The data detective

Anaesthetist John Carlisle has spotted problems in hundreds of research papers — and spurred a leading medical journal to change its practice.

BY DAVID ADAM

If John Carlisle had a cat flap, scientific fraudsters might rest easier at night. Carlisle routinely rises at 4.30 a.m. to let out Wizard, the family pet. Then, unable to sleep, he reaches for his laptop and starts typing up data from published papers on clinical trials. Before his wife's alarm clock sounds 90 minutes later, he has usually managed to fill a spreadsheet with the ages, weights and heights of hundreds of people — some of whom, he suspects, never actually existed.

By day, Carlisle is an anaesthetist working for England's National Health Service in the seaside town of Torquay. But in his spare time, he roots around the scientific record for suspect data in clinical research. Over the past decade, his sleuthing has included trials used to investigate a wide range of health issues, from the benefits of specific diets to guidelines for hospital treatment. It has led to hundreds of papers being retracted and corrected, because of both misconduct and mistakes. And it has helped to end the careers of some large-scale fakers: of the six scientists worldwide with the most retractions, three were brought down using variants of Carlisle's data analyses.

"His technique has been shown to be incredibly useful," says Paul Myles, director of anaesthesia and perioperative medicine at the Alfred hospital in Melbourne, Australia, who has worked with Carlisle to examine research papers containing dodgy statistics. "He's used it to demonstrate some major examples of fraud."

Carlisle's statistical sideline is not popular with everyone. Critics argue that it has sometimes led to the questioning of papers that aren't obviously flawed, resulting in unjustified suspicion.

But Carlisle believes that he is helping to protect patients, which is why he spends his spare time poring over others' studies. "I do it because my curiosity motivates me to do so," he says, not because of an overwhelming zeal to uncover wrongdoing: "It's important not to become a crusader against misconduct."

Together with the work of other researchers who doggedly check academic papers, his efforts suggest that the gatekeepers of science — journals and institutions — could be doing much more to spot mistakes. In medical trials, the kind that Carlisle focuses on, that can be a matter of life and death.

ANAESTHETISTS BEHAVING BADLY

Torquay looks like any other traditional provincial English town, with pretty floral displays on the roundabouts and just enough pastel-coloured cottages to catch the eye. Carlisle has lived in the area for 18 years and works at the town's general hospital. In an empty operating theatre, after a patient has just been stitched up and wheeled away, he explains how he began to look for faked data in medical research.

More than ten years ago, Carlisle and other anaesthesiologists began chattering about results published by a Japanese researcher, Yoshitaka Fujii. In a series of randomized controlled trials (RCTs), Fujii, who then worked at Toho University in Tokyo, claimed to have examined the impact of various medicines on preventing vomiting and nausea in patients after surgery. But the data looked too clean to be true. Carlisle, one among many concerned, decided to check the figures, using statistical tests to pick up unlikely patterns in the data. He showed in 2012 that, in many cases, the likelihood of the patterns having arisen by chance was "infinitesimally small"[1]. Prompted in part by this analysis, journal editors asked Fujii's present and former universities to investigate; Fujii was fired from Toho University in 2012 and had 183 of his papers retracted, an all-time record. Four years later, Carlisle co-published an analysis of results from another Japanese anaesthesiologist, Yuhji Saitoh — a frequent co-author of Fujii's — and demonstrated that his data were extremely suspicious, too[2]. Saitoh currently has 53 retractions (see go.nature.com/2jxtgxf).

"I do it because my curiosity motivates me to do so."

Other researchers soon cited Carlisle's work in their own analyses, which used variants of his approach. In 2016, researchers in New Zealand and the United Kingdom, for example, reported problems in papers by Yoshihiro Sato, a bone researcher at a hospital in southern Japan[3]. That ultimately led to 27 retractions, and 66 Sato-authored papers have been retracted in total.

Anaesthesia had been rocked by several fraud scandals before Fujii and Saitoh's cases — including that of German anaesthetist Joachim Boldt, who has had more than 90 papers retracted. But Carlisle began to wonder whether only his own field was at fault. So he picked eight leading journals and, working in his spare moments, checked through thousands of randomized trials they had published.

In 2017, he published an analysis in the journal *Anaesthesia* stating that he had found suspect data in 90 of more than 5,000 trials published over 16 years[4]. At least ten of these papers have since been retracted and six corrected, including a high-profile study published in *The New England Journal of Medicine* (*NEJM*) on the health benefits of the Mediterranean diet. In that case, however, there was no suggestion of fraud: the authors had made a mistake in how they randomized participants. After the authors removed erroneous data, the paper was republished with similar conclusions[5].

Carlisle has kept going. This year, he warned about dozens of anaesthesia studies by an Italian surgeon, Mario Schietroma at the University of L'Aquila in central Italy, saying that they were not a reliable basis for clinical practice[6]. Myles, who worked on the report with Carlisle, had raised the alarm last year after spotting suspicious similarities in the raw data for control and patient groups in five of Schietroma's papers.

The challenges to Schietroma's claims have had an impact in hospitals around the globe. The World Health Organization (WHO) cited Schietroma's work when, in 2016, it issued a recommendation that anaesthetists should routinely boost the oxygen levels they deliver to patients during and after surgery, to help reduce infection. That was a controversial call: anaesthetists know that in some procedures,

EMLI BENDIXEN FOR NATURE

too much oxygen can be associated with an increased risk of complications — and the recommendations would have meant hospitals in poorer countries spending more of their budgets on expensive bottled oxygen, Myles says.

The five papers Myles warned about were quickly retracted, and the WHO revised its recommendation from ‘strong’ to ‘conditional’, meaning that clinicians have more freedom to make different choices for various patients. Schietroma says his calculations were assessed by an independent statistician and through peer review, and that he purposely selected similar groups of patients, so it’s not surprising if the data closely match. He also says he lost raw data and documents related to the trials when L’Aquila was struck by an earthquake in 2009. A spokesperson for the university says it has left enquiries to “the competent investigating bodies”, but did not explain which bodies those were or whether any investigations were under way.

SPOTTING UNNATURAL DATA

The essence of Carlisle’s approach is nothing new, he says: it’s simply that real-life data have natural patterns that artificial data struggle to replicate. Such phenomena were spotted in the 1880s, were popularized by the US electrical engineer and physicist Frank Benford in 1938, and have since been used by many statistical checkers. Political scientists, for example, have long used a similar approach to analyse survey data — a technique they call Stouffer’s method after sociologist Samuel Stouffer, who popularized it in the 1950s.

In the case of RCTs, Carlisle looks at the baseline measurements that describe the characteristics of the groups of volunteers in the trial, typically the control group and the intervention group. These include height, weight and relevant physiological characteristics — usually described in the first table of a paper.

In a genuine RCT, volunteers are randomly allocated to the control or (one or more) intervention groups. As a result, the mean and the standard deviation for each characteristic should be about the same — but not too identical. That would be suspiciously perfect.

Carlisle first constructs a *P* value for each pairing: a statistical measurement of how likely the reported baseline data points are if one assumes that volunteers were, in fact, randomly allocated to each group. He then pools all these *P* values to get a sense of how random the measurements are overall. A combined *P* value that looks too high suggests that the data are suspiciously well-balanced; too low and it could show that the patients have been randomized incorrectly.

The method isn’t foolproof. The statistical checks demand that the variables in the table are truly independent — whereas in reality, they often aren’t. (Height and weight are linked, for example.) In practice, this means that some papers that are flagged up as incorrect actually aren't — and for that reason, some statisticians have criticized Carlisle's work.

But Carlisle says that applying his method is a good first step, and one that can highlight studies that might deserve a closer look, such as requesting the individual patient data behind the paper.

"It can put up a red flag. Or an amber flag, or five or ten red flags to say this is highly unlikely to be real data," says Myles.

MISTAKES VERSUS MISCREANTS

Carlisle says that he is careful not to attribute any cause to the possible problems he identifies. In 2017, however, when Carlisle's analysis of 5,000 trials appeared in *Anaesthesia* — of which he is an editor — an accompanying editorial by anaesthetists John Loadsman and Tim McCulloch at the University of Sydney in Australia took a more provocative line[7].

It talked of "dishonest authors" and "miscreants" and suggested that "more authors of already published RCTs will eventually be getting their tap on the shoulder". It also said: "A strong argument could be made that every journal in the world now needs to apply Carlisle's method to all the RCTs they've ever published."

This provoked a strongly worded response from editors at one journal, *Anesthesiology*, which had published 12 of the papers Carlisle highlighted as problematic. "The Carlisle article is ethically questionable and a disservice to the authors of the previously published articles 'called out' therein," wrote the journal's editor-in-chief, Evan Kharasch, an anaesthesiologist at Duke University in Durham, North Carolina[8]. His editorial, co-written with anaesthesiologist Timothy Houle at Massachusetts General Hospital in Boston, who is the statistical consultant for *Anesthesiology*, highlighted problems such as the fact that the method could flag up false positives. "A valid method to detect fabrication and falsification (akin to plagiarism-checking software) would be welcome. The Carlisle method is not such," they wrote in a correspondence to *Anaesthesia*[9].

In May, *Anesthesiology* did correct one of the papers Carlisle had highlighted, noting that it had reported "systematically incorrect" *P* values in two tables, and that the authors had lost the original data and couldn't recalculate the values. Kharasch, however, says he stands by his view in the editorial. Carlisle says Loadsman and McCulloch's editorial was "reasonable" and that the criticisms of his work don't undermine its value. "I'm comfortable thinking the effort worthwhile whilst others might not," he says.

THE DATA CHECKERS

Carlisle's isn't the only method to emerge in the past few years for double-checking published data.

Michèle Nuijten, who studies analytical methods at Tilburg University in the Netherlands, has developed what she calls a "spellcheck for statistics" that can scan journal articles to check whether the statistics described are internally consistent. Called statcheck, it verifies, for example, that data reported in the results section agree with the calculated *P* values. It has been used to flag errors, usually numerical typos, in journal articles going back decades.

And Nick Brown, a graduate student in psychology at the University of Groningen, also in the Netherlands, and James Heathers, who studies scientific methods at Northeastern University in Boston, Massachusetts, have used a program called GRIM to double-check the calculation of statistical means, as another way to flag suspect data.

Neither technique would work on papers that describe RCTs, such as the studies Carlisle has assessed. Statcheck runs on the strict data-presentation format used by the American Psychological Association. GRIM works only when data are integers, such as the discrete numbers generated in psychology questionnaires, when a value is scored from 1 to 5.

"Trying to find flaws in other people's work is not something that will make you very popular."

There is growing interest in these kinds of checks, says John Ioannidis at Stanford University in California, who studies scientific methods and advocates for the better use of statistics to improve reproducibility in science. "They are wonderful tools and very ingenious." But he cautions about jumping to conclusions over the reason for the problems found. "It's a completely different landscape if we're talking about fraud versus if we're talking about some typo," he says.

Brown, Nuijten and Carlisle all agree that their tools can only highlight problems that need to be investigated. "I really don't want to associate statcheck with fraud," says Nuijten. The true value of such tools, Ioannidis says, will be to screen papers for problematic data before they are published — and so prevent fraud or mistakes reaching the literature in the first place.

Carlisle says an increasing number of journal editors have contacted him about using his technique in this way. Currently, most of this effort is done unofficially on an ad hoc basis, and only when editors are already suspicious.

At least two journals have taken things further and now use the statistical checks as part of the publication process for all papers. Carlisle's own journal, *Anaesthesia*, uses it routinely, as do editors at the *NEJM*. "We are looking to prevent a rare, but potentially impactful, negative event," a spokesperson for the *NEJM* says. "It is worth the extra time and expense."

Carlisle says he is very impressed that a journal with the status of the *NEJM* has introduced these checks, which he knows at first hand are laborious, time-consuming and not universally popular. But automation would be needed to introduce them on the scale required to check even a fraction of the roughly two million papers published across the world each year, he says. He thinks it could be done. Statcheck works in this way, and is being used routinely by several psychology journals to screen submissions, Nuitjen says. And text-mining techniques have allowed researchers to assess, for instance, the *P* values in thousands of papers as a way to investigate *P*-hacking — in which data are tweaked to produce significant *P* values.

One problem, several researchers in the field say, is that funders, journals and many in the scientific community give a relatively low priority to such checks. "It is not a very rewarding type of work to do," Nuijten says. "It's you trying to find flaws in other people's work, and that is not something that will make you very popular."

Even finding that a study is fraudulent does not always end the matter. In 2012, researchers in South Korea submitted to *Anesthesia & Analgesia* a report of a trial that looked at how facial muscle tone could indicate the best time to insert breathing tubes into the throat. Asked, unofficially, to take a look, Carlisle found discrepancies between patient and summary data, and the paper was rejected.

Remarkably, it was then submitted to Carlisle's own journal with different patient data — but Carlisle recognized the paper. It was rejected again, and editors on both journals contacted the authors and their institutions with their concerns. To Carlisle's astonishment, a few months later the paper — unchanged from the last version — was published in the *European Journal of Anaesthesiology*. After Carlisle shared the paper's dubious history with the journal editor, it was retracted in 2017 because of "irregularities in their data, including misrepresentation of results"[10].

After seeing so many cases of fraud, alongside typos and mistakes, Carlisle has developed his own theory of what drives some researchers to make up their data. "They think that random chance on this occasion got in the way of the truth, of how they know the Universe really works," he says. "So they change the result to what they think it should have been."

As Carlisle has shown, it takes a determined data checker to spot the deception. ■

David Adam *is a science journalist based in London.*

1. Carlisle, J. B. *Anaesthesia* **67**, 521–537 (2012).
2. Carlisle, J. B. & Loadsman, J. A. *Anaesthesia* **72**, 17–27 (2017).
3. Bolland, M. J., Avenell, A., Gamble, G. D. & Grey, A. *Neurology* **87**, 2391–2402 (2016).
4. Carlisle, J. B. *Anaesthesia* **72**, 944–952 (2017).
5. Estruch, R. *et al. N. Engl. J. Med.* **378**, e34 (2018).
6. Myles, P. S., Carlisle, J. B. & Scarr, B. *Anaesthesia* **74**, 573–584 (2019).
7. Loadsman, J. A. & McCulloch, T. J. *Anaesthesia* **72**, 931–935 (2017).
8. Kharasch, E. D. & Houle, T. T. *Anesthesiology* **127**, 733–737 (2017).
9. Kharasch, E. D. & Houle, T. T. *Anaesthesia* **73**, 125–126 (2018).
10. *Eur. J. Anaesthesiol.* **34**, 249 (2017).