LETTER TO THE EDITOR

Poor statistical reporting and spin in neuromuscular fatigue research

Martin E. Héroux[1,2] · Simon C. Gandevia[1,3]

Dear Editor,

Our longstanding interest in neuromuscular fatigue led us to retrieve the paper by Mira et al. (2018). This interest turned to dismay after reading the abstract:

> "[…] RESULTS. Time to EXH was longer in $POST_{ABS}$ than PRE (34 ± 5 vs. 27 ± 4 min, $P < 0.001$), and $POST_{REL}$ tended to be longer than PRE (30 ± 6 min, $P = 0.053$). […] At EXH, MVC and $Db_{10:100}$ were similarly reduced in all sessions (-37 to -42% and -30 to -37%, respectively). Db_{100} tended to be less depressed in $POST_{ABS}$ than PRE (-40 ± 9 vs. $-48 \pm 16\%$, $P = 0.050$) and in $POST_{REL}$ than PRE ($-39 \pm 9\%$, $P = 0.071$). P_t decreased similarly in $POST_{ABS}$ and PRE (-52 ± 16 vs. $-54 \pm 16\%$), but $POST_{REL}$ tended to be less depressed than PRE ($-48 \pm 4\%$, $P = 0.075$). CONCLUSIONS: This study confirms fatigue attenuation at isotime after training.[…]" (our emphasis)

Specifically, we were struck by the four non-significant *p*-values interpreted as scientifically meaningful.

As researchers, we have been conditioned to celebrate when our *p*-values align with our stated hypotheses. However, when *p*-values fall short of the mark, some researchers choose to spin their results (Diong et al. 2018). Unfortunately, this is a slippery slope that is difficult to navigate.

Communicated by Westerterp/Westerblad.

✉ Simon C. Gandevia
  s.gandevia@neura.edu.au

1 Neuroscience Research Australia, Margarete Ainsworth Building, Randwick, Sydney, NSW 2031, Australia

2 School of Medical Sciences, University of New South Wales, Sydney, NSW 2052, Australia

3 Prince of Wales Clinical School, University of New South Wales, Sydney, NSW 2052, Australia

While Mira et al. (2018) imply their results associated with $p = 0.05$, $p = 0.053$, $p = 0.071$ and $p = 0.075$ are scientifically important, the Results section includes the following:

> "However, the $\%VO_{2peak}$ at which GET and RCP occurred were not different between pre and post-training ($P = 0.087$ and $P = 0.308$, respectively)."

The discussion goes on to report:

> "Pt was less depressed in $POST_{ABS}$ than in PRE from 25% TTE_{PRE}, which suggests attenuation of excitation–contraction coupling (ECC) failure. $Db_{10:100}$ also displayed a lower depression in $POST_{ABS}$ vs. PRE, although this did not reach statistical significance (interaction effect, $P = 0.093$)."

It has been said that interpreting *p*-values just above a chosen threshold as statistical trends is inappropriate and actively misleading because more often than not, collecting additional data renders these trends statistically non-significant (Wood et al. 2014). Mira et al. (2018) go one step further and cherry-pick trends that suit their narrative. Nevertheless, the weakness of the evidence generated from such flexibility often reveals itself when trying to discuss such results:

> "Although our findings corroborate the first hypothesis and $POST_{REL}$ tended to be longer than PRE […]"
> "TTE tended to be longer in $POST_{REL}$ than in PRE, which is somehow in agreement with our hypothesis."
> "Although not significantly, MVC and Pt also showed higher values at rest post-training."
> "Db_{100} and Pt tended to be less depressed in $POST_{REL}$ than PRE. As TTE also only tended to be prolonged in POS_{TREL}, this suggests that training tended to attenuate peripheral fatigue at EXH for a slightly higher power output and exercise duration." (our emphasis)

Spin is not new to science. In fact, its incidence is increasing (Boutron and Ravaud 2018). At the same time, there is vigorous debate as to whether the *p*-value and null hypothesis significance testing should be revisited or retired

(Wasserstein et al. 2019). Regardless of whether one follows this debate, no statistician or statistical textbook has ever recommended such flexible interpretations of *p*-values.

Research results should be reported and interpreted carefully, with a focus on the size and certainty of studied effects rather than their fickle *p*-values (Wasserstein et al. 2019). The fact that this paper was approved by eight authors, 2–3 reviewers and 1 editor highlights the endemic nature of the problem. We hope that highlighting the statistical reporting and spin present in Mira et al. (2018) incites a healthy discussion at the *European Journal of Applied Physiology* on ways to improve research reporting and scientific thinking.

Compliance with ethical standards

Conflict of interest The authors have no conflicts of interest in relation to the letter.

References

Boutron I, Ravaud P (2018) Misrepresentation and distortion of research in biomedical literature. Proc Natl Acad Sci USA 115:2613–2619

Diong J, Butler AA, Gandevia SC, Héroux ME (2018) Poor statistical reporting, inadequate data presentation and spin persist despite editorial advice. PLoS ONE 13:e0202121

Mira J, Aboodarda SJ, Floreani M et al (2018) Effects of endurance training on neuromuscular fatigue in healthy active men. Part I: strength loss and muscle fatigue. Eur J Appl Physiol 118:2281–2293

Wasserstein RL, Schirm AL, Lazar NA (2019) Moving to a world beyond "*p* %3c 0.05". Am Stat 73(sup1):1–19

Wood J, Freemantle N, King M, Nazareth I (2014) Trap of trends to statistical significance: likelihood of near significant *P* value becoming more significant with extra data. BMJ 348:g2215

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.