



EDITORIALS

## Increasing the reproducibility of research will reduce the problem of apophenia (and more)

Philip M. Jones, MD, MSc · Janet Martin, PharmD, MSc (HTA&M)

Received: 7 April 2021 / Revised: 7 April 2021 / Accepted: 7 April 2021 / Published online: 7 May 2021  
© Canadian Anesthesiologists' Society 2021

In their thought-provoking article, Hanson *et al.*<sup>1</sup> eloquently and convincingly describe how humans are naturally predisposed to identifying meaningful patterns where none truly exist and tend to prefer positive over negative results—the tendency known as *apophenia*. In the research context, when these two elements are combined, this creates the aggregate issue of *false positive research findings* (hereafter referred to as simply *false positives*).

We agree that false positives are a real problem in the biomedical literature. They pose a serious threat to public health by creating the conditions where a medical intervention is thought to offer a benefit for a patient, but, in reality, it does not. Although it is difficult to quantify the incidence of false positives, we tend to agree with Ioannidis that it is quite possible that most published research findings (including both positive and negative studies) are false.<sup>2</sup>

Inductively, it follows that, if we *minimize* false positives, we will tend to *maximize* true positives (i.e., a study proclaiming it has “found” something will tend to be correct, rather than tending to be a false positive). It also follows that, if we create conditions that favour finding true

positives, then, under similar favourable conditions and in similar research contexts (i.e., similar patients, interventions, outcomes, sample sizes, durations, and settings), we would expect these results to be able to be reproduced by other researchers. Therefore, we believe that the *reproducibility of research* should be inversely proportional to the false positive risk. That is, as the false positive risk goes down, the reproducibility of previous research must go up. Using this line of reasoning, we contend that assessing the reproducibility of research findings may be a potent and direct measure of how likely research findings are to be correct.

In this editorial, we will expound upon the themes raised in the Hanson *et al.* paper to highlight some additional aspects contributing to the “false positive problem” in the biomedical literature and explore some practical steps that could be taken to improve the reproducibility of research.

We assert that, if we could systematically reduce the incidence of false positives (thereby increasing the reproducibility of research), many other beneficial effects would follow, including increased trust in research findings, less research funding wastage, improved clinical care for our patients, and improved net benefits to society. The key question is, therefore, *how can we systematically reduce false positives?*

Before addressing the issue of how false positives may be reduced, let us first address how reproducibility may be defined, and whether low reproducibility of research findings is a real problem.

### What is reproducibility?

Although there are multiple interpretations of the word,<sup>3</sup> we define reproducibility of research as follows:

P. M. Jones, MD, MSc ()  
Departments of Anesthesia & Perioperative Medicine and  
Epidemiology & Biostatistics, University of Western Ontario,  
London, ON, Canada  
e-mail: pjones8@uwo.ca

University Hospital - London Health Sciences Centre, 339  
Windermere Rd, Rm C3-110, London, ON N6A 5A5, Canada

J. Martin, PharmD, MSc (HTA&M)  
Departments of Anesthesia & Perioperative Medicine and  
Epidemiology & Biostatistics, University of Western Ontario,  
London, ON, Canada

*In general, under similar research circumstances (e.g., in clinical research, patients' ages and comorbidities, co-interventions, care providers, etc.), if a particular intervention actually causes a particular outcome (beneficial or adverse), then other investigators would also find similar outcomes.*

By *similar*, we mean:

*Subsequent experiments' results are within each other's 95% prediction intervals (note these prediction intervals are not traditionally computed confidence intervals).*<sup>4,5</sup>

If this does not occur, we may surmise that there is a good chance the research findings in question represent a false positive, or false discovery. Importantly, many authors have not employed a *prediction interval* as their criterion for reproducibility and have solely used statistical significance of study results to declare whether a finding could be reproduced. This technique is flawed because of sampling error.<sup>5</sup> Nonetheless, in many domains, scientists have attempted to reproduce results. These attempts show that, unfortunately, there is strong evidence that the reproducibility of findings in the biomedical literature is very low.

## Is there a crisis of reproducibility in science?

Examples of poor reproducibility of scientific discoveries abound. We will give just a few prominent examples here.

In basic science, a survey of over 1,500 researchers reported that more than 70% of them tried and failed to reproduce another scientist's experiment's results.<sup>6</sup> In an ironic twist, more than half failed to reproduce their own findings.<sup>6</sup>

In psychology, attempts to reproduce the results of 100 experimental and correlational studies published in three psychology journals, closely approximating the original conditions, showed that the subsequent effect sizes were only half the magnitude of the original effects.<sup>7</sup> Moreover, the distribution of effect sizes showed that the original studies favoured large effect sizes (unlikely to occur commonly) and small *P* values, but the subsequent studies had much smaller effect sizes, and, therefore, much larger *P* values.<sup>7</sup> Whereas 97% of the original studies had statistically significant results, only 36% of the subsequent studies were statistically significant.<sup>7</sup> Although it is a somewhat flawed metric,<sup>4</sup> only 47% of the original effect sizes were in the 95% confidence interval (CI) of the subsequent studies. Taken together, the findings suggest many of the original findings were very likely to have been false positives (false discoveries).

In preclinical cancer research, the biotechnology company Amgen attempted to reproduce the results of 53

"landmark" studies — only six (11%) were subsequently confirmed.<sup>8</sup> Similarly, in a study of 67 projects, the drug company Bayer could only validate 25% of their preclinical studies (mostly in oncology).<sup>9</sup>

In critical care, a review determined that, of 158 clinical practices examined in 275 articles, reproduction of results was attempted for 66 practices (42%).<sup>10</sup> Consistent with the psychology examples above, the original studies' effect sizes were systematically overestimated (absolute difference, 16% vs 8%; *P* = 0.003). Fifty-six percent of the reproduction attempts resulted in effects that were inconsistent with the original study, and the majority of these inconsistencies leaned toward the original study concluding efficacy of an intervention while the reproduction attempt concluded a lack of efficacy. Worse, two practices originally reported to be efficacious actually indicated harm after the reproduction study was done.<sup>10</sup> Although a similar review in the general anesthesiology literature has yet to be done, there is ample evidence that anesthesiology has not been spared from the issue of *failure to reproduce*.<sup>1,11</sup>

Non-reproducibility has also been identified in many other non-biomedical research domains such as economics,<sup>12</sup> engineering,<sup>13</sup> and machine learning.<sup>14</sup>

## Why does poor reproducibility happen?

Poor reproducibility is a multi-factorial problem. Many of the salient issues were prominently raised by Ioannidis in a seminal 2005 article provocatively entitled "Why most published research findings are false."<sup>2</sup> In this article, many elements are considered as contributors to the overall problem of false discoveries, including a limited prior probability of a finding being true, bias, limited statistical power from small sample sizes, faulty/manipulated analytical techniques, conflicts of interest (financial, promotion of academic physicians), and misinterpretation. Here, we will invoke some of Ioannidis' elements, and expand upon them, to discuss why poor reproducibility exists specifically in the anesthesiology and critical care literature.

Limited prior probability of an intervention being effective

When reviewing results from, say, a randomized clinical trial (RCT), it is reasonable to ask "Given that the *P* value for the primary outcome was 0.05, what are the chances that the findings reported are not actually real (i.e., the findings are a false discovery)?" Many people would assume that the chances are about 5%. The actual answer, however, is that the chance of the discovery being false is

closer to 36%. If this is the case, it is little wonder that there is a reproducibility problem in medicine. To understand why the false positive risk may be so high despite the *P* value being so low, we need to first start with a discussion of the prevalence of real effects of a given intervention or exposure.

Most preclinical interventions, say, from drug companies, fail in the preclinical stage.<sup>15</sup> Those drugs that do make it to formal human testing in phase 1 and 2 trials also have a very high failure rate of 85–93%.<sup>15</sup> The chances of a drug being truly effective rises somewhat in phase 3 trials because the preliminary work already done has weeded out many (but not even close to all) failures. The Table shows some estimates of the prior probability (i.e., the probability, before a study is done, of a real effect) for various study types.

The reason prior probabilities of real effects are important is because of what is called the “screening problem.” This problem is shown in Fig. 1. In essence, the lower the prior probability of a real effect, the higher the chance that any given study or clinical trial that is “statistically significant” will in fact represent a false positive. This is because many studies will be “screened” for positivity, but relatively few will actually have real effects.

Many people believe that a *P* value can be interpreted as “the probability that the results obtained were due to chance”, or, “the probability of the null hypothesis being true (i.e., no effect of an intervention)”. These interpretations are both wrong. The correct interpretation of a *P* value is “the probability of obtaining the data observed (or more extreme data) if there really was no difference between the groups.”<sup>20</sup> It is critically important to note that the correct *P* value definition says nothing about the probability of the null hypothesis being correct or incorrect, as it holds the null hypothesis (of no effect) as true, while indicating whether the data are compatible with that assumption of no effect. As clinicians, we are not interested in how “rare” the data may be when there is no difference between groups. Instead, we are interested in how likely we are to be correct when we think we have

found something “significantly different”. The latter question is answered directly by the positive predictive value of a hypothesis test as shown in Fig. 1, not by the *P* value. In the example given in Fig. 1 and based on a prevalence of real effects of 10%, the false positive risk is not 5%, but 36%. This high false positive risk is shocking to many and the reason for it is poorly understood by many, if not most, clinicians and researchers. Nonetheless, this concept is critically important when deciding which interventions, based on research, to apply to our patients. When one understands how to determine the false positive risk of clinical studies using the 2x2 table in Fig. 1, it becomes very apparent how reproducibility may be a serious issue, since many studies would have high false positive risks.

#### Fragility of results from clinical trials

Figure 2 shows a fictitious clinical trial of hydroxyethyl starch (HES) compared with saline for perioperative fluid replacement. The first analysis (*panel A*), when eight patients in the saline group and 18 in the HES group experience the adverse outcome, is statistically significant, and the study concludes that HES is harmful. Nevertheless, by shifting just one patient’s outcome in the saline group from a good outcome to an adverse outcome (*panel B*), the subsequent analysis is *not* statistically significant, and the conclusion of the study changes to “there is no evidence of HES causing harm”.

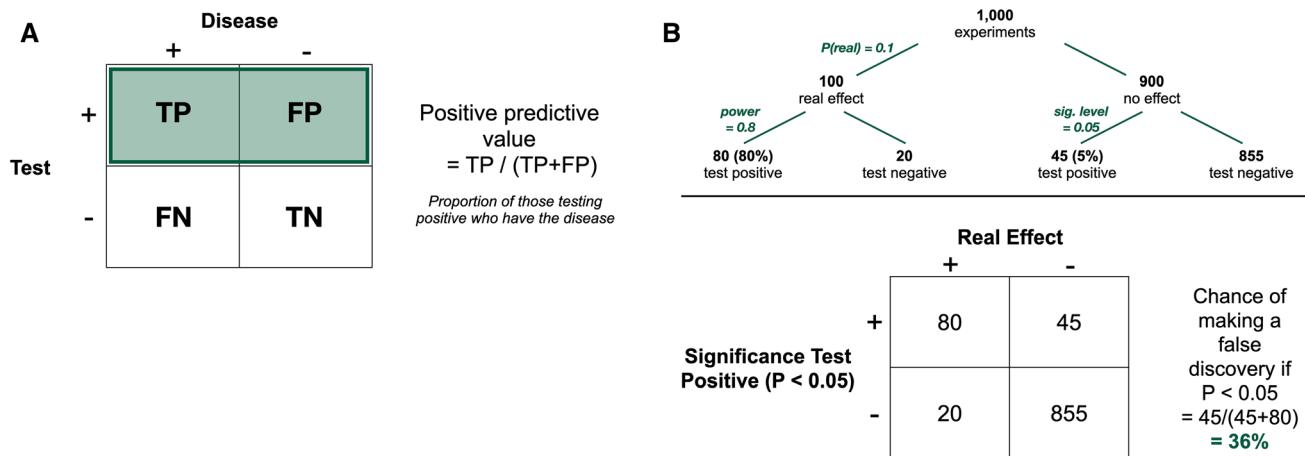
This example shows the concept of *fragility* of study results. Ideally, many outcomes would need to be changed before affecting a study’s conclusion. The fewer the outcomes that need to be altered to change the study’s conclusion, the more fragile are the study’s results. The *fragility index* (FI) can be calculated by changing the status of patients in the treatment group with the fewest events, from not having an event to having an event, until the *P* value exceeds 0.05.<sup>21</sup> For the example in Fig. 2, the FI is only 1 (i.e., extremely fragile).

Unfortunately, in anaesthesiology and critical care RCTs, the median [interquartile range] FI is very small at 3

**TABLE** Types of studies and estimated prior probabilities of real effects

Type of study	Estimated prior probability of real effect	References
Genomic	$10^{-5}$ to $10^{-3}$	16,17
Preclinical	$10^{-3}$ to 0.1	17
Phase 1 and 2 drug trials	0.01 to 0.2	15,18
Phase 3 trials	0.1 to 0.5	15,18,19
Observational studies	? (Probably quite low, perhaps) 0.01 to 0.2	2

The probability scale ranges from 0 (no chance of occurring) to 1 (always occurs)



**Fig. 1** The “screening problem.” *Panel A* shows the 2x2 table commonly used to assess diagnostic test performance. The true disease state forms the columns while the test result forms the rows. The positive predictive value (PPV) is the proportion of true positives (TP) divided by the total number of positive tests (comprising both TPs and false positives [FP]). FN = false negative; TN = true negative. *Panel B* shows the stratification of 1,000 fictitious experiments where the prevalence of a real finding is set at 10% (i.e., a probability [ $P(\text{real})$ ] of 0.1). The power of the “diagnostic test” (here, a statistical hypothesis test), which is the ability to detect TP, is set at 80% (power = 0.8). Since the power is not 100%, 20 of the experiments where there is in fact a real effect are not detected by the hypothesis test and are therefore false negatives. Of the remaining 900 experiments where no real effect is present, 45 (5%) of them will screen as positives because the significance level of the hypothesis test (false positive rate) was fixed at 5% (probability of 0.05). The important thing to understand is that the false positive risk overall is not 5%; rather it is the complement of the positive predictive value of the hypothesis test (in this case, 36%). As the prevalence of real effects approaches zero, the false positive risk approaches 100%. (Copyright of the Figure is retained by the author.) Le « problème de sélection ». *Le panneau A* montre le tableau 2x2 couramment utilisé pour évaluer les performances des tests

[1–7].<sup>21</sup> This means that changing the clinical outcomes of only three patients is enough to completely eliminate statistical significance. What is even more sobering is that the loss to follow-up in these studies often exceeds the FI, suggesting that the fragility might be even worse if the results of losses to follow-up could be known, further jeopardizing the legitimacy of the conclusions. The fact that our studies are so fragile is a major contributing factor to poor reproducibility, especially if we rely on  $P$  values to direct our conclusions.

#### Limited statistical power from small sample sizes

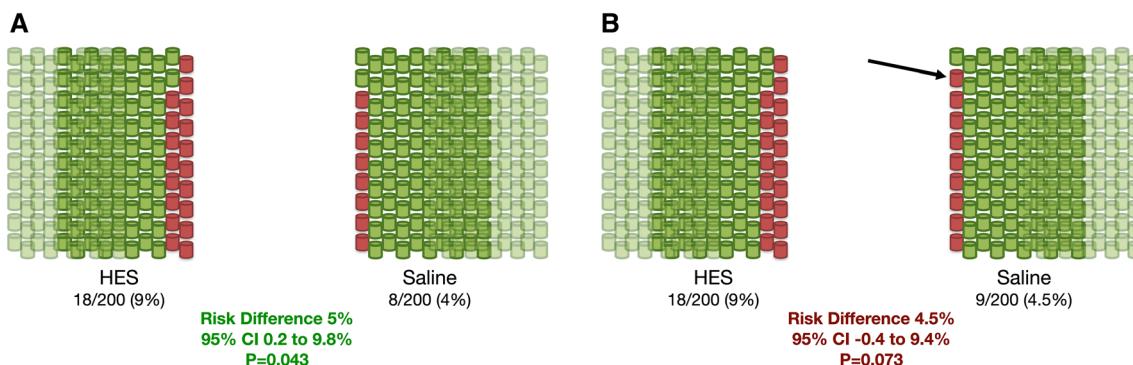
A further contribution to poor reproducibility in the anesthesiology literature, and one that is intimately related to fragility, stems from the fact that anesthesiologists routinely perform small, underpowered RCTs. A study that determined the sample sizes from RCTs published in six general anesthesiology journals in

diagnostiques. Le véritable état de la maladie forme les colonnes tandis que le résultat du test forme les lignes. La valeur prédictive positive (VPP) est la proportion de vrais positifs (VP) divisé par le nombre total de tests positifs (comprenant à la fois des VP et des faux positifs [FP]). FN = faux négatif; VN = vrai négatif. *Le panneau B* montre la stratification de 1000 expériences fictives où la prévalence d'une découverte réelle est fixée à 10 % (c.-à-d. une probabilité [ $P$  (réelle)] de 0,1). La puissance du « test diagnostique » (ici, un test d'hypothèse statistique), qui est la capacité de détecter les VP, est fixée à 80 % (puissance = 0,8). Puisque la puissance n'est pas 100 %, 20 expériences où il y a en réalité un effet réel ne sont pas détectées par le test d'hypothèse et sont donc de faux négatifs. Sur les 900 expériences restantes où aucun effet réel n'est présent, 45 (5 %) seront dépistées comme positives parce que le niveau de signification du test d'hypothèse (taux de faux positifs) a été fixé à 5 % (probabilité de 0,05). En résumé, le risque de faux positif n'est pas de 5 % globalement; il constitue plutôt le complément de la VPP du test d'hypothèse (dans ce cas, 36 %). Au fur et à mesure que la prévalence des effets réels se rapproche de zéro, le risque de faux positif se rapproche de 100 %. (Le droit d'auteur de la figure est conservé par l'auteur.)

2010 and 2016 found that the median [interquartile range] sample size of 112 RCTs published in 2016 was only 80 [52–136].<sup>22</sup> This was only a mild increase over the 143 RCTs published in 2010 (63 [41–101]).<sup>22</sup> Publishing such small studies directly contributes to poor reproducibility, since sampling error and limited statistical power will commonly result in disparate results when another study of similarly low sample size is done in an attempt to reproduce the original study's results.<sup>2</sup>

#### Manipulated analyses, confounding, and selective reporting

The scientific method serves largely as a *constraining process* that exerts tight control over critical aspects of a particular study or experiment. This control permits, in turn, more reproducible conditions for any subsequent studies. Therefore, it follows that the more manipulation or corruption of the scientific method occurs, the more we



**Fig. 2** The concept of fragility of results. In this fictitious 400-patient randomized clinical trial of perioperative fluid replacement using hydroxyethyl starch (HES) vs saline, patients are represented by the cartoon cylinders (200 in each group). Good outcomes are symbolized by green cylinders; adverse outcomes are symbolized by red cylinders. In *panel A*, 18 patients in the HES group and eight patients in the saline group have the adverse outcome, and the calculated risk difference, its 95% CI, and the  $P$  value from the Chi square test are shown. In *panel B*, the effect of switching just one patient (arrow) in the saline arm from a good outcome to an adverse outcome is sufficient to change the  $P$  value from one that was statistically significant to a  $P$  value which is no longer statistically significant. (Copyright of the Figure is retained by the author.) Le concept de fragilité des résultats. Dans cette étude clinique randomisée fictive de 400 patients sur le remplacement liquide

would expect poor reproducibility of study results because of the *researcher degrees of freedom* these manipulations open up. Manipulation of study results is a complex problem, but there are a few common themes we will discuss here.

The statistical analysis of conventional RCTs is relatively straightforward because the baseline differences between groups are, by definition, due to chance alone as long as the randomization was not corrupted. This reduces the need to control for confounding (although we acknowledge that even though confounding is not a concern for adequately powered RCTs, the precision and power of effect size estimates can be improved in small RCTs by conditioning on baseline imbalances).<sup>23,24</sup> In contrast, statistical analysis of observational studies is a major area where researchers can easily manipulate their results. On a modern computer, analyzing a data set with a million patients using most regression techniques takes merely seconds. Covariates for the regression model can be added, deleted, or transformed in seconds and the analysis repeated. Within an hour, an experienced analyst could easily run 50 or more different regression models. Every time this is done, the  $P$  values are displayed and this iterative process with “live” access to results can potentially result in data-driven analysis and conclusions. If the analyst’s goal is to determine a statistically significant relationship between a covariate and an outcome, this can often be achieved in short order. This

périopératoire utilisant l’hydroxyéthylamidon (HES) vs la solution saline, les patients sont représentés par les cylindres dessinés (200 dans chaque groupe). Les devenirs favorables sont symbolisés par des cylindres verts, les devenirs défavorables par des cylindres rouges. Dans le *panneau A*, 18 patients du groupe HES et huit patients du groupe solution saline ont les devenirs indésirables, et la différence de risque calculée, son IC de 95 %, et la valeur  $P$  du test chi carré sont affichés. Dans le *panneau B*, l’effet de changer les résultats d’un seul patient (flèche) du groupe solution saline d’un devenir favorable à un devenir défavorable est suffisant pour changer la valeur  $P$  d’une valeur statistiquement significative à une valeur  $P$  qui n’est plus statistiquement significative. (Le droit d’auteur de la figure est conservé par l’auteur.)

process is known as “ $P$  hacking.”<sup>20</sup> There is strong evidence that manipulation of analyses has resulted in researchers, determined to “achieve” statistical significance, performing successive manipulations until obtaining  $P$  values just below 0.05.<sup>25,26</sup> Worse, there is evidence from a survey of a random sample of 390 consulting biostatisticians (*American Statistical Association* members) that researchers commonly apply pressure to statisticians to manipulate results.<sup>27</sup> The most common inappropriate requests, as reported by more than 20% of the respondents, were “removing or altering some data records to better support the research hypothesis; interpreting the statistical findings on the basis of expectation, not actual results; not reporting the presence of key missing data that might bias the results; and ignoring violations of assumptions that would change results from positive to negative”.<sup>27</sup> Finally, these inappropriate requests were more common in less experienced biostatisticians, raising the disturbing spectre of researchers intimidating more junior analysts into getting their wishes.<sup>27</sup>

Further manipulations are easily achieved by analyzing a study and selecting to report those outcomes that particularly interest the researcher, or outcomes that can “spin” the results of a study the way the researcher desires. There is strong evidence that spin occurs in the anesthesiology literature.<sup>21</sup> There is similarly strong evidence that researchers commonly switch primary and

secondary outcomes from those they originally intended to report to others<sup>28,29</sup> (usually switching from non-statistically significant outcomes to statistically significant outcomes).<sup>28</sup> This problem is exacerbated by the fact that, although the *International Committee of Medical Journal Editors* first mandated RCT registration in 2005, as of 2015 only 38% of anesthesiology RCTs from six journals were adequately registered.<sup>28,A</sup> The resultant freedom to change outcomes without detection contributes to poor reproducibility of results.

Researchers sometimes selectively report entire studies. This is known as publication bias, and it is the failure to publish a study on the basis of the direction or strength of the study findings. Researchers are often tempted to preferentially report “positive” results that align with their expectations, or that make the research findings sound more interesting. Nevertheless, regardless of the motivation behind it, the net effect of publication bias is likely a decrease in the reproducibility of published study results.

Publication bias introduces bias into the evidence base and prevents accurate elicitation of the net effect of an intervention. Without the full body of evidence available, it is difficult to make informed decisions. This is akin to only the tip of the iceberg of information being made available (which is cherry-picked to represent the “positive” results), while the majority of the relevant information lies out of sight beneath the water. To understand the net effect of interventions, we need access to all of the studies, with all of their outcomes transparently reported.

Stellar examples of publication bias exist. For example, of all antidepressant studies reported to the Food and Drug Administration (FDA) for regulatory approval, over 30% were never published.<sup>30</sup> About 94% of the published studies were positive, while only 51% of the total body of studies registered with the FDA were positive.<sup>30</sup> Examples of publication bias in the anesthesiology literature also exist.<sup>31</sup>

A final manipulation of data involves fraud such as fabricating data. If the data are not real, it is not hard to see that it will be very difficult to reproduce the results! We would all like to believe that fraud in anesthesiology research is rare. Unfortunately, we have famous examples of fraudulent research in our speciality.<sup>32</sup> There is also a systematic review of survey results showing that approximately 12% of researchers had personal knowledge of a colleague who had fabricated or falsified data, and 2% admitted to doing it themselves.<sup>33</sup> These

findings, due to the sensitivity of the survey questions, very likely underestimate the true incidence of fraudulent research activities and raise the troubling possibility that research fraud may be more commonplace than we have previously acknowledged.

## How can we improve the reproducibility of research?

Complex problems often involve complex solutions. Nevertheless, we argue that, despite the complexity of the reproducibility problem, several relatively simple and concrete steps could be taken, almost immediately, to improve the rate of reproducibility of research in anesthesiology.

1) We cannot expect high quality, reproducible research with inadequate research training. Not all physicians, even in academic centres, need to be “researchers”. We mandate that medical doctors are trained properly to deliver proper clinical care, but, historically, we have allowed anyone (even those without formal training) to perform “research”. Medical doctors who are identified as researchers must be trained, and, possibly, credentialed. Formal guidelines for what constitutes a minimum of proper training should be developed. As Altman stated in 1994, “Why are [false discoveries] so common? Put simply, much poor research arises because researchers feel compelled for career reasons to carry out research that they are ill equipped to perform, and nobody stops them. Regardless of whether a doctor intends to pursue a career in research, he or she is usually expected to carry out some research with the aim of publishing several papers. The length of a list of publications is a dubious indicator of ability to do good research; its relevance to the ability to be a good doctor is even more obscure.”<sup>34</sup>

2) It follows from the above that universities must stop incentivizing poor-quality research by allowing academic promotion with a few excellent quality studies rather than simply rewarding the quantity of poor-quality studies. Systematically pushing toward higher quality studies will improve their reproducibility.

3) Building from Hanson *et al.*’s suggestion,<sup>1</sup> reproducibility will likely increase if registration occurs for RCTs and observational studies. The latter have traditionally escaped serious scrutiny while RCTs have attracted a lot of standardization in their planning, conduct, and reporting. But, as we have argued, RCTs are usually at a *lower* risk of bias from a flawed or manipulated analysis whereas observational studies are at a very high risk. Having researchers commit, *a priori*, to an analysis plan that defines exposures, inclusion and exclusion criteria, outcomes, and pre-planned data manipulations/transformations will reduce researcher degrees of freedom and

<sup>A</sup> The Canadian Journal of Anesthesia’s Policy Statement on Registration of Randomized Clinical Trials and Systematic Reviews. Available from URL: <https://www.springer.com/journal/12630/submission-guidelines> (accessed April 2021).

improve the reproducibility of findings from observational studies. ClinicalTrials.gov has offered registration of observational studies, for free, since inception.<sup>B</sup>

An extension of the above is that, because many researcher degrees of freedom are contained within a statistical analysis, very detailed statistical analysis plans (SAPs) should be expected for every type of study, and they should be registered before the study begins. Many general study registration sites do not have the ability to upload documents including protocols or SAPs, but fortunately alternatives for registration exist that do permit uploading of protocol/SAP documents.<sup>C</sup> These documents are then immutable and they receive a timestamp and a digital object identifier (DOI). The goal of the SAP is to constrain the researcher into particular decisions before they have a chance to see the data. Notably, this includes detailed lists of variables to be included in regression models.

Having an SAP does not in any way stifle the ability of researchers to explore their data or to identify unexpected findings. Nevertheless, because the researchers did not plan the analysis, they should clearly label the analysis in the published article as “exploratory” or “unplanned”. This provides important transparency and context for a reader.

4) The amount of research funding wasted has been estimated to be as high as 85%.<sup>35</sup> To reduce this waste, we should fund fewer small RCTs since they have often have virtually no potential to change clinical care in a meaningful way. Instead, in our opinion, research funding should be concentrated on early-stage development studies (for novel drugs, devices, and technologies) and very large RCTs (ideally, 1,000 to 50,000 participants). The interventions for these large, pragmatic RCTs should be prioritized by a consensus group of stakeholders (including patients), and they should mostly be multicentre<sup>1</sup> (since the risk of biased effect estimates is likely lower in multicentre RCTs<sup>36</sup> and their external validity is greater). Large, multicentre RCTs will naturally have better reproducibility (i.e., much lower risk of false positive research findings).

5) We echo Hanson *et al.*'s suggestion<sup>1</sup> that journals should create systems whereby they would guarantee publication of papers whose authors have pre-submitted all aspects of a highly-protocolized study before they begin recruitment, irrespective of results, as long as the protocol was followed without significant deviation (also known as “pre-registered reports”). The rigid adherence to a pre-

specified protocol will improve reproducibility of their work.

6) Where statistical methods are complex, which usually occurs in observational studies, authors should be expected to provide multiple sensitivity analyses. If the observed effect is real (and therefore likely reproducible), it should be congruent with the results from most other analytical methods. When the methods are incongruent, this provides valuable information to the reader on the fidelity of the conclusions proffered by the authors, and it provides the authors with a chance to explain why the results were model-sensitive.

7) In addition to the typical frequentist statistical reporting using *P* values and confidence intervals, we should also incorporate Bayesian interpretations of results. This would afford the opportunity to incorporate a *prior probability* of a real effect into a *posterior probability* and more accurately reflect the risk of false discoveries. Said differently, Bayesian interpretations would provide better information about which results are likely to be reproducible and which are not. Multiple ways of accomplishing this Bayesian framework, based on frequentist *P* values, have been published and could be readily incorporated by medical journals.<sup>20,37</sup>

8) In agreement with Hanson *et al.*,<sup>1</sup> we believe serious consideration to using a stricter *P* value threshold as nominal statistical significance (say, < 0.005 instead of < 0.05) should be given. If a true effect exists, and if studies' sample sizes are adequate, a stricter threshold should filter out false positives while allowing most true positives to still be correctly identified.<sup>38-40</sup>

## Conclusions

*Apophenia*, or the tendency of identifying meaningful patterns where none truly exist, is part of the human condition. We cannot surmount it, but we can strive to manage its influence. In the research context, apophenia can lead to a substantial risk of false positive findings. We believe that the issue of false positive research findings and the poor reproducibility of research findings are inextricably linked.

In this editorial, we have outlined that poor reproducibility in the biomedical literature is a serious problem. We have discussed some of the principal reasons why poor reproducibility may be occurring, and we have offered some practical suggestions that may improve rates of reproducibility. We believe that increasing research reproducibility will, in turn, improve the confidence we have in our anesthesiology, perioperative medicine, critical care, and pain interventions and will improve our patients' clinical outcomes. The time for simply being aware of

<sup>B</sup> Registration of Observational Studies at ClinicalTrials.gov. Available from URL: <https://clinicaltrials.gov/ct2/manage-recs/how-register#Considerations> (accessed April 2021).

<sup>C</sup> For example, the Open Science Framework. Available from URL: <https://osf.io> (accessed April 2021).

these issues is long over. It is now time for all of us—researchers, readers, clinicians, peer reviewers, editors, and administrative leaders—to act.

## L'augmentation de la reproductibilité de la recherche réduira le problème de l'apophénie (entre autres)

Dans leur article portant à réflexion, Hanson *et coll.*<sup>1</sup> décrivent de façon à la fois éloquente et convaincante comment l'Homme est naturellement prédisposé à identifier des corrélations significatives là où il n'y en a pas et a tendance à préférer les résultats positifs aux résultats négatifs, une tendance connue sous le nom d'*apophénie*. Dans le contexte de la recherche, lorsque ces deux tendances sont combinées, cela crée le problème cumulé de ce qu'on appelle des *résultats de recherche faussement positifs* (ci-après appelés simplement *faux positifs*).

Nous convenons que les faux positifs constituent un problème bien réel dans la littérature biomédicale. Ils posent un risque important pour la santé publique en créant des conditions dans lesquelles une intervention médicale est considérée bénéfique pour un patient alors qu'en réalité, elle ne l'est pas. Bien qu'il soit difficile de quantifier l'incidence des faux positifs, nous sommes généralement d'accord avec Ioannidis pour dire qu'il est tout à fait possible que la plupart des résultats de recherche publiés (y compris les études positives et négatives) soient faux.<sup>2</sup>

Par raisonnement inductif il s'ensuit que, si nous *minimisons* les faux positifs, nous aurons tendance à *maximiser* les vrais positifs (c.-à-d. une étude proclamant avoir « trouvé » quelque chose aura tendance à être correcte, plutôt que d'avoir tendance à être un faux positif). Il s'ensuit également que, si nous créons des conditions favorables à la recherche de vrais positifs, alors, dans des conditions favorables similaires et dans des contextes de recherche similaires (c.-à-d. des patients, des interventions, des critères d'évaluation, des tailles d'échantillons, des durées et des paramètres similaires), nous nous attendons à ce que ces résultats puissent être reproduits par d'autres chercheurs. Ainsi, nous croyons que la *reproductibilité de la recherche* devrait être inversement proportionnelle au risque de faux positif. En d'autres termes, à mesure que le risque de faux positif baisse, la reproductibilité des recherches antérieures devrait augmenter. En se fondant sur ce raisonnement, nous soutenons que l'évaluation de la reproductibilité des résultats de la recherche peut être une

mesure puissante et directe de la probabilité que des résultats de recherche soient corrects.

Dans cet éditorial, nous discuterons des thèmes abordés dans l'article de Hanson *et coll.* afin de mettre en évidence d'autres éléments contribuant au « problème de faux positifs » dans la littérature biomédicale; en outre, nous explorerons certaines mesures pratiques qui pourraient être posées pour améliorer la reproductibilité de la recherche.

Nous affirmons que, si nous pouvions réduire de manière systématique l'incidence des faux positifs (augmentant ainsi la reproductibilité de la recherche), de nombreux autres effets bénéfiques s'ensuivraient, notamment une confiance accrue dans les résultats de recherche, moins de gaspillage dans le financement de la recherche, une amélioration des soins cliniques pour nos patients et une amélioration des avantages nets pour la société. La question clé est donc de savoir *comment réduire les faux positifs de façon systématique?*

Avant d'aborder la question de savoir comment réduire les faux positifs, penchons-nous en premier lieu sur la façon de définir la reproductibilité, et tentons de déterminer si une faible reproductibilité des résultats de recherche est un véritable problème.

### Qu'est-ce que la reproductibilité?

Bien qu'il existe de multiples interprétations du mot,<sup>3</sup> nous définissons la reproductibilité de la recherche comme suit :

*En générل, dns des circonstncest de recherche semblbles (p. ex., dns l recherche clinique, l'âge et les comorbidités des ptients, les co-interventions, les fournisseurs de soins, etc.), si une intervention en prticulier provoque effectivement un résultat pticulier (bénéfique ou défavorable), d'autres chercheurs prviendrinent également à des résultats semblbles.*

pr semblbles, nous entendons :

*Les résultats d'expériences ultérieures se trouvent dns les intervalles de prédiction de 95 % les uns des utres (notez que ces intervalles de prédiction ne sont ps les intervalles de confince trditionnellement clculés).<sup>4,5</sup>*

Si cel ne se produit ps, nous pouvons supposer qu'il y a de fortes chances que les résultats de la recherche en question représentent un faux positif, ou une fausse découverte. Fait important, de nombreux auteurs n'ont pas utilisé un *intervalle de prédiction* comme critère de reproductibilité et ont uniquement utilisé la signification statistique des résultats de l'étude pour déclarer si leurs conclusions pouvaient être reproduites. Cette technique est incorrecte en raison des erreurs d'échantillonnage.<sup>5</sup> Néanmoins, dans de nombreux domaines, les scientifiques ont tenté de reproduire les résultats. Ces

tentatives montrent que, malheureusement, il existe de solides données probantes que, dans la littérature biomédicale, la reproductibilité des résultats est très faible.

### **Y a-t-il une crise de la reproductibilité en sciences ?**

Les exemples de faible reproductibilité des découvertes scientifiques abondent. Nous ne donnerons ici que quelques exemples marquants.

En science fondamentale, une enquête menée auprès de plus de 1500 chercheurs a indiqué que plus de 70 % d'entre eux avaient essayé de et échoué à reproduire les résultats d'expériences d'un autre scientifique.<sup>6</sup> Ironie du sort, plus de la moitié d'entre eux n'ont pas réussi à reproduire leurs propres observations!<sup>6</sup>

En psychologie, les tentatives de reproduire les résultats de 100 études expérimentales et corrélationnelles publiées dans trois revues de psychologie, en reproduisant aussi étroitement que possible les conditions originales, ont montré que les tailles d'effet subséquentes n'étaient que de la moitié de l'ampleur des effets originaux.<sup>7</sup> En outre, la répartition des tailles d'effet a montré que les études initiales favorisaient les grandes tailles d'effet (peu susceptibles de se produire couramment) et les petites valeurs  $P$ , mais les études subséquentes avaient eu des tailles d'effets beaucoup plus petites, et, par conséquent, des valeurs  $P$  beaucoup plus importantes.<sup>7</sup> Alors que 97 % des études initiales ont rapporté des résultats statistiquement significatifs, seulement 36 % des études subséquentes étaient statistiquement significatives.<sup>7</sup> Bien qu'il s'agisse d'une mesure quelque peu imparfaite,<sup>4</sup> seulement 47 % des tailles d'effet originales se trouvaient dans l'intervalle de confiance (IC) de 95 % des études subséquentes. Pris ensemble, ces résultats suggèrent que bon nombre des résultats originaux étaient très susceptibles d'avoir été de faux positifs (ou de fausses découvertes).

Dans la recherche préclinique sur le cancer, la société de biotechnologie Amgen a tenté de reproduire les résultats de 53 études « phare » — seulement six (11 %) résultats ont été confirmés par la suite.<sup>8</sup> De la même manière, dans une étude portant sur 67 projets, la société pharmaceutique Bayer n'a pu valider que 25% de ses études précliniques (principalement en oncologie).<sup>9</sup>

En soins intensifs, un compte rendu a permis de déterminer que, sur les 158 pratiques cliniques examinées dans 275 articles, la reproduction des résultats avait été tentée pour 66 pratiques (42 %).<sup>10</sup> Conformément aux exemples de psychologie ci-dessus, les tailles d'effet des études originales ont été systématiquement surestimées (différence absolue, 16 % vs 8 %;  $P = 0,003$ ). Cinquante-six pour cent des tentatives de reproduction ont eu comme résultat des effets qui ne correspondaient pas à ceux de

l'étude originale, et la majorité de ces incohérences touchaient au fait que l'étude originale avait conclu à l'efficacité d'une intervention tandis que la tentative de reproduction concluait à un manque d'efficacité. Pire encore, deux pratiques initialement signalées comme efficaces ont été rapportées comme étant nocives après la fin de l'étude de reproduction.<sup>10</sup> Bien qu'un examen similaire dans la littérature générale d'anesthésiologie n'ait pas encore été fait, les données probantes selon lesquelles l'anesthésiologie n'a pas été épargnée en ce qui touche à la question de *défaut de reproduction* abondent.<sup>1,11</sup>

La non-reproductibilité a également été identifiée dans de nombreux autres domaines de recherche non biomédicale tels que l'économie,<sup>12</sup> l'ingénierie<sup>13</sup> et l'apprentissage automatique.<sup>14</sup>

### **Pourquoi une mauvaise reproductibilité se produit-elle?**

La mauvaise reproductibilité est un problème multifactoriel. Bon nombre des problèmes marquants ont été décrits brillamment par Ioannidis en 2005 dans son article novateur au titre provocateur : « Pourquoi la plupart des résultats de recherche publiés sont faux ».<sup>2</sup> Dans cet article, de nombreux éléments sont considérés comme des contributeurs au problème global des fausses découvertes, y compris la probabilité initiale limitée qu'une conclusion soit vraie, les biais, une puissance statistique limitée liée à de petites tailles d'échantillons, des techniques analytiques défectueuses/manipulées, des conflits d'intérêts (financiers, liés à la promotion universitaire)<sup>4</sup> et une mauvaise interprétation. Nous aborderons ici certains des éléments d'Ioannidis et les approfondirons afin de discuter des raisons pour lesquelles une mauvaise reproductibilité existe spécifiquement dans la littérature d'anesthésiologie et des soins intensifs.

La probabilité initiale limitée qu'une intervention soit efficace

Lors de l'examen des résultats d'une étude clinique randomisée contrôlée (ERC), par exemple, il est raisonnable de se demander : « Si la valeur  $P$  pour le critère d'évaluation principal est de 0,05, quelles sont les chances que les résultats rapportés ne soient pas véritablement réels (c.-à-d. que les résultats soient une fausse découverte)? » Nombreux sont ceux qui supposeraient que les chances sont d'environ 5 %. La véritable réponse, cependant, est que la probabilité que la découverte soit fausse est plus proche de 36 %. Si tel est le cas, il n'est alors pas étonnant qu'il y ait un problème de reproductibilité en médecine. Pour comprendre pourquoi le risque de faux positif peut être si élevé malgré une valeur

*P* si faible, nous devons d'abord commencer par une discussion portant sur la prévalence des effets réels d'une intervention ou d'une exposition donnée.

Par exemple, la plupart des interventions précliniques des compagnies pharmaceutiques échouent à l'étape préclinique.<sup>15</sup> Les médicaments qui se rendent aux tests humains officiels dans les essais de phase 1 et 2 ont également un taux d'échec très élevé de 85-93 %.<sup>15</sup> Les chances qu'un médicament soit véritablement efficace augmentent quelque peu dans les essais de phase 3, parce que le travail préliminaire déjà effectué a éliminé de nombreux échecs (mais pas tous). Le tableau montre certaines estimations de la probabilité initiale (c.-à-d. la probabilité, avant qu'une étude ne soit réalisée, d'un effet réel) pour divers types d'étude.

La raison pour laquelle les probabilités initiales d'effets réels sont importantes réside dans ce qu'on appelle le « problème de sélection ». Ce problème est indiqué dans la figure 1. Essentiellement, plus la probabilité initiale d'un effet réel est faible, plus les chances qu'une étude ou un essai clinique donné qui est « statistiquement significatif » soit en fait un faux positif seront élevées. Ceci est dû au fait que de nombreuses études seront « triées » pour leur positivité, mais relativement peu d'études auront véritablement des effets réels.

Beaucoup de gens croient qu'une valeur *P* peut être interprétée comme « la probabilité que les résultats obtenus étaient dus au hasard » ou « la probabilité que l'hypothèse nulle soit vraie (c.-à-d. aucun effet d'une intervention) ». Ces interprétations sont toutes deux erronées. L'interprétation correcte d'une valeur *P* est « la probabilité d'obtenir les données observées (ou des données plus extrêmes) s'il n'y avait réellement aucune différence entre les groupes ».<sup>20</sup> Il est d'une importance cruciale de noter que la définition correcte de la valeur *P* n'a aucun rapport avec la probabilité que l'hypothèse nulle soit correcte ou incorrecte, étant donné qu'elle part du principe que l'hypothèse nulle (sans effet) est vraie, tout en indiquant si les données sont compatibles avec cette hypothèse d'effet nul. En tant que cliniciens, peu nous importe de savoir dans quelle mesure les données sont

« rares » lorsqu'il n'y a pas de différence entre les groupes. Ce qui nous intéresse plutôt, c'est la probabilité que nous ayons raison lorsque nous pensons avoir trouvé quelque chose de « significativement différent ». La réponse directe à cette question réside dans la valeur prédictive positive d'un test d'hypothèse tel que démontré dans la figure 1, et non par la valeur *P*. Dans l'exemple donné à la figure 1 et basé sur une prévalence d'effets réels de 10 %, le risque de faux positifs n'est donc pas de 5 %, mais de 36 %. Ce risque élevé de faux positifs est choquant pour beaucoup et la raison en est mal comprise par de nombreux cliniciens et chercheurs, sinon la majorité. Néanmoins, ce concept est d'une importance cruciale lorsqu'il s'agit de décider quelles interventions, basées sur la recherche, s'appliquent à nos patients. Lorsque l'on comprend comment déterminer le risque de faux positifs des études cliniques à l'aide du tableau 2x2 de la figure 1, il devient très évident à quel point la reproductibilité peut être un problème grave, puisque de nombreuses études souffrent d'un risque élevé de faux positifs.

#### Fragilité des résultats des essais cliniques

La figure 2 montre un essai clinique fictif comparant l'hydroxyéthylamidon (HES) à une solution saline pour le remplacement liquidien périopératoire. La première analyse (*panneau A*), lorsque huit patients du groupe salin et 18 du groupe HES ressentent des devenirs indésirables, est statistiquement significative, et l'étude conclut que le HES est nocif. Néanmoins, en déplaçant le résultat d'un seul patient dans le groupe salin d'un devenir favorable à un devenir défavorable (*panneau B*), l'analyse subséquente *n'est pas* statistiquement significative, et la conclusion de l'étude change à « il n'y a pas de données probantes soutenant que le HES est nocif ».

Cet exemple démontre le concept de *fragilité* des résultats d'une étude. Idéalement, de nombreux résultats devraient être modifiés avant d'affecter la conclusion d'une étude. Moins il y a de résultats à altérer pour modifier la conclusion d'une étude, plus les résultats de l'étude sont fragiles. L'*indice de fragilité* (IF) peut être calculé en

**TABLEAU** Types d'études et probabilités initiales estimées d'effets réels

Type d'étude	Probabilité antérieure estimée d'effet réel	Références
Génomique	$10^{-5}$ à $10^{-3}$	16,17
Préclinique	$10^{-3}$ à 0,1	17
Études médicamenteuse de phase 1 et 2	0,01 à 0,2	15,18
Études de phase 3	0,1 à 0,5	15,18,19
Études observationnelles	? (Probablement assez faible, peut-être) 0,01 à 0,2	2

L'échelle de probabilité varie de 0 (aucune chance de se produire) à 1 (se produit toujours)

modifiant le statut des patients dans le groupe de traitement présentant le moins d'événements, les faisant passer de 0 événement à un événement, jusqu'à ce que la valeur  $P$  dépasse 0,05.<sup>21</sup> Concernant l'exemple présenté dans la figure 2, l'IF n'est que de 1 (c.-à-d. extrêmement fragile).

Malheureusement, dans les ERC en anesthésiologie et en soins intensifs, l'IF médian [écart interquartile] est très faible, à 3 [1-7].<sup>21</sup> Cela signifie qu'il suffit de changer les résultats cliniques de seulement trois patients pour éliminer complètement la signification statistique. Ce qui est encore plus déconcertant, c'est que dans ces études, la perte au suivi dépasse souvent l'IF, ce qui suggère que la fragilité pourrait être encore plus prononcée si les résultats des patients perdus au suivi étaient connus, mettant encore plus en péril la légitimité des conclusions. Le fait que nos études soient si fragiles est un facteur majeur contribuant à une mauvaise reproductibilité, surtout si nous nous appuyons sur les valeurs  $P$  pour tirer nos conclusions.

#### Puissance statistique limitée à cause de petites tailles d'échantillons

Un autre facteur contribuant à la mauvaise reproductibilité dans la littérature d'anesthésiologie, qui est intimement lié à la fragilité, provient du fait que les anesthésiologistes réalisent bien souvent des ERC de petite taille qui manquent de puissance. Une étude qui a déterminé la taille de l'échantillon des ERC publiées dans six revues d'anesthésiologie générale en 2010 et 2016 a révélé que la taille d'échantillon médiane [écart interquartile] de 112 ERC publiées en 2016 n'était que de 80 [52-136].<sup>22</sup> Il ne s'agissait que d'une légère augmentation par rapport aux 143 ERC publiées en 2010 (63 [41-101]).<sup>22</sup> La publication de si petites études contribue directement à une mauvaise reproductibilité, puisque les erreurs d'échantillonnage et la puissance statistique limitée se traduiront souvent par des résultats disparates lorsqu'une autre étude de taille d'échantillon tout aussi faible est réalisée dans le but de reproduire les résultats de l'étude originale.<sup>2</sup>

#### Analyses manipulées, et communication confondante et sélective

La méthode scientifique est en grande partie un *processus contraignant* exerçant un contrôle serré sur les aspects cruciaux d'une étude ou d'une expérience en particulier. Ce contrôle permet à son tour de créer des conditions plus reproductibles pour toute étude ultérieure. Par conséquent il s'ensuit que, plus il y a de manipulation ou de corruption de la méthode scientifique, plus nous devrions nous attendre à une mauvaise reproductibilité des résultats de l'étude en raison du *niveau de liberté du chercheur* que ces manipulations permettent. La manipulation des résultats

d'une étude est un problème complexe, mais il existe quelques thèmes communs dont nous discuterons ici.

L'analyse statistique des ERC conventionnelles est relativement simple parce que les différences de base entre les groupes sont, par définition, dues au hasard, si tant est que la randomisation n'a pas été corrompue. Cela réduit la nécessité de contrôler les facteurs de confusion (bien que nous reconnaissions que même si ce n'est pas un problème pour les ERC ayant suffisamment de puissance, la précision et la puissance des estimations de tailles d'effets peuvent être améliorées dans les petites ERC en manipulant les déséquilibres initiaux).<sup>23,24</sup> En revanche, l'analyse statistique des études observationnelles est un domaine majeur où les chercheurs peuvent facilement manipuler leurs résultats. Sur un ordinateur moderne, l'analyse d'un ensemble de données avec un million de patients utilisant la plupart des techniques de régression ne prend que quelques secondes. Les covariables pour le modèle de régression peuvent être ajoutées, supprimées ou transformées en quelques secondes et l'analyse répétée. En moins d'une heure, un analyste expérimenté pourrait facilement exécuter 50 modèles différents de régression ou plus. À chaque fois, les valeurs  $P$  sont affichées et ce processus itératif avec accès « en direct » aux résultats peut potentiellement donner lieu à des analyses et à des conclusions axées sur les données. Si l'objectif de l'analyste est de déterminer une relation statistiquement significative entre une covariable et un résultat, cela peut souvent être réalisé en un rien de temps. Ce processus est connu sous le nom de piratage de  $P$  ou « *P hacking* ».<sup>27,20</sup> Il existe des données probantes solides que la manipulation des analyses a permis à des chercheurs, déterminés à « atteindre » une signification statistique, d'effectuer des manipulations successives jusqu'à obtenir des valeurs  $P$  juste en dessous de 0,05.<sup>25,26</sup> Pire encore, des données probantes d'une enquête menée auprès d'un échantillon aléatoire de 390 biostatisticiens-conseils (membres de l'*American Statistical Association*) attestent que les chercheurs font souvent pression sur les statisticiens pour qu'ils manipulent les résultats.<sup>27</sup> Les demandes inappropriées les plus courantes, telles que rapportées par plus de 20 % des répondants, étaient « la suppression ou la modification de certaines données afin de mieux appuyer l'hypothèse de recherche; l'interprétation des résultats statistiques en fonction des attentes et non des résultats réels; la non-communication de l'absence de données clés qui pourraient biaiser les résultats; et le déni des violations des hypothèses qui changeraient les résultats de positifs à négatifs ».<sup>27</sup> Enfin, ces demandes inappropriées étaient plus fréquentes auprès des biostatisticiens moins expérimentés, ce qui soulève le problème inquiétant de chercheurs intimidant des analystes plus jeunes afin de parvenir à leurs fins.<sup>27</sup>

D'autres manipulations sont facilement réalisées en analysant une étude et en choisissant de rapporter les résultats qui intéressent particulièrement le chercheur, ou les conclusions qui peuvent « faire parler » les résultats d'une étude comme le souhaite le chercheur. Certaines données probantes convaincantes existent, selon lesquelles ce type d'interprétation est chose commune dans la littérature d'anesthésiologie.<sup>21</sup> Il existe également des données probantes solides que les chercheurs modifient régulièrement les critères d'évaluation primaires et secondaires de ceux qu'ils avaient l'intention de rapporter<sup>28,29</sup> (passant habituellement de résultats non statistiquement significatifs à des résultats statistiquement significatifs).<sup>28</sup> Ce problème est exacerbé par le fait que, bien que le *Comité international des rédacteurs de revues médicales* ait commencé à rendre obligatoire l'enregistrement des ERC en 2005, seulement 38 % des ERC en anesthésiologie dans six revues avaient été correctement enregistrées en 2015.<sup>28A</sup> La liberté de changer les résultats sans détection qui en résulte contribue à une mauvaise reproductibilité des résultats.

Les chercheurs rapportent parfois des études entières de manière sélective. C'est ce qu'on appelle le biais de publication, soit l'omission de publier une étude sur la base de l'orientation ou de la force des résultats de l'étude. Les chercheurs sont souvent tentés de rapporter de préférence des résultats « positifs » qui correspondent à leurs attentes ou qui rendent les résultats de la recherche plus intéressants. Néanmoins, quelle que soit la motivation qui le sous-tend, l'effet net du biais de publication est probablement une diminution de la reproductibilité des résultats publiés de l'étude.

Le biais de publication introduit un biais dans la base de données probantes et empêche la compréhension exacte de l'effet net d'une intervention. Sans avoir accès à l'ensemble des données probantes, il est difficile de prendre des décisions éclairées. Cela s'apparente à n'avoir accès qu'à la pointe de l'iceberg de l'information (qui est choisie avec soin pour représenter les résultats « positifs »), tandis que la majorité des informations pertinentes sont immergées sous l'eau. Pour comprendre l'effet net des interventions, nous avons besoin d'avoir accès à toutes les études, et à tous leurs résultats, communiqués de façon transparente.

Il existe d'excellents exemples de biais de publication. Par exemple, parmi toutes les études sur les antidépresseurs présentées à la *Food and Drug Administration* (FDA) pour approbation réglementaire, plus de 30 % n'ont jamais été

publiées.<sup>30</sup> Environ 94 % des études publiées étaient positives, alors que seulement 51 % du nombre total d'études enregistrées auprès de la FDA étaient positives.<sup>30</sup> Il existe aussi des exemples de biais de publication dans la littérature d'anesthésiologie.<sup>31</sup>

Une dernière manipulation des données implique la fraude, telle que la fabrication de données. Si les données ne sont pas réelles, inutile d'être un génie pour réaliser qu'il sera très difficile de reproduire les résultats! Nous aimerions tous croire que la fraude dans la recherche en anesthésiologie est rare. Malheureusement, il existe des exemples tristement célèbres de recherche frauduleuse dans notre spécialité.<sup>32</sup> En outre, selon une revue systématique des résultats d'un sondage, environ 12 % des chercheurs étaient personnellement au courant qu'un collègue avait fabriqué ou falsifié des données, et 2 % ont admis l'avoir fait eux-mêmes.<sup>33</sup> Ces résultats, en raison de la sensibilité des questions de l'enquête, sous-estiment très probablement l'incidence réelle des activités de recherche frauduleuses et soulèvent la possibilité troublante que la fraude en recherche pourrait être plus courante que nous ne le pensions.

### **Comment pouvons-nous améliorer la reproductibilité de la recherche?**

Les problèmes complexes impliquent souvent des solutions complexes. Néanmoins, nous soutenons que, malgré la complexité du problème de reproductibilité, plusieurs mesures relativement simples et concrètes pourraient d'ores et déjà être prises pour améliorer le taux de reproductibilité de la recherche en anesthésiologie.

- 1) Nous ne pouvons pas nous attendre à une recherche reproductible de haute qualité si la formation en recherche est inadéquate. Tous les médecins, même dans les centres universitaires, n'ont pas besoin d'être des « chercheurs ». Les médecins ont l'obligation d'être formés correctement pour fournir des soins cliniques appropriés; cependant, historiquement, nous avons permis à n'importe qui (même ceux qui n'ont pas de formation officielle) de faire de la « recherche ». Les médecins identifiés comme chercheurs doivent être formés et, possiblement, accrédités. Des lignes directrices officielles sur ce qui constitue un minimum de formation appropriée devraient être élaborées. Comme Altman le déclarait en 1994 : « Pourquoi [les fausses découvertes] sont-elles si courantes? En d'autres termes, beaucoup de mauvaises recherches sont menées parce que les chercheurs se sentent obligés, pour des raisons professionnelles, d'entreprendre des recherches qu'ils ne sont pas bien

<sup>A</sup> The Canadian Journal of Anesthesia's Policy Statement on Registration of Randomized Clinical Trials and Systematic Reviews. Disponible à l'adresse: <https://www.springer.com/journal/12630/submission-guidelines> (consulté avril 2021).

formés pour exécuter, et personne ne les arrête. Peu importe si un médecin a l'intention de poursuivre une carrière en recherche, l'attente générale est qu'il ou elle fasse de la recherche avec pour but de publier plusieurs articles. La longueur d'une liste de publications est un indicateur douteux de la capacité de faire de la recherche de qualité; sa pertinence quant à la capacité d'un individu d'être un bon médecin est encore plus nébuleuse. »<sup>34</sup>

- 2) Il découle de ce qui précède que les universités doivent cesser d'encourager la recherche de mauvaise qualité en favorisant la promotion académique avec quelques études d'excellente qualité plutôt que de simplement récompenser une quantité d'études de mauvaise qualité. C'est en promouvant systématiquement des études de meilleure qualité qu'on améliorera leur reproductibilité.
- 3) En prenant la suggestion de Hanson *et coll.* comme point de départ,<sup>1</sup> la reproductibilité augmentera probablement si les ERC et les études observationnelles sont enregistrées. Traditionnellement, les études observationnelles ont échappé à tout examen rigoureux, alors que les ERC ont fait l'objet de beaucoup de standardisation dans leur planification, leur réalisation et la communication de leurs résultats. Mais, comme nous l'avons mentionné plus haut, les ERC courrent généralement un risque moins élevé de biais lié à une analyse erronée ou manipulée, alors que le risque est très élevé pour les études observationnelles. En demandant aux chercheurs de s'engager, *a priori*, à respecter un plan d'analyse qui définit les expositions, les critères d'inclusion et d'exclusion, les critères d'évaluation et les manipulations/transformations de données pré-planifiées, il sera possible de réduire les degrés de liberté des chercheurs et d'améliorer la reproductibilité des résultats des études observationnelles. Depuis sa création, ClinicalTrials.gov offre l'enregistrement gratuit des études observationnelles.<sup>B</sup>

Par extension à ce qui précède, parce que les chercheurs disposent de nombreux degrés de liberté dans une analyse statistique, des plans d'analyse statistique très détaillés devraient être fournis pour chaque type d'étude, et ces derniers devraient être enregistrés avant le début de l'étude. De nombreux sites d'enregistrement d'études générales n'offrent pas la possibilité de téléverser des documents, y compris des protocoles ou des plans d'analyse statistique, mais heureusement, il existe des solutions de rechange pour

l'enregistrement qui autorisent le téléversement de documents de protocoles/plans d'analyse statistique.<sup>C</sup> Une fois téléchargés, ces documents sont alors immuables et horodatés, et reçoivent un identificateur d'objets numériques (DOI). L'objectif du plan d'analyse statistique est de contraindre le chercheur à prendre des décisions spécifiques avant d'avoir la chance de voir les données. Il s'agit notamment de listes détaillées de variables à inclure dans les modèles de régression.

Le fait de disposer d'un plan d'analyse statistique n'étouffe en rien la capacité des chercheurs à explorer leurs données ou identifier des résultats inattendus. Néanmoins, comme les chercheurs n'ont pas planifié l'analyse, ils devraient clairement qualifier l'analyse dans l'article publié d'*« exploratoire »* ou de *« non planifiée »*. Ainsi, cela donne au lecteur une transparence et un contexte importants.

- 4) Le montant du financement de la recherche gaspillé a été estimé à 85 %.<sup>35</sup> Pour réduire ce gaspillage, nous devrions financer un nombre moindre de petites ERC puisqu'elles n'ont souvent pratiquement aucun potentiel de modifier les soins cliniques de manière significative. Selon nous, le financement de la recherche devrait plutôt se concentrer sur les études de développement à un stade précoce (pour les médicaments, dispositifs et technologies innovants) et les très grandes ERC (idéalement, de 1000 à 50 000 participants). Les interventions évaluées dans ces grandes ERC pragmatiques devraient être priorisées par un groupe de consensus d'intervenants (y compris les patients), et elles devraient être pour la plupart multicentriques<sup>1</sup> (étant donné que le risque d'estimations biaisées des effets est probablement plus faible dans les ERC multicentriques<sup>36</sup> et leur validité externe plus grande). Les ERC multicentriques de grande taille disposeront naturellement d'une meilleure reproductibilité (c.-à-d. d'un risque beaucoup plus faible de résultats de recherche faussement positifs).
- 5) Nous appuyons la suggestion de Hanson *et coll.*<sup>1</sup> selon laquelle les revues devraient créer des systèmes qui garantiraient la publication des articles dont les auteurs ont soumis au préalable tous les aspects d'une étude suivant des protocoles rigoureux avant de commencer le recrutement, quels que soient les résultats, tant que le protocole a été respecté sans écart significatif (ce qu'on appelle également des « comptes rendus pré-enregistrés »). L'adhésion rigide à un protocole

<sup>B</sup> Registration of Observational Studies at ClinicalTrials.gov. Disponible à l'adresse: <https://clinicaltrials.gov/ct2/manage-recs/howregister#Considerations> (consulté avril 2021).

<sup>C</sup> Par exemple, le Open Science Framework. Disponible à l'adresse: <https://osf.io/> (consulté avril 2021).

- prédéterminé améliorera la reproductibilité de leurs travaux.
- 6) Lorsque les méthodes statistiques sont complexes, ce qui se produit habituellement dans les études observationnelles, les auteurs devraient fournir de multiples analyses de sensibilité. Si l'effet observé est réel (et donc probablement reproductible), il devrait être compatible avec les résultats de la plupart des autres méthodes analytiques. Lorsque les méthodes ne correspondent pas, cela fournit alors des informations précieuses au lecteur quant à la fidélité des conclusions atteintes par les auteurs, et cela donne aux auteurs l'occasion d'expliquer pourquoi leurs résultats étaient sensibles au modèle.
  - 7) En plus des rapports statistiques fréquentistes typiques utilisant des valeurs  $P$  et des intervalles de confiance, des interprétations bayésiennes des résultats devraient également être incorporées. Cela donnerait l'occasion d'intégrer une *probabilité initiale* d'un effet réel dans une *probabilité postérieure* et reflèterait avec davantage de précision le risque de fausses découvertes. En d'autres termes, les interprétations bayésiennes fourniraient de meilleures informations quant aux résultats susceptibles d'être reproductibles et à ceux qui ne le sont pas. Plusieurs façons de mettre en œuvre ce cadre bayésien, basé sur des valeurs  $P$ , ont été publiées et pourraient être facilement intégrées par des revues médicales.<sup>20,37</sup>
  - 8) En accord avec Hanson *et coll.*,<sup>1</sup> nous pensons qu'il conviendrait d'envisager sérieusement d'utiliser un seuil de valeur  $P$  plus strict en tant que signification statistique nominale (par exemple,  $< 0,005$  au lieu de  $< 0,05$ ). S'il existe un effet réel, et si les tailles d'échantillon des études sont adéquates, un seuil plus strict devrait écarter les faux positifs tout en permettant d'identifier correctement la plupart des vrais positifs.<sup>38–40</sup>

## Conclusion

*L'apophénie*, ou la tendance à identifier des corrélations significatives là où il n'y en a pas, fait partie de la condition humaine. Nous ne pouvons pas la surmonter, mais nous pouvons nous efforcer de *diminuer son influence*. Dans le contexte de la recherche, l'apophénie peut entraîner un risque considérable de résultats faussement positifs. Nous croyons que le problème des résultats de recherche faussement positifs et la faible reproductibilité des résultats de recherche sont inextricablement liés.

Dans cet éditorial, nous avons souligné qu'une mauvaise reproductibilité dans la littérature biomédicale constitue un

grave problème. Nous avons discuté de certaines des principales raisons pour lesquelles une mauvaise reproductibilité peut survenir, et nous avons formulé quelques suggestions pratiques afin d'améliorer les taux de reproductibilité. Nous croyons que l'augmentation de la reproductibilité de la recherche améliorera, à son tour, la confiance que nous accordons à nos interventions en anesthésiologie, en médecine périopératoire, en soins intensifs et en douleur, et améliorera les devenirs cliniques de nos patients. Le temps de la simple prise de conscience de ces problèmes est révolu depuis longtemps. Il est maintenant temps pour nous tous — chercheurs, lecteurs, cliniciens, réviseurs pairs, rédacteurs en chef et dirigeants administratifs — d'agir.

**Disclosures** Dr. Philip Jones is the Deputy Editor-in-Chief of the *Canadian Journal of Anesthesia*; he had no involvement in the handling of this manuscript.

**Conflicts** Neither author has any conflicts to declare other than those listed under Disclosures above.

**Funding statement** None.

**Editorial responsibility** This submission was handled by Dr. Stephan K.W. Schwarz, Editor-in-Chief, *Canadian Journal of Anesthesia*.

**Déclaration** Dr Philip Jones est rédacteur en chef adjoint du *Journal canadien d'anesthésie*; il n'a été aucunement impliqué dans le traitement de ce manuscrit.

**Conflits** Aucun des auteurs n'a de conflits à déclarer autres que ceux énumérés dans la déclaration ci-dessus.

**Déclaration de financement** Aucune.

**Responsabilité éditoriale** Ce manuscrit a été traité par Dr Stephan K.W. Schwarz, rédacteur en chef, *Journal canadien d'anesthésie*.

## References

1. Hanson NA, Lavallee MB, Thiele RH. Apophenia and anesthesia: how we sometimes change our practice prematurely. Can J Anesth 2021; DOI: <https://doi.org/10.1007/s12630-021-02005-2>.
2. Ioannidis JP. Why most published research findings are false. PLoS Med 2005; DOI: <https://doi.org/10.1371/journal.pmed.0020124>.
3. Vetter TR, McGwin G Jr, Pittet JF. Replicability, reproducibility, and fragility of research findings—ultimately, caveat emptor. Anesth Analg 2016; 123: 244–8.
4. Spence JR, Stanley DJ. Prediction interval: what to expect when you're expecting ... a replication. PLoS One 2016; DOI: <https://doi.org/10.1371/journal.pone.0162874>.
5. Pawel S, Held L. Probabilistic forecasting of replication studies. PLoS One 2020; DOI: <https://doi.org/10.1371/journal.pone.0231416>.

6. Baker M. 1,500 scientists lift the lid on reproducibility. *Nature* 2016; 533: 452-4.
7. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* 2015; DOI: <https://doi.org/10.1126/science.aac4716>.
8. Begley CG, Ellis LM. Drug development: raise standards for preclinical cancer research. *Nature* 2012; 483: 531-3.
9. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets. *Nat Rev Drug Discov* 2011; DOI: <https://doi.org/10.1038/nrd3439-c1>.
10. Niven DJ, McCormick TJ, Straus SE, et al. Reproducibility of clinical research in critical care: a scoping review. *BMC Med* 2018; DOI: <https://doi.org/10.1186/s12916-018-1018-6>.
11. Avidan MS, Ioannidis JP, Mashour GA. Independent discussion sections for improving inferential reproducibility in published research. *Br J Anaesth* 2019; 122: 413-20.
12. Camerer CF, Dreber A, Forsell E, et al. Evaluating replicability of laboratory experiments in economics. *Science* 2016; 351: 1433-6.
13. Han R, Walton KS, Sholl DS. Does chemical engineering research have a reproducibility problem. *Ann Rev* 2019; 10: 43-57.
14. McDermott MB, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: still a ways to go. *Sci Transl Med* 2021; DOI: <https://doi.org/10.1126/scitranslmed.abb1655>.
15. Dowden H, Munro J. Trends in clinical success rates and therapeutic focus. *Nat Rev Drug Discov* 2019; 18: 495-6.
16. Broer L, Lill CM, Schuur M, et al. Distinguishing true from false positives in genomic studies: p values. *Eur J Epidemiol* 2013; 28: 131-8.
17. Hingorani AD, Kuan V, Finan C, et al. Improving the odds of drug development success through human genomics: modelling study. *Sci Rep* 2019; DOI: <https://doi.org/10.1038/s41598-019-54849-w>.
18. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. *Biostatistics* 2019; 20: 273-86.
19. Djulbegovic B, Kumar A, Glasziou P, Miladinovic B, Chalmers I. Medical research: trial unpredictability yields predictable therapy gains. *Nature* 2013; 500: 395-6.
20. Colquhoun D. The reproducibility of research and the misinterpretation of p-values. *R Soc Open Sci* 2017; DOI: <https://doi.org/10.1098/rsos.171085>.
21. Grolleau F, Collins GS, Smarandache A, et al. The fragility and reliability of conclusions of anesthesia and critical care randomized trials with statistically significant findings: a systematic review. *Crit Care Med* 2019; 47: 456-62.
22. Chow JT, Turkstra TP, Yim E, et al. Sample size calculations for randomized clinical trials published in anesthesiology journals: a comparison of 2010 versus 2016. *Can J Anesth* 2018; 65: 611-8.
23. Egbeawale BE, Lewis M, Sim J. Bias, precision and statistical power of analysis of covariance in the analysis of randomized trials with baseline imbalance: a simulation study. *BMC Med Res Methodol* 2014; DOI: <https://doi.org/10.1186/1471-2288-14-49>.
24. Kahan BC, Jairath V, Doré CJ, Morris TP. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials* 2014; DOI: <https://doi.org/10.1186/1745-6215-15-139>.
25. de Winter JC, Dodou D. A surge of p-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ* 2015; DOI: <https://doi.org/10.7717/peerj.733>.
26. Chavalarias D, Wallach JD, Li AH, Ioannidis JP. Evolution of reporting P values in the biomedical literature, 1990-2015. *JAMA* 2016; 315: 1141-8.
27. Wang MQ, Yan AF, Katz RV. Researcher requests for inappropriate analysis and reporting: a U.S. survey of consulting biostatisticians. *Ann Intern Med* 2018; 169: 554-8.
28. Jones PM, Chow JT, Arango MF, et al. Comparison of registered and reported outcomes in randomized clinical trials published in anesthesiology journals. *Anesth Analg* 2017; 125: 1292-300.
29. De Oliveira GS, Jr Jung MJ, McCarthy RJ. Discrepancies between randomized controlled trial registry entries and content of corresponding manuscripts reported in anesthesiology journals. *Anesth Analg* 2015; 121: 1030-3.
30. Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med* 2008; 358: 252-60.
31. De Oliveira GS, Chang R, Kendall MC, et al. Publication bias in the anesthesiology literature. *Anesth Analg* 2012; 114: 1042-8.
32. McHugh UM, Yentis SM. An analysis of retractions of papers authored by Scott Reuben, Joachim Boldt and Yoshitaka Fujii. *Anaesthesia* 2019; 74: 17-21.
33. Fanelli D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One* 2009; DOI: <https://doi.org/10.1371/journal.pone.0005738>.
34. Altman DG. The scandal of poor medical research. *BMJ* 1994; 308: 283-4.
35. Glasziou P, Chalmers I. Research waste is still a scandal—an essay by Paul Glasziou and Iain Chalmers. *BMJ* 2018; DOI: <https://doi.org/10.1136/bmj.k4645>.
36. Unverzagt S, Prondzinsky R, Peinemann F. Single-center trials tend to provide larger treatment effects than multicenter trials: a systematic review. *J Clin Epidemiol* 2013; 66: 1271-80.
37. Goodman SN. Toward evidence-based medical statistics. 2: the Bayes factor. *Ann Intern Med* 1999; 130: 1005-13.
38. Koletsis D, Solmi M, Pandis N, et al. Most recommended medical interventions reach  $P < 0.005$  for their primary outcomes in meta-analyses. *Int J Epidemiol* 2020; 49: 885-93.
39. Ioannidis JP. The proposal to lower P value thresholds to .005. *JAMA* 2018; 319: 1429-30.
40. Wayant C, Scott J, Vassar M. Evaluation of lowering the P value threshold for statistical significance from .05 to .005 in previously published randomized clinical trials in major medical journals. *JAMA* 2018; DOI: <https://doi.org/10.1001/jama.2018.12288>.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.