Author/s:
Davidson, A

Title:
Embracing uncertainty: The days of statistical significance are numbered

Date:
2019-10-01

Citation:
Davidson, A. (2019). Embracing uncertainty: The days of statistical significance are numbered. Paediatric Anaesthesia, 29 (10), pp.978-980. https://doi.org/10.1111/pan.13721.

Persistent Link:
https://hdl.handle.net/11343/286466

PROF ANDREW DAVIDSON (Orcid ID : 0000-0002-7050-7419)

Article type : Editorial

Embracing uncertainty: the days of statistical significance are numbered

Editorial

Andrew Davidson [1,2,3]

1) Department of Anaesthesia, Royal Children's Hospital, Flemington road, Parkville, AUSTRALIA
2) Head of Anaesthesia Research, Murdoch Children's Research Institute, Flemington road, Parkville, ASUTRALIA
3) Department of Paediatrics, University of Melbourne, Parkville, ASUTRALIA

Corresponding address

Prof Andrew Davidson

Department of Anaesthesia, Royal Children's Hospital, 50 Flemington road, Parkville, VIC 3207 AUSTRALIA

Phone +61 3 9345 4008

Email Andrew.davidson@rch.org.au

The interpretation of medical science is going through a fundamental change. This change will have considerable impact because it will change the way we report medical science and transform the way clinicians think and respond to statistical information. It will be a difficult transition, but the rewards are great. The change is nothing to do with genes or cellular activity or new discoveries.

The change is simply getting rid of the phrase "statistical significance".

For years we have known that using a P value of 0.05 as a threshold for statistical significance is an arbitrary choice. There is no scientific or philosophical basis for choosing 0.05. It almost certainly originates from a 1926 paper by Fisher describing an experiment involving manure. He stated no logic for this choice apart from saying "personally, the writer prefers to set a low standard of significance at the 5 percent point".[1] Prominent groups have previously cautioned against its use, but it has persisted.[2] Recently the American Statistical Association again called for it to be abandoned, and in a single issue published 43 articles exploring ways to move to a world without P<0.05.[3]

Some have suggested replacing 0.05 with 0.005 [4]. This may help reduce false assumptions of truth but it ignores the underlying fundamental problem. The problem is the whole concept of "statistical significance". ANY choice of a P value which is "statistically significant" is flawed. Such choices force researchers, authors and commentators to declare whether there is a meaningful effect or not. Such binary decision rules mean studies and findings are categorised as positive or negative; with us or against us. Evidence was either found, or it was not. It is easy to see the problem with this. If we take 0.05 as statistically significant then a finding with 0.05 will be reported as finding evidence which should drive change. A similar study with the *same* effect size, but a P value of 0.06 would be reported as not finding evidence and hence not drive change. This is ludicrous. An editorial in NEJM stated "*the notion that a treatment is effective for a particular outcome if P<0.05 and ineffective if that threshold is not reached is a reductionist view of medicine that does not always reflect reality*".[5] Switching 0.05 to 0.005 doesn't fix the problem. If we decide 0.005 is meaningful, then does that then mean that 0.006 isn't?

Nature has recently published a call for the end of the term "statistical significance". [6] The call has been signed by more than 800 senior medical researchers, biologists and statisticians. The commentary is succinct and convincing; just two pages long. It should be read by all medical

researchers and clinicians. The crux of their argument is to *"never conclude there is 'no difference' or 'no association' just because a P value is larger than a threshold such as 0.05, or equivalently, because a confidence interval includes zero. Neither should we conclude that two studies conflict because one had a statistically significant result and the other did not."*

In such a world without statistical significance, there will always be uncertainty.

How should our journal, *Pediatric Anesthesia,* deal with this new uncertainty? Currently journals are unsure how to do this. Most are likely to embrace uncertainty in stages.

Moving away from a dichotomous finding and embracing uncertainty requires a greater appreciation of the principles underlying the reporting results. This inevitably requires a greater focus on confidence intervals. In any study comparing two groups there will almost inevitably be a difference between the means or medians of the two groups. Similarly risk or odds ratios will almost never be 1. There is nearly always "a difference". When interpreting what this means for changing practice, we first consider two things; the effect size and the degree of precision of the estimate. The effect size is simply how big the difference is; or how large or small the risk or odds ratio. The degree of precision is reflected in the P value or the 95% confidence intervals. The P value indicates the probability of the results (or more extreme results) in the sample if the null hypothesis is true. The 95% confidence intervals are preferred (and indeed mandated in our journal) as they give the reader a better idea of the effect size and the power of the study. But the 95% confidence interval still tempts the author to say the difference is significant or not depending on whether or not zero falls within the interval when comparing means (or 1 falls within the interval if computing a ratio). Insisting that authors report the 95% confidence intervals only partly fixes the problem.

A crucial step is to ask authors not to conclude a result is "positive" or "negative" purely around a P value threshold; *or* bound of a 95% confidence interval. This is going to be a hard habit to break. In some circumstances readers may still want some terminology to describe the size of a P value or strength of evidence for any effect. The optimal terminology to express these levels evidence has yet to be agreed. Some have suggested a P value around 0.05 shows "weak evidence", while a P value less than 0.01 shows "strong evidence" and a P value less than 0.001 shows "very strong evidence". This terminology should not be fixed, and merely provides an example of how authors might provide a qualitative assessment of the strength of evidence.

Another step is to ask the author to address the actual difference between means or medians (or size of the odds or risk ratio), and also highlight the upper and lower bounds of the confidence interval *and* discuss the implications of these. In other words, the author will have to directly address not only the strength of evidence for any effect (the size of the P value), but also the actual

effect size and the precision of the estimated effect. This is important. It will reduce the risk of unwarranted conclusions and make authors and readers think more about what the study actually found.

Thus the *conclusions* drawn about the overall relevance of the results must be based on the P value (*if* they report a P value), the effect size, *and* the bounds of the 95% confidence intervals, rather than simply whether or not the P value is greater or smaller than 0.05. Conclusions will tend to fall into the following. Studies with larger P values and confidence intervals that are wide and span both what may or may not be clinically or scientifically important effects will more likely require tempered conclusions. Studies with confidence intervals that fall largely within a range that may be regarded as clinically or scientifically irrelevant are more likely to conclude there is no meaningful effect, regardless of the size of the P value. Studies whose P values are smaller and 95% confidence intervals span effect sizes that are largely greater than the minimally important clinical or scientific effect size may warrant more definitive conclusions that there is a clinically or scientifically relevant effect. Having to discuss the actual data, rather than focus on the P value, will have the indirect benefit of encouraging researchers to report data that are more easily understood. For example the absolute risk reduction is more easily understood than a risk or odds ratio.

These steps are going to require subjective, honest, judgement on the part of authors and greater vigilance by editors and readers.

Some fear that removing the "gatekeeper" of 0.05 will increase the number of authors or commentators that make claims of certainty when the evidence is particularly weak. This may indeed happen, but such claims are easily refuted by looking at the actual numbers.

Less dichotomous certainty about the relevance of an individual study will also prompt authors to consider their results in a wider context. Good papers already do this. Instead of summarising their results with statements of fact, authors will have to put more emphasis on how their results fit with previous knowledge and hence the overall shift in evidence.

The greatest objection to removing "statistical significance" is the fear that clinicians will not be able to interpret studies. Proponents of "statistical significance" think clinicians need yes-no directives. I think this shows a poor understanding of clinical medicine. Clinicians deal with uncertainty on a daily basis. Biology is complex and clinical medicine is rarely precise. Clinicians are already wary of claims that x causes y. As the medical literature abandons "statistical significance" clinicians will trust that literature to a greater extent because uncertainty is recognised.

Some fear removing “statistical significance” will simply lead to chaos. I think this underestimates the intelligence of our readers. There will inevitably be some initial confusion as we all adjust to an era without dichotomous statements, but the rewards of moving to a more accurate and rational interpretation of data far outweigh the risks of causing some initial confusion. Removing “statistical significance” will certainly spur a whole new area of investigation into how we report and interpret data. That is long overdue.

In summary, *Pediatric Anesthesia* already requires authors to report 95% confidence intervals. We will begin to encourage authors to embrace uncertainty and phase out the concept of dichotomous “statistical significance”. Authors are encouraged to comply with the following recommendations. At this stage they are not mandatory, but authors and papers that embrace uncertainty will be greatly appreciated.

1) Specific P value thresholds such as 0.05 should not drive dichotomous interpretations.
2) If qualitative language is used to describe the strength of evidence for an effect based simply on the size of a P value, the terminology should be justifiable and reasonable.
3) Authors should specifically address the clinical or scientific relevance of the bounds of the 95% confidence intervals.
4) Authors should be appropriately tempered in their conclusions, using language that acknowledges uncertainty where appropriate. The conclusions should be influenced by not only the P value but also the effect size and bounds of the 95% confidence intervals.

Pediatric Anesthesia reports the science behind caring for our most vulnerable cohort, children, at their most vulnerable time, during anaesthesia and critical illness. As such it behoves us to lead and adopt the most rigorous approach to the interpretation of the data we publish.

The way we report and interpret data will evolve substantially over the next decade. We are indeed entering exciting times; albeit somewhat uncertain.

ETHICS - none required

FUNDING - ‘This editorial was funded by departmental resources.'

DISCLOSURES - The author is Editor in Chief of this journal.

1. Fisher RA. The arrangement of field experiments. *Journal of Ministry of Agriculture* 1926; 33: 503-13

2. Wasserstein R, Lazar N. The ASA's statement on p-values: context, process and purpose. *The American Statistician* 2016; 70: 129-33

3. Wasserstein RL, Schrim AL, Lazar NA. Moving to a wold beyond "p<0.05". *The American Statistician* 2019; 73 sup1: 1-19

4. Ioannidis J. The proposal to lower P value thresholds to .005, *JAMA* 2018; 319: 1429-30

5. Harrington D, D'Agostino RB, Gatsonis C, Hogan JW, Hunter DJ, Normand ST, Drazen JM, Hamel MB. New guidelines for statistical reporting in the journal. *New England Journal of Medicine* 2019; 381: 285-6

6. Amrhein V, Greenland S, McShane B. Retire statistical significance. *Nature* 2019; 567: 305-7