CLINICAL TRIALS

To catch a crook, you might try statistics

Gary R. Cutter

A new systematic review detected strong evidence of unreliable results via use of statistical and logical analyses of 33 randomized clinical trials. Our outrage at these rare occurrences of potential fraud could be rooted in our disdain at our failures in peer review, given that this special effort was required to detect long-running potential fraud.

Refers to Bolland, M. J., *et al.* Systematic review and statistical analysis of the integrity of 33 randomized controlled trials. *Neurology* **87**, 1–12 (2016).

In a recent issue of *Neurology*, Mark J. Bolland and colleagues[1] present analysis that provides methodological approaches to reviewing validity of clinical trials and directly impugns a body of literature. The authors' criticism relates to work conducted by Yoshihiro Sato, a bone specialist whose studies of patient groups with various neurological disorders contain improbable results, which Bolland *et al.* argue are likely to be unreliable. The authors form their conclusions on the basis of exceptionally high productivity and recruitment rates, improbably similar randomized treatment groups, remarkably positive outcomes, inconsistent reporting and duplicate data across studies, and questionable ethical oversight.

The authors' analysis considers the 33 randomized controlled trials (RCTs) authored by Sato over 15 years. They note that the high publication rate suggests that Sato and colleagues have had a large and well-funded research network, with access to large numbers of patients with high levels of comorbidity who consent to participate. Bolland and co-investigators point out that a few of the studies by Sato and colleagues had remarkable accrual rates. For example, one of the studies involved recruiting 500 community-dwelling patients with Alzheimer disease (AD) older than 70 years in a few months, with other large studies with monthly follow-ups ongoing simultaneously.

In a clever use of statistical analysis, the authors collected the 513 baseline variables reported across the 33 trials in which Sato was involved, and examined the distribution of the reported *P* values obtained in tests for differences between the treatment groups.

Although Bolland *et al.* ignored the correlation amongst these baseline characteristics in their determination, the unusual *P* value distribution still suggests an improbable similarity of baseline characteristics between the treatment versus no-treatment groups, and the between-group uniformity of baseline characteristics was even greater across articles in which Sato was the first author. In a simple attempt to make results in the treatment group look more significant than they are, one might leave out baseline characteristic values that differ significantly from the mean. However, in the analysed trials, even the nonsignificant *P* values toward the tails of the distribution were under-represented — a finding that might better be explained by fabrication of data rather than by deletion of data points. Bolland and colleagues also show that

> “such a high number of improbable results suggests … falsification of at least some of the results”

in 27 of the 30 two-group trials, sample sizes in each group were identical, which is also highly unlikely.

Even more importantly, Bolland and co-investigators state that Sato and colleagues have reported highly improbable positive outcomes. The studies by Sato *et al.* show remarkably similar results of active intervention (such as vitamin D supplementation or sunlight exposure, vitamin K supplementation, vitamin B_{12} supplementation, or treatment with the osteoporosis drugs) in various populations with differing age ranges and comorbidities. All these studies showed remarkable reduction in the relative risk of hip fractures (mean RR 0.22; varying between 0.15 and 0.31), which exceed those reported by other investigators (residronate, RR 0.58; alendronate, RR 0.62; vitamin D supplementation, RR 1.15[1]). The authors also question the unusually high study completion rates (>91%), given that the risks associated with hip fractures and background comorbidity are associated with high mortality.

Finally, Bolland and colleagues note that Sato's articles show logical inconsistencies between the trials. In one trial, which assessed the benefit of resedronate and calcium supplementation in patients with AD, the baseline hip fracture rate was 86 per 1,000 patient-years, whereas in another trial which evaluated the benefit of risedronate plus calcium and vitamin D supplementation in patients with AD, baseline hip fracture rate was only 15 per 1,000 patient-years. Other problems are also noted; for example in calculating fracture rates, the authors ignore dropouts, deaths and withdrawals, resulting in inaccurately low fracture rates. Moreover, adverse events are inconsistently reported: for example, one article reports no adverse events even though it mentions elsewhere that 10 patients withdrew from the study because of death or intercurrent illness. Misleading text and duplicate data across studies are also observed. Moreover, Bolland and colleagues were unable to identify any other studies that were purportedly approved by the board that approved the studies led by Sato and colleagues, which Bolland *et al.* interpret as irregularity in ethical oversight.

Overall, the Bolland *et al.* article presents quite a few unlikely results published by Sato and colleagues. Although any of the presented

improbable findings in isolation might have a plausible explanation, collectively, such a high number of improbable results suggests — at least in my view — falsification of at least some of the results.

The fact that a lack of variability is often a clue to falsification is not a novel finding. Over a quarter of century ago, Bailey *et al.*[2] pointed out that falsified biomedical results could be detected via lack of variability, an observation that was corroborated and improved upon by findings of DeMets *et al.*[3], who stated that statistical analysis can not only detect fraud, but can also distinguish between fraud, bias, errors, misunderstanding, and incompetence.

Detecting fraud is, however, not trivial. The clinical trials in which Sato was involved appeared in some of the most prestigious journals, yet the issues pointed out by Bolland and colleagues were not detected in peer review, raising the question of whether peer review should be complemented with professional reviewers or other checks run by the journals. Peer review usually focuses on a single submitted article, and is predominantly carried out by volunteer academics who accommodate peer reviews into their already busy schedules. Fraud is not the primary thing peer reviewers are looking for and, viewing at the extensive work by Bolton *et al.*, it is too time-consuming to reliably detect all research misconduct during peer review.

Multicentre clinical trials can partly insulate trials from certain forms of falsification: individual site data can be compared, and a site providing falsified data can be eliminated without compromising the entire trial. This is a major reason for randomization of the treatments by site.

Making trial and other experimental data publicly available have been suggested as one strategy to reduce fraud, as any interested individual could search for evidence, thereby making fraud less tempting. However, the consequences of the false positives should be kept in mind: do we wish to impugn even one honest researcher in our attempts to catch dishonest ones? Moreover, the costs of data sharing are high, and storing the data in repositories cannot prevent uploading aptly falsified data that would escape the statistical techniques used by Bolton *et al.*

To reduce research fraud, we need to understand more deeply why professional researchers and clinicians commit fraud. Certainly, personal interest is a major factor. Review boards thoroughly assess conflicts of interest to protect authors from gaining excess monetary gains via publications, but we have little to protect against advancing one's personal interests — indeed, we reward it. Promotions, grants and grant renewals all depend on publications. Moreover, researchers might commit fraud for more egotistic reasons than mere survival in their chosen career, such as for money or to get a product on the market, to boost the market value of their start-up company, or perhaps even to deliberately sabotage competitors. Focusing on detecting the characteristics of people whose drive to advance their personal interests overwhelm their training and ethics could potentially help reduce the incidence of scientific misconduct.

Finally, the article by Bolland and co-investigators raises a larger question about ethics in science. We live in a postfactual world where individuals are entitled to say anything, whether it is true or not, and display these positions via social media where editorial control is absent. Inaccurate medical information appears widely on the Internet, a phenomenon that researchers and clinicians accept much more calmly than we do instances of fraud in medical research. One possible reason could be that irresponsible statements on social media have become commonplace, whereas integrity is still expected in medical research, making us fundamentally appalled at instances of fraud and the fact that peer review missed them.

Gary R. Cutter is at the UAB School of Public Health, Department of Biostatistics, 1667 University Boulevard, Ryals Building Room 410b, Birmingham, Alabama 35243, USA.

cutterg@uab.edu

doi:10.1038/nrneurol.2016.194
Published online 16 Dec 2016

1. Bolland, M. J., Avenell, A., Gamble, G. & Grey, A. Systematic review and statistical analysis of the integrity of 33 randomized controlled trials. *Neurology* **87**, 1–12 (2016).
2. Bailey, K. Detecting fabrication of data in a multicenter collaborative animal study. *Control. Clin. Trials* **12**, 741–752 (1991).
3. DeMets, D. Distinctions between fraud, bias, errors, misunderstanding, and incompetence. *Control. Clin. Trials* **18**, 637–650 (1997).

Competing interests statement

The author is on Data and Safety Monitoring Boards for AMO Pharma Apotek, Gilead Pharmaceuticals, Horizon Pharmaceuticals, Modigenetech/Prolor, Merck, Merck/Pfizer, Neuren, Opko Biologics, Reata Pharmaceuticals, Receptos/Celgene, Sanofi-Aventis, Teva pharmaceuticals, National Heart Lung and Blood Institute (Protocol Review Committee), National Institute of Child Health and Development (OPRU oversight committee). He runs MS Patient Registry, which receives a grant from Consortium of MS Centers. Moreover, he is on Consulting or Advisory Boards for Cerespir, Genzyme, Genentech, Innate Therapeutics, Jannsen Pharmaceuticals, Klein-Buendel Incorporated, Medimmune, Medday, Nivalis, Novartis, Opexa Therapeutics, Roche, Savara, Somahlution, Teva Pharmaceuticals, Transparency Life Sciences, and TG Therapeutics. He is President of Pythagoras, a private consulting company located in Birmingham, Alabama, USA.