As submitted to BJSM
1 August 2018

Retraction Warranted:
Unreliable Data and Findings in IAAF-sponsored Research in BJSM

Roger Pielke, Jr., University of Colorado Boulder
Ross Tucker, University of Cape Town
Erik Boye, Oslo University Hospital

Background

On 30 April 2018 we requested the performance data reported in Bermon and Garnier (2017, BG17) from Dr. Bermon and the editor of BJSM.[1] We requested 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, based on performance data publicly available from the events that they analyzed.

We consider that independent replication of their results is important because the paper forms an important basis for a recently announced hyperandrogenism policy by the International Association of Athletics Federations (IAAF). The new regulations would compel female participants in certain events to undergo medical treatments to lower their testosterone levels in order to be eligible to compete.[2] Thus, the research reported in BG17 is impactful and policy relevant. As BG17 is both funded and conducted by IAAF in support of its own regulations, it is perfectly reasonable to expect that independent scholars will wish to replicate their work.

On 6 July 2018 we and BJSM received from Dr. Bermon a subset of the original 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 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.[3] On 9 July 2018 we submitted an earlier version of the present manuscript calling for BG17 to be retracted, for reasons discussed below. On 27 July 2018 the editor of BJSM notified us that BJSM would not retract BG17 and would not request further data disclosure from the authors.

In this Discussion we document the unreliable data and findings of BG17, both of which are confirmed by reported results of BHKE18. We also present the retraction policy of the publisher of BJSM and guidelines of the Committee on Publication Ethics (COPE) which are followed by most scientific publishers. We conclude that in rejecting retraction of BG17, BJSM does not follow its own policies or international standards of science publication. In this straightforward case, BJSM has 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 gives an impression that it is protecting the IAAF from the normal expectations of scientific research.

Unreliable Data

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

(a) replication of the overall summary statistics found in Table 3 of BG17, and,
(b) re-creation of the underlying dataset based on reported times from the 2011 (Daegu) and 2013 (Moscow) World Championships (via Wikipedia)

With respect to (a, replication) Table 1 shows that we were able to successfully reproduce the summary statistics with only small differences (emphasized).

	Summary statistics from Table 3 from Bermon & 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***

Table 1. Replication of summary statistics for women's running events from Bermon and Garnier (2017) for women's track events. Small differences in replication emphasized in bold italics.

With respect to (b, re-creation) 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 reported results from the 2011 and 2013 World Championships. We re-created the data for four events (women's 400m, 400mH, 800m and 1500m) [4] because they are central to the new regulations promulgated by the IAAF. According to IAAF, these regulations are based on the results and conclusions of BG17.

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.

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

Table 2. Re-creation of data of BG17 for four events, summarizing total problematic data points identified.

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 as well in the anonymous medical data, which are the basis for the study's main conclusions regarding the performance effects of elevated testosterone levels. Such pervasive errors in the four 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, thus BHKE18 erred in its reporting of errors. Figure 1 shows that the dropped data points can be found in every event.

Figure 1. Dropped data points from BG17 to BHKE18, based on observations reported by event in each paper.

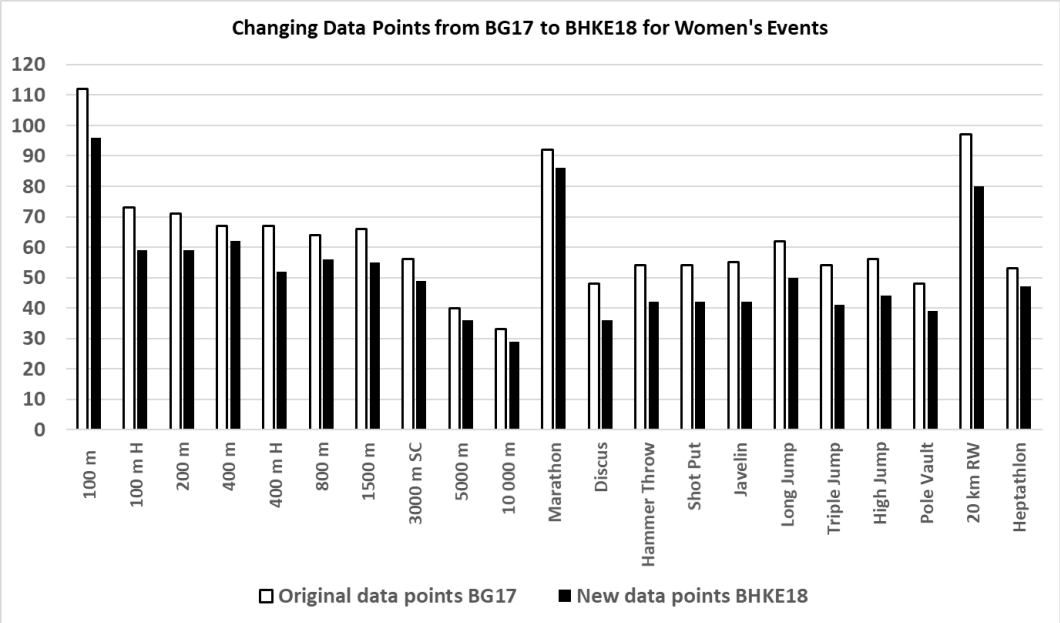


Table 3 shows the number of observations in BG17, the number of observations after data points were dropped in BHKE18, and those in our recreation of the BG17 dataset after identifying erroneous or anomalous data points. The table indicates that there are remaining uncorrected data errors in BHKE18.

Data points for four women's events

	BG17	BHKE18	Our analysis
400 m	67	62	45
400 m H	67	52	48
800 m	64	56	53
1500 m	66	55	51

Table 3. Comparison of data points in two published reports with our analysis of data provided by Bermon after correction for errors

The presence of unreliable data in BG17 is unambiguous: we have documented it empirically, the lead author has admitted to presence of errors and a subsequent analysis has sought to re-do the study after dropping 230 observations, acknowledging "errors."[5] Further, it appears that some amount of unreliable data persists in BHKE18, since the data provided to us does not match

that used in the updated BHKE18 paper. We next show that the unreliable data leads to unreliable results.

Unreliable Results

The problematic data underpinning BG17 are significant and consequential for the results reported for all events, including the four regulated events. Table 4 compares the sample numbers, means and standard deviations we replicated using the full data set provided to us by Dr. Bermon, with the corrected data once we had removed all duplicates, dopers and phantom times, as described previously (Table 2). It reveals that all three outcomes change for all female athletes in the four events upon the elimination of the previously described problematic data points. The change in aggregate times when using corrected data is of a similar magnitude to that of the testosterone effects that the authors seek to identify. Such consequential data errors easily confound identification of the effects that the analysis seeks to quantify.

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

Table 4. Performance changes for all athletes using replicated data (that provided to us of BG17) and corrected data (based on our re-creation).

Because we do not have access to the associated medical data, we cannot know what impact problematic data may have had on the BG17 conclusions. However, we can compare the results reported in BHKE18 with those of BG17 to assess the impact of dropping 220 observations.

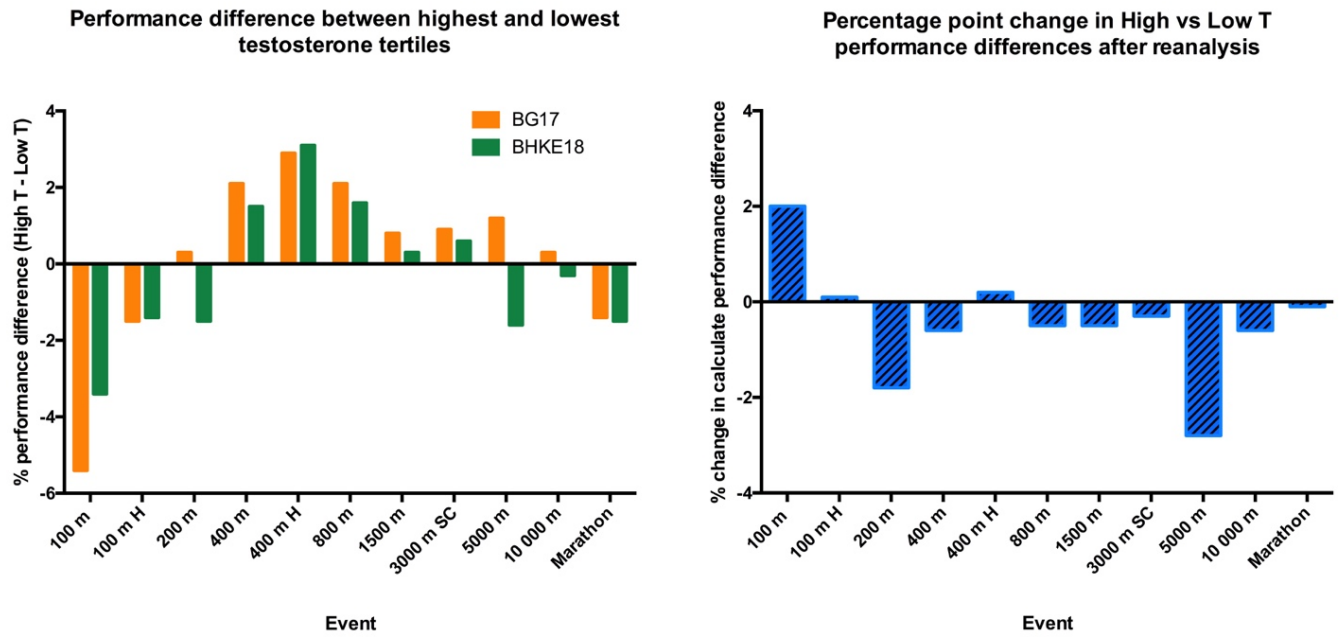
BG17 and BHKE18 both report performances by event for athletes in the top and bottom tertiles of testosterone levels. It is this difference which is argued by IAAF as the basis for regulation of certain events for female athletes, alleging that women with higher testosterone outperform women with lower testosterone.

In comparing the two studies, the reported differences between these tertiles changed dramatically, as shown in Table 5. The size of changes in results from BG17 to BHKE18 are in a majority of instances similar to the magnitude of effect being investigated. Figure 2 shows the differences in results between the two studies, by event.

	BG17				BHKE18				Change in percent difference from BG17 to BHKE18
	Lowest T	Highest T	Difference Between Highest and Lowest (Sec)	Percent difference Highest over Lowest	Lowest T	Highest T	Difference Between Highest and Lowest in Sec	Percent difference Highest over Lowest	
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%

Table 5. Differences in results reported in BG17 and BHKE18

Figure 2: Performance differences between High and Low Testosterone tertiles (left panel) and change in difference between BK17 and BHKE18 (right panel)



Important differences in results between the two studies include:

- For 8 of 11 events the difference in performances between the highest and lowest tertiles decreased (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 6 of 11 events reported in BHKE18, the low T tertile is faster than the high T tertile (compared to 3 of 11 events in BG17);
- In the four regulated events the average difference in times was reduced by 0.4% points (from 2.0% to 1.6%), and only 1 of 4 meets the BHKE18 standard for statistical significance (BG17 reported 3 of 4).

Clearly and unambiguously, the results reported in BG17 change quantitatively in BHKE18 upon removal of 220 data points and introduction of new methods. Without a full replication, it is impossible to know, but it seems plausible that the authors were unable to closely reproduce the earlier results using the same methods reported in BG17. The results of BG17 are clearly unreliable, and those of BHKE18 are of unknown validity.

Without access to the medical data and all linked performances used in BG17, it is impossible to know how or why certain athletes 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 unreliable data, leading consequently to unreliable results.

Retraction Warranted

A strength of science is that it is self-correcting. Errors are inevitable in research and when they are identified, they are corrected. The Committee on Publication Ethics explains that is 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.”[6] 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.”[7]

In various stages, we have identified and documented unreliable data and findings in BG, based on our independent analyses of ~25% of the original data of BG17, as provided to us by Dr. Bermon, and is confirmed by the subsequent effort to re-do BG17 in BHKE18. The revised approach of BHKE18 explicitly acknowledges the removal of 230 data points (itself a likely error, based on comparisons with the data we were provided).

Thus, there can be no doubt that BG17 contains "such seriously flawed or erroneous data that their findings and conclusions cannot be relied upon."

Based on the evidence, BG17 does not present 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 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.

We maintain our call for BG17 to be retracted and suggest that BHKE18 also merits consideration for retraction. BHKE18 is not a peer-reviewed study, and despite the clear differences in results compared to BG17, mischaracterizes its conclusions as providing "consistent and robust results and has strengthened the evidence." The IAAF stated to the New York Times of BHKE18 that "the conclusions remain the same" as BG17. This demonstrable falsehood has been enabled by BJSM, and is sure to propagate further into the scientific literature and policy settings.

Consequently, the unwillingness of BJSM to uphold basic standards of scientific integrity has unnecessarily muddied the evidence base on which sports governance decisions are being made. At worst, there is an appearance that BJSM is shielding IAAF and its researchers from meeting normal expectations in scientific research.

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 (in this case IAAF) is sponsoring research to support a policy that it advocates. Conflicts of interest are best dealt with via transparency and commitments to integrity. We encourage BJSM 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 can't be corrected we are no longer dealing with science, but something else.

[1] 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, *British Journal of Sports Medicine*. http://dx.doi.org/10.1136/bjsports-2017-097792
[2] https://www.iaaf.org/news/press-release/eligibility-regulations-for-female-classifica
[3] Bermon, S., Hirschberg, A. L., 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*, bjsports-2018.
[4] The mile is also included in the regulations but is not an event of the World Championships.
[5] Again, only 220 data points were actually dropped.
[6] https://publicationethics.org/files/retraction%20guidelines.pdf
[7] https://authors.bmj.com/policies/correction-retraction-policies/