The International Sports Law Journal
https://doi.org/10.1007/s40318-019-00143-w

ARTICLE

Scientific integrity and the IAAF testosterone regulations

Roger Pielke Jr.[1] · **Ross Tucker**[2] · **Erik Boye**[3]

Abstract
In April 2018, the International Association of Athletics Federations (IAAF) announced new regulations governing the eligibility of certain female athletes with differences of sexual development accompanied by elevated levels of natural testosterone. Such women with testosterone levels above a specific threshold would be banned from competing as females unless they were to undergo medical intervention. In this paper, we examine key elements of the scientific basis offered by IAAF in support of the regulations, based on a subset of original performance data provided to us by IAAF. We identify significant flaws in the data used by IAAF leading to unreliable results. Further, these failures have not been corrected by IAAF or the academic journal which has published them, leading to a comprehensive failure of scientific integrity. We argue that the IAAF testosterone regulations are based on a flawed scientific foundation and that this case offers more general lessons for the sport governance community on the importance of upholding the standards of scientific integrity expected in other areas of policy and regulation.

Keywords IAAF · CAS · Testosterone · Caster Semenya · Dutee Chand · Scientific integrity

1 Introduction

In April 2018, the International Association of Athletics Federations (IAAF) announced new regulations governing the eligibility of certain female athletes.[1] The regulations are the latest incarnation of "sex testing" in international athletics, an issue that the sport has struggled with for more than a half century (Pielke 2017). In this instance, the IAAF notes explicitly that the 2018 regulations are not intended to police a line between male and female, but to identify and regulate a certain class of females whose participation in competition would be allowed only through medical intervention.

The latest IAAF regulations focus on females with "differences of sexual development" (DSDs) whose natural level of testosterone is elevated. Females with elevated natural testosterone levels that are not associated with DSDs are not covered by the regulations. Further, the regulations apply only to five events: 400 m, 400 m hurdles, 800 m, 1500 m and 1 mile.[2] In order to participate in these events in international IAAF-sanctioned competitions, women who meet the regulatory criteria are required to medically reduce their naturally testosterone levels to below a designated threshold, set at 5 nmol per litre (nmol/L).

The 2018 regulations represent a second attempt by the IAAF to regulate the female competition classification using testosterone levels in female athletes.[3] A first iteration of the regulations was proposed in 2011 following the emergence of South African Caster Semenya as a top international athlete in 2009 and an initial suspension and then reinstatement of Semenya by the IAAF (Pielke 2017). The 2011 testosterone regulations were suspended in 2015 by the Court of Arbitration for Sport (CAS) as a result of a challenge brought by Indian sprinter Dutee Chand, who had been suspended from competing under the regulations.[4]

✉ Roger Pielke Jr.
pielke@colorado.edu

[1] University of Colorado Boulder, Boulder, USA

[2] University of Cape Town, Cape Town, South Africa

[3] Oslo University Hospital, Oslo, Norway

[1] https://www.iaaf.org/news/press-release/eligibility-regulations-for-female-classifica.

[2] And all other events over distances between 400 m and 1 mile.

[3] The first attempt to regulate female classification emphasized hyperandrogenism, whereas the second attempt focuses on "differences of sexual development". In both instances, the emphasis is on regulating female eligibility based on naturally occurring testosterone levels.

[4] CAS operates through arbitration panels, typically comprised of three arbitrators. In this paper, when referring to CAS judgements, these refer to arbitration judgements made by such panels.

Published online: 07 February 2019

In this paper, we provide an independent critique of scientific research conducted by IAAF and subsequently presented as underpinning its 2018 regulations. We find this research to be deeply flawed and uncorrected even after the errors were called to the attention of IAAF and the scientific journal which published them. This situation raises important questions of scientific integrity. The IAAF research has also been critiqued by a number of researchers (e.g. Camporesi 2018; Karkazis and Carpenter 2018; Menier 2018; Sőnksen et al. 2018). We add to these critiques in this analysis of IAAF research based on our unique (partial) access to original IAAF research data.

We proceed in three parts. First, we review the 2015 judgement by CAS in the Chand arbitration, which established a clear role for the use of scientific evidence in any future IAAF testosterone regulations. Second, we evaluate a key part of the scientific evidence offered by the IAAF in support of the new testosterone regulations. Finally, we conclude with more general recommendations for how sports organizations and the sport science community can work to better ensure scientific integrity.

We define "scientific integrity" consistent with Douglas and Bour (2014) to consist "of proper reasoning processes and handling of evidence essential to doing science" and "a respect for the underlying empirical basis of science". It is uncontroversial that policy and regulatory decision making (whether in a sport context or other) should be grounded in evidence produced with scientific integrity. We find that this standard has not been met in this case.

2 The 2015 CAS Chand judgement and IAAF response

In the CAS Chand arbitration judgment of 24 July 2015 that suspended the first iteration of the IAAF testosterone regulations, discussion of the scientific basis for the regulations was important in the decision.[5] The lack of a relevant scientific basis for the regulations was a crucial factor in that panel's decision to suspend the regulations. Below, we distil the panel's discussion of the role of evidence in justifying the regulations. We also summarize the panel's conclusion on what evidence would be necessary to uphold such regulations, setting the stage for the second iteration of the regulations in 2018.

The panel (at paragraph 519) asserted that,

> Once an athlete is legally recognised as female, the Panel considers that an athlete must be permitted to compete in the female category unless her naturally high androgen levels confer a significant performance advantage over other female competitors, comparable to the performance advantage that male athletes enjoy over female athletes.

At 520, the panel is explicit about what was meant by "performance", asking the IAAF to demonstrate a correlation between elevated testosterone levels subject to regulation and "a real competitive advantage". We interpret "real competitive advantage" to mean an advantage observed in actual elite competition among females, and not based on speculation, physical characteristics of female athletes or a comparison of male and female anatomy.

The CAS panel (at 527) concluded that certain scientific evidence in the context of actual elite athletic competition was important to adjudicating this issue (**emphasis** added by us):

> The Panel considers the lack of evidence regarding the quantitative relationship between enhanced levels of endogenous testosterone and enhanced athletic performance to be an important issue. While a 10% difference in athletic performance certainly justifies having separate male and female categories, a 1% difference may not justify a separation between athletes in the female category, given the many other relevant variables that also legitimately affect athletic performance. **The numbers therefore matter.**

The CAS panel (at 528) explained (emphasis in original):

> However, in order to justify excluding an individual from competing in a particular category on the basis of a naturally occurring characteristic such as endogenous testosterone, it is not enough simply to establish that the characteristic has some performance enhancing effect. Instead, the IAAF needs to establish that the characteristic in question confers such a **significant performance** advantage over other members of the category that allowing individuals with that characteristic to compete would subvert the very basis for having the separate category and thereby prevent a level playing field. The **degree or magnitude of the advantage is therefore critical.**

The CAS panel (at 531) concluded that the IAAF had a burden to provide this evidence in support of its regulations:

> In the context of this issue, the onus lies on the IAAF. The IAAF has not established, on the balance of probabilities, that the Hyperandrogenism Regulations apply only to exclude female athletes that are shown to have

[5] CAS 2014/A/3759 Dutee Chand v. Athletics Federation of India (AFI) and The International Association of Athletics Federations (IAAF) http://www.tas-cas.org/fileadmin/user_upload/award_internet.pdf.

> a competitive advantage of the same order as that of a male athlete.

At that time, IAAF was unable to produce the evidence deemed by the CAS panel to be necessary in defending the regulations. The CAS panel (at 548) thus suspended the regulations pending the production of new evidence:

> In these circumstances, the Panel is unable to uphold the validity of the Regulations. The Panel therefore suspends the Hyperandrogenism Regulations for a period of two years, subject to the following provisos. At any time during that two-year period, the IAAF may submit further written evidence to the CAS concerning the magnitude of the performance advantage that hyperandrogenic females enjoy over other females as a result of their abnormally high androgen levels.

Chand was consequently deemed eligible to run, as were other athletes who may have fallen under the 2011 testosterone regulations. The result of the CAS judgement was to give the IAAF a deadline of July 2017 to produce the evidence that the panel had asserted would be necessary to uphold such regulations.

In July 2017, the *British Journal of Sports Medicine* (BJSM) published Bermon and Garnier (2017, hereafter BG17), a study of the relationship of testosterone levels and athletic performances at the 2011 and 2013 IAAF World Championships. When the new IAAF released the new regulations in April 2018, it explicitly stated that the only new published evidence related to the relationship between testosterone and actual performance among elite female athletes was BG17. In context, the BG17 research clearly represents the IAAF's response to the mandate to provide empirical data on the relationship of elite female athlete performance and testosterone that it was given by CAS in the Chand decision. That is, BG17 provides the evidence requested by CAS for the quantification of a performance advantage enjoyed by women with elevated levels of testosterone. The IAAF's 2018 regulations cite BG17 as crucial evidence for performance advantages, stating:

> There is a broad medical and scientific consensus, supported by peer-reviewed data and evidence from the field, that the high levels of endogenous testosterone circulating in athletes with certain DSDs can significantly enhance their sporting performance.

The claim here of "peer-reviewed data" is not correct, as IAAF has refused to release the performance data associated with the cited study to other researchers or even to the journal which published BG17. With the exception of our analysis reported below, based on a subset of the data shared with us, the data have not been reviewed by peers and were not at the time of this claim by the IAAF. Nonetheless, "peer-reviewed data and evidence from the field" is referenced to the following footnote in the regulations:

> Peer-reviewed data from the IAAF World Championships in Daegu (2011) and Moscow (2013) indicate that women in the highest tertile (top 33%) of testosterone levels performed significantly better than women in the bottom tertile (bottom 33%) in the following events: 400 m hurdles (top tertile, with mean T concentration of 1.94 nmol/L, outperformed bottom tertile, with mean T concentration of 0.43 nmol/L, by 3.13%; 400 m (top tertile, with mean T concentration of 7.39 nmol/L, outperformed bottom tertile, with mean T concentration of 0.40 nmol/L, by 1.50%; and 800 m (top tertile, with mean T concentration of 3.28 nmol/L, outperformed bottom tertile, with mean T concentration of 0.39 nmol/L, by 1.60%): Bermon and Garnier (2017), Serum androgen levels and their relation to performance in track and field: mass spectrometry results from 2127 observations in male and female elite athletes, Br J Sports Med 2017;0:1-7, additional material at http://bjsm.bmj.com/content/51/17/1309.

This same footnote appears in the "Explanatory Notes/Q&A" that accompany the regulations.[6] The numbers in these footnotes are from BG17, and the hyperlink in the footnote is to BG17. It is also clear from the regulations that BG17 was the basis for the application of the regulations to events of distances of 400 m to 1 mile. There can be no doubt that BG17 forms the empirical basis for the 2018 IAAF testosterone regulations and is the primary response by IAAF to that CAS ruled must be provided by the IAAF in the 2015 Dutee Chand ruling.

3 Flawed data and unreliable results in BG17

On 30 April 2018, we requested the performance data reported in BG17 from the first author (Dr. S. Bermon) and the editor of BJSM. We requested only aggregate performance data and not any linked medical data that would raise privacy concerns. We made this request after our inability to reproduce sample numbers, means and standard deviations as presented in their Table 3, which was based on performance data publicly available from the events that they analysed.

We consider independent replication of the results of BG17 to be a major concern, not only because the paper formed an important basis for the 2018 regulations, but also

[6] https://www.iaaf.org/news/press-release/eligibility-regulations-for-female-classifica.

Table 1 Replication of summary statistics for women's running events from Bermon and Garnier (2017) for women's track events

	Summary statistics from Table 3 from Bermon and Garnier (2017)			Our replication based on provided data		
	N	Average	SD	*N*	Average	SD
100 m	112	11.88	0.88	112	11.88	0.88
100 m H	73	13.15	0.48	73	13.15	0.48
200 m	71	23.43	0.9	71	24.43	0.90
400 m	67	52.32	2.56	67	*52.19*	*2.59*
400 m H	67	56.34	2.65	67	*56.30*	*2.59*
800 m	64	121.8	5.42	64	121.80	5.42
1500 m	66	250.16	6.42	66	*250.15*	6.42
3000 m SC	56	581.61	17.39	56	581.61	17.39
5000 m	40	932.67	39.73	40	932.67	39.73
10,000 m	33	1912.6	55.6	33	*1912.63*	*55.50*
Marathon	92	9726.6	790.9	96	*9726.63*	*790.87*

Small differences in replication emphasized in italics

N number of observations, *SD* standard deviation, *H* hurdles, *SC* steeplechase

because the paper was produced in-house by IAAF researchers. As BG17 is both funded and conducted by IAAF in support of its own regulations, it is appropriate that independent scholars replicate or examine their work. It is uncommon and unadvisable that IAAF sees its role as serving as both the regulatory body and the primary producer of evidence justifying its own proposed regulations.

On 6 July 2018, more than 2 months after our initial request to see the original data, we and BJSM received from Dr. Bermon a subset of the data of BG17, specifically for the 11 women's running events reported in their Table 3. Unknown to us at this time, and not mentioned to us by either Dr. Bermon or BJSM, on 7 July 2018, BJSM published Bermon et al. (2018, BHKE18), which included the acknowledgment of methodological changes that had resulted in changes to sample sizes and calculated performance differences compared to the original 2017 study. Below, we document the unreliable data and findings of BG17, both of which are confirmed by reported results of BHKE18.

3.1 Flawed data

Upon receiving 25% of the original data from BG17, we first undertook two tasks:

(A) replication of the overall summary statistics found in Table 3 of BG17 and
(B) recreation of the underlying dataset based on reported times from the 2011 (Daegu) and 2013 (Moscow) World Championships (via Wikipedia).

With respect to replication (A), Table 1 shows that we were able to successfully reproduce the summary statistics with only small differences (emphasized). This replication confirms that the data that we were provided by Dr. Bermon were in fact identical (or nearly so, we cannot explain the small differences between reported results and our replication) to that which was used in BG17.

With respect to recreation (B), we found significant anomalies and errors in the underlying data for the four events for which we recreated the data set by cross-checking times provided by Dr. Bermon with publicly available results from the 2011 and 2013 World Championships. We recreated the data for four events (women's 400 m, 400 m H, 800 m and 1500 m) because they are central to the new regulations promulgated by the IAAF.

We have identified three types of anomalies/errors, in addition to the inclusion of times (for several events) for athletes who have been disqualified by IAAF for doping. These are:

- *Duplicated athletes* More than one time is included for an individual. In each of these instances, more than one time from the 2011 and 2013 World Championships is included for the same athlete.
- *Duplicated times* The same time is repeated once or more for an individual athlete, which is clearly a data error.
- *Phantom times* No athlete could be found with the reported time for the event.

Table 2 provides a summary of the problematic data points for the four events.

Problematic data make up between 17 and 33% of the values used in the BG17 analysis for these four events. Given the pervasiveness of these errors, we consider it likely that similar problems might be found in the data for the other 17 women's events and 22 men's events, and perhaps also in the anonymous medical data, which are the basis for the

Table 2 Recreation of data of BG17 for four events, summarizing total problematic data points identified

Event	Original data points	Duplicated athletes	Athletes included who were DQ'ed for doping	Duplicated times	Phantom times	Total problematic data points	Per cent of total (%)
400 m	67	6	0	5	11	22	32.8
400 m H	67	6	0	12	1	19	28.4
800 m	64	8	3	0	0	11	17.2
1500 m	66	10	2	0	3	15	22.7

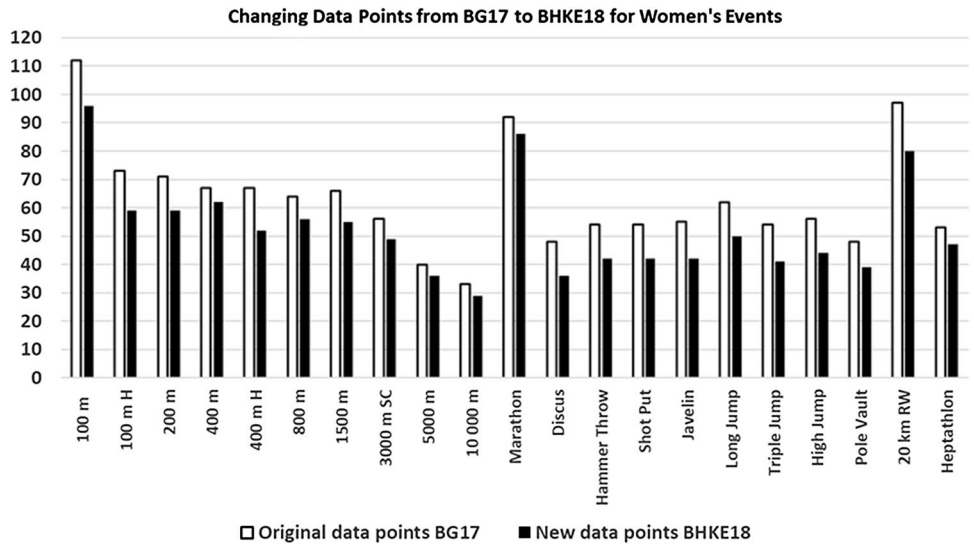

Fig. 1 Total number of data points in BG17 and BHKE18, based on observations reported by event in each paper. The differences between the open bars and black bars represent dropped data points

Table 3 Comparison of numbers of data points in two published reports, provided by Bermon, after correction for errors, with our recreation of data from these events

	BG17	BHKE18	Our analysis
Data points for four women's events			
400 m	67	62	45
400 m H	67	52	48
800 m	64	56	53
1500 m	66	55	51

study's main conclusions regarding the performance effects of elevated testosterone levels. Such pervasive errors in the four regulated events for which we carefully recreated data call into question the fidelity of the entire analysis.

When sharing the partial data, Dr. Bermon notified us that the dataset contained "some errors". This was further confirmed with the publication of BHKE18, which stated: "We have excluded 230 observations, corrected some data capture errors and performed the modified analysis on a population of 1102 female athletes". A comparison of reported observations in BG17 and BHKE18 indicates that only 220 observations were dropped from one study to the next, and thus, BHKE18 erred in its reporting of excluded observations. Figure 1 shows that the dropped data points can be found in every event.

Table 3 shows the number of observations in BG17, the number of observations after data points were dropped in BHKE18, and those in our recreation of the BG17 dataset after identifying erroneous or anomalous data points, for the four running events at the focus of this analysis. The table indicates that there are remaining uncorrected data errors in BHKE18.

The presence of unreliable data in BG17 is unambiguous: we have documented it empirically, the lead author has admitted to the presence of errors, and a subsequent analysis has sought to redo the study after dropping 220 observations, explicitly acknowledging "errors". Further, it appears that some amount of unreliable data persists in BHKE18, since the data provided to us do not match that used in the updated BHKE18 paper. We next show that the unreliable data lead to unreliable results with respect to performance differentiated by testosterone levels.

3.2 Unreliable results

After recreating the data for four events, we next conducted two additional tasks:

Table 4 Performance changes for all athletes using original data (provided to us by BG17) and corrected data (based on our recreation)

EVENT	Original data points	Corrected data points	Original mean	Corrected mean	Replicated SD	Corrected SD
400 m	67	45	52.19	52.85	2.59	2.94
400 m H	67	48	56.30	56.61	2.59	2.97
800 m	64	53	121.80	122.03	5.42	5.76
1500 m	66	51	250.15	245.96	6.42	7.16

(C) We compared means and standard deviations for performance results for all athletes across the four regulated events based on original results reported by BG17 and our calculations after dropping problematic data points;
(D) We compared the performance results reported by testosterone tertile in BG17 and BHKE18 to assess the impact of the 220 dropped data points on results reported in each analysis.

Table 4 compares the sample numbers, means and standard deviations of the original data, with the corrected data once we had removed all erroneous data points, as described above (Table 2). The comparison suggests that the problematic data underpinning BG17 are significant and consequential for the results reported for all events. Table 4 reveals that all three outcomes change for all four events upon the elimination of the problematic data points. The changes in aggregate times when using the corrected data are of a similar magnitude to that of the effects of testosterone that the authors seek to identify. Such consequential data errors confound identification of the effects that the analysis seeks to quantify.

Since we do not have access to the linked medical data for each performance, we cannot know what impact the problematic data may have had on the BG17 conclusions regarding testosterone's influence on performance. However, the large magnitude of performance changes on all athletes when correcting these problematic data points significantly undermines any conclusions that can be drawn even before the medical data are linked to the new performance set.

In a final analytical step, we compare the results reported in BHKE18 with those of BG17 to assess the impact of dropping 220 observations. Note that this analysis compares what was published initially to that published in the updated analysis, notwithstanding that (a) BG17 uses data that contain errors, (b) BHKE18 dropped 220 data points identified as problematic, yet likely still has errors or data points that disagree with what is publicly available for the competitions in question, and (c) aside from the partial BG17 data that were provided to us, the balance of BG17 data and the BHKE18 data has not been made available to peers, and as a letter to BJSM, nor was BHKE18 peer-reviewed prior to publication.

BG17 and BHKE18 both compare performances by athletes in the top and bottom tertiles of testosterone levels across different events. It is this difference which is argued by IAAF as the basis for regulation of certain events for female athletes. In comparing BG17 to BHKE18, the reported differences between these tertiles for all women's running events changed dramatically, as shown in Table 5. The size of changes in results from BG17 to BHKE18 is similar to the magnitude of effect being investigated. Figure 2 shows the differences in results between the two studies, by event.

We highlight some important differences between results reported in the two studies:

- For 8 of 11 running events, the performance difference between the highest and lowest tertiles decreased from BG17 to BHKE18, including in 3 of 4 of the regulated events;
- In three events, the performance difference changed from positive (high T faster than low T) to negative (high T slower than low T);
- In BHKE18, the low T tertile is faster than the high T tertile in 6 of 11 events, compared to 3 of 11 events in BG17.
- In the four regulated events, the average difference in times was reduced by 0.4% in absolute terms (i.e. from 2.0 to 1.6%), and only 1 of 4 meets the BHKE18 standard for statistical significance (BG17 reported 3 of 4).
- In the 100 m sprint, the advantage of low T athletes over high T athletes (5.4% and 3.4%, respectively) is larger than any of the advantages observed within each study of high T over low T across all events.

Clearly and unambiguously, the results reported in BG17 change quantitatively in BHKE18 upon removal of 220 data points and introduction of new methods. The results of BG17 are clearly unreliable, and those of BHKE18 are of unknown validity. Further, without access to the medical data and all linked performances used in BG17, it is impossible to know how or why certain athletes/results were removed and others not. What is unequivocal is that BG17 used unreliable data, and thus, its results are also unreliable. Different data and methods were used in BHKE18, leading to significantly different results, based on the almost certain use of flawed data, leading consequently to unreliable results. The bottom line is that the use of flawed data makes it impossible to know what,

Table 5 Differences between results reported in BG17 and BHKE18

	BG17				BHKE18				
	Lowest T	Highest T	Difference between highest and lowest (s)	Per cent improvement highest over lowest (%)	Lowest T	Highest T	Difference between highest and lowest (s)	Per cent improvement highest over lowest (%)	Change in per cent improvement from BG17 to BHKE18 (%)
100 m	11.44	12.06	0.62	−5.4	11.81	12.21	0.4	−3.4	2.0
100 m H	13.02	13.21	0.19	−1.5	13.05	13.23	0.18	−1.4	0.1
200 m	23.25	23.17	−0.08	0.3	23.28	23.62	0.34	−1.5	−1.8
400 m	52.1	51.02	−1.08	2.1	51.86	51.08	−0.78	1.5	−0.6
400 m H	56.66	55.02	−1.64	2.9	57.5	55.7	−1.8	3.1	0.2
800 m	122	119.4	−2.6	2.1	122.24	120.29	−1.95	1.6	−0.5
1500 m	250	247.9	−2.1	0.8	250.67	249.9	−0.77	0.3	−0.5
3000 m SC	584.5	579.2	−5.3	0.9	581.42	578.17	−3.25	0.6	−0.3
5000 m	928	917.1	−10.9	1.2	924.72	939.9	15.18	−1.6	−2.8
10 000 m	1914	1909	−5	0.3	1907.2	1913.1	5.9	−0.3	−0.6
Marathon	9431	9562	131	−1.4	9619.5	9764.7	145.2	−1.5	−0.1

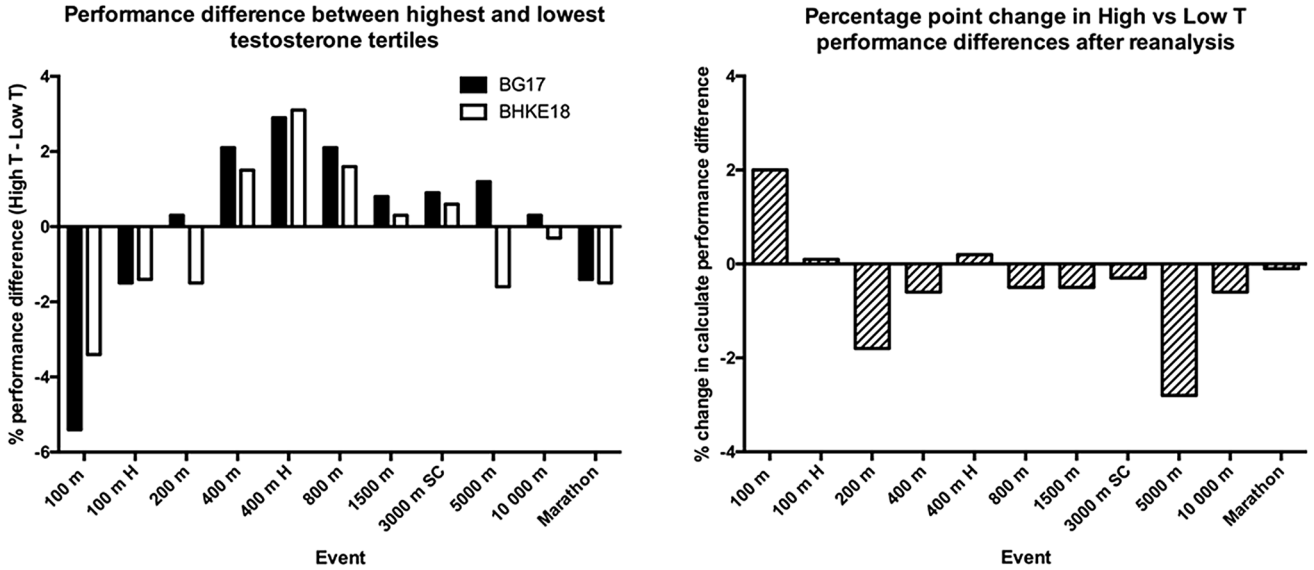

Fig. 2 Performance differences between high and low testosterone tertiles (left panel) and change in difference between BK17 and BHKE18 (right panel)

if any, relationship exists between the variables of BG17 and BHKE18 or to verify the reported results.

4 The importance of upholding scientific integrity in sports governance

The actions of the IAAF and BJSM reflect significant shortfalls in formal processes and norms related to scientific integrity. As an organization pursuing regulations, it is unfortunate but nonetheless common nowadays to observe a conflicted organization depart from conventional standards of scientific integrity. Overcoming such conflicts is one important role for academic journals that publish peer-reviewed research. In this case, inexplicably, BJSM also fell short of conventional standards among scientific publishers. If sports governance is to be based upon robust evidence, organizations in and around sport will need to uphold higher standards that we have observed in this case.

As we explored the issues associated with the data and results of BG17 and BHKE18, we were in extensive contact with the editors of BJSM. Our request that the journal work to secure the release of the balance of performance data of BG17 and that of BHKE18 was refused. Similarly, once we identified errors in the data that were provided to us by Dr. Bermon, the editors of BJSM rejected our call for the paper to be retracted. Further, after reviewing a submission from us on the flaws in BG17 and the irregularities in the process, BJSM refused to publish our analysis after more than 3 months of deliberation—not due to any scientific shortfalls (indeed, the BJSM peer reviewers judged our analysis to be scientifically sound) but because our analysis was critical of how the issue was handled by BJSM.

The actions of BJSM appear to contradict the retraction policy of the publisher of BJSM and guidelines of the Committee on Publication Ethics (COPE), which are followed by most scientific publishers. COPE explains that in some cases, the retraction of a scientific paper may be warranted: "Retraction is a mechanism for correcting the literature and alerting readers to publications that contain such seriously flawed or erroneous data that their findings and conclusions cannot be relied upon. Unreliable data may result from honest error or from research misconduct".[7] COPE further explains: "Publications should be retracted as soon as possible after the journal editor is convinced that the publication is seriously flawed and misleading (or is redundant or plagiarised). Prompt retraction should minimize the number of researchers who cite the erroneous work, act on its findings or draw incorrect conclusions". BJM, the publisher of BJSM, has a retraction policy that, like most scientific publishers, follows the guidelines of COPE: "Retractions are considered by journal editors in cases of evidence of unreliable data or findings, plagiarism, duplicate publication, and unethical research".[8]

We conclude that in rejecting retraction of BG17, BJSM did not follow its own policies or international standards of science publication. In this straightforward case, BJSM compromised its scientific integrity and contributed to what appears to be a highly dubious evidence base for an important policy issue in athletics. Furthermore, the lack of action to uphold its stated policies from BJSM has the effect that the IAAF is protected from the normal expectations of scientific research. A strength of science is that it is self-correcting.

Based on the evidence, BG17 should not have presented an editor or publisher with a complex or difficult situation. Yet, not only were we surprised and disappointed by the BJSM decision not to withdraw BG17 in the light of the evidence, but BJSM has also refused to require the authors of BG17 or BHKE18 to release their data to allow for independent replication. The editorial process used by BJSM to arrive at these judgements is also unknown. We find these issues to be highly problematic for a scientific journal. Errors are inevitable in research, and when they are identified, they are corrected. In this instance, self-correction mechanisms in science have broken down.

Despite the rejection of our request by BJSM, we maintain our call for BG17 to be retracted and suggest that BHKE18 also merits consideration for retraction. Neither BJSM nor the IAAF has made available data from either analysis and cannot be considered peer-reviewed studies. Despite the clear differences between BG17 and BHKE18, the latter erroneously claims that it presents "consistent and robust results and has strengthened the evidence". An IAAF spokesperson stated to the *New York Times* about BHKE18 that "the conclusions remain the same" as BG17, a stance also articulated by Sebastian Coe, the head of the IAAF.[9] This demonstrable falsehood has been enabled by BJSM and is sure to propagate further into the scientific literature and policy settings.[10]

The IAAF set itself up for problems by conducting research on performance effects associated with testosterone using in-house researchers. This creates at a minimum a perception of a conflict of interest that could have been mitigated to some degree by allowing independent researchers access to data and evidence, in order to replicate findings. In this case, such access was not allowed, except for the small amount of data shared with us, which was subsequently found to contain numerous errors. The unwillingness of the IAAF to correct or acknowledge errors highlights its conflict of interest.

An alternative to the approach to science and evidence employed by the IAAF would have been to provide research funding to an independent body which could request proposals from researchers unaffiliated with the IAAF to address the scientific questions at issue.[11] We would not find it appropriate for cigarette companies to provide the scientific basis for the regulation of smoking or oil companies to provide the scientific basis for regulation of fossil fuels. Sport regulation should be held to the same high standards that we

[7] https://publicationethics.org/files/retraction%20guidelines.pdf.

[8] https://authors.bmj.com/policies/correction-retraction-policies/.

[9] https://www.nytimes.com/2018/07/12/sports/iaaf-caster-semenya.html and https://magazin.spiegel.de/SP/2018/33/158846325/index.html.

[10] As of early January 2019, BG17 and BHKE18 have been together cited more than 40 times (Google Scholar), providing a case study in the propagation of flawed scientific results.

[11] A discussion of evidence-based policy and scientific integrity goes well beyond the scope of this paper, but we point readers to Pielke

expect of researchers in other settings where science informs regulation and policy.

More generally, this case illustrates clearly the importance of data sharing in science as well as the role of independent checks on data with policy or regulatory significance. This is especially the case when an interested party is sponsoring research to support a policy that it advocates. Conflicts of interest are best dealt with via transparency and commitments to scientific integrity. We encourage the sports governance community and scientific bodies, including journals that may be closely associated with them, to adopt a more rigorous policy on procedures for retraction and ensuring data availability consistent with best practices among scientific publishers. Mistakes happen. Science is robust because mistakes can be corrected. When mistakes cannot be corrected, we are no longer dealing with science, but something else.

References

Bermon S, Garnier PY (2017) Serum androgen levels and their relation to performance in track and field: mass spectrometry results from 2127 observations in male and female elite athletes. Br J Sports Med. https://doi.org/10.1136/bjsports-2017-097792

Bermon S, Hirschberg AL, Kowalski J, Eklund E (2018) Serum androgen levels are positively correlated with athletic performance and competition results in elite female athletes. Br J Sports Med. https://doi.org/10.1136/bjsports-2018-099700

Camporesi S (2018) A question of 'fairness': why ethics should factor in the court of arbitration for sport's decision on the IAAF hyperandrogenism regulations. Br J Sports Med. https://doi.org/10.1136/bjsports-2018-099387

Douglas HE, Bour E (2014) Scientific integrity in a politicized world. In: Logic, methodology, and philosophy of science: proceedings of the 14th international congress, pp 253–268

Karkazis K, Carpenter M (2018) Impossible "choices": the inherent harms of regulating women's testosterone in sport. J Bioethical Inq 4:579–587

Menier A (2018) Use of event-specific tertiles to analyse the relationship between serum androgens and athletic performance in women. Br J Sports Med 52(23):1540

Parkhurst Justin (2017) The politics of evidence: from evidence based policy to the good governance of evidence. Routledge, London

Pielke RA Jr (2007) The honest broker: making sense of science in policy and politics. Cambridge University Press, Cambridge

Pielke R Jr (2017) Sugar, spice and everything nice: how to end 'sex testing' in international athletics. Int J Sport Policy Polit 9(4):649–665

Pielke R Jr, Boye E (2019) (under review) Scientific integrity and anti-doping regulation. Int J Sport Policy Polit

Sőnksen PH, Bavington LD, Boehning T, Cowan D, Guha N, Holt R, Bőhning D (2018) Hyperandrogenism controversy in elite women's sport: an examination and critique of recent evidence. Br J Sports Med 52:1481–1482

Footnote 11 (continued)

and Boye (under review), Pielke (2007), Parkhurst (2017) for further discussion of the use of science in decision making.