduction

Statistical work was undertaken in connection with several aspects of the marker test for GH abuse by elite athletes. No new statistical developments were required; the challenge was to consider statistical features of the problem of identifying doping amongst athletes, and to devise appropriate responses.

In an ideal case, the statistical process to determine whether an athlete shows evidence of doping would be as follows. First, an indicator of doping is identified. A large random sample of elite athletes is then selected as a peer group. For each member of this peer group, the value of this indicator is found from assays, and an adverse finding should be recorded if the value for an athlete being tested is so far away from those of the peer group so as to leave no reasonable doubt that the athlete has been doping. If appropriate, the peer group may be subdivided (e.g. by sex) or the indicator may be adjusted to take account of other relevant information (e.g. age).

This shows that a two-step process is needed; identification of the doping indicator, and calibration using the peer group. In practice, further steps are also needed. In later sections, we therefore discuss statistical problems stemming from requirements laid down by the World Anti-Doping Agency (WADA) for two different assays to be used to measure concentrations for any given marker, problems of assay harmonisation and problems stemming from measurement uncertainty associated with assay results.

2. The main statistical processes

2.1. Identification of suitable indicators of doping

In a (statistically) ideal world, one would undertake an administration study by taking a random sample from the population (that is, the group) of all elite athletes, allocating some to placebo and the rest to the active drug. Since this would be quite unethical, the GH-2000 placebo controlled, GH administration double blind study had to use healthy volunteer recreational athletes as a proxy group for elite athletes. This enabled the GH-2000 team to identify a marker or markers capable of distinguishing between active and placebo subjects. While one can always hope that a single marker might work adequately, statistical theory indicates that a combination of two or more markers is always likely to distinguish more clearly between two groups; the technique of discriminant analysis was used [1] to distinguish between the active and placebo

* Corresponding author. Tel.: +44 1227 823798.
E-mail address: E.E.Bassett@kent.ac.uk (E.E. Bassett).

1096-6374/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ghir.2009.04.024

subjects in the GH-2000 double blind study. The main statistical issue involved concerned the choice of data for this discriminant analysis technique.

In this trial, subjects were divided into three groups; placebo, low dose GH, high dose GH, and the chosen treatment was administered daily for 28 days. Blood samples were taken on Days 0 (i.e. immediately preceding the first dose), 21, 28, 30, 33, 42 and 84, so that the treatment and washout periods were both covered. From these samples, values were obtained from eight possible markers of GH doping. Since half-lives of different markers will not all be identical, it is to be expected that different combinations of markers will discriminate best during administration and post-administration periods.

There were some apparent compliance problems during the trial; in several cases blood samples were taken on a day close to, but not exactly on, the intended day, and in some cases a target day was missed altogether. The statistical issues in these cases were resolved in a common-sense way. More seriously, some clients failed to complete the 28-day course of treatment, so that anomalous values were obtained for several markers on Day 28.

There is clearly no single statistical approach which can obtain optimum discrimination between treatment groups on all Days. Apart from one attempt [2] to model a trajectory of markers in 8-dimensional space over the various time points used, the resolution used for this statistical issue was to pick a day for which it was considered that the data were relatively stable, and to use stepwise linear discriminant analysis to identify an appropriate marker score; that is, a combination of the markers which best discriminated between the placebo and active drug groups. The form of variability in marker levels made it clear that working with the logarithms of marker concentrations would be appropriate, and logged values were used consistently.

It turned out [1] that, once markers IGF-I and P-III-P were included in the function, there was no benefit from including further markers; moreover, the discrimination between active drug and placebo groups was good [1]. However, it is possible that other markers could have been included had one concentrated on an alternative day. Even if one were convinced that an athlete being tested was guilty of GH doping, one would not know whether the athlete was currently (Day 21?) doping or whether the athlete had recently (Day 33?) entered the washout phase. It is well known [1] that different markers have different half-lives, so those markers with short half-lives will tend to add 'noise' rather than 'signal' in the washout phase, while they could be of great value in detecting those currently doping.

It was also clear that the pattern of response to GH administration differed in males and females. The result of this analysis was therefore the production of two doping indicators, EM1 for males and EF3 for females. Both were linear combinations of logarithms of IGF-I and P-III-P; in essence, what is important in both cases is the relative weighting of the two markers in the indicators. P-III-P was in both cases given slightly greater weight than IGF-I.

A consequence of the choice of data (Day 21) used in this statistical analysis is that the tests based on EM1 and EF3 are optimised for use in athletes in an ON period (currently doping with GH). Since IGF-I has a shorter half-life than P-III-P, increasing the weight given to the latter would be expected to give greater sensitivity for testing athletes in the OFF period (recently stopped doping).

Another statistical approach which could be used would be to examine the values of these two markers separately rather than in combination. For example, one could consider judging a current doper using IGF-I alone, and a recent doper using P-III-P alone. While this might seem statistically acceptable, the method is impractical since the markers (especially P-III-P) increase in response to an injury, so could be open to misinterpretation. However, in principle one might be able to use two marker combinations, using different weightings of IGF-I and P-III-P to reflect the need to cover the ON and OFF periods in different ways.

2.2. Calibration of doping indicators

Levels of the chosen markers are known [3] to vary with age; in the case of IGF-I and P-III-P they decrease with age after peaking around puberty. Since one needs to ensure that no athlete has an unacceptable risk of a false positive, there are only two ways of calculating cut-off points beyond which an adverse finding should be recorded. One can take the 'worst case' scenario; that is, use the minimum age in the sample, and calculate the cut-off point suitable for that, accepting that older athletes will have an even smaller risk of a false positive finding. Alternatively, and much more satisfactorily, one can adjust the marker score for age. It has been found [3] that, for these two markers, a model in which the log of the marker level decreased linearly with the reciprocal of age fitted the data on elite athlete marker levels well, over the range of ages studied. Accordingly, age-adjusted versions of EM1 and EF3 have been devised [2]; these (once standardised – see below) are what have been termed GH-2000 marker scores: EM1b for males, EF3b for females.

To declare an adverse finding, the marker score for a tested athlete must be compared with scores for his or her peers. The cross-sectional study undertaken as part of the GH-2000 project was used to determine the distributions of male and female marker scores. In (statistical) principle, one can choose to assess how consistent a score is with peer group scores either in a parametric or in a non-parametric way. However, for legal purposes one needs to establish that an athlete's score is so extreme that it can be viewed as inconsistent with the natural range of scores, and so indicates doping 'beyond reasonable doubt'. The probability level associated with such doubt is a matter for legal argument rather than statistical assessment, but we have used 0.0001 (=1/10,000) as a value which will be in the right region.

To use a non-parametric argument is impractical for such an extreme value, since one needs data on a random sample of substantially more than 10 000. One must therefore work with a parametric argument; this requires one to fit a distribution to the sample values, and then to estimate the 99.99% point of the distribution. There are two statistical issues associated with this process; while one might hope that some tractable distribution fits the data well, one can never formally conclude scientifically that the data do in fact follow that distribution; all one can say is that the data are consistent with that distribution. The second problem is that estimating a percentage point requires one to use values for the various parameters of the selected distribution; since these are only estimates, subject to random errors, it follows that the same will hold for any estimate of the 99.99% point.

A statistical issue which in the event caused no difficulty, but which needed to be borne in mind, is the handling of assay results viewed by the lab as lying below the limit of detection (LOD). We have heard arguments that these should be excluded from the database, on the grounds that the values are unreliable. However, omitting values of this type from a database will bias the results, since there is relevant information contained in them; namely, that the value is small. (In statistical language, such data are viewed as censored, rather than omitted.) The presence of individuals in the elite athlete database with low values for IGF-I or P-III-P causes no practical problem, as long as suitable methods are used to assess these values; it is only high values for the markers which would cause an adverse finding to be drawn. However, the problem would be much more awkward for tests for which an adverse finding could stem from low assay values.

For the age-adjusted versions of EM1 and EF3, the distributions (over the sample of elite athletes used in the GH-2000 study) both