ILLUSTRATION BY PIOTR KOWALCZYK

HOW MANY CLINICAL TRIALS CAN'T BE TRUSTED?

Investigations suggest that, in some fields, at least one-quarter of clinical trials might be problematic or fake, warn researchers. They urge stronger scrutiny. **By Richard Van Noorden**

How many clinical-trial studies in medical journals are fake or fatally flawed? In October 2020, John Carlisle reported a startling estimate[1].

Carlisle, an anaesthetist who works for England's National Health Service, is renowned for his ability to spot dodgy data in medical trials. He is also an editor at the journal *Anaesthesia*, and in 2017, he decided to scour all the manuscripts he handled that reported a randomized controlled trial (RCT) — the gold standard of medical research. Over three years, he scrutinized more than 500 studies[1].

For more than 150 trials, Carlisle got access to anonymized individual participant data (IPD). By studying the IPD spreadsheets, he judged that 44% of these trials contained at least some flawed data: impossible statistics, incorrect calculations or duplicated numbers or figures, for instance. And 26% of the papers had problems that were so widespread that the trial was impossible to trust, he judged — either because the authors were incompetent, or because they had faked the data.

Carlisle called these 'zombie' trials because they had the semblance of real research, but closer scrutiny showed they were actually hollow shells, masquerading as reliable information. Even he was surprised by their prevalence. "I anticipated maybe one in ten," he says.

When Carlisle couldn't access a trial's raw data, however, he could study only the aggregated information in the summary tables. Just 1% of these cases were zombies, and 2% had flawed data, he judged (see 'The prevalence of 'zombie' trials'). This finding alarmed him, too: it suggested that, without access to the IPD — which journal editors usually don't request and reviewers don't see — even an experienced sleuth cannot spot hidden flaws.

"I think journals should assume that all submitted papers are potentially flawed and editors should review individual patient data before publishing randomised controlled trials," Carlisle wrote in his report.

Carlisle rejected every zombie trial, but by now, almost three years later, most have been published in other journals — sometimes with different data to those submitted with the manuscript he had seen. He is writing to journal editors to alert them, but expects that little will be done.

Do Carlisle's findings in anaesthesiology extend to other fields? For years, a number of scientists, physicians and data sleuths have argued that fake or unreliable trials are frighteningly widespread. They've scoured RCTs in various medical fields, such as women's health, pain research, anaesthesiology, bone health and COVID-19, and have found dozens or hundreds of trials with seemingly statistically impossible data. Some, on the basis of their personal experiences, say that one-quarter of trials being untrustworthy might be an underestimate. "If you search for all randomized trials on a topic, about a third of the trials will be fabricated," asserts Ian Roberts, an epidemiologist at the London School of Hygiene & Tropical Medicine.

The issue is, in part, a subset of the notorious paper-mill problem: over the past decade, journals in many fields have published tens of thousands of suspected fake papers, some of which are thought to have been produced by third-party firms, termed paper mills.

But faked or unreliable RCTs are a particularly dangerous threat. They not only are about medical interventions, but also can be laundered into respectability by being included in meta-analyses and systematic reviews, which thoroughly comb the literature to assess evidence for clinical treatments. Medical guidelines often cite such assessments, and physicians look to them when deciding how to treat patients.

Ben Mol, who specializes in obstetrics and gynaecology at Monash University in Melbourne, Australia, argues that as many as 20–30% of the RCTs included in systematic reviews in women's health are suspect.

Many research-integrity specialists say that the problem exists, but its extent and impact are unclear. Some doubt whether the issue is as bad as the most alarming examples suggest. "We have to recognize that, in the field of high-quality evidence, we increasingly have a lot of noise. There are some good people championing that and producing really scary statistics. But there are also a lot in the academic community who think this is scaremongering," says Žarko Alfirević, a specialist in fetal and maternal medicine at the University of Liverpool, UK.

This year, he and others are conducting more studies to assess how bad the problem is. Initial results from a study led by Alfirević are not encouraging.

Laundering fake trials

Medical research has always had fraudsters. Roberts, for instance, first came across the issue when he co-authored a 2005 systematic review for the Cochrane Collaboration, a prestigious group whose reviews of medical research evidence are often used to shape clinical practice. The review suggested that high doses of a sugary solution could reduce death after head injury. But Roberts retracted it[2] after doubts arose about three of the key

trials cited in the paper, all authored by the same Brazilian neurosurgeon, Julio Cruz. (Roberts never discovered whether the trials were fake, because Cruz died by suicide before investigations began. Cruz's articles have not been retracted.)

A more recent example is that of Yoshihiro Sato, a Japanese bone-health researcher. Sato, who died in 2016, fabricated data in dozens of trials of drugs or supplements that might prevent bone fracture. He has 113 retracted papers, according to a list compiled by the website Retraction Watch. His work has had a wide impact: researchers found that 27 of Sato's retracted RCTs had been cited by 88 systematic reviews and clinical guidelines, some of which had informed Japan's recommended treatments for osteoporosis[3].

Some of the findings in about half of these reviews would have changed had Sato's trials been excluded, says Alison Avenell, a medical researcher at the University of Aberdeen, UK. She, along with medical researchers Andrew Grey, Mark Bolland and Greg Gamble, all at the University of Auckland in New Zealand, have pushed universities to investigate Sato's work and monitored its influence. "It probably diverted people from being given more effective treatment for fracture prevention," Avenell says.

The concerns over zombie trials, however, are beyond individual fakers flying under the radar. In some fields, swathes of RCTs from different research groups might be unreliable, researchers worry.

During the pandemic, for instance, a flurry of RCTs was conducted into whether ivermectin, an anti-parasite drug, could treat COVID-19. But researchers who were not involved have since pointed out data flaws in many of the studies, some of which have been retracted. A 2022 update of a Cochrane review argued that more than 40% of these RCTs were untrustworthy[4].

"Untrustworthy work must be removed from systematic reviews," says Stephanie Weibel, a biologist at the University of Wuerzberg in Germany, who co-authored the review.

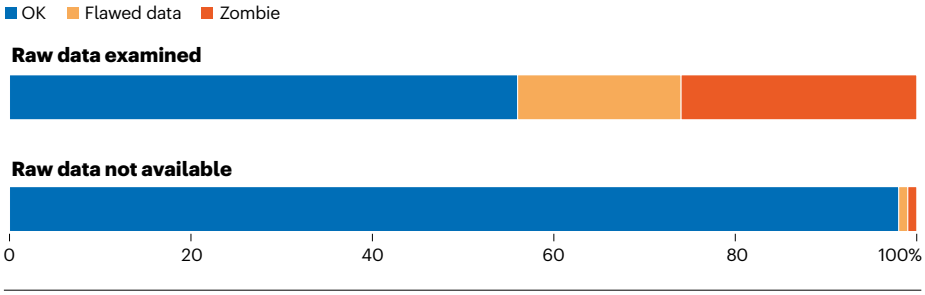
In maternal health — another field seemingly rife with problems — Roberts and Mol have flagged studies into whether a drug called tranexamic acid can stem dangerously heavy bleeding after childbirth. Every year, around 14 million people experience this condition, and some 70,000 die: it is the world's leading cause of maternal death.

In 2016, Roberts reviewed evidence for using tranexamic acid to treat serious blood loss after childbirth. He reported that many of the 26 RCTs investigating the drug had serious flaws. Some had identical text, others had data inconsistencies or no records of ethical approval. Some seemed not to have adequately randomized the assignment of their participants to control and treatment groups[5].

When he followed up with individual authors to ask for more details and raw data, he generally got no response or was told that records were missing or had been lost because of computer theft. Fortunately, in 2017, a large, high-quality multi-centre trial, which Roberts helped to run, established that the drug was effective[6]. It's likely, says Roberts, that in these and other such cases, some of the dubious trials were copycat fraud — researchers saw that a large trial was going on and produced small, substandard copies that no one would question. This kind of fraud isn't a victimless crime, however. "It results in narrowed confidence intervals such that the results look much more certain than they are. It also has the potential to amplify a wrong result, suggesting that treatments work when they don't," he says.

THE PREVALENCE OF 'ZOMBIE' TRIALS

More than one-quarter of a subset of manuscripts describing randomized clinical trials submitted to the journal *Anaesthesia* between 2017 and 2020 seemed to be faked or fatally flawed when their raw data could be examined, editor John Carlisle reported. He called these 'zombies'. But when their raw data could not be obtained, Carlisle could label only 1% as zombies.



REF. 1

> "UNTRUSTWORTHY WORK MUST BE REMOVED FROM SYSTEMATIC REVIEWS."

That might have happened for another question: what if doctors were to inject the drug into everyone undergoing a caesarean, just after they give birth, as a preventative measure? A 2021 review[7] of 36 RCTs investigating this idea, involving a total of more than 10,000 participants, concluded that this would reduce the risk of heavy blood loss by 60%.

Yet this April, an enormous US-led RCT with 11,000 people reported only a slight and not statistically significant benefit[8].

Mol thinks problems with some of the 36 previous RCTs explains the discrepancy. The 2021 meta-analysis had included one multi-centre study in France of more than 4,000 participants, which found a modest 16% reduction in severe blood loss, and another 35 smaller, single-centre studies, mostly conducted in India, Iran, Egypt and China, which collectively estimated a 93% drop. Many of the smaller RCTs were untrustworthy, says Mol, who has dug into some of them in detail.

It's unclear whether the untrustworthy studies affected clinical practice. The World Health Organization (WHO) recommends using tranexamic acid to treat blood loss after childbirth, but it doesn't have a guideline on preventive administration.

From four trials to one

Mol points to a different example in which untrustworthy trials might have influenced clinical practice. In 2018, researchers published a Cochrane review[9] on whether giving steroids to people due to undergo caesarean-section births helped to reduce breathing problems in their babies. Steroids are good for a baby's lungs but can harm the developing brain, says Mol; benefits generally outweigh harms when babies are born prematurely, but the balance is less clear when steroids are used later in pregnancy.

The authors of the 2018 review, led by Alexandros Sotiriadis, a specialist in maternal–fetal medicine at the Aristotle University of Thessaloniki in Greece, analysed the evidence for administering steroids to people delivering by caesarean later in pregnancy. They ended up with four RCTs: a British study from 2005 with more than 940 participants, and three Egyptian trials conducted between 2015 and 2018 that added another 3,000 people into the evidence base. The review concluded that the steroids "may" reduce rates of breathing problems; it was cited in more than 200 documents and some clinical guidelines.

In January 2021, however, Mol and others, who had looked in more depth into the papers, raised concerns about the Egyptian trials. The largest study, with nearly 1,300 participants, was based on the second author's thesis, he noted — but the trial end dates in the thesis differed from the paper. And the reported ratio of male to female babies was an impossible 40% to 60%. Mol queried the other papers, too, and wrote to the authors, but says he did not get satisfactory replies. (One author told him he'd lost the data when moving house.) Mol's team also reported statistical issues with some other works by the same authors.

In December 2021, Sotiriadis's team updated

its review[10]. But this time, it adopted a new screening protocol. Until that year, Cochrane reviews had aimed to include all relevant RCTs; if researchers spotted potential issues with a trial, using a 'risk of bias' checklist, they would downgrade their confidence in its findings, but not remove it from their analysis. But in 2021, Cochrane's research-integrity team introduced new guidance: authors should try to identify 'problematic' or 'untrustworthy' trials and exclude them from reviews. Sotiriadis's group now excluded all but the British research. With only one trial left, there was "insufficient data" to draw firm conclusions about the steroids, the researchers said.

By last May, as Retraction Watch reported, the large Egyptian trial was retracted (to the disagreement of its authors). The journal's editors wrote in the retraction notice that they had not received its data or a satisfactory response from the authors, adding that "if the data is unreliable, women and babies are being harmed". The other two trials are still under investigation by publisher Taylor & Francis as part of a larger case of papers, says Sabina Alam, director of publishing ethics at the firm. Before the 2018 review, some clinical guidelines had suggested that administering steroids later in pregnancy could be beneficial, and the practice had been growing in some countries, such as Australia, Mol has reported. The latest updated WHO and regional guidelines, however, recommend against this practice.

Overall, Mol and his colleagues have alleged problems in more than 800 published medical research papers, at least 500 of which are on RCTs. So far, the work has led to more than 80 retractions and 50 expressions of concern. Mol has focused much of his work on papers from countries in the Middle East, and particularly in Egypt. One researcher responded to some of his e-mails by accusing him of racism. Mol, however, says that it's simply a fact that he has encountered many suspect statistics and refusals to share data from RCT authors in countries such as Iran, Egypt, Turkey and China — and that he should be able to point that out.

Screening for trustworthiness

"Ben Mol has undoubtedly been a pioneer in the field of detecting and fighting data falsification," says Sotiriadis — but he adds that it is difficult to prove that a paper is falsified. Sotiriadis says he didn't depend on Mol's work when his team excluded those trials in its update, and he can't say whether the trials were corrupt.

Instead, his group followed a screening protocol designed to check for 'trustworthiness'. It had been developed by one of Cochrane's independent specialist groups, the Cochrane Pregnancy and Childbirth (CPC) group, coordinated by Alfirević. (This April, Cochrane formally dissolved this group and some others, as part of a reorganization strategy.) It provides a detailed list of criteria that authors should follow to check the trustworthiness of an RCT — such as whether a trial is prospectively registered and whether the study is free of unusual statistics, such as implausibly narrow or wide distributions of mean values in participant height, weight or other characteristics, and other red flags. If RCTs fail the checks, then reviewers are instructed to contact the original study authors — and, if the replies are not adequate, to exclude the study.

"We're championing the idea that, if a study doesn't pass these bars, then no hard feelings, but we don't call it trustworthy enough," Alfirević explains.

For Sotiriadis, the merit of this protocol was that it avoided his having to declare the trials faulty or fraudulent; they had merely failed a test of trustworthiness. His team ultimately reported that it excluded the Egyptian trials because they hadn't been prospectively registered and the authors didn't explain why.

Other Cochrane authors are starting to adopt the same protocol. For instance, a review[11] of drugs aiming to prevent pre-term labour, published last August, used it to exclude 44 studies — one-quarter of the 122 trials in the literature.

What counts as trustworthy?

Whether trustworthiness checks are sometimes unfair to the authors of RCTs, and exactly what should be checked to classify untrustworthy research, is still up for debate. In a 2021 editorial[12] introducing the idea of trustworthiness screening, Lisa Bero, a senior research integrity editor at Cochrane, and a bioethicist at the University of Colorado Anschutz Medical Campus in Aurora, pointed out that there was no validated, universally agreed method.

"Misclassification of a genuine study as problematic could result in erroneous review conclusions. Misclassification could also lead to reputational damage to authors, legal consequences, and ethical issues associated with participants having taken part in research, only for it to be discounted," she and two other researchers wrote.

For now, there are multiple trustworthiness protocols in play. In 2020, for instance, Avenell and others published REAPPRAISED, a checklist aimed more at journal editors. And when Weibel and others reviewed trials investigating ivermectin as a COVID-19 treatment last year, they created their own checklist, which they call a 'research integrity assessment'.

Bero says some of these checks are more labour-intensive than editors and systematic reviewers are generally accustomed to. "We need to convince systematic reviewers that

EMLI BENDIXEN

Anaesthetist John Carlisle at work.

this is worth their time," she says. She and others have consulted biomedical researchers, publishers and research-integrity experts to come up with a set of red flags that might serve as the basis for creating a widely agreed method of assessment.

Despite the concerns of researchers such as Mol, many scientists remain unsure how many reviews have been compromised by unreliable RCTs. This year, a team led by Jack Wilkinson, a health researcher at the University of Manchester, UK, is using the results of Bero's consultation to apply a list of 76 trustworthiness checks to all trials cited in 50 published Cochrane reviews. (The 76 items include detailed examination of the data and statistics in trials, as well as inspecting details on funding, grants, trial registration, the plausibility of study methods and authors' publication records – but, in this exercise, data from individual participants are not being requested.)

The aim is to see how many RCTs fail the checks, and what impact removing those trials would have on the reviews' conclusions. Wilkinson says a team of 50 is working on the project. He aims to produce a general trustworthiness-screening tool, as well as a separate tool to aid in inspecting participant data, if authors provide them. He will discuss the work in September at Cochrane's annual colloquium.

Alfirević's team, meanwhile, has found in a study yet to be published that 25% of around 350 RCTs in 18 Cochrane reviews on nutrition and pregnancy would have failed trustworthiness checks, using the CPC's method. With these RCTs excluded, the team found that one-third of the reviews would require updating because their findings would have changed. The researchers will report more details in September.

In Alfirević's view, it doesn't particularly matter which trustworthiness checks reviewers use, as long as they do something to scrutinize RCTs more closely. He warns that the numbers of systematic reviews and meta-analyses that journals publish have themselves been soaring in the past decade – and many of these reviews can't be trusted because of shoddy screening methods. "An untrustworthy systematic review is far more dangerous than an untrustworthy primary study," he says. "It is an industry that is completely out of hand, with little quality assurance."

Roberts, who first published in 2015 his concerns over problematic medical research in systematic reviews[13], says that the Cochrane organization took six years to respond and still isn't taking the issue seriously enough. "If up to 25% of trials included in systematic reviews are fraudulent, then the whole Cochrane endeavour is suspect. Much of what we think we know based on systematic reviews is wrong," he says.

Bero says that Cochrane consulted widely to develop its 2021 guide on addressing problematic trials, including incorporating suggestions from Roberts, other Cochrane reviewers and research-integrity experts.

Asking for data

Many researchers worried by medical fakery agree with Carlisle that it would help if journals routinely asked authors to share their IPD. "Asking for raw data would be a good policy. The default position has just been to trust the study, but we've been operating from quite a naive position," says Wilkinson. That advice, however, runs counter to current practice at most medical journals.

In 2016, the International Committee of Medical Journal Editors (ICMJE), an influential body that sets policy for many major medical titles, had proposed requiring mandatory data-sharing from RCTs. But it got pushback – including over perceived risks to the privacy of trial participants who might not have consented to their data being shared, and the availability of resources for archiving the data. As a result, in the latest update to its guidance, in 2017, it settled for merely encouraging data sharing and requiring statements about whether and where data would be shared.

> "JOURNALS SHOULD ASSUME THAT ALL SUBMITTED PAPERS ARE POTENTIALLY FLAWED."

The ICMJE secretary, Christina Wee, says that "there are major feasibility challenges" to be resolved to mandate IPD sharing, although the committee might revisit its practices in future. Many publishers of medical journals told *Nature*'s news team that, following ICMJE advice, they didn't require IPD from authors of trials. (These publishers included Springer Nature; *Nature*'s news team is editorially independent.)

Some journals, however – including Carlisle's *Anaesthesia* – have gone further and do already require IPD. "Most authors provide the data when told it is a requirement," Carlisle says.

Even when IPD are shared, says Wilkinson, scouring it in the way that Carlisle does is a time-consuming exercise – creating a further burden for reviewers – although computational checks of statistics might help.

Besides asking for data, journal editors could also speed up their decision-making, research-integrity specialists say. When sleuths raise concerns, editors should be prepared to put expressions of concern on medical studies more quickly if they don't hear back from authors, Avenell says. This April, a UK parliamentary report into reproducibility and research integrity said that it should not take longer than two months for publishers to publish corrections or retractions of research when academics raise issues.

And if journals do retract studies, authors of systematic reviews should be required to correct their work, Avenell and others say. This rarely happens. Last year, for instance, Avenell's team reported that it had carefully and repeatedly e-mailed authors and journal editors of the 88 reviews that cited Sato's retracted trials to inform them that their reviews included retracted work. They got few responses – only 11 of the 88 reviews have been updated so far – suggesting that authors and editors didn't generally care about correcting the reviews[3].

That was dispiriting but not surprising to the team, which has previously recounted how institutional investigations into Sato's work were opaque and inadequate. The Cochrane collaboration, for its part, stated in updated guidance in 2021 that systematic reviews must be updated when retractions occur.

Ultimately, a lingering question is – as with paper mills – why so many suspect RCTs are being produced in the first place. Mol, from his experiences investigating the Egyptian studies, blames lack of oversight and superficial assessments that promote academics on the basis of their number of publications, as well as the lack of stringent checks from institutions and journals on bad practices. Egyptian authorities have taken some steps to improve governance of trials, however; Egypt's parliament, for instance, published its first clinical research law in December 2020.

"The solution's got to be fixes at the source," says Carlisle. "When this stuff is churned out, it's like fighting a wildfire and failing."

Richard Van Noorden is a features editor at *Nature* in London.

1. Carlisle, J. B. *Anaesthesia* **76**, 472–479 (2021).
2. Roberts, I., Smith, R. & Evans, S. *BMJ* **334**, 392 (2007).
3. Avenell, A., Bolland, M. J., Gamble, G. D. & Grey, A. *Account. Res.* https://doi.org/10.1080/08989621.2022.2082290 (2022).
4. Popp, M. *et al. Cochrane Database Syst. Rev.* **6**, CD015017 (2022).
5. Ker, K., Shakur, H. & Roberts, I. *BJOG* **123**, 1745–1752 (2016).
6. WOMAN Trial Collaborators. *Lancet* **389**, 2105–2116 (2017).
7. Bellos, I. & Pergialiotis, V. *Am. J. Obstet. Gynecol.* **226**, 510–523 (2021).
8. Pacheco, L. D. *et al. N. Engl. J. Med.* **388**, 1365–1375 (2023).
9. Sotiriadis, A. *et al. Cochrane Database Syst. Rev.* **8**, CD006614 (2018).
10. Sotiriadis, A. *et al. Cochrane Database Syst. Rev.* **8**, CD006614 (2021).
11. Wilson, A. *et al. Cochrane Database Syst. Rev.* **8**, CD014978 (2022).
12. Boughton, S. L., Wilkinson, J. & Bero, L. *Cochrane Database Syst. Rev.* **6**, ED000152 (2021).
13. Roberts, I. *et al. BMJ* **350**, h2463 (2015).