Sharing Data and Code: A Comment on the Call for the Adoption of More Transparent Research Practices in Sport and Exercise Science

Received: 28 March 2020

Supplementary materials:

[https://github.com/SciBorgo/Open-
data-in-sports-science](https://github.com/SciBorgo/Open-data-in-sports-science)

For correspondence:

david.borg@griffith.edu.au

David N Borg¹, Joshua J Bon^{2,3}, Kristin L Sainani⁴, Brenton J Baguley⁵, Nicholas J Tierney^{3,6}, and Christopher Drovandi^{2,3,7}

¹The Hopkins Centre, Menzies Health Institute Queensland, Griffith University, Brisbane, Australia; ²School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia; ³Australian Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), Australia; ⁴Department of Epidemiology and Population Health, Stanford University, Stanford, CA; ⁵Institute for Physical Activity and Nutrition, Deakin University, Geelong, Australia; ⁶Department of Econometrics and Business Statistics, Monash University, Melbourne, Australia; ⁷Centre for Data Science, Queensland University of Technology, Brisbane, Australia.

Citation: Borg DN, Bon JJ, Sainani KL, Baguley BJ, Tierney NJ, and Drovandi C. (2020). Sharing Data and Code: A Comment on the Call for the Adoption of More Transparent Research Practices in Sport and Exercise Science. *SportRxiv*. <https://doi.org/10.31236/osf.io/ftdgi>.

Excerpt

The recent article by Caldwell et al. [1] calls for improved transparency in sport and exercise science research, with a focus on registered reports [2]. The importance of the article, and similar articles from other disciplines (e.g., [3–6]), cannot be overstated. In our opinion, the ‘call for action’ could be era-defining for open science practices in our field. Caldwell et al. [1] eloquently explain registered reports, a central feature of open science and the open science framework (OSF) [2]. An example publication model is also provided, showing readers how registered reports are implemented—a process that may be unfamiliar to many researchers. In support of the ‘call to action’, we wish to highlight another significant component of open science—the release of analytical materials (i.e., data and code used for analysis) alongside a manuscript—with the hope to further improve transparency and reproducibility in sport and exercise science research.

Sharing data and code alongside a manuscript allows reviewers (pre-publication) and readers (post-publication) to verify the implemented analysis [7]. Releasing data and code can increase the impact of research, by allowing others to update and/or extend statistical models to other contexts and identify any errors [8]. Further, open data is a critical step toward data aggregation, and meta-analysis—strategies that may help alleviate sample size concerns and improve sample-bias corrections [9]. Critically, ensuring data and code are publicly available could prevent wasted research endeavours and accelerate scientific discoveries.

The aim of open science is to provide a progressive, publicly available workflow of the entire research process, from project conception to publication [2]. As per the open science framework, four main stages facilitate this process: search and discover, study design, data collection, and analysis and publication [10]. An important element of the data collection and analysis stage is the storage of “*data, code, and other materials...for citing and sharing*” [10]. Ideally, authors would share sufficient data and code to allow replication of the analysis. Unfortunately, such practices are uncommon in sport and exercise science. To establish this claim, we investigated the state of the field, in terms of data and code sharing. We sampled 400 articles, at random, published in quartile one ‘sports science’ journals [11] in 2019, to determine the proportion sharing data and/or code (Supplementary material, available at <https://github.com/SciBorgo/Open-data-in-sports-science>). Of the 299 eligible articles, only 4.3% [95% confidence interval; 2.3 to 7.3] shared data, with an additional 1.7% [0.6 to 4.0] stating data were available upon request (Table 1). No surveyed articles shared code or computer-generated code—often referred to as syntax (Table 1). These findings are eye opening, but not surprising to the authors. Preferably, researchers would be more forthcoming with the tools and processes used to generate inference from data.

Table 1. The proportion of articles published in 2019 in quartile one ‘sports science’ journals sharing data, code and stating data were available upon request. Proportions are also reported for the subdisciplines of sport and exercise science, sports medicine and rehabilitation.

Discipline/ subdiscipline	All eligible articles (n = 299)	Sport and exercise sci- ence † (n = 148/299)	Sports medicine (n = 137/299)	Rehabilitation (n = 14/299)
Sharing data	13/299 4.3% [2.3 to 7.3]	8/148 5.4% [2.4 to 10.4]	5/137 3.6% [1.2 to 8.3]	0/14 0.0%
Sharing code	0/298 0.0%	0/148 0.0%	0/136 0.0%	0/14 0.0%
Data available upon request ‡	5/286 1.7% [0.6 to 4.0]	2/140 1.4% [0.2 to 5.1]	3/132 2.3% [0.5 to 6.5]	0/14 0.0%

Note. Proportions are reported as the mean and 95% confidence interval.

† Includes research from the subdisciplines of physiology and exercise physiology

‡ Data availability statements were considered not applicable for articles that shared data.

There are legitimate reasons that prevent data from being released (e.g., intellectual property or privacy concerns). However, in such circumstances, alternatives exist. For example, authors could: (a) release a limited set of data that excludes variables that threaten privacy; (b) randomly perturb values by a small amount, to alleviate privacy concerns; or (c) share a simulated dataset, generated from the fitted

model. If authors are unsure of how to simulate data from a fitted model, sharing code will allow others to do so.

Caldwell et al. [1] acknowledged barriers that may discourage authors from using registered reports. Similarly, authors may be deterred from releasing data/code due to a fear of scrutiny and a lack of incentive. There is a need for a culture change in sport and exercise science research, where errors are viewed as part of the normal scientific process [12]. For science to be truly self-correcting, researchers must be encouraged to be active participants in the correction process, and importantly, stigma minimised when honest errors are made [12]. There is also a need to incentivise research transparency. Without greater incentive for data/code to be housed on publicly available platforms, researchers are unlikely to change existing practices. Therefore, journals must provide greater incentive for authors to release their analytical materials at manuscript submission, by implementing such a mandatory policy.

Researchers may be unsure of how to make their data and code publicly available. Most, if not all, journals allow data/code to be submitted as supplementary material, making it easily accessible to the reader. Equally, authors can share data and code in open repositories, including GitHub, OSF projects, Mendeley, and institutional databases. Links to these resources can be specified in the main or supplemental text of a paper. Data sharing and its associated challenges are discussed in recent work from Tierney and Ram [13], who provide an in depth, practical guide for making data freely available. For individuals more comfortable with point-and-click software, the integration of statistical programs with the OSF (e.g., JASP) and secondary platforms (e.g., jamovi, which integrates with R) provide avenues to easily share reproducible code, at no financial expense.

The adoption of open science practices cannot be facilitated by individuals alone. Individuals can advocate for open science when involved with projects and during the peer-review process. However, to facilitate widespread change, sports science journals need to adapt their policies to enable and incentivise transparent research. Like other fields (e.g., psychology), the lack of transparency in sport and exercise science research strongly suggests the discipline is also experiencing a ‘reproducibility crisis’ [5,6]. Other issues are currently receiving greater attention, such as the creation of controversial statistical tests [14,15]. We believe that there is a pressing need to focus on research transparency [1]. Registered reports and the mandatory release of analytical material are key steps in the right direction. These practices would improve the quality and reproducibility of published work in our discipline and could restore trust in sport and exercise science research.

Funding and competing interests

This work was not financially supported, and the authors have no conflicts of interest to declare.

Data and code availability

The data and code that support this article are available at <https://github.com/SciBorgo/Open-data-in-sports-science>. Details of the systematic search can also be found at the same address.

References

- [1] Caldwell AR, Vigotsky AD, Nuckols G, Boardley I, Schmidt J, Tenan M, et al. Moving Sport and Exercise Science Forward: A Call for the Adoption of More Transparent Research Practices. *Sports Med.* 2020;50(3):449–459. <https://doi.org/10.1007/s40279-019-01227-1>
- [2] Foster ED, Deardorff A. Open Science Framework (OSF). *J Med Libr Assoc.* 2017;105:203–206. <https://doi.org/10.5195/jmla.2017.88>
- [3] Stodden V, Seiler J, Ma Z. An empirical analysis of journal policy effectiveness for computational reproducibility. *PNAS.* 2018;115:2584–2589. <https://doi.org/10.1073/pnas.1708290115>
- [4] Rowhani-Farid A, Barnett AG. Has open data arrived at the British Medical Journal (BMJ)? An observational study. *BMJ Open.* 2016;6(10):e011784. <https://doi.org/10.1136/bmjopen-2016-011784>
- [5] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science.* 2015;349. <https://doi.org/10.1126/science.aac4716>
- [6] Shrout PE, Rodgers JL. Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. *Ann Rev Psychol.* 2018;69:487–510. <https://doi.org/10.1146/annurev-psych-122216-011845>
- [7] Wallach JD, Boyack KW, Ioannidis JPA. Reproducible research practices, transparency, and open access data in the biomedical literature, 2015–2017. *PLoS Biol.* 2018;16:e2006930. <https://doi.org/10.1371/journal.pbio.2006930>
- [8] Allen C, Mehler DMA. Open science challenges, benefits and tips in early career and beyond. *PLOS Biology.* 2019;17:e3000246. <https://doi.org/10.1371/journal.pbio.3000246>
- [9] Choudhury S, Fishman JR, McGowan ML, Juengst ET. Big data, open science and the brain: lessons learned from genomics. *Front Hum Neurosci* 2014;8. <https://doi.org/10.3389/fnhum.2014.00239>
- [10] Open Science Framework (OSF). Accessed March 2020. Available from: <https://osf.io/>
- [11] Journal Rankings on Sports Science. Accessed March 2020. Available from: <https://www.scimagojr.com/journalrank.php?category=3699&year=2018>
- [12] Brown AW, Kaiser KA, Allison DB. Issues with data and analyses: Errors, underlying themes, and potential solutions. *PNAS. National Academy of Sciences;* 2018;115:2563–2570. <https://doi.org/10.1073/pnas.1708279115>
- [13] Tierney NJ, Ram K. A Realistic Guide to Making Data Available Alongside Code to Improve Reproducibility. *arXiv.* 2020; arXiv:200211626
- [14] Sainani KL, Lohse KR, Jones PR, Vickers A. Magnitude-based Inference is not Bayesian and is not a valid method of inference. *Scandinavian J Med Sci Sports.* 2019;29:1428–1436. <https://doi.org/10.1111/sms.13491>
- [15] Tenan MS, Vigotsky A, Caldwell AR. On the Statistical Properties of the Dankel-Loenneke Method. *SportRxiv;* 2019. <https://doi.org/10.31236/osf.io/8ndhg>