BJA Open

BJA Open, 1 (C): 100003 (2022)

doi: 10.1016/j.bjao.2022.100003
Advance Access Publication Date: 1 March 2022
Original Research Article

ORIGINAL RESEARCH ARTICLE

False-positive and false-negative risks for individual multicentre trials in critical care

David Sidebotham[1,2,*] and C. Jake Barlow[1,2]

[1]Department of Anaesthesia, Auckland City Hospital, Auckland, New Zealand and [2]Cardiothoracic and Vascular Intensive Care Unit, Auckland City Hospital, Auckland, New Zealand

*Corresponding author. E-mail: dsidebotham@adhb.govt.nz

Abstract

Background: In medical research, null hypothesis significance testing (NHST) is the dominant framework for statistical inference. NHST involves calculating *P*-values and confidence intervals to quantify the evidence against the null hypothesis of no effect. However, *P*-values and confidence intervals cannot tell us the probability that the hypothesis is true. In contrast, false-positive risk (FPR) and false-negative risk (FNR) are post-test probabilities concerning the truth of the hypothesis, that is to say, the probability a real effect exists.
Methods: We calculated the FPR or FNR for 53 individual multicentre trials in critical care based on a pretest probability of 0.5 that the hypothesis was true.
Results: For trials reporting statistical significance, the FPR varied between 0.1% and 57.6%. For trials reporting non-significance, the FNR varied between 1.7% and 36.9%. Twenty-six of 47 trials (55.3%) reporting non-significance provided strong or very strong evidence in favour of the null hypothesis; the remaining trials provided limited evidence. There was no obvious relationship between the *P*-value and the FNR.
Conclusions: The FPR and FNR showed marked variability, indicating that the probability of a real or absent treatment effect differed substantially between trials. Only one trial reporting statistical significance provided convincing evidence of a real treatment effect, and nearly half of all trials reporting non-significance provided limited evidence for the absence of a treatment effect. Our findings suggest that the quality of evidence from multicentre trials in critical care is highly variable.

Keywords: Bayes factor; Bayes theorem; critical care; false-negative risk; false-positive risk; research design; sample size; significance testing

In medical research, the default method of statistical inference is null hypothesis significance testing (NHST). The basic premise of NHST is simple, although it is widely misunderstood; *P*-values and confidence intervals (CIs) are calculated to quantify the extent to which the sample data provide evidence against the null hypothesis of no effect in the population from which the samples are drawn.[1] If $P \leq \alpha$ (the significance threshold) or the $1-\alpha$ CI excludes the null value, the null hypothesis is rejected and the result is considered statistically significant.

The problems associated with NHST are well known.[1,2] *P*-values and CIs are probabilities concerning the data, not the hypothesis. A *P*-value of 0.05 means there is a 5% (1 in 20) chance of observing data at least as extreme as that observed under a true null hypothesis, which may be interpreted as providing moderate evidence against the null hypothesis. However, as clinicians and researchers, we are primarily interested in whether an effect is real, not how probable the data are. Rejection of the null hypothesis means we accept the alternative hypothesis, but non-rejection does not mean we accept the null hypothesis. Finally, the rejection decision forces an arbitrary dichotomisation of results into significant or non-significant, despite the fact that there is little to distinguish the data when *P* is just below or just above the

Received: 23 December 2021; **Accepted:** 27 January 2022

significance threshold. Recognising these issues, in 2015, the American Statistical Association issued a statement outlining the misunderstandings and misuses of *P*-values and suggesting some alternative approaches.[3]

Accepting that NHST is well established, one approach is to evaluate the outcome of NHST using metrics derived from Bayesian inference.[4–6] In this article, we calculate Bayes factors (BFs) and use Bayes theorem to estimate post-test probabilities for the hypothesis for individual multicentre trials in critical care.

A BF is the ratio of how probable the data are under two competing models (Supplementary Appendix 1). Using standard notation

$$\mathrm{BF}_{A:B} = \frac{p(data|Model\ A)}{p(data|Model\ B)}$$

where each term is a conditional probability indicating the likelihood of the data given one or the other model is true. Here, a $\mathrm{BF}_{A:B}$ of 5 means the data are five times more probable under Model A than Model B, and a $\mathrm{BF}_{A:B}$ of 0.5 means the data are twice as likely under Model B than Model A. A $\mathrm{BF}_{A:B}$ of 1 means the data are equally likely under each model.

For our purposes, the two models are the null and alternative hypotheses. BFs can be constructed to indicate the strength of evidence favouring either the null or the alternative hypothesis. A BF is determined by the sample size and the difference between the groups, but not by the direction of the difference. Thus, it is immaterial which group is assigned as the control. For a trial with two groups reporting a binary outcome (e.g. mortality), BFs can be calculated from a two-by-two contingency table containing the raw trial data (Table 1).

In conjunction with an estimate of the pretest probability, we can use the BF to calculate post-test probabilities for the hypothesis. For a statistically significant result, the false-positive risk (FPR) is the probability that the null hypothesis is true, and for a non-significant result, the false-negative risk (FNR) is the probability that the alternative hypothesis is true. For the pretest probability, we can assign a value of 0.5, consistent with equipoise. As highlighted in the Discussion, assigning a particular value to the pretest probability does not limit the applicability of the method.

The FPR and FNR provide a useful insight into trial data that are not available from traditional *P*-values and CIs. We created a web-based calculator for estimating the BF and either the FPR or FNR (available at https://chrisjake.github.io/bayes-trial-eval.html).

Table 1 Two-by-two contingency table for the outcome from a trial with an intervention and a control group reporting a mortality outcome. Cell counts (a–d) represent the numbers in each group by outcome. Marginal row totals are designated n_1 and n_2. Marginal column totals are designated m_1 and m_2. The total sample size is designated *N*.

	Dead	Alive	Total
Intervention	a	b	n_1
Control	c	d	n_2
Total	m_1	m_2	*N*

Methods

Data source

Data were obtained from a previously reported structured review of critical care trials.[7] The data set includes all multicentre, superiority trials reporting mortality outcomes published in the *New England Journal of Medicine*, the *Journal of the American Medical Association*, and the *Lancet* between January 1, 2010 and March 31, 2020. In addition, we separately analysed data from the corticosteroid arms of the Randomised Evaluation of COVID-19 Therapy (RECOVERY) and Randomized, Embedded, Multifactorial Adaptive Platform Trial for Community-Acquired Pneumonia (REMAP-CAP) trials.[8] The RECOVERY and REMAP-CAP trials are ongoing platform trials evaluating mortality outcomes for interventions for COVID-19.

Data analysis

Trials were categorised as significant or non-significant for the primary mortality outcome on the basis of the statistical model described in the published report.

For each trial, we created a two-by-two contingency table from the raw data (Table 1). Data were independently extracted by both authors. From the contingency tables, we calculated a BF for each trial using the contingencyTableBF function of the BayesFactor package (v 0.9.12-4.2) in R (R Core Team 2020, v 4.03, R Studio v 1.3.959; RStudio, PBC, Boston, MA, USA). The contingencyTableBF function calculates BFs using the method of Gunel and Dickey.[9] We assumed independent multinominal sampling with fixed row margins, as this sampling method most closely resembles that used in randomised trials. With this sampling strategy, the marginal rows of the contingency table (i.e. the number of participants randomised to each group) are assumed to be fixed and the marginal column totals (i.e. the number of participants who survive or die) are free to vary.

For trials reporting statistical significance for the primary mortality outcome, we defined

$$\mathrm{BF}_{1:0} = \frac{p(data|H_1)}{p(data|H_0)} \quad (1)$$

where $p(\mathrm{data}|H_1)$ is the probability of the data given the alternative hypothesis is true, and $p(\mathrm{data}|H_0)$ is the probability of the data given the null hypothesis is true.

For trials reporting non-significance for the primary mortality outcome, we defined

$$\mathrm{BF}_{0:1} = \frac{p(data|H_0)}{p(data|H_1)} \quad (2)$$

The null hypothesis is that the rows in the contingency table are not associated (i.e. there is no effect of the intervention), and the alternative hypothesis is that the rows are associated. This arrangement is equivalent to standard null and two-sided alternative hypotheses. With this arrangement, the two hypotheses are mutually exclusive and their probabilities sum to 1.

Assuming a pretest probability of 0.5 for both the null and alternative hypotheses, we used Bayes theorem to calculate the FPR or the FNR. For trials reporting significance, we calculated the FPR as

$$FPR = \frac{1}{BF_{1:0} + 1} \quad (3)$$

Here, the FPR is the probability that the null hypothesis is true given the data (i.e. $p[H_0|\text{data}]$). The positive predictive value (PPV) is the probability the alternative hypothesis is true given the data (i.e. $p[H_1|\text{data}]$) and is equivalent to 1−FPR.

For trials reporting non-significance, we calculated the FNR as

$$FNR = \frac{1}{BF_{0:1} + 1} \quad (4)$$

Here, the FNR is the probability that the alternative hypothesis is true given the data (i.e. $p[H_1|\text{data}]$). The negative predictive value is the probability that the null hypothesis is true given the data (i.e. $p[H_0|\text{data}]$) and is equivalent to 1−FNR.

A descriptive classification was used to categorise the evidence in favour of the null or alternative hypothesis based on the BF and the associated FPR or FNR (Table 2).[9,10] To facilitate our analysis, for trials in the data set not reporting *P*-values, we calculated a two-sided *P*-value from the raw data using the prop.test function in R. Data from the RECOVERY and REMAP-CAP trials were analysed in isolation to the main data set.

A detailed explanation of conditional probability, BFs, Bayes theorem, and the assumptions used for developing equations (3) and (4) are provided in Supplementary Appendix 1.

Results

Six of 53 (11.3%) trials reported a statistically significant difference between the intervention and control groups for the primary mortality outcome. This value differs from that which we reported previously (5/54; 9.3%).[7] One trial reported mortality as a secondary outcome[11] and was included in error in our previous analysis. One trial reported an adjusted *P*-value of 0.04 and an unadjusted *P*-value of 0.08 for the primary outcome.[12] We previously classified this result as non-significant, but it is counted as significant here, as it was significant under the model used by the researchers. The trial by Papazian and colleagues[12] compared neuromuscular block to placebo in patients with severe acute respiratory distress syndrome (ARDS). The $BF_{1:0}$ was 0.83 and the FPR was 54.8%.

A table summarising the 53 trials and a spreadsheet containing the full data set is provided in Supplementary Appendix 2.

Table 2 Descriptive classification system for interpreting Bayes factors (BFs) and the false-positive risk (FPR) or false-negative risk (FNR). The FPR and FNR assume a pretest probability of 50:50. For a statistically significant result, we calculated $BF_{1:0}$ and the FPR. $BF_{1:0}$ quantifies the evidence in favour of the alternative hypothesis, and the FPR quantifies the probability the alternative hypothesis is false. For a statistically non-significant result, we calculated $BF_{0:1}$ and the FNR. $BF_{0:1}$ quantifies the strength of evidence in favour of the null hypothesis, and the FNR quantifies the probability the null hypothesis is false.

Evidence category	BF	FPR or FNR (%)
Ambiguous	<3	>25
Moderate	3–10	9.1–25
Strong	10–30	3.2–9.1
Very strong	30–100	1.0–3.2
Extreme	>100	<1.0

Trials reporting significance

For trials reporting a significant mortality difference between the intervention and control groups, the $BF_{1:0}$ varied between 0.48 and 737.1 and the FPR varied between 67.7% and 0.1%. The $BF_{1:0}$ of 0.48 (FPR 67.7%) was calculated from the trial by Cavalcanti and colleagues,[13] which was an event-driven trial with no pre-planned sample size.

Of the six trials reporting significance, four provided ambiguous evidence, one provided moderate evidence, and one provided extreme evidence in favour of the alternative hypothesis. The trial by Guerin and colleagues,[14] which compared prone with supine positioning in patients with severe ARDS, had a $BF_{1:0}$ of 737.1 and an FPR of 0.1%, consistent with extreme evidence in favour of the alternative hypothesis. Excluding the trial by Cavalcanti and colleagues,[13] for the three trials providing ambiguous evidence, the reported *P*-value was between 0.03 and 0.05, and the FPR ranged from 44.0% to 57.6%.[12,15,16]

One trial by Ferguson and colleagues,[17] which compared high-frequency oscillation ventilation with conventional ventilation in patients with severe ARDS, provided moderate evidence in favour of the alternative hypothesis ($BF_{1:0}$ 5.18; FPR 16.2%). However, another trial in the data set investigating the same hypothesis, by Young and colleagues,[18] reported non-significance and provided strong evidence in favour of the null hypothesis ($BF_{0:1}$ 11.3; FNR 8.2%).

Trials reporting non-significance

For trials reporting a non-significant mortality difference between the intervention and control groups, the $BF_{0:1}$ varied between 1.7 and 56.3 and the FNR varied between 36.9% and 1.7%. Of the 47 trials reporting non-significance, 25 (53.2%) provided strong evidence and one provided very strong evidence in favour of the null hypothesis. Thus, 26 of 47 (55.3%) trials provided strong or very strong evidence. For five trials (10.6%), the data were ambiguous, favouring neither hypothesis. The remaining 16 trials (34.0%) provided moderate evidence in favour of the null hypothesis.

Figure 1 shows the relationship between the reported *P*-value, the $BF_{0:1}$, and the FNR. Figure 2 shows the relationship between the sample size, the $BF_{0:1}$, and the FNR.

All trials with a sample size less than 500 provide limited (ambiguous or moderate) evidence in favour of the null hypothesis. Two trials had a sample size greater than 20 000. The trial by Young and colleagues,[19] which compared proton pump inhibitors with histamine antagonists for stress ulcer prophylaxis, reported a *P*-value of 0.054 and provided strong evidence in favour of the null hypothesis ($BF_{0:1}$ 15.7; FNR 6.0%). The trial by Heddle and colleagues,[20] which compared short- and long-term blood storage, reported a *P*-value of 0.340 and provided very strong evidence in favour of the null hypothesis ($BF_{0:1}$ 56.3; FNR 1.7%).

Corticosteroids and survival for patients with COVID-19

The RECOVERY trial, which included 1007 participants in the steroid arm, reported a significant 12.1% reduction in mortality for patients treated with dexamethasone (odds ratio [OR] 0.59; 95% CI: 0.44–0.78).[8] The data provided very strong evidence in favour of the alternative hypothesis ($BF_{1:0}$ 87.1; FPR 1.1%).

The REMAP-CAP trial, which included 197 participants in the steroid arm, was terminated once the data from RECOVERY were made available.[8] The REMAP-CAP trial reported a

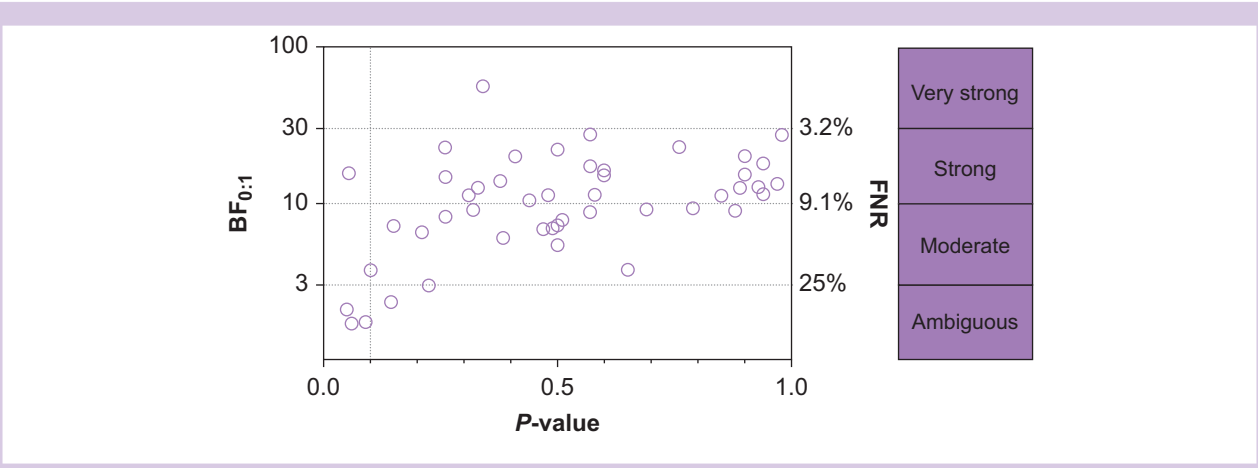

Fig 1. Relationship between the P-value, Bayes factor ($BF_{0:1}$), and the false-negative risk (FNR) for trials reporting a non-significant difference between the intervention and control groups for the primary mortality outcome. For clarity, the $BF_{0:1}$ is shown on a logarithmic scale. The horizontal grey lines distinguish the categories of evidence in favour of the null hypothesis. The vertical grey line indicates a P-value of 0.1.

non-significant 6.8% reduction in mortality in patients treated with hydrocortisone (OR 0.71; 95% CI: 0.38–1.33). The data provide moderate evidence in favour of the null hypothesis ($BF_{0:1}$ 3.6; FNR 21.6%).

Discussion

There are several benefits to reporting the BF and either the FPR or FNR. First and most obviously, unlike a P-value or a CI, the FPR and FNR quantify the probability that a real effect exists, which is the primary interest of clinicians wishing to translate research into evidence-based practice.

Second, for P-values close to the significance threshold, the FPR and FNR demonstrate how misleading the results of NHST can be. It is a common misunderstanding that when α is 0.05, there is a 5% chance that no real effect exists. As our results show, for P-values in the range 0.03–0.05, the FPR varied between 44.0% and 57.6%.[12,15,16] Note, the range does not include the trial by Cavalcanti and colleagues,[13] as independent multinominal sampling may not be valid for an event-driven trial.

Third, the FPR and FNR quantify how much a trial result informs our prior belief that an intervention is effective. Consider the trial by Combes and colleagues,[21] which compared veno-venous extracorporeal membrane oxygenation (VV ECMO) with lung-protective ventilation in patients with severe ARDS. The P-value was 0.09, close to the significance threshold. Under conditions of equipoise (i.e. a pretest probability of 0.5), the post-test probability of a real effect is 0.365 (i.e. FNR 36.5%). The data have only shifted our pretest probability by 13.5%, and we have learnt little from the trial.

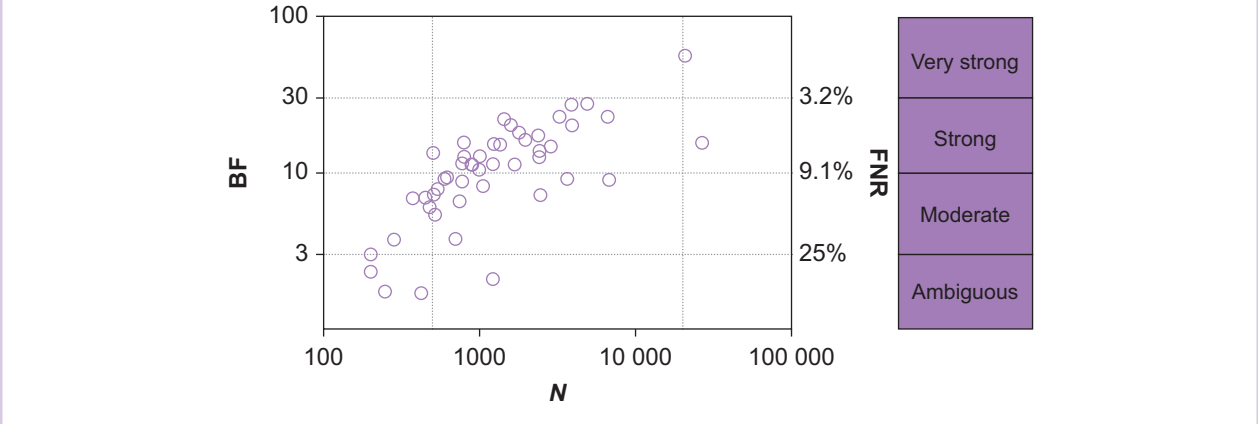

Fig 2. Relationship between the sample size, Bayes factor ($BF_{0:1}$), and the false-negative risk (FNR) for trials reporting a non-significant difference between the intervention and control groups for the primary mortality outcome. For clarity, the $BF_{0:1}$ and the sample size are shown on logarithmic scales. The horizontal grey lines distinguish the categories of evidence in favour of the null hypothesis. The two vertical grey lines indicate sample sizes of 500 and 20 000.

Given the significant result from an earlier multicentre trial,[22] we might consider a pretest probability of 0.8 more realistic for VV ECMO. For a pretest probability of 0.8, the FNR for the trial of Combes and colleagues[21] increases to 69.7% (see Supplementary Appendix 1). The data have shifted our pretest probability from 0.8 to 0.697, a 10.3% difference. Again, we have learnt very little from the trial. Indeed, when the FPR or FNR is high, we learn very little from the trial, irrespective of the pretest probability that an intervention is effective. In contrast, the trial by Guerin and colleagues,[14] comparing prone positioning with supine positioning in patients with severe ARDS, has a PPV (1—FPR) of 99.9%. For all but the most enthusiastic prior belief, the data have substantially increased the pretest probability. We have learnt a great deal from the trial.

Lastly, the BF and FNR are useful for interpreting non-significant results, which are not easily deciphered with NHST.[1] In our data set, 56.3% of non-significant results provided strong or very strong evidence in favour of the null hypothesis. Thus, more than half of all non-significant results are meaningfully interpretable from the BF. Our finding is consistent with that of Hoekstra and colleagues,[10] who found that 28 of 43 (65.1%) non-significant results reported in the *New England Journal of Medicine* in 2015 had a BF consistent with at least strong evidence in favour of the null hypothesis.

Figure 1 shows no obvious relationship between the *P*-value and the strength of the evidence favouring the null hypothesis, particularly for $P>0.1$, highlighting that *P*-values do not inform our thinking on the likely truth of the hypothesis.

Overall, larger sample sizes were associated with a higher $BF_{0:1}$ and a lower FNR (Fig. 2). The trial by Young and colleagues[19] is particularly interesting, as despite reporting a *P*-value very close to the significance threshold (0.054), the $BF_{0:1}$ was 15.7 and the FNR was 6.0%, indicating strong evidence in favour of the null hypothesis.

The RECOVERY trial provides very strong evidence of a mortality-sparing effect of dexamethasone in patients with COVID-19.[8] Given the data, the chance that there is no mortality-sparing effect of dexamethasone (i.e. the FPR) is only 1.1%. The RECOVERY trial is one of the few examples in critical care research, where the data are overwhelmingly convincing of a therapeutic effect. The REMAP-CAP data are less conclusive. However, the comparison is unfair, given that the steroid arm of the REMAP-CAP trial was terminated early and was therefore underpowered to demonstrate a mortality-sparing effect.

It is important to emphasise that the BF and the FPR or FNR only apply to the populations in which the hypothesis is tested and within the design parameters of the trial. In our previous publication from this data set, we demonstrated that low trial participant susceptibility is potentially an important contributor to the high proportion of multicentre trials in critical care reporting non-significance for mortality outcomes.[7] If few trial participants are realistically susceptible to the intervention (e.g. because of a low 'dose' of the intervention) or the outcome (e.g. if most participants are 'always survivors' or 'never survivors'), there will be little separation between the study groups. In this circumstance, the trial will inevitably report non-significance, and the BF will always be high and the FNR will always be low. It is plausible that the high proportion of trials providing strong support for the null hypothesis is partly because of low trial participant susceptibility. Conversely, underpowered trials because of small sample sizes will tend to have a high FPR (if reporting significance) or high FNR (if reporting non-significance), rendering the result uninterpretable. This is plausibly the case for half the trials in the data set.

The FPR and FNR do not completely resolve the problems associated with NHST. First, like *P*-values, the FPR and FNR do not quantify the effect size, only the probability that a non-zero effect exists. Second, calculating the FPR and FNR requires estimating the pretest probability that a real effect exists. The values for the FPR and FNR reported here apply to a pretest probability of 0.5. It might be reasonably argued that for multicentre trials, which are typically based on significant results reported from single-centre trials or meta-analyses, a pretest probability of 0.5 is too low. Choosing a pretest probability greater than 0.5 reduces the FPR and increases the FNR (see Supplementary Appendix 1). Nevertheless, as pointed out earlier, irrespective of the pretest probability, when the FPR and FNR are high we learn very little from the trial.

Our investigation builds on the work of others. As noted earlier, Hoekstra and colleagues[10] used the BF, calculated identically to here, to quantify the evidence in favour of the null hypothesis for trials reporting non-significance. Colquhoun[4] reported a method for calculating the FPR from the *P*-value for trials reporting continuous outcome variables. Other investigators have estimated the overall FPR and FNR for different populations of studies using a variety of assumptions and modelling strategies.[23–25]

The FPR and FNR provide an intuitive insight into published trial data that is not possible with traditional *P*-values and CIs. The FNR is particularly useful for interpreting non-significant results. Our analysis indicates that the quality of evidence from multicentre trials in critical care varies greatly between trials. Only one trial reporting statistical significance provided convincing evidence of a real treatment effect, and nearly half of all trials reporting non-significance provided limited evidence for the absence of a treatment effect.

Authors' contributions

Project conception: DS.
Data extraction/analysis/interpretation: both authors.
Creation of web-based calculator: CJB.
Drafting of paper: DS.
Editing/approval of paper: both authors.

Declarations of interest

The authors declare that they have no conflicts of interest.

Funding

Institutional or departmental sources.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bjao.2022.100003.

References

1. Sidebotham D. Understanding significance testing. *Anaesthesia* 2021; **76**: 1659—64
2. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle *P* value generates irreproducible results. *Nat Methods* 2015; **12**: 179—85

3. Wasserstein RL, Lazar NA. The ASA statement on *p*-values: context, process, and purpose. *Am Stat* 2016; **70**: 129–33
4. Colquhoun D. The false positive risk: a proposal concerning what to do about *p*-values. *Am Stat* 2019; **73**: 192–201
5. Matthews RAJ. Moving towards the post $p < 0.05$ era via the analysis of credibility. *Am Stat* 2019; **73**: 202–12
6. Gannon MA, de Braganca CA, Polpo A. Blending Bayesian and classical tools to define optimal sample-size-dependent significance levels. *Am Stat* 2019; **71**: 213–22
7. Sidebotham D, Popovich I, Lumley T. A Bayesian analysis of mortality outcomes in multicentre clinical trials in critical care. *Br J Anaesth* 2021; **127**: 487–94
8. Sterne JAC, Murthy S, Diaz JV, et al. Association between administration of systemic corticosteroids and mortality among critically ill patients with COVID-19: a meta-analysis. *JAMA* 2020; **324**: 1330–41
9. Jamil T, Ly A, Morey RD, Love J, Marsman M, Wagenmakers EJ. Default "Gunel and Dickey" Bayes factors for contingency tables. *Behav Res Methods* 2017; **49**: 638–52
10. Hoekstra R, Monden R, van Ravenzwaaij D, Wagenmakers EJ. Bayesian reanalysis of null results reported in medicine: strong yet variable evidence for the absence of treatment effects. *PLoS One* 2018; **13**, e0195474
11. van den Boogaard M, Slooter AJC, Bruggemann RJM, et al. Effect of haloperidol on survival among critically ill adults with a high risk of delirium: the REDUCE randomized clinical trial. *JAMA* 2018; **319**: 680–90
12. Papazian L, Forel JM, Gacouin A, et al. Neuromuscular blockers in early acute respiratory distress syndrome. *N Engl J Med* 2010; **363**: 1107–16
13. Cavalcanti AB, Suzumura EA, Laranjeira LN, et al. Effect of lung recruitment and titrated positive end-expiratory pressure (PEEP) vs low PEEP on mortality in patients with acute respiratory distress syndrome: a randomized clinical trial. *JAMA* 2017; **318**: 1335–45
14. Guerin C, Reignier J, Richard JC, et al. Prone positioning in severe acute respiratory distress syndrome. *N Engl J Med* 2013; **368**: 2159–68
15. Annane D, Renault A, Brun-Buisson C, et al. Hydrocortisone plus fludrocortisone for adults with septic shock. *N Engl J Med* 2018; **378**: 809–18
16. Gao Smith F, Perkins GD, Gates S, et al. Effect of intravenous beta-2 agonist treatment on clinical outcomes in acute respiratory distress syndrome (BALTI-2): a multicentre, randomised controlled trial. *Lancet* 2012; **379**: 229–35
17. Ferguson ND, Cook DJ, Guyatt GH, et al. High-frequency oscillation in early acute respiratory distress syndrome. *N Engl J Med* 2013; **368**: 795–805
18. Young D, Lamb SE, Shah S, et al. High-frequency oscillation for acute respiratory distress syndrome. *N Engl J Med* 2013; **368**: 806–13
19. Young PJ, Bagshaw SM, Forbes AB, et al. Effect of stress ulcer prophylaxis with proton pump inhibitors vs histamine-2 receptor blockers on in-hospital mortality among ICU patients receiving invasive mechanical ventilation: the PEPTIC randomized clinical trial. *JAMA* 2020; **323**: 616–26
20. Heddle NM, Cook RJ, Arnold DM, et al. Effect of short-term vs. long-term blood storage on mortality after transfusion. *N Engl J Med* 2016; **375**: 1937–45
21. Combes A, Hajage D, Capellier G, et al. Extracorporeal membrane oxygenation for severe acute respiratory distress syndrome. *N Engl J Med* 2018; **378**: 1965–75
22. Peek GJ, Mugford M, Tiruvoipati R, et al. Efficacy and economic assessment of conventional ventilatory support versus extracorporeal membrane oxygenation for severe adult respiratory failure (CESAR): a multicentre randomised controlled trial. *Lancet* 2009; **374**: 1351–63
23. Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005; **2**: e124
24. Jager LR, Leek JT. An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics* 2014; **15**: 1–12
25. Sidebotham D. Are most randomised trials in anaesthesia and critical care wrong? An analysis using Bayes' theorem. *Anaesthesia* 2020; **75**: 1386–93

Handling editor: Phil Hopkins