
Article:
Bramley, P. orcid.org/0000-0001-6209-6640, Hulman, J. and Wanstall, H. (2024) Risk of bias and problematic trials: characterising the research integrity of trials submitted to Anaesthesia. Anaesthesia, 79 (12). pp. 1309-1316. ISSN 0003-2409

https://doi.org/10.1111/anae.16411

Original Article

Risk of bias and problematic trials: characterising the research integrity of trials submitted to *Anaesthesia*

Paul Bramley[1,2] **Joshua Hulman**[1] and **Helen Wanstall**[3]

1 Department of Anaesthesia, 3 Emergency Department, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK
2 School of Medicine and Population Health, University of Sheffield, Sheffield, UK

Summary

Background There is some evidence for systematic biases and failures of research integrity in the anaesthesia literature. However, the features of problematic trials and effect of editorial selection on these issues have not been well quantified.

Methods We analysed 209 randomised controlled trials submitted to *Anaesthesia* between 8 March 2019 and 31 March 2020. We evaluated the submitted manuscript, registry data and the results of investigations into the integrity of the trial undertaken at the time of submission. Trials were labelled `concerning´ if failures of research integrity were found, and `problematic´ if identified issues would have warranted retraction if they had been found after publication. We investigated how `problematic´ trials were detected, the distribution of p values and the risk of outcome reporting bias and p-hacking. We also investigated whether there were any factors that differed in problematic trials.

Results We found that false data was the most common reason for a trial to be labelled as `concerning´, which occurred in 51/62 (82%) cases. We also found that while 195/209 (93%) trials were preregistered, we found adequate registration for only 166/209 (79%) primary outcomes, 100/209 (48%) secondary outcomes and 11/209 (5%) analysis plans. We also found evidence for a step decrease in the frequency of p values > 0.05 compared with p values < 0.05. `Problematic´ trials were all single-centre and appeared to have fewer authors (incident risk ratio (95%CI) 0.8 (0.7–0.9)), but could not otherwise be distinguished reliably from other trials.

Conclusions Identification of `problematic´ trials is frequently dependent on individual patient data, which is often unavailable after publication. Additionally, there is evidence of a risk of outcome reporting bias and p-hacking in submitted trials. Implementation of alternative research and editorial practices could reduce the risk of bias and make identification of problematic trials easier.

Correspondence to: Paul Bramley
Email: paul.e.bramley@googlemail.com
Accepted: 17 July 2024
Keywords: publication bias; questionable research practices; randomised controlled trials; research integrity

Introduction

Randomised controlled trials are a key component of evidence-based medicine and can provide high quality evidence to assess the efficacy and safety of interventions. They provide the best causal evidence for an effect, are synthesised into meta-analyses and used to form clinical guidelines. However, there are growing concerns about the replicability of research across scientific disciplines [1, 2].

There are multiple potential mechanisms for poor rates of replication, including well-described issues such as publication bias [3], outcome reporting bias [4], p-hacking [5, 6] and fraud or fabrication [7]. Randomised controlled trials are vulnerable to each of these to varying degrees. Accordingly, there has been interest in designing tools and processes to limit failures of research integrity and to detect them when they occur pre- and post-publication [8].

The fallibility of randomised controlled trials should not be surprising, given there have been several high profile cases of anaesthetists having multiple published trials retracted [9]. Additionally, we know that trials submitted to journals contain false results [10]; however, it has not been explored how these trials are detected, and whether they differ from other submitted trials. Additionally, most of the evidence regarding bias in the anaesthetic literature examines articles already published and those included in systematic reviews. We are therefore limited in our ability to assess the impact biases from trials submitted, but ultimately not published, could have had. This in turn constrains our ability to determine the mechanism(s) driving the risk of various biases and the potential benefits of introducing new research processes to detect them.

To address this problem, we aimed to analyse an existing cohort of randomised controlled trials submitted for publication to *Anaesthesia* to quantify how trials with failures of research integrity were identified and to quantify the risk of outcome reporting bias, p-hacking and evidence for selection based on statistical significance. Furthermore, we aimed to investigate whether there were any common features of trials which were associated with failures of research integrity and determine if they could be used to detect problematic trials.

Methods

As detailed previously [10], for a period between 8 March 2019 and 31 March 2020, all randomised controlled trials submitted to *Anaesthesia*, which were not subject to a desk rejection, were evaluated for potentially problematic features. Summary baseline data were assessed for excessive group similarity or difference. Presented results were assessed for implausible or impossible results. Additionally, individual participant data (IPD) was requested routinely for all trials which were submitted from the countries with the highest rates of submission during this period, as well as all trials which were accepted provisionally for publication.

After evaluating available documents, trials were either labelled as `not concerning´ or `concerning´, where there was evidence of problems with research integrity. The `concerning´ trials were subclassified as `problematic´ (if it was felt the findings were so compromised as to have warranted retraction had the paper been published, also known as a `zombie´ trial) or `questionable´ (if there were substantial problems with research integrity, but not so egregious as to warrant retraction). While the categorisation was subjective, `questionable´ trials tended to have fewer concerns, and those concerns had plausible benign explanations, particularly where key results in the trial were not at risk of being invalidated. `Problematic´ trials tended to have convincing evidence of fraud or fabrication in the IPD, failures of research integrity without plausible benign explanations or large portions of results which were highly likely to be false.

For all `concerning´ trials, we reviewed the results of the original investigation to categorise the type of research integrity failures found. The categories used were: false data (problems detected within the IPD), which were typically issues such as repeated patient data, impossible values and patterns suggestive of falsified data; false results (presented results were incorrect), which was based on re-analysing IPD or by finding impossible results in the manuscript; plagiarism; discrepancies between manuscript and other materials, such as finding that the IPD was incompatible with the reported methods or important and unexplained differences between pre-registration and manuscript; and evidence of changes when resubmitted (either to *Anaesthesia* or elsewhere).

We extracted further information about all submitted trials by reviewing the first submitted manuscript for each trial before revision, and any public trial registration, but not IPD. To investigate the risk of outcome reporting bias and p-hacking (performing multiple statistical analyses and reporting only those which are statistically significant), we checked whether trials were preregistered on a public registry. Specifically, we reviewed the manuscript for any mention of registration, then also searched the World Health Organization clinical trials registry (which draws from all major national trials registries) with an additional search of clinicaltrials.gov since it is the most widely used and has better search functionality. We then compared whether outcomes and analyses reported in the manuscript matched registration. Secondary outcomes were coded as matching if all registered outcomes were reported, with no additional outcomes added. We summarised these variables and quantified their association with `problematic´ trials and publication.

To investigate trends in statistical significance we extracted the p value for the primary outcome in each trial (if a logical statement such as `< *0.01*´ was reported, we

recorded this as `0.01´), as well as the direction of the outcome with reference to the hypothesis (adjudicated by the study team based on registration, hypothesis framework and further trial information). Additionally, we extracted the date of the end of recruitment (as reported either in the manuscript or registration) and date of submission for publication. We also extracted if there was any statement made about availability of IPD, code or other research documents. These variables were also summarised and assessed for their association with `problematic´ trials and publication.

We examined several additional variables for their association with `problematic´ trials, chosen as they are widely reported and have a plausible association with fraudulent studies. These included: study size (number of patients randomised to all groups); number of centres in which the trial was conducted; number of authors; reported funding source; presence of a clear conflict of interest; whether the trial was clinical or laboratory-based; and if any authors had any papers retracted previously. If the manuscript listed more than seven authors, only the first, second, last authors and any author listed as having contributed to data handling were checked. We chose to compare the differences between `problematic´ and `not problematic´ (the combination of `not concerning´ and `questionable´) trials as we felt that identifying studies that would have been retracted (and therefore added minimal information to the literature) was of more value than combining these with trials with less severe issues with potentially benign explanations. Additionally, for all trials labelled as `concerning´ we undertook a risk of bias assessment using the Cochrane Risk of Bias-2 tool [11]. The author undertaking risk of bias assessment was blinded to whether the trial being assessed was listed as `questionable´ or `problematic´.

Data were assessed using summary statistics and statistical tests were undertaken using linear regression for data where residuals were approximately normal; logistic regression for binary variables; Cox proportional hazards modelling for survival data; zero-truncated Poisson regression for number of authors; χ^2 tests (with Yates' continuity correction) for independence tests with categorical data and where there were no events in one arm to do logistic regression; and Mann–Whitney U tests for p value distributions and number of centres. All tests were exploratory in nature and no measures were taken to control for multiple testing. Data cleaning and analysis was performed in R version 4.3.1 (R Foundation, Vienna, Austria).

After reviewing the project protocol, this project was approved by the Editorial Board of *Anaesthesia*, and the host institution confirmed that ethical approval was not required as IPD were not available to us.

Results

Between 8 March 2019 and 31 March 2020, 212 randomised controlled trials were submitted to *Anaesthesia* that were not desk rejected. One row appeared to be duplicated in the analysis dataset and for two papers the original manuscript was not available leaving 209 trials with manuscripts and analysis results available. Of 209 trials, 138 (66%) had IPD available, 62 (30%) were categorised as `concerning´ and 35 (17%) were classified as `problematic´.

Of the 62 trials identified as `concerning´, false data was the issue identified most frequently (Table 1). There were 22 `concerning´ trials in which false data was the only problem identified, of which 14 were found to be `problematic´. Fourteen trials were identified as `concerning´ with false data and false results only, of which seven were identified as `problematic´. No `concerning´ trials were identified using plagiarism or discrepancies between manuscript and protocol alone. All 35 `problematic´ trials had IPD available; only two trials were found to have false results without IPD available.

The quality of preregistration appeared to differ between groups (Table 2). Primary outcomes in manuscripts matched registration more in `not problematic´ trials, where a test for independence was not significant when including failure to register a primary outcome as a category (df = 2, p = 0.11) with an odds ratio (95%CI) of 0.5 (0.2–1.1) for successful primary outcome registration (vs. not registered)

Table 1 Findings of investigations into `concerning´ trials, split into `questionable´ trials and `problematic´ trials. Values are number (proportion).

	Questionable n = 27	Problematic n = 35	Concerning n = 62
Discrepancies	4 (15%)	5 (14%)	9 (15%)
False results	15 (56%)	14 (40%)	29 (47%)
False data	18 (67%)	33 (94%)	51 (82%)
Changed submission			
No	8 (30%)	17 (49%)	25 (40%)
Yes	5 (19%)	6 (17%)	11 (18%)
Not resubmitted	14 (52%)	12 (34%)	26 (42%)
Plagiarism	1 (4%)	6 (17%)	7 (11%)
Overall risk of bias			
High	12 (44%)	12 (34%)	24 (39%)
Some	15 (56%)	21 (60%)	36 (58%)
Low	0 (0%)	2 (6%)	2 (3%)

Table 2 Registration and study document availability of submitted trials split by `problematic´ and `not problematic´ (`concerning´ and `questionable´) groups. Values are number (proportion).

	Not problematic n = 174	Problematic n = 35	Overall n = 209
Pre-registered	162 (93%)	33 (94%)	195 (93%)
Registered primary outcome			
Yes	142 (82%)	24 (69%)	166 (79%)
No	21 (12%)	9 (26%)	30 (14%)
Missing	11 (6%)	2 (6%)	13 (6%)
Registered secondary outcomes			
Yes	86 (49%)	14 (40%)	100 (48%)
No	60 (35%)	13 (37%)	73 (35%)
Missing	28 (16%)	8 (23%)	36 (17%)
Registered analysis			
Yes	11 (6%)	0 (0%)	11 (5%)
No	1 (1%)	0 (0%)	1 (0%)
Not enough detail	162 (93%)	35 (100%)	197 (94%)
Data availability			
No	150 (86%)	25 (71%)	175 (84%)
Open	4 (2%)	2 (6%)	6 (3%)
On request	20 (11%)	8 (23%)	28 (13%)
Code not available	174 (100%)	35 (100%)	209 (100%)
Study documents available			
No	162 (93%)	31 (89%)	193 (92%)
Open	1 (1%)	1 (3%)	2 (1%)
On request	11 (6%)	3 (9%)	14 (7%)

with `problematic´ trials. A similar pattern was seen in registration of secondary outcomes, which were registered less in `problematic´ trials (df = 2, p = 0.50); OR (95%CI) 0.7 (0.3–1.4) for `problematic´ trials registering secondary outcomes successfully vs. not. Successful preregistration of primary and secondary outcomes occurred in 13/65 (20%) `problematic´ trials and 83/174 (48%) `non-problematic´ trials (OR 0.6 (95%CI) 0.3–1.4). Successful registration was more common in accepted studies (online Supporting Information Table S1) and 10/18 accepted trials had successfully registered primary and secondary outcomes. No `problematic´ trial had a pre-registered analysis with enough detail to match to their presented analysis, though 162/174 (93%) `not problematic´ trials also did not provide enough detail. A test for independence was not significant when including `missing´ as a category (df = 3, p = 0.28), and when comparing successful analysis registration to not (df = 1, p = 0.27).

p Values were reported for the primary outcome in 204/209 (98%) manuscripts. The median p value for the primary outcome was < 0.05 in all trials (Table 3), with no clear difference in p value distribution between `problematic´ and `not problematic´ trials (Fig. 1 and online Supporting Information Figure S1). Figure 1 shows that p values immediately > 0.05 were substantially less frequent than those immediately < 0.05 in all submitted trials, where this was a step change rather than part of a trend. This is suggestive of biases which select for statistically significant results, including p-hacking and outcome reporting bias.

Being categorised as `problematic´ was associated with a lower rate of having a p value < 0.05 (OR 0.7 (95%CI) (0.3–1.6)). The rate of publication for trials was 13/133 (10%) for p values < 0.05, compared with 5/71 (7%) for those with p values ≥ 0.05. This gives an odds ratio of publication with a p value < 0.05 of 1.4 ((95%CI 0.5–4.6); p = 0.51). However, the rate of having the expected primary outcome was lower for published trials (10/18, 56%) than for unpublished (129/191, 68%) (online Supporting Information Table S1). There did not appear to be a relationship between trial outcome being expected or not based on problematic or non-problematic groups (df = 2, p = 0.73). However, eight trials which declared a significant result was found (which was expected) only stated a p value of `< *0.05*´, which we encoded as not significant in our other analysis. After excluding trials which listed end dates of recruitment after submission, the delay between end of recruitment and submission was shorter for trials with a p value < 0.05 (hazard ratio (95%CI) 1.2 (0.8–1.6)), and there was a shorter delay to submission for `problematic´ vs. `not problematic´ trials (hazard ratio (95%CI) 1.7 (1.1–2.6)). However, there was a substantial amount of missing data; 9/35 problematic trials did not provide enough information about recruitment to calculate the delay in submission.

The median sample size was similar between `problematic´ and `not problematic´ studies (difference (95%CI) -10 (-70–49) participants). However, there were zero `problematic´ studies with > 500 participants, while there were four in the `not problematic´ group. `Problematic´ studies also had fewer authors (incident risk ratio (95%CI) 0.8 (0.7–0.9)). No trial found to be `problematic´ was conducted in more than one centre, though the distribution of number of centres was not significantly different (p = 0.27). The rate of possible conflicts of interest reported was higher in `non-problematic´ trials (df = 1, p = 0.05). Less `problematic´ trials were grant funded, but an overall test for independence was non-significant (df = 2, p = 0.53). Risk of bias assessment showed that `problematic´ trials and `questionable´ trials were scored

Table 3 Reported characteristics of submitted trials split by `problematic´ and `not problematic´ groups. Values are median (IQR [range]) or number (proportion).

	Not problematic n = 174	**Problematic n = 35**	**Overall n = 209**
Primary outcome p value	0.017 (0.001–0.189 [0.0001–1.000])	0.0235 (0.001–0.109 [0.0007–0.990])	0.0195 (0.001–0.168 [0.0001–1.000])
Missing	4 (2%)	1 (3%)	5 (2%)
Results significant			
Yes	113 (65%)	20 (57%)	133 (64%)
No	57 (33%)	14 (40%)	71 (34%)
Missing	4 (2%)	1 (3%)	5 (2%)
Primary outcome direction			
Expected	115 (66%)	24 (69%)	139 (67%)
Null	54 (31%)	11 (31%)	65 (31%)
Unexpected	3 (2%)	0 (0%)	3 (1%)
Missing	2 (1%)	0 (0%)	2 (1%)
Number of centres			
≥ 2	6 (3%)	0 (0%)	6 (3%)
1	168 (97%)	35 (100%)	203 (97%)
Number of authors	7 (5–8 [2–15])	5 (4–7 [1–10])	6 (5–8 [1–15])
Days delay in submission	309 (146–582 [14–3850])	149 (79–360 [10–1050])	298 (124–565 [10–3850])
Missing	29 (17%)	10 (29%)	39 (19%)
Reported funding			
Grant	67 (39%)	10 (29%)	77 (37%)
Industry	5 (3%)	1 (3%)	6 (3%)
None	102 (59%)	24 (69%)	126 (60%)
Conflict of interest	22 (13%)	0 (0%)	22 (11%)
Study size; n	89 (56–143 [16–1930])	80 (63–147 [24–398])	88 (59–144 [16–1930])

approximately equally (Table 1 and online Supporting Information Table S2), though `problematic´ trials were rated as high risk overall less frequently than `questionable´ trials.

Discussion

We performed an analysis of a cohort of randomised controlled trials submitted for publication to *Anaesthesia* and made two broad findings. First, `problematic´ clinical trials differ in some ways from `not problematic´ studies, but not consistently or substantially enough to identify them, which normally requires IPD. This is of concern given the low rate of data availability stated in submissions. Second, trials appear to be at risk of biases which affect the literature systematically; lack of adequate preregistration of outcomes introduces a risk of outcome reporting bias and lack of preregistration of analysis plans introduces a risk of p-hacking. Additionally, we found longer submission times for results which were not significant, which is a mechanism that can introduce publication bias.

The academic anaesthesia community has been active in investigating failures of research integrity [12–14], and there have also been attempts to assess the anaesthesia literature for risk of bias systematically. Jones et al. found that the majority of randomised controlled trials published in six anaesthesia journals were not preregistered adequately, and of those that were, approximately 40% showed discrepancies between the registered and reported primary outcome, and approximately 90% showed discrepancies between the registered and reported secondary outcomes [15]. We have shown substantial improvements in these rates, even in submitted rather than published studies, which is likely due to the increased uptake of preregistration since 2015 the last year studied by Jones et al. Okonya et al. found similarly low levels of statements of data, code and study documentation availability in a sample of published studies in anaesthesia

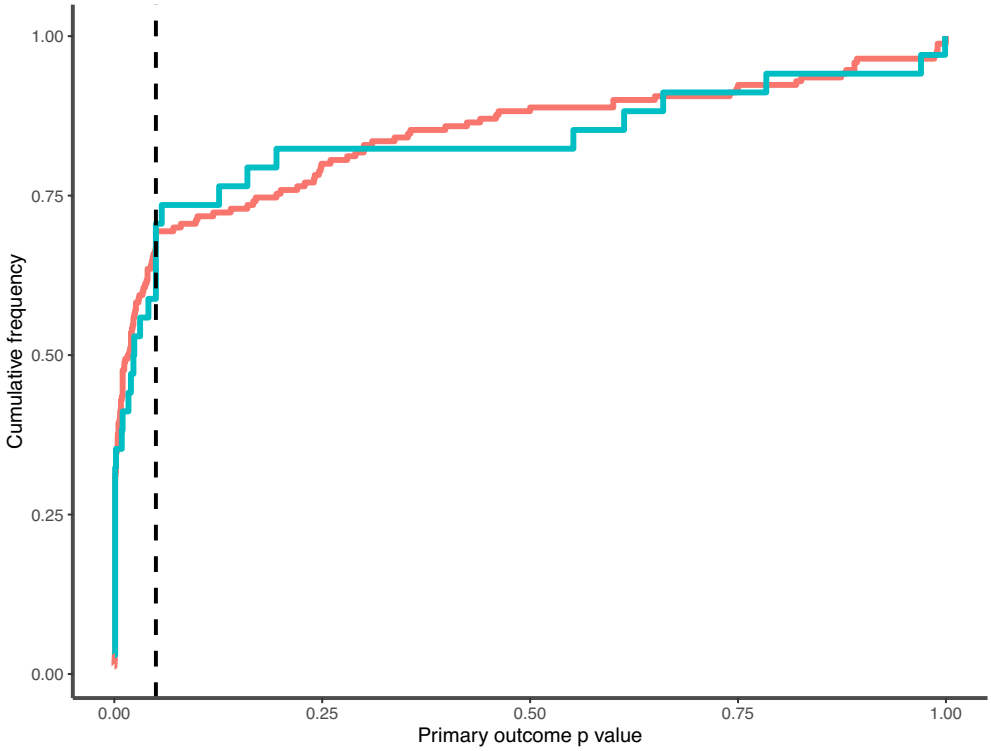

Figure 1 Cumulative frequency of primary outcome p values, split between `problematic´ trials (blue) and `not problematic´ trials (red). Dashed line represents a p value of 0.05.

journals [16], though different rates of preregistration, likely due to not studying randomised controlled trials exclusively. Chuang et al. also found a discontinuity in the frequency of p values greater and less than 0.05 in published randomised controlled trials, which supports our evidence for selection based on statistical significance [17]. In addition, some statistical evidence of publication bias has been found in the majority of analysed reviews [18], which is not an effect we could assess directly. Together these findings suggest that trials are likely to have a degree of bias, which reduces the replicability of results.

Alternative editorial methods have been proposed which better align the incentives of journals and authors to reduce the risk of bias. Registered reports (where submissions to journals are made based on a study proposal before data are collected and accepted or rejected on this basis) in other fields have been found to be associated with less significant results [19] but papers being rated as higher quality [20].

By examining trials at the point of submission to a journal, we were able to investigate the effects of editorial selection on bias; we found there was no strong evidence for editorial preference for significant results, which is consistent with results outside of the anaesthesia literature [21, 22]. The higher rate of registration in published, rather than submitted, randomised controlled trials is also a novel finding within anaesthesia, though the cause for this cannot be identified.

There has been no previous systematic investigation in the anaesthesia literature (or, to our knowledge, elsewhere in the clinical literature) about the ways in which problematic randomised controlled trials are detected, whether they differ from those with no concerns about their results, and if these differences could be used to distinguish between the two. Our results highlight the importance of making IPD available to identify potentially problematic trial results. While most trials in this cohort had no statement about data availability, work in other fields has shown that even accessing data that is labelled `available on request´ can be challenging [23]. In this context identifying problematic trials may be impossible in many cases after publication. The *BMJ* has recently changed its editorial policy to require data and statistical code and both are made available in a public repository for all research it publishes [24]. It would be possible for anaesthesia journals to consider implementing similar policies for IPD or even statistical code, given the history of retractions in the field.

This analysis is limited by the fact that it only examines a snapshot of submissions from a single journal, in a single academic domain. We could speculate that there are common features of untrustworthy trials across fields and over time, but these data cannot provide evidence for this. The categorisation of trials as `questionable´ or `problematic´ is entirely dependent on previous analysis, with no further checking provided as part of this work. This categorisation was a subjective one, and it is likely that some concerning trials were not detected, and possible that some trustworthy trials were categorised falsely as such. For example, the focus on countries which were the most prolific in rates of submission likely increased the yield of detecting false data but may have resulted in missing `problematic´ studies from other countries. There are also several variables in this analysis which are at least partially subjective, and it is possible that an alternative set of authors could have categorised values differently, which limits the robustness of our results. Additionally, we can offer no insight into the mechanisms which incentivise questionable research practices, or how problematic studies are generated. Finally, we consider all analyses presented here to be exploratory, and further investigation in other datasets would be required to confirm any associations we have reported.

We conclude that the most common way to identify `problematic´ trials is the analysis of IPD, but that very few trials state that IPD will be available after publication. Additionally, we find that despite improving rates of preregistration there is a risk of outcome reporting bias and p-hacking in trials submitted for publication. We propose that journals and authors should consider implementing novel research practices which increase transparency and reduce the risk of bias.

Acknowledgements

This research was funded by a grant from the Association of Anaesthetists/*Anaesthesia* (NIAA22R206) and hosted by Sheffield Teaching Hospitals NHS Foundation Trust. We would like to thank Dr J Carlisle for generously allowing us access to the results of his analyses. No data are available. Statistical code used to generate the results can be found on GitHub (https://github.com/2ae487be22/Problematic-Trials-Public-Code). PB was previously an Editorial Fellow of *Anaesthesia*. No other competing interests declared.

References

1. Klein RA, Vianello M, Hasselman F, et al. Many labs 2: investigating variation in replicability across samples and settings. *Adv Methods Pract Psychol Sci* 2018; **1**: 443–90. https://doi.org/10.1177/2515245918810225.
2. Errington TM, Mathur M, Soderberg CK, et al. Investigating the replicability of preclinical cancer biology. *eLife* 2021; **10**: e71601. https://doi.org/10.7554/eLife.71601.
3. Kicinski M, Springate DA, Kontopantelis E. Publication bias in meta-analyses from the Cochrane Database of Systematic Reviews. *Stat Med* 2015; **34**: 2781–93. https://doi.org/10.1002/sim.6525.
4. Goldacre B, Drysdale H, Dale A, et al. COMPare: a prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials* 2019; **20**: 118. https://doi.org/10.1186/s13063-019-3173-2.
5. Simonsohn U, Nelson LD, Simmons JP. P-curve: a key to the file-drawer. *J Exp Psychol Gen* 2014; **143**: 534–47. https://doi.org/10.1037/a0033242.
6. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 2011; **22**: 1359–66. https://doi.org/10.1177/0956797611417632.
7. Simonsohn U, Nelson L, Simmons J. [109] Data Falsificada (Part 1): ``Clusterfake.´´ *Data Colada* 2023. https://datacolada.org/109 (accessed 22/03/2024).
8. Bradley SH, DeVito NJ, Lloyd KE, et al. Reducing bias and improving transparency in medical research: a critical overview of the problems, progress and suggested next steps. *J R Soc Med* 2020; **113**: 433–43. https://doi.org/10.1177/0141076820956799.
9. Retraction Watch. The Retraction Watch Leaderboard. 2015. https://retractionwatch.com/the-retraction-watch-leaderboard/ (accessed 22/03/2024).
10. Carlisle JB. False individual patient data and zombie randomised controlled trials submitted to Anaesthesia. *Anaesthesia* 2021; **76**: 472–9. https://doi.org/10.1111/anae.15263.
11. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019; **366**: l4898. https://doi.org/10.1136/bmj.l4898.
12. Shafer SL. Tattered threads. *Anesth Analg* 2009; **108**: 1361–3. https://doi.org/10.1213/ane.0b013e3181a16846.
13. Carlisle JB. The analysis of 168 randomised controlled trials to test data integrity. *Anaesthesia* 2012; **67**: 521–37. https://doi.org/10.1111/j.1365-2044.2012.07128.x.
14. Wise J. Boldt: the great pretender. *BMJ* 2013; **346**: f1738. https://doi.org/10.1136/bmj.f1738.
15. Jones PM, Chow JTY, Arango MF, Fridfinnson JA, Gai N, Lam K, Turkstra TP. Comparison of registered and reported outcomes in randomized clinical trials published in Anesthesiology journals. *Anesth Analg* 2017; **125**: 1292–300. https://doi.org/10.1213/ANE.0000000000002272.
16. Okonya O, Rorah D, Tritz D, Umberham B, Wiley M, Vassar M. Analysis of practices to promote reproducibility and transparency in anaesthesiology research. *Br J Anaesth* 2020; **125**: 835–42. https://doi.org/10.1016/j.bja.2020.03.035.
17. Chuang Z, Martin J, Shapiro J, Nguyen D, Neocleous P, Jones PM. Minimum false-positive risk of primary outcomes and impact of reducing nominal *P*-value threshold from 0.05 to 0.005 in anaesthesiology randomised clinical trials: a cross-sectional study. *Br J Anaesth* 2023; **130**: 412–20. https://doi.org/10.1016/j.bja.2022.11.001.
18. Hedin RJ, Umberham BA, Detweiler BN, Kollmorgen L, Vassar M. Publication bias and nonreporting found in majority of systematic reviews and meta-analyses in Anesthesiology journals. *Anesth Analg* 2016; **123**: 1018–25. https://doi.org/10.1213/ANE.0000000000001452.
19. Scheel AM, Schijen MRMJ, Lakens D. An excess of positive results: comparing the standard psychology literature with registered reports. *Adv Methods Pract Psychol Sci* 2021; **4**: 25152459211007467. https://doi.org/10.1177/25152459210074.

20. Soderberg CK, Errington TM, Schiavone SR, et al. Initial evidence of research quality of registered reports compared with the standard publishing model. *Nat Hum Behav* 2021; **5**: 990–7. https://doi.org/10.1038/s41562-021-01142-4.
21. van Lent M, Overbeke J, Out HJ. Role of editorial and peer review processes in publication bias: analysis of drug trials submitted to eight medical journals. *PLoS One* 2014; **9**: e104846. https://doi.org/10.1371/journal.pone.0104846.
22. Olson CM, Rennie D, Cook D, et al. Publication bias in editorial decision making. *JAMA* 2002; **287**: 2825–8. https://doi.org/10.1001/jama.287.21.2825.
23. Naudet F, Sakarovitch C, Janiaud P, Cristea I, Fanelli D, Moher D, Ioannidis JPA. Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in *The BMJ* and *PLOS Medicine*. *BMJ* 2018; **360**: k400. https://doi.org/10.1136/bmj.k400.
24. Loder E, Macdonald H, Bloom T, Abbasi K. Mandatory data and code sharing for research published by *The BMJ*. *BMJ* 2024; **384**: q324. https://doi.org/10.1136/bmj.q324.

Supporting Information

Additional supporting information may be found online via the journal website.

Figure S1. Frequency of primary outcome p values, split between `problematic´ trials and `non-problematic´ trials.

Table S1. Study information split by whether the trial was accepted for publication.

Table S2. Risk of bias assessment results for `questionable´ and `problematic´ trials.