Institution: University of Southampton		
Unit of Assessment: 10 Mathematical Sciences		
Title of case study: 10-06 Statistical Methodology for Growth Hormone Misuse in Competing Athletes		
Period when the underpinning research was undertaken: 2013 – 2018		
Details of staff conducting the underpinning research from the submitting unit:		
Name(s):	**Role(s) (e.g. job title):**	**Period(s) employed by submitting HEI:**
Dankmar Böhning Wei Liu	Professor in Medical Statistics Professor in Statistics	October 2011 – present September 1990 – present
Period when the claimed impact occurred: July 2015 – July 2020		
Is this case study continued from a case study submitted in 2014? N		

1. Summary of the impact

Statistical methodologies developed at the University of Southampton and led by Professor Dankmar Böhning have refined a diagnostic biomarker test to detect growth hormone (GH) misuse. Improvements to the statistical methodology of the original test, particularly for age and gender adjustments, have sustained the confidence and application by the World Anti-Doping Agency (WADA). The improved test is one of only two tests approved by WADA to measure GH misuse and is used in all major sporting events that fall under the governance of the International Athletics Association, from the Olympics to the World Athletics Championships.

The test has provided continued reassurance of fairness in sport by ensuring the statistical methodology meets the challenging WADA requirement of a 1 in 10,000 false-positive rate. Since the test was introduced in 2015, no positive tests (laboratory adverse analytical findings) have occurred, despite it being applied on more than 30,000 occasions worldwide in 31 WADA-accredited laboratories in 103 sporting disciplines. The refined GH test has made a significant contribution towards the goal of GH-doping-free sport.

2. Underpinning research

“Fairness” in competitive sports is a widely accepted paradigm. The International Fair Play Committee defines fairness in the context of the misuse of drugs: “Fair play means not cheating by taking drugs or doping. Anyone who does this ruins the game for everyone else.” Athletes competing with their naturally given resources is widely accepted, whereas the application of stimulating substances to gain advantages in competitions is considered as unfair. This research concentrates on growth hormone (GH), which can have advantageous effects in competition if it occurs in elevated levels. The scientific challenge is to separate naturally occurring GH from external injection (doping).

The current GH test, **GH-2000**, was developed 1996-1999 in a multinational project led by Professor Peter Sönksen at St Thomas' Hospital, London. Sönksen became a visiting Professor at Southampton’s Faculty of Medicine in 2002 and established a further project (GH-2004) with Professor Richard Holt (Medicine) to validate the GH-2000 test and achieve robustness against the effects of injury and ethnicity on test results. Subsequently, GH-2000 was first used at the 2012 London Olympic and Paralympic Games, increasing the window of GH detection after administration from around 12-24 hours to several weeks. Two Russian powerlifters returned a positive test and were disqualified, having never been caught by the previous GH test.

The GH-2000 score combines two assays of GH-sensitive biomarkers as well as a correction term for age and is given by the generic formula: $GH\text{-}2000\ Score = b0 + b1\ log\ IGF1 + b2\ log\ PIIINP + b3/age$, where $b0$, $b1$, $b2$, $b3$ are gender-specific constants, and $IGF1$ and $PIIINP$ are insulin-like growth factor-I (IGF-I) and the amino-terminal pro-peptide of type III collagen (P-III-NP). Each of these are measured by assays and currently, for each biomarker, three different assays are available. In consequence, a variety of GH scores exist for which decision limits need to be determined; these are set by WADA to maintain a **1 in 10,000 false positive rate** which,

while seemingly high, is seen as a robust and appropriate level of reliability in situations where legal conflicts with athletes under suspicion of GH misuse arise.

For the GH-2000 test used in the 2012 Games, and for the existing GH test, the decision limits had been developed using a relatively small dataset of 352 serum samples. A successful 2013 appeal to the Court of Arbitration in Sport against a positive result to the existing GH test indicated that the numbers of samples used to develop the test's decision limits should be increased. In a project supported by Partnership for Clean Competition (PCC) [**G1**], Professor Dankmar Böhning of Southampton's School of Mathematical Sciences joined Sönksen and Holt in 2013 to lead the statistical modelling and inference to develop decision limits for the GH-2000 detection methodology based on a much larger dataset.

Serum samples were collected for anti-doping purposes from 859 female athletes and 971 male athletes representing 200 countries and competing in 26 different track and field events at the 2011 World Athletics Championships in Daegu, South Korea. For his new dataset, GH-2000 scores were calculated separately for male and female athletes using combinations of 2 different assays for PIIINP and 3 different assays for IGF1; for all these 6 GH scores new decision limits were developed by Böhning and his group [**3.1**], which were then adopted by WADA.

With the collection and establishment of an increasingly large database it became apparent that the original score showed a positive age effect in the male athlete population, which could potentially place older male athletes at a disadvantage. Böhning led the development of novel statistical methods to refine the tests that included a much-needed adjustment for age [**3.2, 3.3**].

In addition, with Professor Wei Liu, Böhning developed exact decision limits in the following sense. Previously, decision limits were developed using an approximation by the normal distribution. In this work decision limits were developed using results from distribution theory which allowed use of the exact t-distribution, thus avoiding the normal approximation. As a result the decision limits now meet precisely the 1 in 10,000 false-positive rate required by WADA [**3.4**].

In 2017, with the support of WADA [**G2**], these decision limits were further investigated on the basis of routinely collected growth hormone data from competing athletes by the world-wide WADA accredited laboratories. The motivation behind this was to see if the decision limits could be confirmed if the biomarker test is applied routinely. Here, the motivation was to see if the decision limits remain stable if recalculated on a big data set. In addition, the intention was to make them more robust and less likely to be challenged legally because of increased numbers of athletes and broader diversity of sporting disciplines. In essence, the previously developed decision limits are currently being reconfirmed with ongoing scientific investigations with the expectation of them being implemented in 2021 [**3.5**].

3. References to the research

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

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

3.3 D. Böhning, W. Böhning, N. Guha, D.A. Cowan, C. Bartlett, P.H. Sönksen, and R.I.G. Holt, (2018). A Correction to the Age-Adjustment of the GH-2000 Score used in the Detection of Growth Hormone Misuse. *BMC Research Notes* 11: 650. https://bmcresnotes.biomedcentral.com/articles/10.1186/s13104-018-3741-7

3.4 D. Böhning, **W. Liu**, R.I.G. Holt, W. Böhning, N. Guha, P. Sönksen, D. Cowan, and T. Liang (2017). Exact statistical calculation of the uncertainty term in the decision limits of the GH2000-score for growth hormone misuse detection (doping). *Statistical Methods in Medical Research*. https://doi.org/10.1177%2F0962280217739452

3.5 R.I. Holt, W. Böhning, N. Guha, C. Bartlett, D.A. Cowan, P.H. Sönksen, and **D. Böhning** (2018). Analysis of doping control hGH biomarker normative data for review of test decision limits. Report on WADA grant *Recalculation of the decision limits on the basis of world-wide routinely collected GH data of competing athletes*. Available on request.

Grants

G1 Partnership for Clean Competition (PCC); GH-2004: Novel biomarkers for the detection of IGF-I abuse; PI Professor Holt; Mar 2011 – Feb 17; £182,056

G2 WADA; Recalculation of the decision limits on the basis of world-wide routinely collected GH data of competing athletes; PI Professor Holt; Jan 2017 – August 2017; £7,651

4. Details of the impact

"Diagnosing growth hormone misuse is an important issue in competitive sports. For this we need reliable and valid decision limits which identify growth hormone dopers beyond reasonable doubt. These decision limits cannot be provided without solid statistical work. In recent years, Professor Böhning and his statistical colleagues in Southampton have provided strong and excellent contributions to the development of statistical methodology in the construction of the decision limits for growth hormone misuse… We are very happy to be able to cooperate with Professor Böhning and his colleagues." – Scientific Director, **World Anti-Doping Agency (WADA)** [**5.1**]

The new decision limits based on Böhning's analysis of the increased dataset were first incorporated into WADA guidelines in July 2015 [**5.2**]. Since then, according to WADA's 2019 annual report [**5.3**], the refined GH-2000 test has been used on 30,716 occasions. An annual breakdown is given in Figure 1, demonstrating strong growth in the years of the 2016 Summer and 2018 Winter Olympic and Paralympic Games, with a peak of 8,755 tests in 2018. The test has been carried out in 31 WADA-accredited Laboratories including Ankara (Turkey), Barcelona (Spain), Dresden (Germany), Ghent (Belgium), Lausanne (Switzerland), London (UK), Los Angeles (USA), Montreal (Canada), Oslo (Norway), Paris (France), Rio de Janeiro (Brazil), Rome (Italy), Seibersdorf (Austria), Stockholm (Sweden), Sydney (Australia), Salt Lake City (USA) and Warsaw (Poland). [**5.4**]

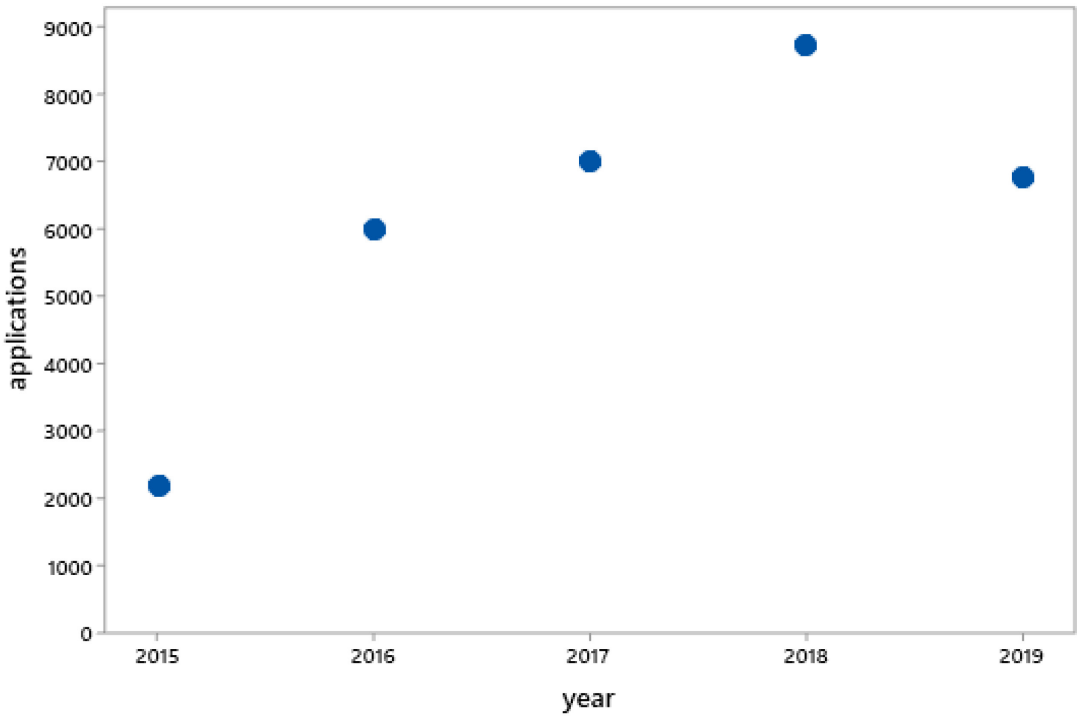

Figure 1: Applications of the biomarker test by year

Figure 2 shows the growth in the number of sporting disciplines involved, reaching a peak of 103 in 2019. Clearly there is no limitation of which discipline can be involved other than whether it makes sense to do so (it would make no sense to apply the test to snooker players, for example).

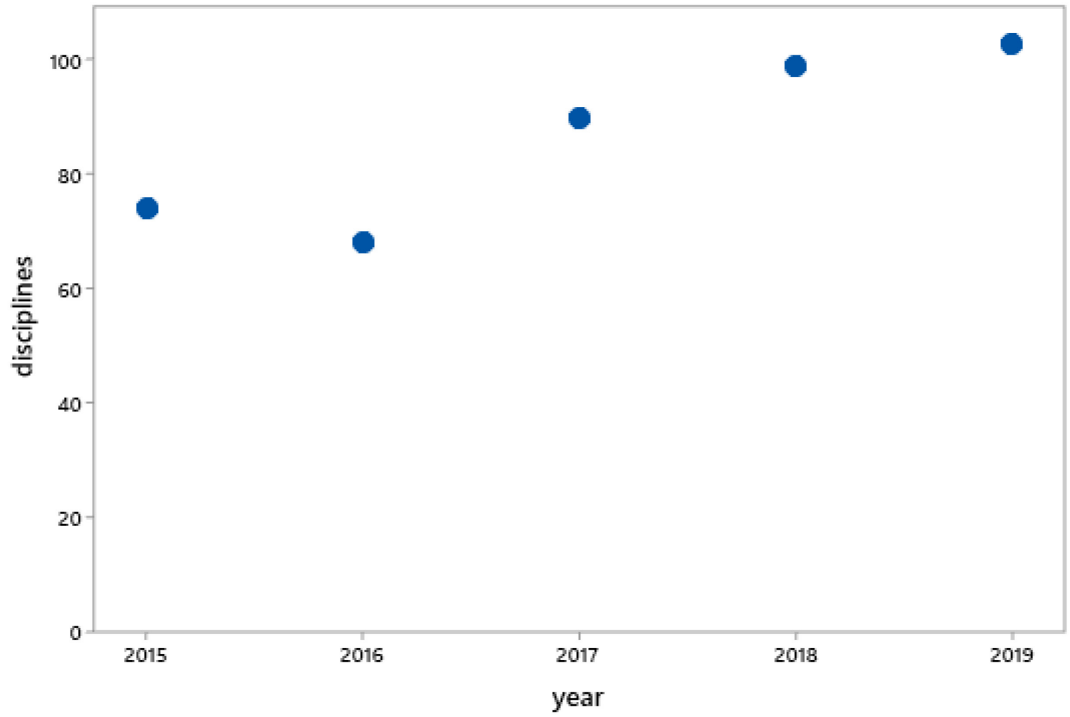

Figure 2: Number of sport disciplines involved in the application of the biomarker test

Since the new decision limits were introduced in 2015 there have been no positive test results. This is a clear indication of the influential effect of the test in preventing GH-doping. Additionally, the statistical improvements to counteract the positive age effect in males was introduced in 2016, a year of strong growth in use of the test, demonstrating further the effectiveness of Böhning's work.

These refinements have helped to sustain the high level of reliability of this test and its continued use in all world sports subject to the WADA code. Furthermore, the statistical methodology can be applied to other tests, not just GH-2000.

The Partnership for Clean Competition in the USA can further attest to the crucial nature of the work at Southampton. This non-profit organisation was founded in 2008 with a mission to fund scientific research globally which advances knowledge in the anti-doping sphere and contributes meaningfully to the protection of clean athletes. Its Executive Director stated: "In recent years, Professor Böhning and his statistical colleagues based at the University of Southampton have provided further refinements to the development of statistical methodology in the construction of the decision limits for growth hormone misuse. There have been modifications that have improved the accuracy of this test in relation to gender differences and help to sustain the 1 in 10000 false positive rate required by WADA. Reliable, valid, and scientifically defendable male and female decision limits accurately identify growth hormone dopers to prevent false positives and false negatives in this test. The work of Professor Böhning and colleagues at Southampton have been of the utmost importance to identify the appropriate statistical methods in growth hormone misuse detection." [**5.5**]

The Head of Science and Medicine at **UK Anti-Doping**, the government-sponsored organisation responsible for protecting sport in the UK from doping, stated: "Professor Böhning's contribution might not be as visible (or possibly as recognised) as the bioanalytical research that underpins the test, but the statistical refinements made in recent years have been just as critical to its continued use in a sector highly vulnerable to legal challenge and reputational damage." [**5.6**]

5. Sources to corroborate the impact

5.1 Letter from Scientific Director, WADA, 28 May 2020

5.2 WADA Guidelines: Human Growth Hormone Biomarkers Test for Doping Control Analyses v1.0 July 2015, with [**3.1**] cited as reference [25] with the new decision limits on p11: https://www.wada-ama.org/sites/default/files/resources/files/wada-guidelines-for-hgh-biomarkers-test-v1.0-2015-en.pdf

5.3 WADA 2019 Annual Report, Towards a World of Clean Sport https://www.wada-ama.org/sites/default/files/resources/files/ar2019_single_08102020_digital.pdf

5.4 List of WADA accredited laboratories. https://www.wada-ama.org/en/what-we-do/science-medical/laboratories/accredited-laboratories

5.5 Letter from Executive Director, Partnership for Clean Competition (PCC), 8 July 2020

5.6 Letter from Head of Science and Medicine, UK Anti-Doping, 21 July 2020.