S Afr J Surg. 2024;62:239-241. Online first
https://doi.org/10.36303/SAJS.00608

S Afr J Surg
ISSN 038-2361
© 2024 The Author(s)
CHARLES SAINT LECTURE

Can we trust the data? – the data detective

J Carlisle

Perioperative Medicine, Anaesthesia and Intensive Care, Torbay and South Devon NHS Foundation Trust, United Kingdom

Corresponding author, email: *john.carlisle@nhs.net*

Despite the title and content of my talk, I am optimistic for the future of healthcare research. I will return to that sense of optimism in my conclusion. But to cheer you up at the end of my talk I first must depress you. I have included one intentional lie in my talk. See if you can spot it.

"Can we trust the data?"

You already know that I am going to tell you that you cannot trust the data. I have encountered many varieties of false data over the past two decades, some of which I will recount to you. Before I do so, I would like you to consider whether true data ensures reliable research.

Again, you might anticipate that I will tell you that true data does not ensure reliable research. On the contrary, but unfortunately, the interpretation of true data is reliable – the findings are reliably false. How can true data, reported accurately, reliably generate false research findings?

In 2005, John Ioannidis – a professor in four disciplines at Stanford University – published what is now the most cited paper in PLoS Medicine, entitled "Why most research findings are false".[1] The crux of the problem discussed by John is that most healthcare staff, including researchers, do not understand *p*-values.

The *p*-value is only true when there is zero effect. Declarations by researchers that an intervention has a non-zero effect based upon a *p*-value is a non-sequitur. It is a bit like opposition politicians claiming that they will save the country because the government has failed to do so. We do know what *p*-values mean when there is zero effect. We know that a *p*-value of 0.001 is as likely as a *p*-value of 0.301, 0.763 or 0.999. When an intervention has a non-zero effect, we do not know what the *p*-value means – *p*-values closer to zero are more likely than *p*-values closer to 1 for interventions that have more effect, but one cannot conclude from the *p*-value that the intervention has any non-zero effect, including the difference found in the experiment.

A justified assertion then, is that all research findings are false if based upon a *p*-value because they are based upon faulty logic. However, John Ioannidis was not making this general assertion. His specific claim, captured in the title of his paper, was that more than half of papers are wrong to declare an effect when the *p*-value is < 0.05. It is incorrect to assume that a *p*-value threshold of 0.05 limits false findings to 1 in 20.

John listed six corollaries consequent on his mathematical analysis. Research findings are more likely to be false – in small studies, for small effect sizes, when more variables are tested, when more tests are used, when prejudiced by financial and other interests and for novel exciting fields.

With John's assessment we have already chalked up more than 50 false findings per 100 and I have yet to talk to you about false data!

By the way – had you been wondering – John Ioannidis' paper has garnered 8 000 citations by papers indexed in PubMed and 13 000 citations in all. His paper has had a big impact, perhaps even more than indicated by the number of citations. But it is not the most cited paper. Heading the leaderboard of most cited papers is one written by Oliver Lowry and colleagues in 1951.[2] It has the snappy title "Protein measurement with the Folin Phenol reagent". It has been cited by 300 000 papers. John Ioannidis' paper, as yet, barely makes the top 100 most-cited papers.

"Can we trust the data?"

Another leaderboard that is relevant to my talk is called "The Retraction Watch Leaderboard".[3] Retraction Watch is a website, set up in 2010 by two American healthcare journalists, Ivan Oransky and Marcus Adam. Their strapline is "Tracking retractions as a window into the scientific process". And that is exactly what the website does – it logs retracted scientific papers and investigates the stories behind those retractions. Despite logging over 45 000 retractions, there remain 499 scientific papers that have not been retracted for every one that has. Their window on the scientific process is narrow, providing a glimpse into 0.2% of papers, but nevertheless it is an interesting glimpse.

The leaderboard on their website lists the 30 authors who have had the most papers retracted. The author at the bottom of the list, with 29 retractions, is Dr Tongtau Liu, affiliated with Shandong University in China. None of the 30 authors specialise in gastrointestinal medicine or surgery, although some of the retracted papers concern genetic expression in hepatic and pancreatic tumours.

I have become familiar with the papers of some of the authors in that list, which range from diverse scientific fields including the engineering of concrete, the properties of nanoparticles and the psychology of beauty.

My foray into the world of detecting false data started with nausea and vomiting. One of the first signs of the pregnancy was being sick – the mother decided that she would orphan the review that she had registered with Cochrane so that she could focus on being a parent. I was asked to foster the review in 2003, as a recently appointed editor for the

Cochrane Anaesthesia Review Group. The title of the review was – ironically – “Drugs for Preventing Postoperative Nausea and Vomiting”[4] {note this is the retraction notice, the reason for which will become apparent}.

I included 737 of the thousands of randomised controlled trials that I reviewed. For two years papers photocopied in the library covered the dining room table at home – digital formats of scientific papers were not widely accessible at the time, first appearing three years earlier in 2000. 2004 was the midst of my review and a very important year for anaesthetists, as the Times published its first Sudoku. Had I not been staring hard at tables and figures about patients throwing up, I would have been counting from 1 to 9.

2006 saw the publication of my Cochrane review, in which I discussed 68 of the 737 trials – these had been authored by a Japanese anaesthetist called Yoshitaka Fujii. Rates of secondary outcomes reported by Fujii were often uniform – for instance, 3 of 40 participants had a headache after placebo, and so too did 3 of 40 participants in two groups that had had an antiemetic drug. This pattern of uniformity was repeated across many of his 68 trials. In a letter that had been published in 2000, a German anaesthetic professor, Peter Kranke, had reported the unlikely statistics associated with this pattern, but his letter was insufficient to trigger any investigation or retraction.[5] Excluding trials by Fujii from the review altered the effect of a drug called granisetron. However, the lawyers at Cochrane judged that I should include his studies in my primary analysis because of the lack of formal investigation or retraction and their concern about litigation.

How then is Fujii today listed at second place on the Retraction Watch leaderboard, with 172 retracted papers?

Despite getting on with my life, I couldn’t quite let go of the apparent problems with Fujii’s papers. In 2010, I wrote a letter to Steve Yentis, the editor in chief of the UK journal, *Anaesthesia*. I challenged him to investigate Fujii’s work, perhaps by assembling a team of people who knew what they were doing. Instead of assembling the Avengers, Steve challenged me straight back – “John, why don’t you have a go at analysing Fujii’s work for false data?”

Rather than tell Steve that I didn’t know what I was doing I decided to think. I remembered that the *p*-value only has meaning when there is no underlying difference between groups. We understand what *p*-values mean for only one part of a randomised controlled trial – the baseline differences documented in table one. We know that the distribution of *p*-values from table one should be uniform – half of values between 0 and 0.5 and half between 0.5 and 1; one tenth of values between 0 and 0.1, one tenth between 0.1 and 0.2 and so on; one hundredth of values between 0 and 0.01 and so on up to one hundredth between 0.99 and 1. Et cetera.

The distribution of *p*-values from table one in Fujii’s papers was very different to uniform.[6] One might expect to encounter such deviation from the uniform in fewer than one analysis in a decillion analyses, a rate equivalent to one atom selected from all the atoms in all the human bodies on earth (plus the few in orbit). I published my analysis, an investigation ensued, Fujii went from Tokyo to Fukushima and for a time he reigned supreme on Retraction Watch’s leaderboard.

Techniques other than analysing table one help identify false data in published research. Discrepancies between protocols and published papers; recycling of the same graphs, numbers and photographs; incorrect statistics – for instance *p*-values discrepant by many orders of magnitude compared with the correct value; incredible results; impossible patient characteristics; unfeasible rates of recruitment – for instance rare diseases by a single hospital; and so on.

However, on the whole, published papers offer little opportunity to identify false data. Hundreds of individual values are summarised as a single mean, and complex data patterns in the individual patient data spreadsheet never see the light of day. In 2017, I published an analysis of baseline data from over 5 000 published randomised controlled trials.[7] I detected aberrant data in about 1 in 20 trials (or 5%). With a few noticeable exceptions, such as the retraction from the *New England Journal of Medicine* of the largest trial on diet, my analysis of these 5 000 trials had little impact.[8]

Last week I completed my 12 years’ term as an editor for the journal *Anaesthesia*. On occasion I had asked authors of submitted papers to provide spreadsheets of individual patient data. Sometimes these spreadsheets contained false data, for instance discrepancies with the summary data in the paper, incorrect calculation of a derivative – such as body mass index from height and weight, incorrect categorisation of continuous variables, and so on. Sometimes spreadsheets had whole rows copied. Sometimes fragments of rows or columns were duplicated. Sometimes there was a surplus or deficit of certain numerals. Sometimes the hard drive holding the spreadsheet had been corrupted, damaged in a fire, confiscated by border control or squashed in an earthquake. I kid you not.

Rarely was the false data in these spreadsheets detectable in the paper. In 2019, I decided that authors who submitted randomised controlled trials to *Anaesthesia* should also send us spreadsheets of individual patient data. The results were striking. When I analysed 153 spreadsheets, I found false data in 67 (44%).[9] The false data was bad enough in 40 trials (26%) that the results of the trial were compromised, as was my belief that the trial was credible. I think it reasonable to assume that *Anaesthesia* is not a special case journal. I think that at least one quarter of papers submitted to any journal are sufficiently compromised to invalidate their credibility.

So, “Can we trust the data?”

No. Consider 100 trials. My work suggests that 30 are literally incredible. John Ioannidis’ work suggests that the results of 40 of the remaining 70 trials are misinterpreted due to the illogic of frequentist statistics. And how many of the remaining 30 trials tell you something useful?

Now for a dash of optimism. Publishers, editors, researchers and readers now know that there is a problem. Recognising a problem is the start of finding solutions. Rates of retraction are increasing – perhaps Chinese papermills are flooding us with fantasies; or perhaps we are getting better at detecting fraud. More papers written by anaesthetists have been retracted than any other medical specialty. There are four anaesthetists in the Retraction Watch leaderboard of 30 authors, in first, second, third and eighth position. They are responsible for 546 of the 1 697 retracted papers in that list. I am unsure of whether I should be proud because I helped detect these fraudulent studies, or ashamed for being an anaesthetist, who either lie a lot or lie unconvincingly.

Of course, fabricators of submitted articles and submitted spreadsheets know that there is an arms race. We find solutions – they improve subterfuge. Artificial intelligence will make it easier to generate papers and spreadsheets that evade the most effective defences. Unfortunately, researchers, editors and readers will remain unable to reliably distinguish fact from fiction. I think that we will have to depend upon the institutions that generate the research to find solutions. Solutions might include cryptographed blockchain ledgers that track each datum, from acquisition, for instance the end of an endoscope or the dispensing of a drug, to their analysis, without the intervention of bad actors, whether human or machine. Institutions could be graded by the quality of their data probity and its resistance to cyber-attack. 'Read only' data will be accessible to independents, both humans and machines.

And what of the problem John Ioannidis publicised? We need to rethink statistics so that we generate and use information most efficiently – no threshold hypothesis tests; no declarations of non-zero effect; no negative studies; no positive studies. I commend you to the online resource "Statistical rethinking" by Richard McElreath; it is liberating.[10] I can also recommend the work by Frank Harrell on efficient trial design, regression modelling strategies and Bayesian inference.[11]

Humans have evolved subterfuge to survive in our social species – too much lying and too little lying lands us in trouble. I might have made many false statements in my talk, but only one was intentional.

I hope by reading this I have provided you with food for thought, if not for statistics.

PS. The intentional lie is itself. I did not otherwise include an intentional lie, so my statement that I did was untrue.

ORCID

J Carlisle https://orcid.org/0000-0003-0420-5241

REFERENCES

1. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2(8):e124. https://doi.org/10.1371/journal.pmed.0020124.
2. Lowry OH, Rosebrough NJ, Farr AL, Randall RJ. Protein measurement with the Folin phenol reagent. J Biol Chem. 1951;193(1):265-75. https://doi.org/10.1016/S0021-9258(19)52451-6.
3. https://retractionwatch.com/the-retraction-watch-leaderboard/. Accessed 18 May 2024.
4. Carlisle J, Stevenson CA. Withdrawn: Drugs for preventing postoperative nausea and vomiting. Retraction notice. https://pubmed.ncbi.nlm.nih.gov/28715610/.
5. Kranke P, Apfel CC, Roewer N, Fujii Y. Reported data on granisetron and postoperative nausea and vomiting by Fujii et al. Are incredibly nice! Anesth Analg. 2000;90(4):1004-7. https://doi.org/10.1213/00000539-200004000-00053.
6. Carlisle JB. The analysis of 168 randomised controlled trials to test data integrity. Anaesthesia. 2012;67(5):521-37. https://doi.org/10.1111/j.1365-2044.2012.07128.x.
7. Carlisle JB. Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. Anaesthesia. 2017;72(8):944-52. https://doi.org/10.1111/anae.13938. Erratum in: Anaesthesia. 2018;73(9):1176. 28580651. https://doi.org/10.1111/anae.13938.
8. Estruch R, Ros E, Salas-Salvado J, et al. Retraction and republication: Primary prevention of cardiovascular disease with a Mediterranean diet. N Engl J Med. 2018;378:2441-2. https://doi.org/10.1056/NEJMc1806491.
9. Carlisle JB. False individual patient data and zombie randomised controlled trials submitted to Anaesthesia. Anaesthesia. 2021;76(4):472-9. https://doi.org/10.1111/anae.15263.
10. McElreath R. "Statistical Rethinking". https://www.youtube.com/watch?v=FdnMWdICdRs. Accessed 18 May 2024.
11. Harrell F. https://hbiostat.org/. Accessed 18 May 2024.