Errors and Integrity in Seeking and Reporting Apparent Research Misconduct

Evan D. Kharasch, M.D., Ph.D., Timothy T. Houle, Ph.D.

THE responsible conduct of research is the bedrock on which the scientific enterprise rests. Scientific integrity is indispensable for preserving the public trust and the trust of the scientific community, protecting those who participate in research (investigators and research subjects), and safeguarding the health and safety of patients who depend on scientific knowledge for their diagnosis and care. When the scientific content of ANESTHESIOLOGY and the integrity of our authors are publically questioned, it is the journal's responsibility to evaluate such claims and to inform our readers about the claims, the pertinent issues, and the results of our evaluation of the claims.

Research misconduct has become an all too familiar stain on the research and clinical landscape, and it is occurring with increasing frequency and public awareness. Variably complicit in research misconduct, explicitly and/or implicitly, are investigators, authors, sponsors, academic institutions, and journals. Revelations of scientific misconduct have been punctuated by several well-known instances of gross fabrication and falsification in our specialty. Never tolerable, misconduct is perhaps even more egregious when occurring in randomized clinical trials, which are a foundational element informing clinical practice.

The scientific community is ever watchful for fabrication and falsification—journal editors included. It is the bane of our existence. Were there only some valid algorithm or software that could detect fabrication and falsification (akin to that which detects duplication and plagiarism with remarkable alacrity and accuracy).

"…ANESTHESIOLOGY assures our readers that we honor and value the responsible conduct of research, and we are constantly vigilant to identify any possible violations of scientific integrity…"

A recent issue of the journal *Anaesthesia* contains an article by Dr. John Carlisle on "data fabrication and other reasons" for nonrandom sampling in randomized clinical trials in the specialty of anesthesia and in two general medical journals.[1] It was accompanied by an editorial that celebrates the article, the statistical method used, and the application thereof to detecting scientific misconduct.[2] We were particularly concerned because the Carlisle article identified 12 research investigations published in ANESTHESIOLOGY with purportedly problematic results (at a *P* value cut-off less than 0.0001) that "might benefit from further investigation"..."to correct and if necessary retract,"[1] and which might be "corrupted."[2]

In brief, the Carlisle Method[3] evaluates the baseline variables (*e.g.*, age, weight) in a manuscript and identifies those with more or less balance than would be expected by chance. Less balance is akin to data not being consistent with random sampling. The Carlisle Method therefore deems such data suspicious.[1,4,5]

The Method assumes that every individual baseline demographic variable in a randomized controlled trial (*i.e.*, what is typically presented in a manuscript's "table 1") is randomly assigned across treatment groups. Using this assumption, the Method then examines the observed distribution of individual *P* values that result from testing the null hypothesis that the randomized groups do not differ. The observed distributions are then compared to the expected distribution that should be observed when the null hypothesis is true (*e.g.*, uniform distribution).

Image: J. P. Rathmell.

Timothy J. Brennan, Ph.D., M.D., served as Handling Editor for this article.

Submitted for publication July 24, 2017. Accepted for publication August 10, 2017. From the Department of Anesthesiology and Department of Biochemistry and Molecular Biophysics, Washington University in St. Louis, St. Louis, Missouri, and Center for Clinical Pharmacology, St. Louis College of Pharmacy and Washington University in St. Louis, St. Louis, Missouri (E.D.K.); and the Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts (T.T.H.).

 Anesthesiology 2017; 127:733-7

<zdoi;10.1097/ALN.0000000000001875>

With some degree of equipoise, the Carlisle article attributes observed deviations from these expected distributions to typographical error, unintentional error, correlation, stratified allocation, poor methodology, fabricated data, or fraud.[1] It politely suggests that the articles in question "might benefit from further investigation," but then asserts that "association of extreme distributions with trial retraction suggests that further investigation of uncorrected unretracted trials and their authors will result in most trials being corrected and some retracted."[1] Put more succinctly, a probability is extrapolated to a fact, which then becomes the basis for judgement and action (*vide infra*).

The Carlisle Method has been touted as the long-sought "Holy Grail" for detecting fabrication and falsification in scientific publishing. *Anaesthesia* "can rightly claim with vicarious pride that one of its own, John Carlisle, is at the forefront of this effort," and that Carlisle's "statistical exposé…made the research world stand up and take notice" in an effort to prevent publication of fraudulent material.[2] *Anaesthesia* declared that it "has decided to screen all randomized trials submitted to the journal using the Carlisle Method" and "will (summarily) reject any that fall foul due to suspicious data that are not consistent with random sampling."[3] Furthermore, it "hopes that other journals will follow suit and also screen submissions."[3] Loadsman and McCulloch proclaim that "a strong argument could be made that every journal in the world now needs to apply Carlisle's method to all the randomized clinical trials they've ever published."[2] All these emphatic exhortations are made while acknowledging that "clearly there is still a lot of work to do to validate the Carlisle Method" and asking "other editors, statisticians, authors and readers to apply the Carlisle Method for themselves and help validate it."[3]

Whereas Carlisle somewhat dispassionately offers several possible explanations for unusually distributed data detected with his method (while still stating that it is "likely that it will lead to the identification, correction and retraction of hitherto unretracted randomised, controlled trials"), the accompanying editorial by Loadsman and McCulloch eschews any pretense of neutrality.[2] Nonrandom distribution of baseline data is deemed evidence of "fraudulently concocted data," "fraudulent material," "misconduct," a "statistical smoking gun," and those in association are called "miscreants." More specifically, it "means that the body of randomized clinical trials in the journals cited by Carlisle is, essentially without question, corrupted." It entreats that "editors of each of the journals included in Carlisle's study urgently need to follow up the randomized clinical trials that have been identified as most likely problematic, whether due to error or less innocent reasons," and that "with the proven utility of the Carlisle method, no doubt more authors of already published randomized clinical trials will eventually be getting their tap on the shoulder." And finally, the Loadsman and McCulloch editorial forewarns, "we have not yet heard the last word from John Carlisle!"

We decry the fabrication and falsification of research, including and not limited to randomized clinical trials. A valid and validated detector for fabrication and falsification would be welcome.

As a responsibility to our readers, we have carefully examined the article by Carlisle and the editorial by Loadsman and McCulloch, and we have heeded the call of *Anaesthesia* to test the Carlisle Method and to "follow up the randomized clinical trials that have been identified as most likely problematic."[2] Unfortunately, we find that both the article and the editorial are deeply flawed, in many aspects, and they cry out for rectification. The Carlisle article is factually incorrect. The Carlisle Method is methodologically flawed and misleading as applied by the author. The Carlisle article is ethically questionable and a disservice to the authors of the previously published articles "called out" therein. The editorial by Loadsman and McCulloch amplifies the significance of these errors, as it incorrectly vilifies the authors of the questioned articles. We will explain each of these problems in the paragraphs that follow.

First, the Carlisle article is factually incorrect. It identified 12 randomized clinical trials published in Anesthesiology that "might benefit from further investigation…to correct and if necessary to retract" because of unusual data distributions. In fact, six of those "randomized clinical trials" were actually studies in animals (three in rats and one each in rabbits, pigs, and nonhuman primates). They were not clinical trials at all. Clinical trials are research studies in which one or more *human subjects* are assigned prospectively to one or more interventions. That the six articles reported animal studies was clearly identified in the title and/or abstract of every article. Such lack of attention to detail, in an article criticizing lack of attention to detail and issuing harsh judgement, is disappointing, at a minimum, and raises concerns about the care with which the analysis was conducted.

Second, the Carlisle Method is methodologically flawed and misleading as applied by the author. It leads to incorrect interpretations about the probability that groups were inadequately randomized into balanced groups. The flaws in the method arise from two main problems with examining the imbalances between randomized groups using statistical inferences (*P* values). The first flaw is that the very idea of using statistical inferences to formally test the balance between randomized groups has been eschewed by statisticians because the balance between two randomized groups is a property of that sample, not a property of the population from which the samples were drawn.[6–8] Stated simply, when examining group balance in randomized clinical trials the null hypothesis being tested is not logically sound. Null hypothesis testing infers a parameter about the population from which the sample was drawn and does not provide a direct inference about the sample under study. Carlisle is clearly aware of this issue and appears to be using the distribution of calculated *P* values as an indirect inference of the imbalance between two groups across many variables. This is

a clever idea, but there are more direct indices of balance for this task that do not rely on the use of an illogic foundation.[9]

The second flaw in the Carlisle Method is much more consequential. It arises from its treatment of correlated baseline variables in "table 1" as independent variables. Each of the methods used by Carlisle to visualize and estimate the degree of violation from expectancy relies on the assumption that the variables in a "table 1" are independent of one another. This might be the case if each variable was randomized by itself, but in clinical trials, *we randomize people, not variables* into treatment groups. Figure 1 illustrates the difference between truly independent variables and variables that are nested within individuals. Because variables are nested within people (*i.e.*, people have an age, body mass index, blood pressure, etc.), slight imbalances in a sample (*e.g.*, one group is slightly older than the other) could lead to other variables also being imbalanced (*e.g.*, older people have higher body mass indices). In a large population, each of these variables is expected to be balanced through randomization, but the variables are likely to be correlated in any given sample. This nesting creates a correlation among the variables that is problematic for Stouffer's method and each of the other techniques employed by Carlisle. Far from being benign, ignoring the fact that variables in "table 1" can be correlated leads to massive bias in the expected distribution of how even well-randomized trials should be distributed.

To illustrate this problem, we evaluated what would happen in a simulation study where there were no problems in the randomization of groups (*i.e.*, there was no fraud or randomization failures), but the variables were correlated. We simulated N = 22,500 clinical trials with groups randomized based on a well-established random number generator (mvrnorm from the MASS package in R). For the simulations, we allowed sample sizes of the trials to vary from 50 to 1,000, the number of items in "table 1" to vary from 5 to 25, and the average correlation between variables in "table 1" to vary from $r = 0.0$ to $r = 0.40$. Each condition was replicated 500 times to measure the variability in each condition. Figure 2 displays the results of the simulations for several of the conditions. The panels illustrate that as the average correlation between items in "table 1" increases, the more likely the Carlisle Method is to wrongly identify the trial as aberrant. This bias was even greater as the number of items in "table 1" increased. Carlisle used a threshold of $P < 0.00001$ to identify that 82 of 5,015 (1.6%) unretracted articles were aberrant randomizations. Using the same threshold, the simulations found that a similar proportion (1 to 3%) of completely sound randomizations would be flagged as aberrant *by chance alone,* if the average correlation among the 20 to 25 items in "table 1" was $r \sim 0.30$. Thus, even if there were no problems with randomization, the Carlisle Method would falsely indicate that many of the trials were problematic, based solely on them having several moderately correlated items or many items that had only minor correlations. Clearly, this degree of false positives that are produced by an ignored assumption precludes the use of the Carlisle Method in common practice.

There is insufficient evidence of sensitivity and specificity for "every journal in the world…to apply Carlisle's method to all the randomized clinical trials they've ever published."[2]

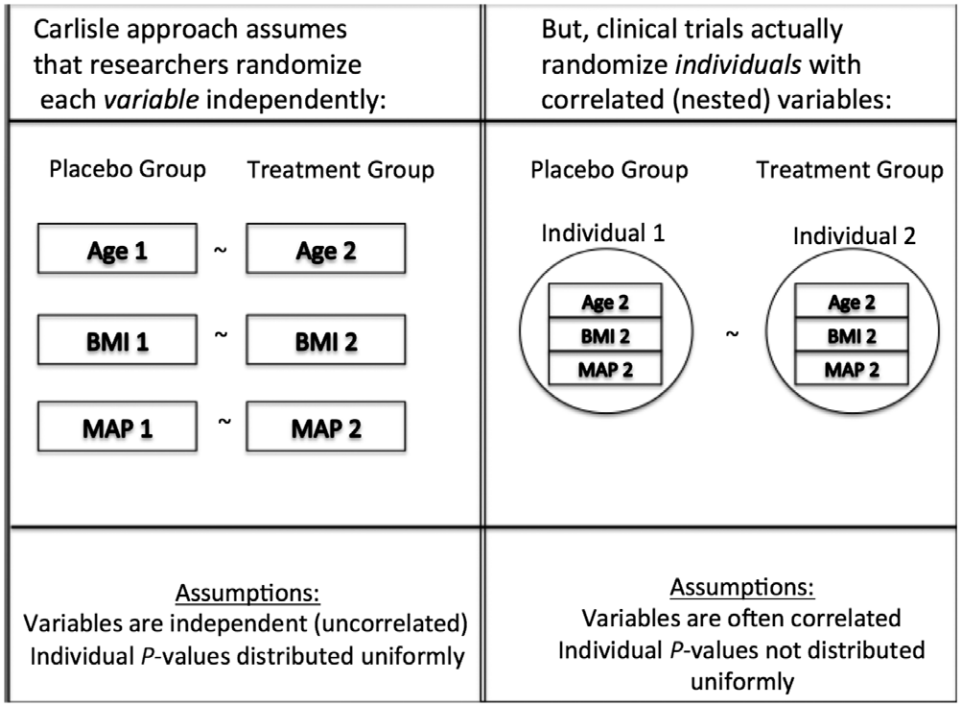

Fig. 1. Illustration of the assumptions required by the Carlisle approach *versus* the conditions that are likely to be observed. Because randomized controlled trials randomize people, not individual variables, the distribution of *P* values is not likely to follow the distributions assumed by Carlisle. BMI = body mass index; MAP = mean arterial pressure.

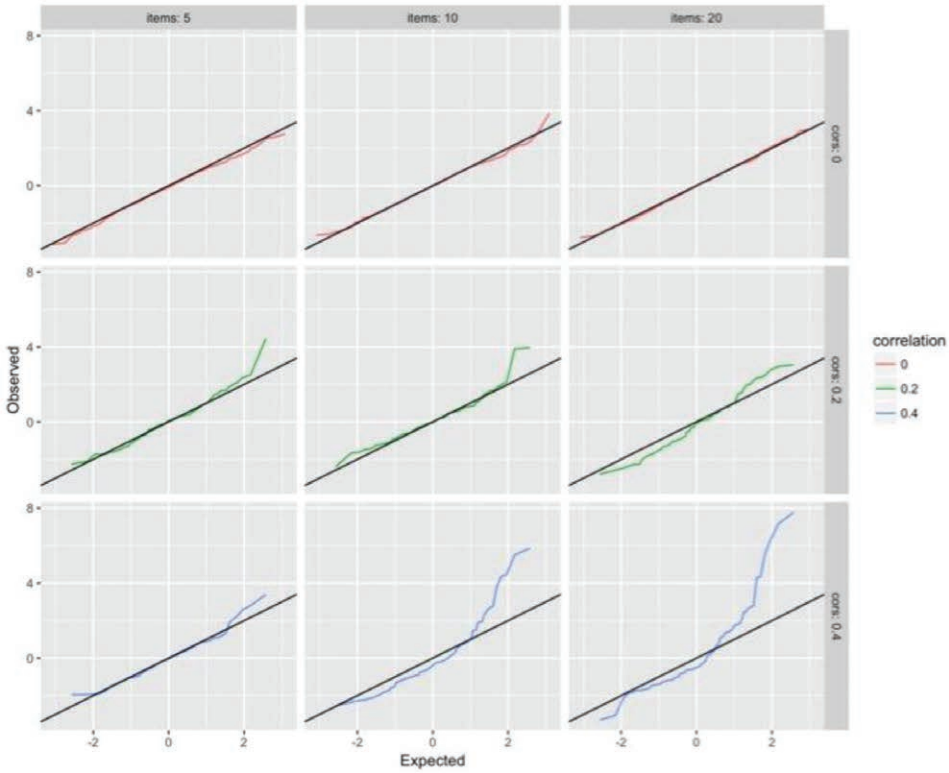

Fig. 2. Selected results from the simulations used to evaluate the Carlisle Method. If the expected distribution from the sum of *P* values from a trials report's "table 1" (*x*-axis) is the same as the observed distribution of *P* values (*y*-axis), the test values (*colored lines*) will fall directly on the *diagonal black line*. Departures from the *black line* indicate false-positives as the observed distributions exceed what is expected by chance. When the correlation between items is 0 (*top row*), the Carlisle Method performs well. However, as the degree of correlation (*rows*) increases, many false positives are observed. This effect is amplified with increasing number of items (columns) in the table.

It is difficult to evaluate the specificity of the Method because the number of true cases of fraud is not known with any certainty. The Method certainly lacked specificity with respect to the articles in Anesthesiology. And it is clear that the Method lacks sensitivity. Carlisle's own analysis showed that it did not work well in the subset of articles that are known to be fraudulent.[1] Specifically, of these, it identified only 8 of 72 (11%) retracted articles—an 89% false negative rate.

Third, the Carlisle article is ethically questionable and a disservice to the authors of the Anesthesiology if not other published articles "called out" therein. Guidelines from the Committee on Publication Ethics (COPE, a forum on publication ethics for editors and publishers of peer-reviewed journals that advises journal editors on how to handle cases of research and publication misconduct) state that in a case of suspected fabricated data in a published article, the first step is to "contact the author to explain concerns but do not make direct accusations" (https://publicationethics.org/resources/flowcharts). Carlisle, or *Anaesthesia*, could have first contacted the author(s) and/or journals of the articles found in question before any publication of the recent article.[1] The Carlisle article could have de-identified the journals involved and not identified the individual articles it questioned. It could have identified the journals while privately notifying the relevant authors and journal editors of the questioned articles. Remarkably, however, the Carlisle article, as the *first action*, *publically identified* (in an online appendix) the authors, titles, journals, and citations of the 78 anesthesia articles that it deemed suspect. Neither the journals nor the authors of these 78 articles were notified, or given a chance to respond, prior to being "outed" by Carlisle's article. The Carlisle article appears to directly contravene the COPE guidance on suspected fabricated data.

Lastly, the editorial by Loadsman and McCulloch amplifies the consequence of these previous errors, as it incorrectly vilifies the authors of the questioned articles. Equating a statistical improbability (using a "smoking gun" method that is both incorrect and known to have "a problem of both sensitivity and specificity illustrated by Carlisle's own data"[2]) with "fraudulently concocted data," "misconduct," "corruption," and "miscreants" should not be acceptable to a learned profession, and certainly not to those authors blindsided and unjustly accused.

In dissecting Carlisle's approach to identifying problems in published research, we are not refuting the idea that articles published in Anesthesiology, or any other journal, may unfortunately contain errors.[10] Nor are we diminishing the vast importance of postpublication peer review for detecting errors and improving the quality of science. In fact, when we re-reviewed

the six clinical research articles in Anesthesiology identified by Carlisle as possessing problematic randomizations, we did discover errors in reporting that were *unrelated* to this data distribution issue. For example, we found a blatantly obvious, typographical error.[11] At the time of this writing we are working with authors to clarify and remedy any such reporting errors. However, as stated by Allison *et al.*, "mistakes in peer-reviewed papers are easy to find but hard to fix."[10] In that regard, Anesthesiology is constantly working to improve our peer review process and standards. What we are criticizing is the use of a method that has core fundamental problems and that leads to an expected array of false-positives and undue stigmatization.

What Next?

Despite the zeal of *Anaesthesia*, the Carlisle Method does not appear to be valid, nor to have the claimed "proven utility" to screen all randomized trials submitted for publication, and certainly not to reject *a priori* those with nonrandomly distributed variables. Similarly unfounded is the call that "other journals will follow suit and also screen submissions."[3] The Carlisle Method is not the "Holy Grail" for detecting fabrication and falsification in scientific publishing.

Authors whose manuscripts have been rejected for nonrandomly distributed baseline variables may wish to contact the journals and seek reconsideration. Perhaps better, the journals may proactively wish to re-examine and reconsider all those manuscripts thus rejected, and to communicate proactively with the authors. Most certainly, *Anaesthesia* owes an apology to the authors of the six "randomized clinical trials" in Anesthesiology that were "outed," when in fact these were animal studies.[1,2] In addition, an indexed erratum correcting the errors in the Carlisle article,[1] published in *Anaesthesia,* would be appropriate. Furthermore, we call on Carlisle and *Anaesthesia* to issue a press release announcing the errors in the Carlisle Method and the recent publication,[1] and to do so with the same distribution path and vigor of the press release publicizing that article (http://www.aagbi.org/news/report-says-dozens-medical-trials-may-contain-inaccurate-data). Lastly, COPE guidelines state that "journal editors should consider retracting a publication if they have clear evidence that the findings are unreliable, either as a result of misconduct (*e.g.*, data fabrication) or honest error (*e.g.*, miscalculation or experimental error)" (https://publicationethics.org/files/retraction%20guidelines_0.pdf). Should this apply to the article by Carlisle and the editorial by Loadsman and McCulloch? Should this be the "last word"?

The leadership of Anesthesiology assures our readers that we honor and value the responsible conduct of research, and we are constantly vigilant to identify any possible violations of scientific integrity in the manuscripts submitted for consideration of publication in our journal. To that goal, we will use any means that have been validated for such use. Our readers and our authors should understand that the Carlisle Method is not and will not be one of them.

Competing Interests

Dr. Kharasch is the Editor-in-Chief of Anesthesiology and his institution receives salary support from the American Society of Anesthesiologists for this position. Dr. Houle is the Statistical Editor of Anesthesiology.

Correspondence

Address correspondence to Dr. Kharasch: editor-in-chief@asahq.org

References

1. Carlisle JB: Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. Anaesthesia 2017; 72:944–52
2. Loadsman JA, McCulloch TJ: Widening the search for suspect data—is the flood of retractions about to become a tsunami? Anaesthesia 2017; 72:931–5
3. Klein AA: What Anaesthesia is doing to combat scientific misconduct and investigate data fabrication and falsification. Anaesthesia 2017; 72:3–4
4. Carlisle JB, Dexter F, Pandit JJ, Shafer SL, Yentis SM: Calculating the probability of random sampling for continuous variables in submitted or published randomised controlled trials. Anaesthesia 2015; 70:848–58
5. Carlisle JB, Loadsman JA: Evidence for non-random sampling in randomised, controlled trials by Yuhji Saitoh. Anaesthesia 2017; 72:17–27
6. Senn SJ: Covariate imbalance and random allocation in clinical trials. Stat Med 1989; 8:467–75
7. Senn S: Testing for baseline balance in clinical trials. Stat Med 1994; 13:1715–26
8. Nguyen TL, Collins GS, Lamy A, Devereaux PJ, Daurès JP, Landais P, Le Manach Y: Simple randomization did not protect against bias in smaller trials. J Clin Epidemiol 2017; 84:105–13
9. Hansen BB, Bowers B: Covariate balance in simple, stratified and clustered comparative studies. Stat Sci 2008; 23:219–36
10. Allison DB, Brown AW, George BJ, Kaiser KA: Reproducibility: A tragedy of errors. Nature 2016; 530:27–9
11. Frölich MA: Role of the atrial natriuretic factor in obstetric spinal hypotension. Anesthesiology 2001; 95:371–6