J Extra Corpor Technol. 2022;54:324–9
The Journal of ExtraCorporeal Technology

Fooled by Significance Testing: An Analysis of the LOVIT Vitamin C Trial

David Sidebotham, MB ChB, FANZCA*

**Department of Anaesthesia and the Cardiothoracic and Vascular Intensive Care Unit, Auckland City Hospital, Auckland, New Zealand*

Presented at the Perfusion Downunder Winter Meeting, Queenstown, New Zealand, August 2022.

Abstract: In this article, I discuss the potential pitfalls of interpreting *p* values, confidence intervals, and declarations of statistical significance. To illustrate the issues, I discuss the LOVIT trial, which compared high-dose vitamin C with placebo in mechanically ventilated patients with sepsis. The primary outcome – the proportion of patients who died or had persisting organ dysfunction at day 28 – was significantly higher in patients who received vitamin C ($p = .01$). The authors had hypothesized that vitamin C would have a beneficial effect, although the prior evidence for benefit was weak. There was no prior evidence for a harmful effect of high-dose vitamin C. Consequently, the pretest probability for harm was low. The sample size was calculated assuming a 10% absolute risk difference, which was optimistic. Overestimating the effect size when calculating the sample size leads to low power. For these reasons, we should be skeptical that vitamin C causes harm in septic patients, despite the significant result.

p-values and confidence intervals are probabilities concerning the chance of obtaining the observed data. However, we are more interested in the chance the intervention has a real effect on the outcome. That is to say, we are more interested in whether the hypothesis is true. A Bayesian approach allows us to estimate the false positive risk, which is the post-test probability there is no effect of the intervention. The false positive risk for the LOVIT trial (calculated from the published summary data using uniform priors for the parameter values) is 70%.

Most likely, high-dose vitamin C does not cause harm in septic patients. Most likely it has no effect at all. If there is an effect, it is probably small and most likely beneficial. **Keywords:** Randomized trials, Null hypothesis significance testing, Bayes' theorem, Bayes factor, False positive risk. *J Extra Corpor Technol. 2022;54:324–9*

"Don't believe half of what you see and none of what you hear."

—Lou Reed (or possibly Edgar Allen Poe or Thomas Jefferson)

INTRODUCTION

In this article, I review the principles and pitfalls of null hypothesis significance testing, the most popular method of statistical inference in medical research. To illustrate some of the issues, I will try and answer a simple question that arises from a recently published randomized trial: is high-dose vitamin C harmful in septic patients?

A CASE STUDY: THE LOVIT TRIAL

In the LOVIT trial, 827 patients with sepsis were randomized to receive either intravenous vitamin C (50 mg/kg 6 hourly for 96 hours) or placebo (1). The primary outcome was the proportion of patients who died or had persisting organ dysfunction at day 28. The authors estimated the sample size for the trial based on an anticipated event rate in the control group of 50% and an absolute risk difference of 10%. With this, they calculated that they would require 385 patients per group to achieve 80% power at a two-sided type I error rate (alpha) of 5%. Enrolment was subsequently increased to ensure that sufficient COVID-19 patients were included.

Received for publication May 28, 2020; accepted December 3, 2020.
Address correspondence to: David Sidebotham, MB ChB, FANZCA, Anesthesiologist and Intensivist, The Cardiothoracic and Vascular Intensive Care Unit (Ward 48), Building 32, Auckland City Hospital, 2 Park Road, Grafton, Auckland 1023, New Zealand. E-mail: dsidebotham@adhb.govt.nz
The senior author has stated that the authors have reported no material, financial, or other relationship with any healthcare-related business or other entity whose products or services are discussed in this paper.

At day 28, the primary outcome had occurred in 191 of 429 patients (44.5%) in the vitamin C group and in 167 of 434 patients (38.5%) in the control group (risk ratio, 1.21; 95% confidence interval, 1.04–1.40; $p = .01$). The authors concluded that, "In adults with sepsis …, the receipt of intravenous vitamin C resulted in a higher risk of death or persistent organ dysfunction at 28 days than the receipt of placebo."

SO, IS VITAMIN C HARMFUL IN SEPTIC PATIENTS?

The LOVIT trial was done by a well-respected group of international researchers and was almost certainly conducted to an extremely high standard, free from avoidable bias and data manipulation. There are, however, a few things to unpick.

Prior Evidence for a Beneficial Effect of Vitamin C

The finding of worse outcome in the vitamin C group was unexpected. In their introduction, the authors state, "… we tested the hypothesis that a high dose of vitamin C would *reduce* the risk of death or persistent organ dysfunction …" (emphasis mine). The observed outcome was in the opposite direction to that hypothesized.

In justifying the trial, the authors quoted several previous studies, including two reporting improved outcome with vitamin C (2,3). One of the positive studies—published in the well-respected journal *Chest* in 2017—was led by Paul Marik, an intensivist formally affiliated with the East Virginia Medical School (EVMS) (2). The Marik study was not a randomized trial but a before and after comparison. A total of 47 septic patients receiving a combination of vitamin C, hydrocortisone, and thiamine (the "after" group) were compared to an earlier cohort of 47 patients receiving standard care. Mortality in the "after" group was 8.5% compared to 40.4% in the "before" group, which amounts to an eye-watering 31.9% absolute risk difference. The magnitude of this effect size has been much discussed within the critical care community, where it has been greeted with both amazement and skepticism. In March 2022, the website Medpage Today reported an analysis of the Marik study by Kyle Sheldrick, an Australian physician and statistician. On the basis of marked similarities in the baseline characteristics between the two groups, Sheldrick concluded that the data were most likely fabricated (4).

The second article showing a benefit for vitamin C was the CITRIS-ALI trial, published in the *Journal of the American Medical Association* in 2019 (3). In the CITRIS-ALI trial, 167 patients with sepsis and acute respiratory distress syndrome were randomized to receive either placebo or vitamin C (at the same dose used in the LOVIT trial). No difference was observed in the primary outcome ($p = .86$), which was a change in organ support score. There were 46 prespecified secondary outcomes, of which 43 were not significant. One of the three significant secondary outcomes was all-cause mortality at day 28, which occurred in 38 of 82 (46.3%) patients in the placebo group and 25 of 84 (29.8%) patients in the vitamin C group ($p = .03$). However, the study was not powered to detect a mortality difference and there was no adjustment for multiple comparisons. When there is no adjustment for multiple comparisons, we expect about 5% of results to be significant due to *random chance* alone (when alpha is set to .05 or the confidence level set to 95%—see in the following). Thus, 3 from 46 (6.5%) significant secondary outcomes is unsurprising. However, when viewed in isolation, a 16.5% reduction in mortality is striking. In fairness to the authors and the journal, little was made of the observed mortality difference, which was not even mentioned in the abstract.

Statistical Models for the Equivalence of Two Proportions

The LOVIT data was analyzed using a "generalized linear mixed model with binomial distribution and a log-link function, with trial site considered as a random effect." Generalized linear models (GLMs) are a modern approach to regression analysis and have the advantage over other statistical models of accounting for confounding variables, such as differences among study sites. A GLM is the most appropriate way to analyze multicenter data, such as in the LOVIT trial. As a curious reader, all we have to go on are the observed event rates in the groups as reported in the published trial. For randomized data with minimal confounding, we can test for the equivalence of two proportions using a Fisher or Chi-squared test—both simple models for comparing two proportions that are taught in high school statistics courses. The *p*-value for the LOVIT data obtained with the Chi-squared test is .04 and for the Fisher test, it is .07. Although we do not know the extent to which the GLM used by the authors accounted for confounders, we see from the Fisher or Chi-squared test that the raw data do not provide convincing evidence of a real effect. Indeed, the Fisher test does not result in a significant *p*-value at the 5% significance threshold.

SIGNIFICANT VERSUS NOT SIGNIFICANT

A key element of null hypothesis significance testing is dichotomizing results into "significant" and "not significant." A threshold value is chosen for the *p*-value, termed alpha. Alpha is not actually a probability but the upper bound on the type I error rate. A type I error occurs when the null hypothesis is rejected despite the fact that it is true (i.e., a false positive result).

A test of significance requires formulating two competing hypotheses, the null hypothesis (H_0) and the alternative hypothesis (H_1). In randomized trials, the hypothesis structure is rarely stated explicitly, but is usually of the form:

- H_0: treatment A is no different from treatment B
- H_1: treatment A is different from treatment B

When formulated in this way, the null hypothesis is termed a "point null" and the alternative hypothesis is unspecified, free to take any nonzero value.

When $p \leq$ alpha, the null hypothesis is rejected and the alternative hypothesis is accepted. The result is "significant." We can make statements like, "the receipt of intravenous vitamin C resulted in a higher risk of death or persistent organ dysfunction in septic patients." When $p >$ alpha, the situation is a little more nuanced. We do not reject the null hypothesis but neither can we accept it as true (5). We find ourselves in a no man's land, unable to draw any firm conclusions from the data. That is the meaning of "not significant" and that is the status of most multicenter trials in critical care (6–9).

Why choose an alpha value of 5%? Why not 1 or 10%? Is there some inflection point on some curve that affords special meaning to the 5% threshold? Does an alpha value of .05 relate to some obscure aspect of probability theory? The answer to both questions is no. The 5% threshold arose as a quirk of history. In fact, there is a groundswell of support for reducing the default threshold from .05 to .005 (10).

CONFIDENCE INTERVALS VERSUS *p*-VALUES

Many researchers and clinicians think that reporting a point estimate of effect along with a confidence interval makes more sense than reporting a *p*-value. I agree. *p*-values are slippery concepts that are frequently misunderstood (11). Furthermore, confidence intervals provide additional information to *p*-values: a plausible range for the effect size. However, there are a few points worth mentioning.

First, a 95% (i.e., 1–alpha) confidence interval does not provide a range within which there is a 95% chance of finding the true effect size. Rather, if an identical experiment were repeated many times and a 95% confidence interval calculated on each occasion, then 95% of those intervals would contain the true effect size. While subtle, this distinction is important as it confirms for confidence intervals a limitation that exists for *p*-values: both are probabilities concerning the likelihood of obtaining the *data*. They tell us nothing about the probability the (alternative) hypothesis is true (i.e., the intervention has a real effect). Using mathematical notation, a *p*-value can be written as $p(\text{data}|H_0)$. Here, the letter "*p*" refers to "the probability of" and the symbol "|" means "given that." Thus, a *p*-value is the probability of obtaining sample data at least as extreme as that observed given the null hypothesis is true. For a two-sided test, the observed *p*-value is the area bounded by the absolute value of the test statistic and the two tails of the null distribution (Figure 1).

A confidence interval is more complicated than a *p*-value to understand as a probability, but may be written as $p(\text{interval contains } X|X = x)$. In words, we can say that a confidence interval is the probability of obtaining an interval over repeated sampling (i.e., of obtaining multiple sets of sample data) that, with probability 1–alpha, contains the effect size (X) given that X takes a particular value (x). Whew! The point is that both a *p*-value and a confidence interval tell us nothing about the probability an effect is real and are a function of the sample data.

A second issue with confidence intervals is that they are often reported as a risk ratio (relative risk) or an odds ratio, which makes it hard to determine the actual range for the effect size. For instance, the 95% confidence interval for the LOVIT data, calculated using the mixed linear model, is reported as a relative risk of 1.04–1.40. What does that mean? It means that the plausible range for the adverse outcome associated with vitamin C most likely lies somewhere between 1.04 and 1.4 times the baseline risk. In terms of absolute risk difference, the interval is approximately 1.5–16.0%. However, calculating a 95% confidence using the proportions provided in the trial summary using standard formulae yields the interval −.5 to 12.6%. Since the interval includes the null value (zero risk difference), the result would be classified as not significant.

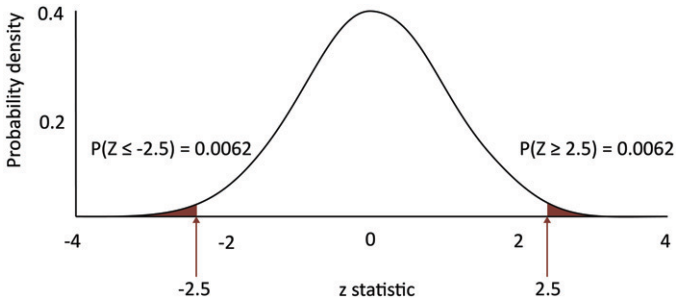

Figure 1. Probability of the test statistic under the null distribution. Data have been collected and a test statistic (*z*-test in this case) has been calculated. The value of the *z*-statistic is +2.5. The curve shows the null distribution, which is the probability density function of the test statistic assuming that the null hypothesis is true. We are doing a two-sided test, where we make no assumptions as to the direction of any difference. The *p*-value is the probability that *z* is <-2.5 or $>+2.5$. The *p*-value is shown as the shaded areas, and in this case is $2 \times .0062 = .0124$. Since $.01 < p \leq .05$, we would reject the null hypothesis at the 5% significance threshold (but not at the 1% threshold).

TWO OTHER CONSIDERATIONS

When a result is extreme or—as is the case with the LOVIT study—or unexpected, two issues should be considered that are not accounted for by the *p*-value or the confidence interval: the pretest probability of the observed result and the probability the alternative hypothesis is true.

Pretest Probability

How plausible is the finding of increased harm in the LOVIT study? Afterall, we are talking about a commonly used, water soluble vitamin, not cyclophosphamide or Novichok. While the dose used in the LOVIT study was high (roughly 20 g/day) compared to normal over-the-counter vitamin C supplements (1–2 g/day), there is no evidence of severe toxicity, even at high doses. Reported side-effects of high-dose vitamin C include diarrhea, nausea, and vomiting. So, could high-dose vitamin C plausibly cause severe toxicity, including death in patients with sepsis. Perhaps. Probably not.

Consider a study done by Rupert Sheldrake and reported in the *Journal of Scientific Exploration in 2003* (12). Doctor Sheldrake was interested in whether a talking African Grey parrot, N'kisi, could read the mind of his owner. After extensive testing of the parrot and its owner, Sheldrake declared that the data were consistent with telepathic powers. In support of his claim, Sheldrake reported an extremely small *p*-value. However, if we a priori consider that it is impossible for parrots (verbal or otherwise) to read the thoughts of human beings, then—irrespective of the *p*-value—the posttest probability that N'kisi had psychic powers is zero. One *p*-value and one parrot are insufficient reasons to upend ones framework for how the universe is constructed (13). While the pretest probability of a harmful effect of vitamin C is not zero, it is probably not high. In fact, it is probably quite low. The pretest probability has a major effect on the chance that a significant finding is associated with a real effect (11,14). Furthermore, *p*-values reported in prestigious journals like the *New England Journal of Medicine* are not immune from the problems that bedevil *p*-values reported in less prestigious journals (15).

The Alternative Hypothesis

The second consideration concerns the alternative hypothesis. As previously noted, a *p*-value relates to the probability of the data under a true *null hypothesis*. What about the probability of the data under a true alternative hypothesis? The hypothesis that states vitamin C does have an effect (either harmful or beneficial) in patients with sepsis. To answer this question (and with apologies to the *Fall*), we must enter the wonderful and frightening world of Bayesian statistics.

THE WONDERFUL AND FRIGHTENING WORLD OF BAYESIAN STATISTICS

Bayesian theory is conceptually demanding, involving conditional probabilities and complicated integrals. There is, however, one saving grace: the results of a Bayesian analysis are intuitive and easily comprehended by nonexperts. Here, I am going to focus one just aspect of Bayesian inference, the Bayes factor.

Bayes Factors

A Bayes factor is a probability ratio concerning the data under two competing models:

$$BF_{A:B} = \frac{p(\text{data} \mid \text{model A})}{p(\text{data} \mid \text{model B})}$$

The statement $p(\text{data}|\text{model A})$ means "the probability of obtaining the sample data given model A is true." The subscript A:B refers to the fact the Bayes factor is the ratio of A to B (as opposed to B to A). So, if the probability of the data under model A is .5 and the probability under model B is .05, the $BF_{A:B}$ is 10, meaning that the data are 10 times more probable under model A than model B.

If we substitute the alternative and null hypothesis for models A and B, we have:

$$BF_{1:0} = \frac{p(\text{data} \mid H_1)}{p(\text{data} \mid H_0)}$$

Unlike a *p*-value, which prioritizes the null hypothesis, a Bayes factor affords equal weight to both hypotheses. Another advantage of a Bayes factor is that it can be used to quantify the evidence of absence. Thus, unlike a *p*-value, a Bayes factor (specifically $BF_{0:1}$) quantifies the evidence in favor of the null hypothesis.

To be clear, Bayes factors are not a panacea for the problems associated with null hypothesis significance testing. When calculating the Bayes factor, prior distributions must be specified for the parameter values, which in this case is the plausible range of values for the control and the intervention event rates. The thorny topic of assigning priors to the parameters is the subject of much debate in the statistical literature. Second, there is no "correct" method for calculating Bayes factors. Both the method and the values assigned to the priors impact upon the end result.

The Bayes factor favoring the alternative hypothesis (i.e., $BF_{1:0}$) calculated for the LOVIT data is .42 (16).* To get the Bayes factor favoring the null hypothesis (i.e., $BF_{0:1}$), we take the reciprocal of .42, which is 2.4. A $BF_{0:1}$ of 2.4

*The Bayes factor was calculated using the Gunnel and Dickey method using default (uninformative priors) for the parameter values. See reference (16). Jamil T, Ly A, Morey RD, et al. Default "Gunel and Dickey" Bayes factors for contingency tables. Behav Res Methods. 2017;49:638–52.

indicates the data are 2.4 times more probable under the *null hypothesis* than the alternative hypothesis. A Bayes factor <3 is considered ambiguous, providing little evidence in favor of either hypothesis. If instead of .42, the $BF_{1:0}$ for the LOVIT data had been 42, we might reasonably claim the data provide very strong evidence in favor of the alternative hypothesis. For instance, the $BF_{1:0}$ calculated form the RECOVERY trial on the mortality-sparing effect of dexamethasone in mechanically ventilated patients with COVID-19 is 87.1 (17). There is very little doubt that dexamethasone saves lives in mechanically ventilated patients with COVID-19. By contrast, the Bayes factor on the LOVIT data provide no meaningful evidence for an effect—either harmful or beneficial for vitamin C in septic patients.

The False Positive Risk

Armed with a Bayes factor and a simplifying assumption, we can estimate posttest probabilities concerning the truth of the hypothesis (17,18). The false positive risk (FPR) is the probability the null hypothesis is true given the test is significant, which can be written using mathematical notation as $p(H_0|\text{data})$.

If we assume that the pretest probabilities associated with the null and the alternative hypotheses are both .5, then the FPR is given by:

$$FPR = \frac{1}{BF_{1:0}+1}$$

For the LOVIT trial, the FPR is .7 (70%), meaning the chance that there is a real effect (either harmful or beneficial) of vitamin C is only 30%. Notice that the pretest probability of a real effect for vitamin C was .5 and the posttest probability is .3. The data have reduced our belief in a real effect of vitamin C by a modest 20%. We have learnt little from the trial. By contrast, the posttest probability of a mortality-sparing effect of dexamethasone from the RECOVERY data is 98.8% (17,19). The RECOVERY data have increased our belief in a beneficial effect of dexamethasone by nearly 50%. We have learnt a great deal from the RECOVERY trial but very little from the LOVIT trial.

Alpha Versus the False Positive Risk

Observant readers may have noticed something puzzling. If the upper bound on the type I error rate (i.e., alpha) is 5%, how can the FPR be 70%? To explain this apparent discrepancy requires a brief foray into conditional probabilities.

The type I error rate is the probability of getting a significant test result given that there is no effect. Using mathematical notation, we can write $p(T^+|H_0)$, where T^+ means "test is significant." By contrast, the FPR is the probability that there is no effect given the test result is significant, which may be written as $p(H_0|T^+)$. Notice that the type I error rate and the FPR are inverse conditional probabilities, where the term inverse refers to the order of the terms. The two quantities are not the same and can take very different values.

In fact, for trials reporting weakly significant *p*-values (.01–.05), the FPR typically exceeds 25%, which is much higher than the 5% value for alpha (10,17,18). In our review multicenter trials in critical care, the FPR for trials reporting statistical significance varied between .1 and 67% (17).

POWER

The power of a study is the probability that the test will be significant given a real effect exists. Power depends on the sample size, the effect size, and alpha. Studies typically report power of at least 80%. However, in reality, the achieved power of a study is usually much <80%. Recently, van Zwet and colleagues estimated that the median achieved power of more than 20,000 trials in the Cochrane Database of Systematic reviews was only 13% (20). Thirteen percent! How could such a situation arise? How could achieved power be less than one-fifth of design power?

When calculating the sample size, authors estimate the effect size that might exist in the population. If the hypothesized effect size is larger than the true effect size, the sample size will be too small to reliably detect the effect and achieved power will be less than design power. Overestimating the true population effect size is termed "delta inflation." Delta inflation is extremely common (6,7). In our review of multicenter trials in critical care, the median hypothesized effect size (i.e., the effect size used to calculate the sample size) was four times higher than the observed effect size (6).

When power is low, several things happen—all of them bad (13). Most studies report nonsignificant results. Of the small proportion of studies that do report significance, a high proportion are false. That is to say, the FPR is high and concordance across studies investigating the same intervention is low. In fact, when power is <20%, the distribution of *p*-values becomes relatively uniform between 0 and 1, meaning that a *p*-value of .87 is about as likely as a *p*-value of .01, *irrespective of the efficacy of the intervention*. The reason most studies report nonsignificant results is that there is a greater proportion of the probability density lies between .05 and 1.0 than between 0 and .05.

Which brings us back to the LOVIT trial. As previously noted, the observed effect size was in the opposite direction to the hypothesized effect size. The data on which the hypothesized effect size was based was at best limited and at worst suspect. We might reasonably argue

that anticipating a 10% difference in the rate of death and organ dysfunction was … optimistic. If that is indeed the case, then the power of the LOVIT trial is probably well below 80%.

CONCLUSION

Hypothesis testing is a tricky business. It is easy to be fooled. As previously noted, *p*-values are slippery and confidence intervals are not quite what they seem. In a previous article, I wrote a checklist, posed as a series of questions, which may help readers to avoid some of the pitfalls (13)

1. Is the effect size that was used to calculate the sample size implausibly large?
2. Were the trial participants realistically susceptible to the intervention?
3. Was there a clinically meaningful treatment separation between the intervention and the control groups?
4. Does the result make sense in terms of prior knowledge and biological plausibility?

And finally, returning to the original question: does vitamin C cause harm in septic patients? Although it is impossible to be certain, the most likely answer is: no. Probably it has no meaningful effect at all. In the unlikely event that vitamin C does have a nonzero effect, it is probably beneficial and probably small. Almost certainly smaller than a 10% absolute risk difference.

REFERENCES

1. Lamontagne F, Masse MH, Menard J, et al. Intravenous vitamin C in adults with sepsis in the intensive care unit. N Engl J Med. 2022;386: 2387–98.
2. Marik PE, Khangoora V, Rivera R, et al. Hydrocortisone, vitamin C, and thiamine for the treatment of severe sepsis and septic shock: A retrospective before-after study. Chest. 2017;151:1229–38.
3. Fowler AA 3rd, Truwit JD, Hite RD, et al. Effect of vitamin C infusion on organ failure and biomarkers of inflammation and vascular injury in patients with sepsis and severe acute respiratory failure: The CITRIS-ALI randomized clinical trial. JAMA. 2019; 322:1261–70.
4. Fiore K. Infamous vitamin C study may rely on fraudulent data, 2022. Available at: https://www.medpagetoday.com/special-reports/exclusives/97865. Accessed June 18, 2022.
5. Alderson P. Absence of evidence is not evidence of absence. BMJ. 2004;328:476–7.
6. 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.
7. Aberegg SK, Richards DR, O'Brien JM. Delta inflation: A bias in the design of randomized controlled trials in critical care medicine. Crit Care. 2010;14:R77.
8. Laffey JG, Kavanagh BP. Negative trials in critical care: Why most research is probably wrong. Lancet Respir Med. 2018;6:659–60.
9. Santacruz CA, Pereira AJ, Celis E, et al. Which multicenter randomised controlled trials in critical care medicine have shown reduced mortality? A systematic review. Crit Care Med. 2019;47:1680–91.
10. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. Nat Hum Behav. 2018;2:6–10.
11. Sidebotham D. Are most randomised trials in anaesthesia and critical care wrong? An analysis using Bayes' Theorem. Anaesthesia. 2020;75: 1386–93.
12. Sheldrake R. Testing a language-using parrot for telepathy. J Sci Explor. 2003;17:601–16.
13. Sidebotham D. Understanding significance testing. Anaesthesia. 2021;76:1659–64.
14. Vail EA, Avidan MS. Trials with "non-significant" results are not insignificant trials: A common significance threshold distorts reporting and interpretation of trial results. Br J Anaesth. 2022; 129: 643–6.
15. Wasserstein RL, Lazar NA. The ASA statement on *p*-values: Context, process, and purpose. Am Stat. 2016;70:129–33.
16. Jamil T, Ly A, Morey RD, et al. Default "Gunel and Dickey" Bayes factors for contingency tables. Behav Res Methods. 2017;49: 638–52.
17. Sidebotham D, Barlow CJ. The false-positive and false-negative risks for individual multicentre trials in critical care. BJA Open. 2022;1: 100003.
18. Colquhoun D. The false positive risk: A proposal concerning what to do about *p*-values. Am Stat. 2019;73:192–201.
19. 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.
20. van Zwet E, Schwab S, Senn S. The statistical properties of RCTs and a proposal for shrinkage. Stat Med. 2021;40:6107–17.