Make journals report clinical trials properly

There is no excuse for the shoddy practice of allowing researchers to change outcomes and goals without saying so, says **Ben Goldacre**.

Science is in flux. The basics of a rigorous scientific method were worked out many years ago, but there is now growing concern about systematic structural flaws that undermine the integrity of published data: selective publication, inadequate descriptions of study methods that block efforts at replication, and data dredging through undisclosed use of multiple analytical strategies. Problems such as these undermine the integrity of published data and increase the risk of exaggerated or even false-positive findings, leading collectively to the 'replication crisis'.

Alongside academic papers that document the prevalence of these problems, we have seen a growth in 'technical activism': groups creating data structures and services to help find solutions. These include the Reproducibility Project, which shares out the work of replicating hundreds of published papers in psychology, and Registered Reports, in which researchers can specify their methods and analytical strategy before they begin a study.

These initiatives can generate conflict, because they set out to hold individuals to account. Most researchers maintain a public pose that science is about healthy, reciprocal, critical appraisal. But when you replicate someone's methods and find discrepant results, there is inevitably a risk of friction.

Our team in the Centre for Evidence-Based Medicine at the University of Oxford, UK, is now facing the same challenge. We are targeting the problem of selective outcome reporting in clinical trials.

At the outset, those conducting clinical trials are supposed to publicly declare what measurements they will take to assess the relative benefits of the treatments being compared. This is long-standing best practice, because an outcome such as 'cardiovascular health' could be measured in many ways. So researchers are expected to list the specific blood tests and symptom-rating scales that they will use, for example, alongside the dates on which measurements will be taken, and any cut-off values they will apply to turn continuous data into categorical variables.

This is all done to prevent researchers from 'data-dredging' their results. If researchers switch from these pre-specified outcomes, without explaining that they have done so, then they break the assumptions of their statistical tests. That carries a significant risk of exaggerating findings, or simply getting them wrong, and this in turn helps to explain why so many trial results eventually turn out to be incorrect.

You might think that this problem is so obvious that it would already be competently managed by researchers and journals. But that is not the case. Repeatedly, academic papers have been published showing that outcome-switching is highly prevalent, and that such switches often lead to more favourable statistically significant results being reported instead. This is despite numerous codes of conduct set up to prevent such switching, most notably the widely respected CONSORT guidelines, which require reporting of all pre-specified outcomes and an explanation for any changes. Almost all major medical journals supposedly endorse these guidelines, and yet we know that undisclosed outcome-switching persists.

AUDIT AND ACCOUNTABILITY ARE THE BREAD AND BUTTER OF GOOD MEDICINE, AND GOOD SCIENCE.

Our group has taken a new approach to trying to fix this problem. Since last October, we have been checking the outcomes reported in every trial published in five top medical journals against the pre-specified outcomes from the registry entries or protocols. Most had discrepancies, many of them major. Then, crucially, we have submitted a correction letter, on every trial that misreported its outcomes, to the journal in question. (All of our raw data, methods and correspondence with journals are available on our website at COMPare-trials.org.)

We expected that journals would take these discrepancies seriously, because trial results are used by physicians, researchers and patients to make informed decisions about treatments. Instead, we have seen a wide range of reactions. Some have demonstrated best practice: the *BMJ*, for instance, quickly published a correction on one misreported trial we found, within days of our letter being posted.

Other journals have not followed the *BMJ*'s lead. The editors at *Annals of Internal Medicine*, for example, have responded to our correction letters with an unsigned rebuttal that, in our view, raises serious questions about their commitment to managing outcome-switching. For example, they repeatedly (but confusedly) argue that it is acceptable to identify "prespecified outcomes" from documents produced after a trial began; they make concerning comments that undermine the crucial resource of trial registers; and they say that their expertise allows them to permit — and even solicit — undeclared outcome-switching. Furthermore, they have declined to publish our response to their 850-word letter in the journal.

In our view, this is troubling. *Annals*' response helps to explain why studies repeatedly find outcome-switching to be hugely prevalent, despite policies to prevent it. But journal editors now need to engage in a serious public discussion on why this is still happening. We are providing specific worked examples to facilitate this discussion, and if our project is regarded as provocative, then that is misguided. Audit and accountability are the bread and butter of good medicine, and good science. Lives are at stake when subtle statistical signals of benefit and risk are sought in noisy, messy trial data. We hope that the structures of science really are in a state of flux, and still changing. ■

NATURE.COM
Discuss this article online at:
go.nature.com/8wdqhd

Ben Goldacre *is a physician, author and senior clinical research fellow at the University of Oxford, UK.*
e-mail: ben.goldacre@phc.ox.ac.uk

COMMENT

EXHIBITION Adolf Fleischmann, pathology sculptor and abstract artist **p.30**

INTERDISCIPLINARITY Can architecture catalyse creativity at the Crick? **p.32**

CONSERVATION Legal loophole allows mango farmers to cull fruit bats in Mauritius **p.33**

GEOLOGY Deep-drilling pioneers: what were they drinking? **p.33**

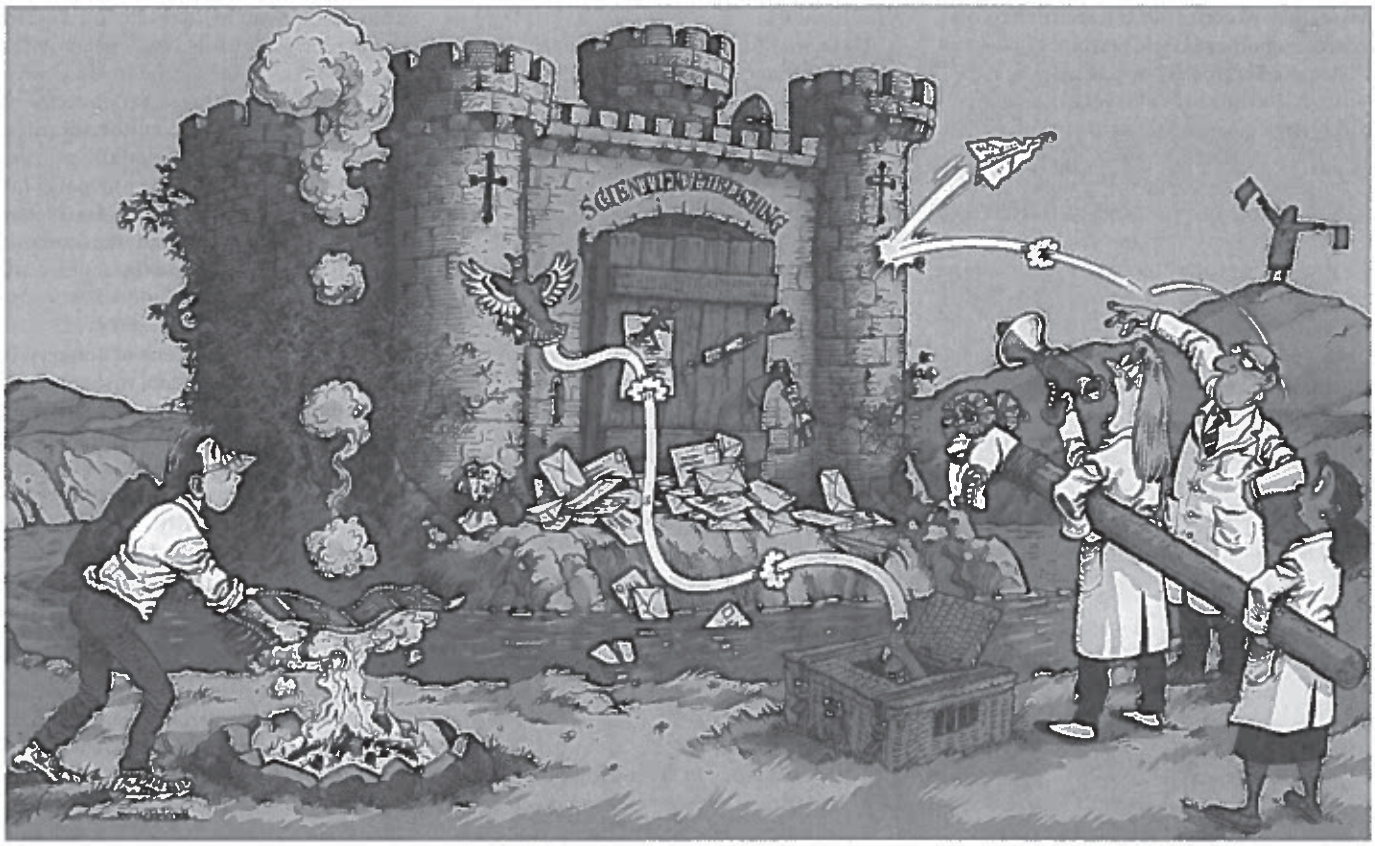

ILLUSTRATION BY DAVID PARKINS

A tragedy of errors

Mistakes in peer-reviewed papers are easy to find but hard to fix, report **David B. Allison** and colleagues.

Just how error-prone and self-correcting is science? We have spent the past 18 months getting a sense of that.

We are a group of researchers working on obesity, nutrition and energetics. In the summer of 2014, one of us (D.B.A.) read a research paper in a well-regarded journal estimating how a change in fast-food consumption would affect children's weight, and he noted that the analysis applied a mathematical model that overestimated effects by more than tenfold. We and others submitted a letter[1] to the editor explaining the problem. Months later, we were gratified to learn that the authors had elected to retract their paper. In the face of popular articles proclaiming that science is stumbling, this episode was an affirmation that science is self-correcting.

Sadly, in our experience, the case is not representative. In the course of assembling weekly lists of articles in our field, we began noticing more peer-reviewed articles containing what we call substantial or invalidating errors. These involve factual mistakes or veer substantially from clearly accepted procedures in ways that, if corrected, might alter a paper's conclusions.

NATURE.COM
For *Nature*'s special collection on reproducibility, see:
go.nature.com/huhbyr

After attempting to address more than 25 of these errors with letters to authors or journals, and identifying at least a dozen more, we had to stop — the work took too much of our time. Our efforts revealed invalidating practices that occur repeatedly (see 'Three common errors') and showed how journals and authors react when faced with mistakes that need correction.

We learned that post-publication ▶

▶ peer review is not consistent, smooth or rapid. Many journal editors and staff members seemed unprepared or ill-equipped to investigate, take action or even respond. Too often, the process spiralled through layers of ineffective e-mails among authors, editors and unidentified journal representatives, often without any public statement added to the original article. Some journals that acknowledged mistakes required a substantial fee to publish our letters: we were asked to spend our research dollars on correcting other people's errors.

As academics who publish, review papers or serve as editors, we appreciate that these issues are complicated. And we feel that journal editors are dedicated and sincere in their efforts. Nevertheless, the scientific community must improve.

Science relies essentially but complacently on self-correction, yet scientific publishing raises severe disincentives against such correction. One publisher states that it will charge the author who initiates withdrawal of a published paper US$10,000.

Here we summarize our experience, the main barriers we encountered, and our thoughts on how to make published science more rigorous. (Details of other resolved issues are available on request.)

SIX PROBLEMS

Editors are often unable or reluctant to take speedy and appropriate action. For one paper, we obtained raw data deposited online, received institutional approval to reanalyse the data, and submitted a letter to the editor (through the manuscript-submission system) describing a need for correction within two weeks. After nine months, we asked the journal why, at minimum, an expression of concern had not been posted. An editor admitted that they had not anticipated the process taking as long as it had. The journal communicated its decision to accept our letter and retract the article 11 months after our submission. The letter and retraction have yet to be published.

Where to send expressions of concern is unclear. Journals rarely state whom to contact about potentially invalidating errors. We had to guess whether to send letters to a staff member or editor, formally submit the letter as a manuscript, or contact the authors of a paper directly. On a few occasions, we opted to contact authors when an apparent invalidating error may have merely been an ambiguous description. In unequivocal cases, we usually contacted the journal. Often, journals provided no way to contact editors directly, and editorial staff corresponded without identifying themselves; we were unsure whether editors were involved.

Journals that acknowledged invalidating errors were reluctant to issue retractions. In one case, we and others found that a paper had mistakenly argued that a statistical adjustment introduced bias, and we submitted a letter to the editor through the journal's submission system[2]. An external statistical review subsequently commissioned by the journal confirmed the error. The authors were asked to retract the article, but they refused. The journal ultimately posted the authors' response to our letter and a summary of commissioned reviewers' criticism. An accompanying editorial published[3] by the journal stated that "it is each author's responsibility to make sure that statistical procedures are correctly used and valid for the study submitted".

Journals charge authors to correct others' mistakes. For one article that we believed contained an invalidating error, our options were to post a comment in an online commenting system or pay a 'discounted' submission fee of US$1,716. With another journal from the same publisher, the fee

STATISTICAL ANALYSIS

Three common errors

A. BARRINGTON BROWN, GONVILLE AND CAIUS COLLEGE/SPL

As the influential twentieth-century statistician Ronald Fisher (**pictured**) said: "To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."

Too many of our post-publication reviews were indeed post mortems. Some studies used inappropriate or non-randomization methods, despite stating that their studies were randomized (see, for example, ref. 5 and go.nature.com/x2i9zz). Others described mathematically or physiologically impossible results: *p*-values greater than 1, or an average height change of about 7 centimetres in adults over 8 weeks[4,6].

Frequent errors, once recognized, can be kept out of the literature with targeted education and policies. Three of the most common are outlined below. These and others are described in depth in an upcoming publication[7].

1 Mistaken design or analysis of cluster-randomized trials. In these studies, all participants in a cluster (for example, a cage, school or hospital) are given the same treatment. The number of clusters (not just the number of individuals) must be incorporated into the analysis. Otherwise, results often seem, falsely, to be statistically significant[8,9]. Increasing the number of individuals within clusters can increase power, but the gains are minute compared with increasing clusters. Designs with only one cluster per treatment are not valid as randomized experiments, regardless of how many individuals are included.

2 Miscalculation in meta-analyses. Effect sizes are often miscalculated when meta-analysts are confronted with incomplete information and do not adapt appropriately. Another problem is confusion about how to calculate the variance of effects. Different study designs and meta-analyses require different approaches. Incorrect or inconsistent choices can change effect sizes, study weighting or the overall conclusions[4].

3 Inappropriate baseline comparisons. In at least six articles, authors tested for changes from the baseline in separate groups; if one was significant and one not, the authors (wrongly) proposed a difference between groups. Rather than comparing 'differences in nominal significance' (the DINS error) differences between groups must be compared directly. For studies comparing two equal-sized groups, the DINS error can inflate the false-positive rate from 5% to as much as 50% (ref. 10).

FIXING POST-PUBLICATION REVIEW
Publishers, editors and researchers must all up their game.

	How to prevent statistical errors in submissions	How to streamline post-publication corrections
Research teams	Tap statistical expertise in the design and analysis of studies from the start. Describe analyses thoroughly.	Curate data and computer code so that they can be made easily available (for a registry of public data repositories, see www.re3data.org).
Manuscript editors	Create protocols to identify papers that need statistical scrutiny and send them to qualified reviewers.	Address readers' concerns swiftly. Use formal expressions of concern as an alert that work is under scrutiny — rather than for condemnation.
Journals and publishers	Require raw data and analysis code to be made available during review.	Create protocols to manage expressions of concern. State clearly who readers should contact and train editors to navigate protocols. Waive publication fees and paywalls for expressions of concern and retractions.

was £1,470 (US$2,100) to publish a letter. Letters from the journal advised that "we are unable to take editorial considerations into account when assessing waiver requests, only the author's documented ability to pay". The Committee on Publication Ethics, an independent body that provides advice on how to handle research misconduct, asserts that readers should not have to pay to read retractions. To our knowledge, no authority has discussed whether third parties should be charged to correct errors.

No standard mechanism exists to request raw data. When we were able to access data online, we could quickly confirm suspected errors. In at least two cases, we requested data from the authors but received summaries of calculations instead. Sometimes we received no data at all, at which point it was not clear whether journal staff should step in. One journal did retract a paper when its authors refused to show their data or explain discrepancies that we had identified and alerted the journal to in a letter[4].

Working directly with authors can delay correction. After we contacted authors about another paper, they offered to reanalyse the data to address our concerns. After a month with no response, we submitted a letter of concern to the journal. The letter was peer-reviewed and accepted within three weeks. The authors, when made aware of the pending publication of our letter, e-mailed us to state that they would prepare a reply, and we asked the journal not to publish our letter so that we could collaborate with the original authors. That process is ongoing, ten months after we identified the error.

Informal expressions of concern are overlooked. Although online platforms such as PubMed Commons offer a convenient way to comment on published papers, they do not include a mediating role for journal editors, and the comments are not incorporated into the literature. Posted concerns are rarely prominent on journals' websites and are not cross-referenced in any useful way. As a result, readers may assume that a flawed paper is correct, potentially leading to misinformed decisions in science, patient care and public policy.

In one case, we chose to post a comment on the journal website and on PubMed Commons after months of private correspondence, in which the authors shared some supplementary data and said that they were preparing a full response. The concerns have been acknowledged but remain unaddressed 15 months after we contacted authors and the journal, and 6 months after we posted our comment (see go.nature.com/fv8tr2).

WHAT CAN BE DONE?

Journals have guidelines for paper submissions and peer review. The Committee on Publication Ethics has outlined recommendations for journals to address problems in areas such as authorship and review. But there is little formal guidance for post-publication corrections. (For our recommendations, see 'Fixing post-publication peer review'.)

"Authors and journals should share data and code quickly when questions arise."

Journals, publishers and scientific societies should standardize, streamline and publicize these processes. Authors and journals should share data and code quickly when questions arise. Researchers can aid this process by accessing statistical expertise for experimental design and analysis.

Ideally, anyone who detects a potential problem with a study will engage, whether by writing to authors and editors or by commenting online, and will do so in a collegial way. Scientists who engage in post-publication review often do so out of a sense of duty to their community, but this important work does not come with the same prestige as other scientific endeavours. Recognizing and incentivizing such activities could go a long way to cleaning up the literature.

Our work was not a systematic search; we simply looked more closely at papers that caught our eye and that we were prepared to assess. We do not know the rate of errors or the motivations behind them (that is, whether they are honest mistakes or a 'sleight of statistics'). But we showed that a small team of investigators with expertise in statistics and experimental design could find dozens of problematic papers while keeping abreast of the literature. Most were detected simply by reading the paper.

A more formal survey would help to determine whether our experiences reflect science in general and whether our recommendations are feasible or effective. Others working to correct the scientific record have encountered similar challenges. Ben Goldacre, a physician and campaigner who is leading COMPare, a project that checks that clinical trials report the outcomes they said they would, told Retraction Watch: "This is a phenomenally laborious process. Not a week goes by that we don't curse the day we set out to do this."

Robust science needs robust corrections. It is time to make the process less onerous. ■

David B. Allison *is a distinguished professor in the Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Alabama, USA.*
Andrew W. Brown *is a scientist in the Office of Energetics and the Nutrition Obesity Research Center, University of Alabama at Birmingham, Alabama, USA.*
Brandon J. George *is a statistician in the Office of Energetics, University of Alabama at Birmingham, Alabama, USA.*
Kathryn A. Kaiser *is an instructor in the Office of Energetics and the Nutrition Obesity Research Center, University of Alabama at Birmingham, Alabama, USA.*
e-mail: dallison@uab.edu

1. Brown, A. W. *et al. Child. Obes.* **10**, 542–545 (2014).
2. Li, P. *et al. Obes. Facts* **8**, 127–129 (2015).
3. Hauner, H. *Obes. Facts* **8**, 125–126 (2015)
4. George, B. J., Brown, A. W. & Allison, D. B. *J. Paramedical Sci.* **6**, 153–154 (2015).
5. George, B. J., Goldsby, T. U., Brown, A. W., Li, P. & Allison, D. B. *Int. J. Yoga* **9**, 87–88 (2016).
6. Thomas, D. M. *et al. World J. Acupunct. Moxibustion* **25**, 66–67 (2015).
7. George, B. J. *et al. Obesity* (in the press).
8. Brown, A. W. *et al. Am. J. Clin. Nutr.* **102**, 241–248 (2015).
9. *Obesity* **23**, 2522 (2015).
10. Bland, J. M. & Altman, D. G. *Am. J. Clin. Nutr.* **102**, 991–994 (2015).

D.B.A. declares competing financial interests: see go.nature.com/hshkkk for details.