EDITORIAL

Imagine a research world without the words *"statistically significant"*. Is it really possible?

Lisa A. Harvey[1] · Martin W. G. Brinkhof[2,3]

To some it may seem unthinkable or even outrageous to suggest that the words *"statistically significant"* be removed from the research vocabulary, but this is exactly what a group of leading statisticians are recommending [1]. They have together just published a landmark series of 43 papers in The American Statistician (the official journal of the American Statistical Association) titled:

> *Statistical Inference in the 21st Century: A World Beyond p < 0.05* [2].

In addition, 800 statisticians (including many of the authors of the 43 papers) have all signed a recommendation published in *Nature* calling on all researchers to stop dichotomising results as significant or not. The provocative title of their call to arms is:

> *Retire statistical significance* [1].

Our two-page editorial cannot do justice to the complexities underlying such a bold recommendation or the nuances of the various arguments and recommendations outlined in the 43 papers. However, our editorial can draw everyone's attention to the issues, and hopefully stimulate a discussion within the spinal cord injury research community about the need to listen to the world's leading statisticians and to develop a much more sophisticated approach to the analyses and interpretation of data than is currently happening.

Some others will no doubt be aware that these issues have been debated ever since the term *"statistical significance"* was introduced by Fisher early last century. There was a public, vigorous and often mocking debate between Fisher, and two other famous statisticians, Pearson and Neyman, during the 1930s and 1940s over this issue and the underlying scientific reasoning (see [3] for a brief overview). Nonetheless, we all ended up inheriting, learning and worshipping the phrase *"statistically significant"*. There have been many pushes by the scientific community to right the wrongs of the past particularly by authors such as Altman and Gardner in the 1980s [4–7], and then more recently by Sterne and others [3, 8]. *Spinal Cord* has also done its small bit over the years to try and raise awareness about this issue. For example, see one of our previous editorials from 2014 titled:

> *Statistical power calculations reflect our love affair with P-values and hypothesis testing: time for a fundamental change* [9].

Yet not much has changed. We still see the phrases *"statistically significant"* or *"not statistically significant"* put forward as though these two phrases say it all. There is hope however that finally things will change. The current very public push in the world's leading multidisciplinary science journal *Nature* is unique, and it might just be the impetus required to make everyone sit up and give this issue some serious consideration. It will require effort on everyone's behalf to come to terms with the alternatives and it will require a lot of researchers to move outside their comfort zones.

There are many places where the novice can read up on the issues. The American Statistician editorial preceding the special collection of 43 papers would be a very good place to start [2]. It begins by providing a list of all the things researchers **should not do**. The list is so important we have included it ad verbatim here:

✉ Lisa A. Harvey
    spinalcord@iscos.org.uk

[1] University of Sydney, Sydney, Australia

[2] Swiss Paraplegic Research, Nottwil, Switzerland

[3] Department of Health Sciences and Health Policy, University of Lucerne, Lucerne, Switzerland

- Don't base your conclusions solely on whether an association or effect was found to be "statistically significant" (i.e., the *p*-value passed some arbitrary threshold such as $p < 0.05$).
- Don't believe that an association or effect exists just because it was statistically significant.
- Don't believe that an association or effect is absent just because it was not statistically significant.
- Don't believe that your *p*-value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.
- Don't conclude anything about scientific or practical importance based on statistical significance (or lack thereof).
- And most importantly….do not say "*statistically significant*" or use any variant, words, asterisks or other statistical trickery to convey the same message. (pages 1–2, [2]).

The American Statistician editorial then goes to great length to say what we **should do** instead. For example, we should report the size (and associated uncertainty) of effects, associations and anything else we measure. We should interpret results in a thoughtful way taking into account context and prior evidence. We should acknowledge that there is uncertainty associated with all estimates, and the list goes on—please read.

We will see over the coming years a fundamental change in the way we all think about statistics, and *Spinal Cord* wants to ensure that it helps to facilitate this change. For now, we won't be banning the phrase "*statistically significant*" but we certainly won't be encouraging it. Instead, we want authors to embrace the reform, and move beyond merely dichotomising results based on some arbitrary *p*-value. We need to also bring readers with us. After all, they are the consumers of research. So everyone has a role to play in ensuring the words "*statistically significant*" are forever removed from our vocabularies. Can you imagine such a research world?

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

References

1. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019;567:305–7.
2. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "p < 0.05". Am Stat. 2019;73:1–19.
3. Sterne JA. Teaching hypothesis tests—time for significant change? Stat Med. 2002;21:985–94.
4. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ. 1995;311:485.
5. Altman DG, Gardner MJ. Statistics with Confidence: Confidence Intervals and Statistical Guidelines. 2nd edn. London: BMJ; 2000.
6. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. BMJ. 1986;292:746–50.
7. Altman D, Machin D, Bryant T, Gardner M (eds). Statistics with confidence. Confidence intervals and statistical guidelines. 2nd edn. (London, BMJ Books, 2000).
8. Sterne JAC, Davey Smith G. Sifting the evidence—what's wrong with significance tests? BMJ. 2001;322:226.
9. Harvey L. Statistical power calculations reflect our love affair with P-values and hypothesis testing: time for a fundamental change. Editorial. 2014;52.