SUPPLEMENTARY MATERIAL FOR WEB CONTENT

1. WHAT CAN WE ACTUALLY CONCLUDE FROM THE RESULTS IN BERMON AND GARNIER 2017?

1.1 Robustness and replicability

Statistical analyses of observational data are subject to numerous choices, also known as 'researchers' degrees of freedom': researchers must choose which test to apply, how to define outcome variables, how to divide up the data, and which statistical threshold to use. In Bermon and Garnier (2017) [1] the authors use a single test, applied to each event. The authors divide the sample into tertiles by testosterone levels (three equally sized groups with low-, intermediate- and high- levels of free testosterone (fT), respectively) and compare the average performance in the high group to those average performance in the low group.

This particular choice of statistical analysis is not conventional, and is one of many ways that it could have been done. We would like to see numerous additional robustness tests, especially since, to the best of our knowledge, this tertile comparison test is not standard in this literature. A more natural test would be to report the raw correlation between fT levels and performance in the sample as a whole.

In addition, we are left to wonder if the results in Bermon and Garnier (2017) tell us anything about the performance of extreme outliers in terms of fT (ie. athletes near the 10nmol/L threshold previously imposed by the IAAF) relative to the average athlete. After all, the results only tell us something about the performance of the highest tertile relative to the lowest tertile. The authors are unable to say anything conclusive about the effect of having testosterone levels above the 10 nmol/L threshold, since they are likely to have small samples at this level. This is surely a point worth mentioning, as the relevant question in the CAS hearings relates to athletes above the 10 nmol/L threshold.

1.2 Multiple hypothesis testing and interpretation.

Can the data used in Bermon and Garnier 2017 tell us something about whether testosterone confers an advantage in particular events, individually, or just whether there is an overall correlation across all events? We argue that, given the sample sizes

used for each event, and the number of statistical tests conducted by the authors, any particular significant result in an event is more likely to arise by chance. Given the number of tests performed, the few significant findings detected could have arisen without there being a true correlation between testosterone and performance for female athletes. To avoid these chance findings (also known as 'false positives') appropriate multiple hypothesis testing corrections ought to be applied. Alternatively, a single test of correlation across all events should be conducted. Our analysis of the data suggests that either of these approaches seems unlikely to yield a robust and significant correlation.

First, we note that two different types of hypotheses could be tested with the data used in Bermon and Garnier 2017, as one could test a) for an effect of fT on performance in each event separately, and draw separate conclusions specific to each event (as is done in this paper); or b) one could test for an effect of fT on performance in athletics events, in general. We address these two types of arguments in turn:

a) Independent tests across multiple events:

Because there are many hypotheses tested in Bermon and Garnier 2017, we are concerned that one or more of their five reported successes (five events) are likely to be false positives, i.e. that the null hypothesis has been rejected when in fact it is true (in other words, the authors conclude that there is a relationship, when in fact there is none and they are detecting only noise in the data). If one significant test is performed and found to be significant, there is <0.05 chance that the null hypothesis should not have been rejected. But as the number of tests increases, the probability that at least one of the rejected null hypotheses should not have been rejected starts to increase. For instance, if the true effect of fT on performance was zero across all events, and tests are conducted for 43 events (men and women), we'd expect two to be significant by chance, on average. If we just looked among the 21 women's events, we'd expect one to be true by chance, on average. So it seems likely that at least one of the significant results reported in the paper is a false positive. For instance, when the authors claim that there is an advantage in the hammer throw, we cannot be sure that this is a robust finding. This is a well-studied problem in statistical sciences, for which there are many proposed solutions.[2] Note that this problem arises regardless of whether the tests are independent or not.

We have hence applied a correction procedure which controls the false discovery rate across the tests conducted in the paper: that is, the overall proportion of reported significant findings that are likely to be false positives. We apply the procedure suggested by Benjamini et al. (2006)[2] to the p-values calculated in the first part of this section. If we apply this correction we find that the lowest corrected p-value (for the event with the largest difference) is 0.239.[1] This means that, given the number of tests performed, there is at least a 24% chance that we could have found such a large correlation simply by chance. This correction is far less harsh than other corrections that control the family-wise error-rate, such as the very conservative Bonferroni correction,[3] which uses a stricter threshold of statistical significances (that means that, if we apply the Bonferroni correction, we are going to find a higher chance that the results in Bermon and Garnier 2017 have arisen simply by chance). Given these findings, we believe that it is scientifically incorrect to draw the conclusions about the specific five events claimed in the paper, as we are unable to state with any confidence that each result is not a false-positive.

b) Testing for an advantage across all events

Given multiple hypothesis testing problems, the dataset used in Bermon and Garnier 2017 is not large enough to conclusively identify those events that exhibit a correlation, and those that do not. Might one argue, instead, that there is evidence for an average effect of fT on performance, across all events? After all, the regulations covering testosterone among female athletes would likely apply to all athletes, not just the events highlighted in the conclusion of this paper. Having established whether there is a significant treatment effect on average, a researcher might then test whether there is evidence for heterogeneity in the effect size across different events.

Unfortunately, without publically available raw data, it is not possible to perform all of the desired robustness checks on the data. Ideally, researchers with access to the raw data should calculate standardized performance scores for each observation, pool those data across all events, and look for an average effect across the full sample. In lieu of access to such data we performed a Fisher's combination test using the p-

[1] There are some cases where the tables in the original paper leave some ambiguity about the underlying data structure (not least, some ambiguity about the total sample size in each tertile). In all such cases we have made assumptions that would bias us in favour of lower p-values, so if anything, the corrections we have applied are conservative.

values calculated from the published data. After performing such test we are unable to reject the "global" null hypothesis that all null hypotheses are true, i.e. the pattern of p-values found in the paper, across all women's events, is not inconsistent with there being no advantage to high fT women, in any of the events. In simpler terms: it is possible that the correlations presented in the paper occurred simply by chance. Therefore, researchers and practitioners should interpret the statistical results in Bermon and Garnier 2017 with extreme caution.

2. DO THE BEST ESTIMATES OF THE ADVANTAGE TO HIGH TESTOSTERONE WOMEN MEET THE THRESHOLD SET BY CAS?

In CAS Interim Award of July 2015,[4] the Panel accepted the evidence that male athletes have a competitive advantage over female athletes on the order of 10-12 %. As further evidence for the re-opening of the *Dutee Chand vs AFI & IAAF* case, we would like to point to some original data collected by Ospina Betancurt[5,6] as part of a doctoral thesis, now partially published. The objective of this doctoral study was to investigate the difference in athletic performance between women with and without hyperandrogenism competing at an elite level in track and field, in order to establish whether sportswomen with hyperandrogenism obtained a performance of 10-12 % and hence whether there is a justification in the regulation of the eligibility of female athletes with hyperandrogenism as required by CAS. The study, recently published for the *Journal of Sport Sciences*[4], confirms that the threshold set by CAS, with respect to the difference between male and female athletes, is indeed correct.

The CAS Panel also noted that the assumption underlying the Hyperandrogenism Regulations is that "endogenous testosterone level within the male range + virilisation (indicating sensitivity to the high levels of testosterone) = a degree of competitive advantage over non-hyperandrogenic females *of commensurate significance to the competitive advantage that male athletes enjoy over female athletes*" [emphasis added]. [4 p.154] The Panel was not satisfied that this assumption was proven valid, hence the suspension of the Regulations. It is important to note that once the requested further evidence was put forward establishing a certain degree of advantage of hyperandrogenic athletes over non-hyperandrogenic athletes, the Panel would have to consider "if the degree of advantage were well below 12 %, whether that justified

excluding women with that advantage from the female category". [4 p.155] As demonstrated by our reanalysis, we have serious reasons to question whether the true effect in Bermon and Garnier 2017 falls in the reported range of 1.8-4.5 %.

3. CONCLUSIONS

The finding of a significant difference in performance in a small subset of statistical tests does not mean a correlation exists for each of them. We have argued that Bermon and Garnier 2017's analysis is vulnerable to problems of false positives, and requires the application of multiple hypothesis test corrections. In other words, the statistical tables presented in a paper such as Bermon and Garnier 2017 cannot "speak for themselves": both the selection of statistical tests and the interpretation of particular statistical differences require important statistical robustness checks.

Further, we noted how the authors' choice to make claims about specific events (i.e. that high testosterone confers an advantage in five specific events alone and, presumably, not the other events) is not obvious, and likely to be inappropriate, since it is understood that their paper is likely to influence the policy discussion about whether testosterone levels should be regulated across *all* female events. In our paper we have argued that the sample size, and effect sizes, in Bermon and Garnier 2017, are simply not large enough to make robust claims about each event separately.

A more reasonable, and policy-appropriate, use of the data would be to establish an *average* correlation between fT and performance across events. However, our reanalysis of the results in Bermon and Garnier 2017 suggest that there is no evidence for such an overall correlation, and allows us to conclude that it is not only inaccurate for the authors to state that there is an advantage of between 1.8% to 4.5% for high fT female athletes in general, it is in fact inaccurate to conclude that is the size of advantage in these specific 5 events. After correcting for multiple hypothesis testing, no differences would be individually significant. And, across all events, the average advantage to high testosterone women is estimated to be 0.7%, with a minimum of -2.6% and a maximum of 4.5% across the 21 events, and only in 12 (57%) of the events do higher fT athlete perform better on average.

In light of our re-analysis, we conclude first that raw data used in such studies, that will have direct implications for real world outcomes, should be made publically

available for other researchers to analyse. In a field where such data is often propriety and difficult for independent researchers to access, efforts should be made to anonymize such data and make it publicly available. Second, we conclude that the interpretation of estimated correlations should also be conducted with great caution, and be referred to independent statisticians.

While do not claim to play the role of such an independent statistical arbitrator in this case (especially since we have not had access to the raw data), our statistical analysis already allows one to conclude that the article by Bermon and Garnier 2017 does not meet the standard of proof set by the CAS, without further analysis, and provides the argument that this independent analysis is necessary in this situation, and others like it.

Reference List

1. Bermon, S., & Garnier, P. Y. (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 ; 51:1309-1314.
2. Benjamini, Y., A. M. Krieger, and D. Yekutieli. (2006) Adaptive Linear Step-up Procedures that Control the False Discovery Rate. Biometrika 93(3) , 491–507.
3. Hochberg, Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75 (1988), 800–803.
4. Court of Arbitration for Sport. Interim Arbitral award: CAS2014/A/3759 Dutee Chand v. Athletics Federation of India (AFI) & The International Association of Athletics Federations. 2015 July 24. Available from: http://www.tas-cas.org/fileadmin/user_upload/award_internet.pdf (Accessed February19, 2018)
5. Ospina Betancurt, J; Zakynthinaki, M; Martinez-Patiño, MJ; et al. (2017a). Sex-differences in elite-performance track and field competition from 1983 to 2015. *J Sport Sci*, 1-7. https://doi.org/10.1080/02640414.2017.1373197
6. Ospina Betancurt, J.(2017b) Controles de sexo, género, hormonales y la inelegibilidad de las mujeres con hiperandrogenismo en el deporte femenino de alto nivel [dissertation]. Madrid: Universidad Politécnica de Madrid;. doi10.20868/UPM.thesis.47157