Discussion:

Collider bias in observational studies: Why the effects of hyperandrogenism in elite women's sport are likely underestimated

N. T. Borgen, Postdoctoral researcher at the Department of Sociology and Human Geography, University of Oslo, Norway.

Corresponding author's contact details:
P.O. Box 1096 Blindern, 0317 OSLO, Norway;
E-mail: n.t.borgen@sosgeo.uio.no
Phone: +47-22855248

The medical condition of hyperandrogenism has entered common parlance because the International Association of Athletics Federation (IAAF) decided to ban women with the condition from competing against women who have 'normal' levels of testosterone (see box 1). This paper argues that the two articles substantiating the IAAF's regulation [1 2] fail to prove that elevated levels of natural testosterone are causally linked to better sporting performance. By studying elite athletes both papers induce a collider stratification bias that attenuates the competitive advantage of testosterone – a little-known bias that often plagues studies of highly selective samples.

Box 1 Background

- In April 2018, IAAF introduced a new eligibility regulation for female athletes, requiring women with Differences of Sexual Development to reduce their blood testosterone levels to below 5 nmol per litre.
- IAAF builds its case on two articles in BJSM from 2017, by authors with declared connections to IAAF, suggesting that testosterone levels in female athletes are positively correlated with athletic performance [1 2]. Several studies have criticized this research[3-6], highlighting the lack of correlation analysis [4], the focus on free testosterone [4], and the problem of type I error in statistical tests [3], all of which the authors of the original articles have attempted to address [7]. There has also been calls to retract the study because of serious flaws in the data [8].
- The regulation was challenged in the Court of Arbitration for Sport by Caster Semenya, 800m Olympic and world champion. The ruling upheld the regulation, amidst further disagreement between medical professional bodies.

Causality

Do elevated levels of natural testosterone cause women to run faster or throw further? And if so – how much faster or further? The size of the competitive advantage of testosterone is a *causal* research question. Yet, the two key papers in BJSM [1 2] have neither sufficiently discussed the challenges related to casual inference with observational data nor used methods that allows for causal interpretation of the results. Consequently, the two studies reveal that elite female athletes with high testosterone levels have better athletic performance (association), but not that these elite athletes have better athletic performance *because* of high testosterone level (causal effect).

The gold standard for medical evidence on cause and effect is randomized controlled trials, but that avenue is clearly not straightforward in the case of hyperandrogenism in elite female athletes. Even in the absence of ethical concerns, randomization is rare in studies of elite athletes and – like most of the sports medicine literature – studies of hyperandrogenism must use observational data.

Associations between variables in observational data can be used to draw conclusions on cause and effect. However, when the researcher does not assign the treatment (testosterone), identification of causal effects requires detailed knowledge about the data-generating process. Although rarely acknowledged, a crucial aspect of this process is which athletes become elite athletes [9]. Restricting the sample to elite athletes amounts to conditioning on prior athletic performance, which has important consequences for the estimated size of correlations in this subgroup.

There's bias in the results

Bias introduced through sample selection is known variously as collider stratification bias, Berkson's paradox, M-bias, sample selection bias, and endogenous selection bias. To illustrate collider stratification bias, I will build on the hypothetical data generating process encoded in the Directed Acyclic Graph (DAG) in figure 1. Let us assume that testosterone levels are randomly distributed at birth and there is a substantial competitive advantage of testosterone. This would mean that women with high levels of testosterone are more likely to become elite athletes despite lacking other favorable traits such as high VO_{2max}. Women with low levels of testosterone on the other hand, must compensate for their low levels of testosterone with other traits; if they do not, they will not become elite athletes.

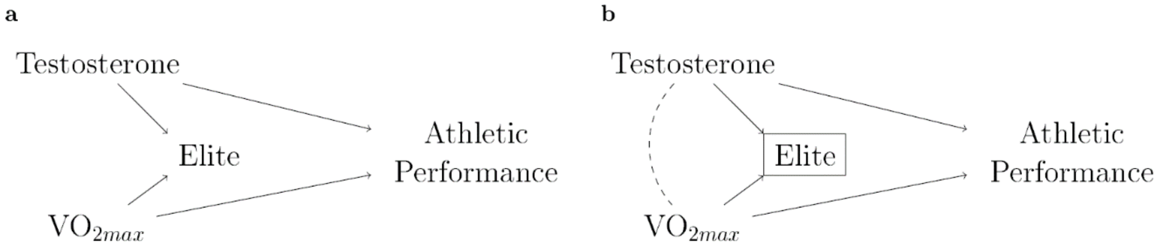

Figure 1: A DAG that encode the hypothetical data generating process and the problem of conditioning on a collider variable. Studying elite athletes amounts to conditioning on elite status, and is shown graphically by drawing a box around the conditioned variable. (a) Testosterone is assumed to be independent of other causes of athletic performance (VO_{2max}), and the marginal association between testosterone and athletic performance identifies the causal effect. (b) Elite is a collider variable on the path Testosterone → Elite ← VO_{2max}, and conditioning on this collider variable induces a spurious association between Testosterone and VO_{2max}. The conditional association between testosterone and athletic performance given elite status does not identify the causal effect of testosterone.

Consequently, elite women athletes with low levels of testosterone are likely to have other exceptional attributes and cannot be compared to those with high testosterone levels. Figure 2 illustrates that restricting the sample to elite athletes induces a spurious negative association between testosterone and other favorable traits (here VO_{2max}). In the simulated data, the correlation between testosterone and VO_{2max} is approximately zero in the full sample ($\rho = 0.0066$, panel A), while the correlation is negative and strong in the elite sample ($\rho = -0.7603$, panel B).

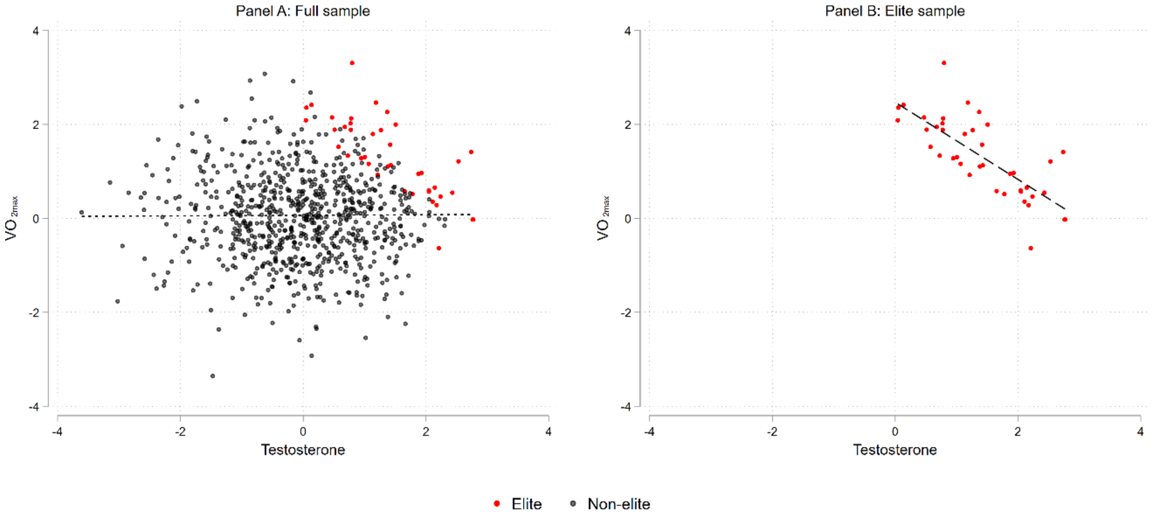

Figure 2: Simulated data to illustrate that restricting the sample to elite athletes induces a negative association between testosterone and other favorable attributes such as VO_{2max}.
Note: The data generating process consists of one random draw of 800 women with $VO_{2max} \sim N(0,1)$, $Testosterone \sim N(0,1)$, and the overall correlation between VO_{2max} and $Testosterone$ equal to 0. Athletic performance at time point 1 (Y_1) and time point 2 (Y_2) is generated as $VO_{2max} * 2 + Testosterone * 2 + \varepsilon$, where $\varepsilon \sim N(0,1)$. The women with 5 percent highest value on Y_1 is classified as elite athletes and the Y_2 variable is considered the outcome variable. The data is generated using Stata 15.1 and random-number seed 1256981. See Supplementary Online Appendix for Stata syntax to replicate the results.

The spurious negative association between testosterone levels and VO_{2max} (collider bias) in turn results in an underestimate of the competitive advantage of testosterone (Figure 3). To identify the competitive advantage, we would have to compare elite athletes with different levels of testosterone but who are otherwise equal. The attenuation occurs because comparing the athletic performance of elite athletes with low and high levels of testosterone implies comparing women that differs in other ways, too (high VO_{2max} vs. medium to high VO_{2max}). The size of the bias depends on the actual effect of testosterone and other traits, and on how elite the sample is. This is discussed at length elsewhere using another sports medicine example. [9]

Collider bias is widespread in much of the sports medicine literature, resulting in either overestimation or underestimation of causal effects. Whenever the sample is (highly) selective (e.g., elite athletes or obese individuals), bivariate correlations within that sample is likely biased, sometimes to a considerable degree. Failure to adequately account of this bias leads to some apparent paradoxes, such as the inverse association between VO_{2max} and running efficiency [9], but also underestimation of effects, as is the case with the competitive advantage of testosterone.

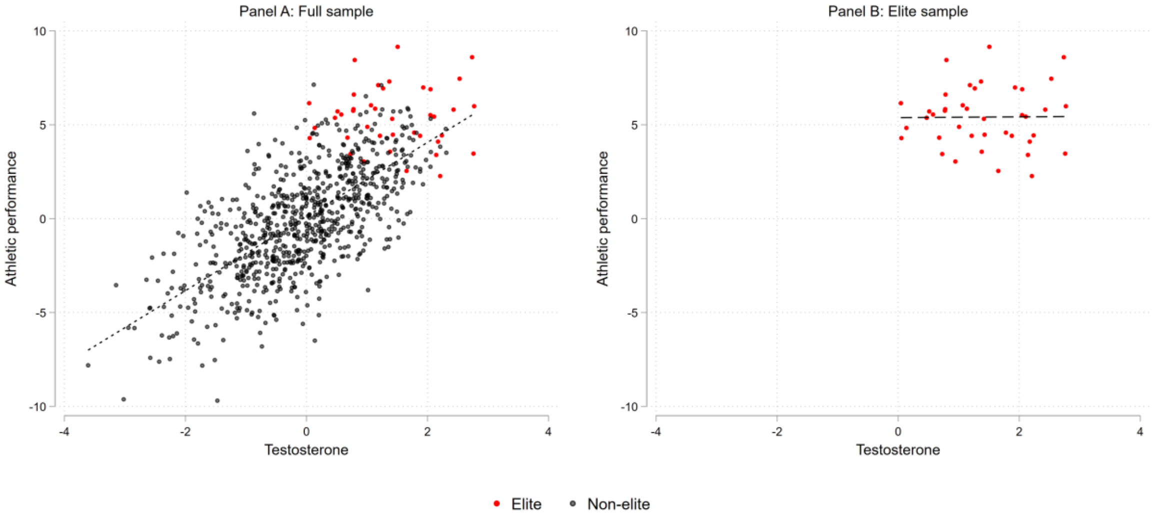

Figure 3: Simulated data to illustrate that restricting the sample to elite athletes attenuates the competitive advantage of testosterone.
Note: See Figure 2 for description of the data generating process.

Box 2 Testosterone levels and collider bias

- The size of the competitive advantage of testosterone is a causal research question.
- Restricting the sample to elite athletes likely induces a spurious negative association between testosterone and other favorable traits; a collider bias.
- Considered in isolation, the collider bias attenuates the competitive advantage of testosterone.

We must get it right

Research aiming to underpin policy and regulations largely addresses causal questions. When investigating causal questions, it is essential to discuss challenges related to causal inference and, if possible, use methods that convincingly address these challenges. Currently, the presence of collider bias in research on hyperandrogenism in women athletes does not allow these studies to sufficiently prove a causal link between elevated levels of natural testosterone and sporting performance.

References

1. Bermon S, Garnier P-Y. 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;51:1309-14.
2. Eklund E, Berglund B, Labrie F, et al. Serum androgen profile and physical performance in women Olympic athletes. *Br J Sports Med* 2017;51:1301-08.
3. Franklin S, Betancurt JO, Camporesi S. What statistical data of observational performance can tell us and what they cannot: the case of Dutee Chand v. AFI & IAAF. *Br J Sports Med* 2018;52(7):420-21.
4. Sőnksen PH, Bavington LD, Boehning T, et al. Hyperandrogenism controversy in elite women's sport: an examination and critique of recent evidence. *Br J Sports Med* 2018; Ahead of print.
5. Menier A. Use of event-specific tertiles to analyse the relationship between serum androgens and athletic performance in women. *Br J Sports Med* 2018; Ahead of print.
6. Camporesi S. A question of 'fairness': Why ethics should factor in the Court of Arbitration for Sport's decision on the IAAF Hyperandrogenism Regulations. 2018; Ahead of print.
7. Bermon S, Hirschberg AL, Kowalski J, et al. Serum androgen levels are positively correlated with athletic performance and competition results in elite female athletes. *Br J Sports Med* 2018; Ahead of print.
8. Pielke R, Tucker R, Boye E. Scientific integrity and the IAAF testosterone regulations. *The International Sports Law Journal* 2019 doi: 10.1007/s40318-019-00143-w
9. Borgen NT. Running Performance,VO2max, and Running Economy: The Widespread Issue of Endogenous Selection Bias. *Sports Medicine* 2018;48(5):1049-58.

Acknowledgements

I thank the Editor in Chief Dr. Karim Khan, Rasmus Østergaard Nielsen, the anonymous reviewers, and Solveig T. Borgen for comments and suggestions on an earlier version of this manuscript.

Supplementary Online Appendix

```
* Stata code to replicate Figure 2 and Figure 3 *

* Title: Collider bias in observational studies: Why the effects of
* hyperandrogenism in elite women´s sport are likely underestimated

* Date: 18 February 2019
* Author: Nicolai Topstad Borgen
* Journal: British Journal of Sports Medicine

* Optional: Install user written commands
ssc install blindschemes, replace all
net install grc1leg, from(http://www.stata.com/users/vwiggins)

set scheme plotplainblind

version 15.1

clear
set seed 1256981
set obs 800

gen t=rnormal()
gen v=rnormal()
gen y1=t*2 + v*2 + rnormal()
gen y2=t*2 + v*2 + rnormal()
cor

su y1, d
gen elite=y1>r(p95)

* Figure 2

cor t v
cor t v if elite

tw      (scatter v t if elite==1, mcolor(red) ms(o))            ///
        (scatter v t if elite==0, mcolor(black%50) ms(o))       ///
        (lfit v t, lcolor(black))                               ///
        , ytitle(VO{subscript: 2max}) xtitle(Testosterone)      ///
        legend(order(1 "Elite" 2 "Non-elite") rows(1))          ///
        name(Fig2a, replace) title(Panel A: Full sample)        ///
        xlabel(#6) ylabel(#6) nodraw

tw      (scatter v t if elite==1, mcolor(red) ms(o))            ///
        (lfit v t if elite==1, lcolor(black))                   ///
        , ytitle(VO{subscript: 2max}) xtitle(Testosterone)      ///
        legend(order(1 "Elite") rows(1))                        ///
        name(Fig2b, replace) title(Panel B: Elite sample)       ///
        xlabel(#6) ylabel(#6) nodraw

grc1leg Fig2a Fig2b, ycommon xcommon name(figure2, replace)
graph display, ysize(7) xsize(15)
graph export Figure2.pdf, replace

* Figure 3

cor y2 t
cor y2 t if elite

tw      (scatter y2 t if elite==1, mcolor(red) ms(o))           ///
        (scatter y2 t if elite==0, mcolor(black%50) ms(o))      ///
        (lfit y2 t, lcolor(black))                              ///
        , ytitle(Athletic performance) xtitle(Testosterone)     ///
        legend(order(1 "Elite" 2 "Non-elite") rows(1))          ///
        name(Fig3a, replace) title(Panel A: Full sample)        ///
        xlabel(#6) ylabel(#6) nodraw

tw      (scatter y2 t if elite==1, mcolor(red) ms(o))           ///
```

```
        (lfit y2 t if elite==1, lcolor(black))                  ///
        , ytitle(Athletic performance) xtitle(Testosterone)     ///
        legend(order(1 "Elite") rows(1))                        ///
        name(Fig3b, replace) title(Panel B: Elite sample)       ///
        xlabel(#6) ylabel(#6) nodraw

grc1leg Fig3a Fig3b, ycommon xcommon name(figure3, replace)
graph display, ysize(7) xsize(15)
graph export Figure3.pdf, replace
```