Serious Problems Found in a Partial Replication of Bermon and Garnier (2017)
9 July 2018

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

On 30 April 2018 we requested the performance data reported in Bermon and Garnier (2017).[1] We did so 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 Bermon and Garnier (2017) is impactful and policy relevant.

On 6 July 2018 we received from Dr. Bermon a subset of this data, specifically for the 11 women's running events reported in their Table 3. We used the provided data to:

> (a) replicate the overall summary statistics found in Table 3 of Bermon and Garnier (2017), and,
> (b) recreate the underlying dataset based on reported times from the 2011 (Daegu) and 2013 (Moscow) World Championships.

With respect to (a, replication) we were able to reproduce the summary statistics with only small differences (highlighted in yellow below).

Table 3 from Bermon & Garnier (2017)			REPLICATION		
	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 (0.90)	71	24.43	0.90
400 m	67	52.23 (2.56)	67	52.19	2.59
400 m H	67	56.34 (2.65)	67	56.30	2.59
800 m	64	121.80 (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

[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

With respect to (b, re-creation) we have found significant anomalies and errors in the underlying data for four events for which we have recreated the data set by cross-checking times provided by Dr. Bermon with reported results from the 2011 and 2013 World Championships. We selected these four events to recreate because they are the focus of the new regulations promulgated by the IAAF based, in part, on the analysis in Bermon and Garnier (2017). These four events are women's 400m, 400mH, 800m and 1500m.

The presence of errors has been confirmed to us via email in general terms by Dr. Bermon. We have identified three types of anomalies/errors, as well as the inclusion of times (for several events) that 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, contrary to the paper's stated methods.
- *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.

Here is 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%**

Problematic data make up between 17% and 33% of the values used in the analysis for these four events. Given the pervasiveness of these errors, we hypothesize that similar data problems might be found in the data for the other 17 women's events and 22 men's events, as well as 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 dataset.

The problematic data are significant and consequential for the results reported for these four events. The table below shows how the replicated sample numbers, means and standard deviations change for all female athletes in the four events upon the elimination of the problematic data points.

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

Because we do not have access to the associated medical data, we cannot know what impact problematic data may have had on the paper's conclusions. However, it seems possible (if not likely) that the presence of as much as one third problematic data (i.e., in 400m) could have a meaningful impact on the paper's quantitative results.

Due to the pervasiveness of problematic data we are calling for Bermon and Garnier (2017) to be retracted immediately by the authors and by BJSM. If a new analysis is subsequently completed and submitted for publication, we request that it be done so only with a full, independent audit of the underlying data and results by a team committed to keeping private the associated medical data. Further, upon publication, any such analysis should also in parallel publish performance data (i.e. not the medical data with privacy concerns) such that replication of this part of the analysis is possible by any independent scholar.

This case illustrates the importance of data sharing in science as well as the role of independent checks on data with policy or regulatory significance. We encourage BJSM to adopt immediately a more rigorous policy on data availability consistent with best practices among scientific publishers.[3] Mistakes happen. Science is robust because they can be corrected.

[3] https://www.nature.com/authors/policies/availability.html