European Spine Journal (2018) 27:975–976
https://doi.org/10.1007/s00586-018-5566-4

EDITORIAL

Is it statistically significant?

Kimberley L. Edwards[1,2]

Received: 4 March 2018 / Accepted: 5 March 2018 / Published online: 17 April 2018

The Editor-in-Chief forwarded me a "Letter to the Editor" and asked me to give some comments. You will find this Letter to the Editor immediately after this Editorial.

Many authors undertake entirely appropriate statistical analyses. However, sometimes analyses can be incomplete, which can lead the author (and reader) to incorrect conclusions about the underlying research question. This can have serious implications for the health of our patients given that we follow an evidenced-based approach to their care.

"Is it statistically significant?" This is a commonly asked question when analysing data. However, we have to look beyond *p* values. First and foremost, is the result clinically meaningful? Then, we need to look at whether one study's result would persist in different samples, because what we often actually want to know is "is Surgery A going to be better or worse than Surgery B for my patient?".

Suppose a new surgical treatment appears to outperform the standard therapy in a research study, then we are interested in assessing whether this apparent effect is likely to be real or could just be a chance finding: *p* values help us to do this. As a reminder, a *p* value is the probability of finding your observed value if the Null Hypothesis (H_0) (that is, a hypothesis of no difference) is true. If you have a large *p* value, then it is increasingly likely that the H_0 was true, i.e., no difference between study groups. Conversely, if you have a very small *p* value (conventionally taken as <0.05), this means it is very unlikely that the result you found was due to chance (assuming that the study was well-conducted).

A *p* value on its own, regardless of size (whether bigger or smaller than the cabalistic 0.05 cut-off), tells the reader nothing about the size of the difference the study found. It does not tell us whether the results are medically important. Rather, it relied solely on the statistical convention of 'if $p<0.05$ then that is good'; but interpreting results is not as simple as that. Rather, in medicine, we need to move away from this dichotomous stance of statistical significance (or not) if we want to optimise patient care [1]. That is, when looking at medical research, we seek to understand not only is, say, surgery A better/worse than surgery B, but rather how much better/worse is it. (Incidentally note how I look both ways, like crossing a road—there is rarely value in undertaking one-sided testing in medicine as that would assume the direction was not important, and clearly whether one treatment is better or worse than the other is vital information).

Instead, we should be focusing on establishing the magnitude of the difference between the groups. I am not the first author (nor will I be the last) to point out this issue. Rothman highlighted this decades ago—a small difference resulting from a study may well be statistically significant simply due to a large sample size; conversely a large, clinically meaningful, difference may be statistically insignificant due to a small sample size [2].

Let's say our study was comparing two different surgical methodologies in terms of the number of days it takes the patient to return to their sport post-operatively. What the surgeon/patient really needs to know is not whether the difference between the two surgeries was statistically significant but how many days difference there was. Consider this—is it better to have 1 day mean difference ($p<0.001$) or 10 days difference ($p=0.04$)? Alternatively, what about 15 days difference ($p=0.06$)?

In determining whether a difference is important, I often recommend that clinical judgement is used. Also, referring to the minimally clinically important difference (MCID), which is often reported for medical outcomes, can be helpful. So if a result is less than the MCID, it does not really matter how small the *p* value is, as the result is not clinically meaningful.

A further problematic area is that of generalisation and whether the results for the described sample would likely be

✉ Kimberley L. Edwards
Kimberley.Edwards@nottingham.ac.uk

[1] Orthopaedics, Trauma and Sports Medicine, School of Medicine, Faculty of Medicine and Health Sciences, University of Nottingham, W/C 1374, West Block, C Floor, Queen's Medical Centre, Nottingham NG7 2UH, UK

[2] Arthritis Research UK Centre for Sport, Exercise and Osteoarthritis, Nottingham University Hospitals Trust, Nottingham, UK

seen in a different sample, that is our own patients. Quantitative research usually uses data from a sample population to seek to be able to know the true value for the population. However, the difference obtained in one study is only an estimate of the population value, and yet that is the figure that we really need. Necessarily, different samples will produce different results. So, what is the real value had we investigated all eligible participants? That is, we need to determine the size of the difference for the population (not just the sample). This is why we need confidence intervals.

A confidence interval tells us the range of the values where we can be confident that the population value (the 'true' value) will lie. So, using a 95% confidence interval, if we were to repeat our experiment 100 times, each time calculating the confidence interval, we would expect 95 of the corresponding confidence intervals to contain the 'true' population mean. (Note, other percentages could be used, but convention means 95% is the norm).

That said, what value you are calculating the confidence interval for is also important. As nicely articulated by Saltychev and Eskola [3] in this issue of European Spine Journal, studies which simply report the confidence intervals for the two mean values separately are not informative. This is because these studies are not telling us about the difference between the two treatments' results, whether or not those results are precise or statistically significant. Thus, it is important that confidence intervals are constructed around the observed difference between the groups.

It is also helpful for studies to provide a measure of risk, for example, by reporting the relative risk (or odds ratio, depending on study type and prevalence of the condition in question) between the groups. Note, when the prevalence of the outcome is rare, the relative risk and odds ratio values are close. However, when prevalence is not rare, the odds ratios tend to produce a more extreme value and its use should be avoided if possible. These ratios are descriptive, not inferential, statistics as they do not determine statistical significance. But, hopefully by now, I have convinced you that statistical significance is not everything.

Relative risk requires the examination of two dichotomous variables, where one variable measures the event (occurred vs. not occurred) and the other variable measures the groups (group 1 vs. group 2). If the relative risk figure equals 1, there is no difference in risk between the groups. If greater than 1, then the event is more likely to occur in the group used as the numerator (the risk figure above the line), and vice versa. For example, if 25/30 patients incurred a postoperative infection following Surgery A and 10/30 for Surgery B, then the relative risk of infection for Surgery A compared to Surgery B is 2.5× (with Surgery A as the numerator). That is, Surgery A has 2.5 times the risk of infection compared to Surgery B. Okay, none of us are having either surgery as the infection risk seems very high, but the data were conjured to illustrate the point!

Hopefully, I have given some food for thought (and maybe a few readers are rushing to edit some previously final drafts of papers!).

Compliance with ethical standards

Conflict of interest The author has no conflicts of interest.

References

1. Gardner MJ, Altman DG (1986) Confidence intervals rather than *P* values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed) 292:746
2. Rothman K (1978) A show of confidence. N Engl J Med 299:1362–1363
3. Saltychev M, Eskola M (2018) Generalising the results—how can we improve our reports? Eur Spine J. https://doi.org/10.1007/s00586-018-5558-4