Journal of Intensive Care Medicine
2022, Vol. 37(7) 954-964

DOI: 10.1177/08850666211053793
journals.sagepub.com/home/jic


Reverse Bayesian Implications of p-Values Reported in Critical Care Randomized Trials

Sarah Nostedt, MD, FRCPC[1,2] and Ari R. Joffe, MD, FRCPC[1,2]

Abstract

Background: Misinterpretations of the p-value in null-hypothesis statistical testing are common. We aimed to determine the implications of observed p-values in critical care randomized controlled trials (RCTs).

Methods: We included three cohorts of published RCTs: Adult-RCTs reporting a mortality outcome, Pediatric-RCTs reporting a mortality outcome, and recent Consecutive-RCTs reporting p-value ≤.10 in six higher-impact journals. We recorded descriptive information from RCTs. Reverse Bayesian implications of obtained p-values were calculated, reported as percentages with inter-quartile ranges.

Results: Obtained p-value was ≤.005 in 11/216 (5.1%) Adult-RCTs, 2/120 (1.7%) Pediatric-RCTs, and 37/90 (41.1%) Consecutive-RCTs. An obtained p-value .05–.0051 had high False Positive Rates; in Adult-RCTs, minimum (assuming prior probability of the alternative hypothesis was 50%) and realistic (assuming prior probability of the alternative hypothesis was 10%) False Positive Rates were 16.7% [11.2, 21.8] and 64.3% [53.2, 71.4]. An obtained p-value ≤.005 had lower False Positive Rates; in Adult-RCTs the realistic False Positive Rate was 7.7% [7.7, 16.0]. The realistic probability of the alternative hypothesis for obtained p-value .05–.0051 (ie, Positive Predictive Value) was 28.0% [24.1, 34.8], 30.6% [27.7, 48.5], 29.3% [24.3, 41.0], and 32.7% [24.1, 43.5] for Adult-RCTs, Pediatric-RCTs, Consecutive-RCTs primary and secondary outcome, respectively. The maximum Positive Predictive Value for p-value category .05–.0051 was median 77.8%, 79.8%, 78.8%, and 81.4% respectively. To have maximum or realistic Positive Predictive Value >90% or >80%, RCTs needed to have obtained p-value ≤.005. The credibility of p-value .05–.0051 findings were easy to challenge, and the credibility to rule-out an effect with p-value >.05 to .10 was low. The probability that a replication study would obtain p-value ≤.05 did not approach 90% unless the obtained p-value was ≤.005.

Conclusions: Unless the obtained p-value was ≤.005, the False Positive Rate was high, and the Positive Predictive Value and probability of replication of "statistically significant" findings were low.

Keywords

Bayesian, critical care, false positive rate, p-value, positive predictive value, randomized controlled trial

In 2005 Ioannidis claimed "most published research findings are false," pointing out that "the high rate of nonreplication… is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than .05."[1(p0696)] In 1999 Goodman discussed the "p-value fallacy… the illusion that conclusions can be produced with certain "error rates' without consideration of information from outside the experiment [ie, biological plausibility and prior evidence]…."[2(p995)] The p-value is the probability of observing the experimental data or more extreme data under the assumption that the null hypothesis (Ho) is correct.[3] Misinterpretations of the p-value as used in null-hypothesis statistical testing (NHST) are common, including that the p-value is the probability the Ho is true (rather, it assumes the Ho is true), that chance alone produced the observed association (rather, it assumes chance was operating alone), or that if you reject Ho because P ≤ .05, the chance you are in error is 5% (rather, this chance is much higher).[3] The only way to determine hypotheses probabilities is by using Bayesian inference methods.[6] Table 1 gives definitions of terms and methods used in this study.

Bayes theorem links the posterior probability of the alternative hypothesis H1, (Pr[H1] | data) to the prior Pr(H1) using the Bayes Factor (BF), the probability of the observed data given the alternative hypothesis H1 divided by the probability of the observed data given the null hypothesis (Ho), that is Pr(data | H1)/

[1]University of Alberta, Edmonton, Alberta, Canada
[2]Stollery Children's Hospital, Edmonton, Alberta, Canada

Received April 27, 2021. Received revised September 29, 2021. Accepted September 30, 2021.

Corresponding Author:
Ari R. Joffe, MD, 4-546 Edmonton Clinic Health Academy, 11405 112 Street, Edmonton, Alberta, Canada T6G 1C9.
Email: ari.joffe@ahs.ca

Table 1. Glossary of Terminology and Methods Used.

Analysis of Credibility (AnCred) for statistically "significant" finding: asks how skeptical one would need to be to find the obtained effect estimate nevertheless unconvincing of an effect.
- **Critical Prior Interval (CPI)**: an interval derived (with values [-S, S]) so that the posterior credible interval, based on the data from the experiment, just includes zero (no effect).
- **Skepticism Limit (SL)**: if external evidence or insight suggests effect sizes are unlikely to exceed the effect size of "S', the finding being statistically significant lacks credibility.

Analysis of Credibility (AnCred) for statistically "non-significant" finding: asks how much of an advocate one would need to be to find the effect estimate nevertheless convincing of an effect.
- **Critical Prior Interval (CPI)**: an interval derived (with interval 1 to AL for odds ratios, or 0 to AL for d) so that the posterior credible interval, based on the data from the experiment, no longer includes zero (no effect).
- **Advocacy Limit (AL)**: if external evidence or insight suggests effect sizes are unlikely to exceed the effect size of "AL", the finding being statistically non-significant lacks credibility.

Bayes Factor$_{01}$ (BF): The probability of the observed data given the null hypothesis Ho divided by the probability of the observed data given the alternative hypothesis H1, formally as Pr(data | Ho) / Pr(data | H1). This is a measure of how much the obtained data supports the one hypothesis versus the other.

Bayes Factor Bound$_{01}$ (BFB): The smallest odds against the Ho (or the highest odds in favor of H1) given the data obtained from the experiment, that can be generated by any reasonable choice of the prior distribution for Ho. This is the Bayes Factor that most favors the rejection of the Ho.
- **Calculated according to Berger et al.**: this can be calculated as -eplog(p), where p is the obtained p-value in an experiment, and $p < 1/e$, that is, p is <.37 (otherwise the Bayes Factor Bound is 1, and the obtained data provides no evidence to change the prior Pr[Ho]).

Bayes's Theorem: The theorem links the posterior probability of H1 given the obtained data (Pr[H1] | data) to the prior probability of H1 (Pr[H1]) using the Bayes Factor (BF). Formally, the Posterior Odds of H1 = BF X Prior Odds of H1, that is: Pr(H1 | data)/Pr(Ho | data) = BF X Pr(H1)/Pr(Ho).

Confidence Interval (95% CI) of effect size: how often 95% confidence intervals computed from very many studies would contain the true effect size (ie, 95% is the frequency with which other unobserved intervals will contain the true effect). Even under ideal conditions the chance that a future estimate will fall within the current interval will usually be much <95% [eg, at best 83%]. The width of the 95% CI is a measure of the precision of the effect size estimate given the obtained data.

False Positive Risk or Rate (FPR): When a statistical test comes out positive (according to the chosen statistical significance threshold), the probability that you have a false positive (ie, that there is no real effect and the results have occurred by chance). Formally, Pr(Ho | significant p-value), the posterior probability of Ho given the data from the experiment (represented by the obtained p-value).
- **Calculated using Colquhoun's calculator**: assuming the experiment had power of 78%. The calculations were derived in the context of testing a precise hypothesis [Ho, that the effect size is zero] and comparing means of two independent samples each of n normally-distributed observations; however, the results should be similar in other situations of obtained p-values [8]. The calculator is used by specifying the prior probability one assigned to the hypothesis, and the data obtained from the experiment (given by the obtained p-value).
- **Calculated using Berger et al.'s method**: using the Bayes Factor Bound given the data obtained from the experiment (ie, the obtained p-value), and specifying the prior probability one assigned to the hypothesis.
- **Positive Predictive Value (PPV)**: Pr(H1 | significant p-value), the highest posterior probability of H1 given the data from the experiment, after specifying the prior probability one assigned to the hypothesis Pr(H1).

Odds: probability/(1-probability)

p-value: The probability, under the assumption of no association [no effect; Ho], of obtaining a result equal to or more extreme that what was actually observed. This can be put as Pr(data or more extreme data | Ho). The p-value is *not* Pr(Ho | data): that is the False Positive Risk.

p-value 80% interval: an interval with an 80% chance of including the p value given by a replication study using the same sample size. This is reported by giving the 10th to the 90th percentile of expected replication p-values.

Probability: odds/(1 + odds)

Reverse-Bayes Approach: Reversing the conventional direction of Bayes's Theorem, which can be used to determine the level of prior belief (ie, Pr[H1]) necessary to reach a desired level of posterior belief (ie, Pr[H1 | data]), given the evidence that has been observed in the data (ie, the Bayes Factor, BF). Formally, Prior Odds of H1 = Posterior Odds of H1 / BF, that is:
[the necessary Pr(H1)/Pr(Ho)] = [the desired Pr(H1 | data)/Pr(Ho | data)] / BF.

Standardized Mean Difference (d): Uses the amount of variation in scores (Standard Deviation) to contextualize the difference between groups in a continuous outcome. Calculated as mean difference between groups / standard deviation in control group.

Pr(data | Ho).[7] Formally, the Posterior Odds of H1 = BF X Prior Odds of H1, that is: Pr(H1 | data)/Pr(Ho | data) = BF X Pr(H1)/Pr(Ho).[12] Bayesian inference has been criticized because the Prior can be subjective (although this is not necessarily so), informed by researchers' beliefs, scientific consensus, and evidence from similar research questions in the same field.[13] Approaches to help resolve this subjectivity "problem" have been suggested that involve so-called "reversing of the Bayesian argument", including: one can convert the observed p-value to a minimum BF (the strongest possible evidence against Ho given the data obtained), calculate the Prior Pr(H1) necessary for the observed p-value to indicate a desired false positive rate (FPR), or calculate the minimum FPR given a specified Prior Pr(H1).[7,14] The approach is based on the fact that "it is entirely justifiable to "flip" Bayes's Theorem around;"[12(p4)] "the basic idea is to invert Bayes' theorem: a specified posterior is

combined with the data to obtain the Reverse-Bayes prior, which is then used for further inference,"[12(p21)] eg, this "allows the assessment of new findings [ie, the data] on the basis of whether the resulting prior is reasonable in the light of existing knowledge [ie, prior evidence]."[12(p2)]

We aimed to describe the implications of observed p-values in published critical care randomized controlled trials (RCTs) in order to demonstrate the importance of the p-value fallacy. We find that, in three cohorts of published RCTs, only those RCTs that obtained a p-value ≤.005 might be considered reliable findings.

Materials and Methods

Included Randomized Trials

As only publicly available published data was recorded, this study did not require ethics board approval.

First, we examined the cohort of 216 human adult multicenter RCTs reviewed by others ["Adult-RCTs"].[17] We included all 57 RCTs that obtained p-value ≤.10. For Analysis of Credibility (see below) we included another 14 RCTs that obtained p-value .11–.20.

Second, we searched the cohort of human pediatric RCTs at https://picutrials.net using search words "mortality" or "multicenter."[18] We examined the abstracts (and full text if necessary) to include any RCT with a reported obtained p-value for a mortality outcome ["Pediatric-RCTs"]. Of 120 eligible RCTs, we included all 25 RCTs that obtained a p-value ≤.10. For Analysis of Credibility we included another 13 RCTs that obtained p-value .11–.20.

Third, we screened the title and abstract of all publications in 6 journals (NEJM, JAMA, Critical Care, Critical Care Medicine, Pediatric Critical Care Medicine, and Intensive Care Medicine) starting backwards from January 2019 for eligibility, until 15 publications were included from each journal. We used a detailed study instruction manual to guide screening and data collection (Supplemental Material 1). Eligibility was defined as: topic involves solely or predominantly (>80%) critically ill patients; RCT comparing groups with respect to some interventional exposure to find an outcome effect size with an explicitly reported p-value; and full publication. We excluded studies if the primary outcome had p-value ≥.10, or an exact p-value for the primary outcome was not reported (often because only a "less than" p-value was provided). This resulted in 90 "Consecutive-RCTs."

From each included study we obtained descriptive information, more detailed for the Consecutive-RCTs, and calculated outcomes as described in the detailed instruction manual (see Supplemental Material 1). We recorded the category of primary outcome; study size; study outcomes; effects sizes (ES); power calculation numbers if reported; and the obtained p-value. In the Consecutive-RCTs we also recorded the main secondary outcome and associated information as above. When an ES was not reported, we calculated this based on the reported values in the published study.

Table 2. P-Value Categories Obtained in the Cohorts of Critical Care RCTs.

P-value category	P = .051 to .10	p = .05 to .0051	p ≤ .005
Adult RCTs			
n studies	13	33	11
% of studies with p ≤ .10	22.8%	57.9%	19.3%
% of all studies	6.0%	15.3%	5.1%
Pediatric RCTs			
n studies	13	10	2
% of studies with p ≤ .10	52%	40%	8%
% of all studies	10.8%	8.3%	1.7%
Consecutive RCTs			
Primary outcome n studies	7	46	37
% of studies with p ≤ .10	7.8%	51.1%	41.1%
Secondary outcome n studies	17	44	26
% of studies	19.5%[a]	50.6%	29.9%

For Consecutive RCTs, only studies with p ≤ .10 for the primary outcome were screened for inclusion.
a.For secondary outcomes, 14 (16%) of p-values were >.10.

Outcomes

First, we described the categorization of obtained p-values into 1) ≤.005, 2) .0051 to .05, and 3) .051 to .10. This was done because there have been calls to lower the threshold for "statistical significance" to p ≤ .005.[19,20] A two-sided p-value .05 corresponds to a BF in favor of H1 ranging from 2.5 to 3.4 "under reasonable assumptions about H1", which is "weak to very weak" evidence.[19(p7)] A two-sided p-value .005 corresponds to a BF in favor of H1 ranging from 14 to 26, which is "substantial to strong evidence."[19(p7)] Regardless of power, with a prior Pr(H1) 10% and p-value threshold .05 the FPR is >33%; reducing the p-value threshold to .005 "would reduce this minimum FPR to 5%… over a wide range of statistical powers."[19(p7)] The false negative rate does not increase if sample sizes are increased to keep power constant; to maintain "80% power would require an increase in sample sizes of about 70%."[19(p7)]

Second, we calculated reverse Bayesian argument values suggested by Colquhoun, assuming study power was ∼80%, based on the evidence provided by the experiment (ie, the observed p-value), and using the online calculator.[8,9] 1) The likelihood ratio of H1/Ho is the relative likelihood of two hypotheses (ie, a BF). 2) The prior Pr(H1) required to have a FPR (the probability that a result which is "statistically significant" at a specified p-value is a false positive result) of 5%. 3) The minimum FPR, assuming a prior Pr(H1) 50% (ie, equipoise). 4) The realistic FPR, assuming a prior Pr(H1) 10%.

Third, we calculated the reverse Bayesian argument values suggested by Berger et al. based on the evidence provided by the experiment (ie, the observed p-value).[10,11] 1) The Bayes Factor Bound (BFB) is an upper bound on the BF (ie, the strongest case for the H1 relative to the Ho, given the data obtained). 2) The highest possible posterior Pr(H1 | data), Pr^U(H1 | obtained p), assuming the prior Pr(H1) 50% (ie, equipoise). 3) The realistic

Table 3. Reverse Bayesian Implications of the Obtained p-Values ≤.10 in Adult RCTs in the Field of Critical Care Research.

Study group	Overall (n = 57)	p > .05 to .10 (n = 13)	p = .05 to .0051 (n = 33)	p ≤ .005 (n = 11)
Reverse Bayesian Argument (Colquhoun)				
Likelihood Ratio: Pr(data | H1)/Pr(data | Ho)	5.0 [2.8, 14.0] (1.1-107.2)	1.8 [1.4, 2.2] (1.1-2.2)	5.0 [3.6, 7.9] (2.8-21.3)	107.2 [43.6, 107.2] (28.6-107.2)
Prior Pr(H1) to have FPR 5%	79.2 [58.0, 87.3] (15.1-94.5)	91.3 [89.6, 93.1] (89.6-94.5)	79.2 [70.6, 84.1] (47.2-87.3)	15.1 [15.1, 28.6] (15.1-39.9)
Minimum FPR (using Prior Pr[H1] of 0.5)	16.7 [7.0, 26.6] (0.9-47.3)	35.7 [31.3, 41.8] (31.3-47.3)	16.7 [11.2, 21.8] (4.5-26.6)	0.9 [0.9, 2.1] (0.9-3.4)
Realistic FPR (using Prior Pr[H1] of 0.1)	64.3 [39.8, 76.6] (7.7-89.0)	83.3 [80.4, 86.6] (80.4-89.0)	64.3 [53.2, 71.4] (29.7-76.6)	7.7 [7.7, 16.0] (7.7-24.0)
Berger Bayesian Calculations				
BF Bound [upper bound for Pr(data | H1)/ Pr(data | Ho)]	3.5 [2.5, 7.4] (1.6-58.3)	2.0 [1.8, 2.2] (1.6-2.2)	3.5 [2.9, 4.8] (2.5-10.6)	58.3 [21.1, 58.3] (13.9-58.3)
PPV: Highest Posterior Pr^U(H1 | data) using Prior Pr(H1) of 0.5.	77.8 [71.1, 87.9] (61.5-98.3)	66.4 [63.7, 68.5] (61.5-68.5)	77.8 [74.1, 82.8] (71.1-94.4)	98.3 [95.5, 98.3] (93.3-98.3)
PPV: Realistic Posterior Pr^R(H1 | data) using Prior Pr(H1) of 0.1	28.0 [21.4, 44.9] (15.1-86.6)	18.0 [16.4, 19.5] (15.1-19.5)	28.0 [24.1, 34.8] (21.4-54.1)	86.6 [70.1, 86.6] (60.7-86.6)
P-intervals implications				
10^{th} percentile of p-value prediction interval	0.0001 [<0.0001, 0.0002]	0.0003 [0.0002, 0.0004]	0.0001 [<0.0001, 0.0001]	<0.0001
90^{th} percentile of p-value prediction interval	0.72 [0.47, 0.88] (0.13-0.99)	0.99 [0.95, 0.99] (0.95-0.99)	0.72 [0.60, 0.81] (0.38-0.88)	0.13 [0.13, 0.24] (0.13-0.32)
Probability a replication study will have p ≤ 0.05	58.3 [50.0, 71.7] (37.6-91.3)	44.1 [40.7, 46.8] (37.6-46.8)	58.3 [53.7, 64.7] (50.0-76.9)	91.3 [84.3, 91.3] (80.2-91.3)

Values as: median [IQR] (range). BF: Bayes factor; FPR: false positive rate; Ho: null hypothesis; H1: alternative hypothesis; Pr: probability; PPV: positive predictive value. We report values to one decimal place for consistency of presentation; this is not intended to suggest the values can be known with such precision given the small sample sizes.

posterior Pr(H1 | data), Pr^R(H1 | obtained p), assuming the prior Pr(H1) 10%. These Pr(H1 | p) can be thought of as positive predictive values (PPV) of the statistically significant finding.

Fourth, we calculated the reverse Bayesian argument values suggested by Matthews, called "Analysis of Credibility."[15,16] This determined whether the observed data provided Bayesian credible evidence for the H1 of a nonzero effect. The range of prior effect sizes (critical prior interval, CPI) were calculated which, when combined with the likelihood (based on the obtained 95% CI of effect size in the study), lead to a posterior range of effect sizes that just excluded no effect at the Bayesian 95% credibility level. 1) For obtained p ≤ .05, the skepticism limit (SL) was calculated; if prior evidence supported (a lower bound of) effect sizes ≥SL, the study provided credible evidence for a nonzero effect (ie, was enough to defeat the skepticism CPI limit). 2) For obtained p > .05, the advocacy limit (AL) was calculated; if prior evidence supported effect sizes lying wholly within the advocacy CPI (from no effect to AL), the study provided credible evidence of a nonzero effect (despite statistical non-significance in the particular study).

Fifth, we determined the implications of the obtained p-values for a replication study.[21,22] 1) The 80% P-interval, ie, the 10^{th} and 90^{th} percentile of expected replication p-values given by a replication study using the same sample size. 2) The probability that a replication study using the same sample size would obtain a p ≤ .05.

Statistics

We presented descriptive results using counts and percentages, median, interquartile range [IQR], and range (minimum to maximum) as appropriate. We explored predictors of p-value category using univariate and multiple variable logistic regressions for each cohort of RCTs. The possible predictors for univariate analyses were pre-specified: field of sepsis (the most common category for Adult-RCTs and Pediatric-RCTs) or respiratory (the most common category for Consecutive-RCTs), mortality as primary outcome, study year 2011 to 2019 (for Adult-RCTs and Pediatric-RCTs), multicenter study or number of centers (>20 for Adult-RCTs, >10 for Pediatric-RCTs), number of patients, mortality in control group (for Adult-RCTs and Pediatric-RCTs), study RR (for studies with p ≤ .10 in Adult-RCTs, p ≤ .20 in Pediatric-RCTs, and with categorical outcome in Consecutive-RCTs), and higher mortality in intervention group (for Adult-RCTs with p ≤ .05). For the Consecutive-RCTs we added study continent of Europe (47.8% of included studies), higher impact journal (NEJM or JAMA), species non-human-animal (NHA), and standardized mean difference (d; in studies with continuous outcomes). In

the multiple regressions variables were included if their p-value in the univariate regression was <.10 (a common yet somewhat arbitrary method used for exploratory analyses). We planned to force into the multiple regressions "multicenter study" (for Pediatric-RCTs and Consecutive-RCTs), and NHA (for Consecutive-RCTs). We considered $p \leq .05$ as statistically suggestive in the multiple regressions.

There were no consistent predictors of p-value category in Adult-RCTs, Pediatric-RCTs, and Consecutive-RCTs [E-Tables 5–7, Supplemental Material 2]. Factors consistently not associated with p-value category included field of study, multicenter or number of centers, number of patients, study year, and for Consecutive-RCTs, high-impact journal, NHA subjects, and observed ES.

Results

The 216 Adult-RCTs and 120 Pediatric-RCTs are described in E-Tables 1 and 2 (Supplemental Material 2). We screened 269 studies in 6 journals, and after exclusions (E- Table 3, Supplemental Material 2) included 90 Consecutive-RCTs (21 [23%] in NHA) described in E- Table 4 (Supplemental Material 2).

Outcome: P-Value Categories

The obtained p-value was ≤.005 in 11/216 (5.1%) of Adult-RCTs, 2/120 (1.7%) of Pediatric-RCTs, and 37/90 (41.1%) of Consecutive-RCTs (that were screened to have a p-value ≤.10) (Table 2). Of RCTs having p-value ≤.05, the proportions that had p-value ≤.005 were 25%, 17%, and 45% respectively.

Outcome: Reverse Bayesian Implications

Reverse Bayesian implications according to Colquhoun were similar for all cohorts of RCTs (Tables 3–6). As expected, the likelihood ratio increased, and the prior Pr(H1) necessary to have a FPR of 5%, minimum FPR, and realistic FPR decreased as the p-value category was more stringent. An obtained p-value .051–.10 did not reflect a "trend" to statistical significance; for example, in Adult-RCTs, the likelihood ratio was 1.8 [1.4, 22], prior Pr(H1) to have a FPR of 5% was 91.3% [89.6, 93.1], the minimum and realistic FPR were 35.7% [31.3, 41.8] and 83.3% [80.4, 86.6] (Table 2). An obtained p-value .05–.0051 had high FPR; for example, in Adult-RCTs, minimum and realistic FPR of 16.7% [11.2, 21.8] and 64.3% [53.2, 71.4] respectively. Only having an obtained p-value ≤.005 had high likelihood ratio and low FPR (eg, in Adult-RCTs the realistic FPR was 7.7% [7.7, 16.0]).

Table 4. Reverse Bayesian Implications of the Obtained p-Values ≤0.10 in Pediatric RCTs in the Field of Critical Care Research.

Study group	Overall (n = 25)	p > 0.05 to 0.10 (n = 13)	p = 0.05 to 0.0051 (n = 10)	p ≤ 0.005 (n = 2)
Reverse Bayesian Argument				
Likelihood Ratio: Pr(data | H1)/Pr(data | Ho)	2.2 [1.4, 7.4] (1.1-59.9)	1.5 [1.3, 1.8] (1.1-2.2)	6.1 [4.9, 16.6] (3.1-24.4)	59.9
Prior Pr(H1) to have FPR 5%	89.6 [72.1, 93.1] (24.1-94.5)	92.6 [91.3, 93.8] (89.6-94.5)	75.9 [53.6, 79.6] (43.8-86.2)	24.1
Minimum FPR (using Prior Pr[H1] of 0.5)	31.3 [12.0, 41.6] (1.6-47.3)	39.8 [35.7, 44.3] (31.3-47.3)	14.4 [5.8, 17.1] (3.9-24.7)	1.6
Realistic FPR (using Prior Pr[H1] of 0.1)	80.4 [55.1, 86.5] (13.1-89.0)	85.6 [83.3, 87.8] (80.4-89.0)	59.9 [34.5, 64.9] (27.0-74.7)	13.1
Berger Bayesian Calculations				
BF Bound [upper bound for Pr(data | H1)/Pr(data | Ho)]	2.2 [1.8, 4.5] (1.6-29.6)	1.8 [1.7, 2.0] (1.6-2.2)	4.0 [3.4, 8.5]] (2.6-12.0)	29.6
PPV: Highest Posterior Pr^U(H1 | data) using Prior Pr(H1) of 0.5.	68.5 [63.8, 82.0] (61.5-96.7)	64.5 [62.7, 66.4] (61.5-68.5)	79.8 [77.5, 89.4] (72.2-92.3)	96.7
PPV: Realistic Posterior Pr^R(H1 | data) using Prior Pr(H1) of 0.1	19.5 [16.4, 33.5] (15.1-76.7)	16.8 [15.8, 18.0] (15.1-19.5)	30.6 [27.7, 48.5] (22.4-57.1)	76.7
P-intervals implications				
10^{th} percentile of p-value prediction interval	0.0002 [<0.0001, 0.0004]	0.0004 [0.0003, 0.0005]	0.0001 [<0.0001, 0.0001]	<0.0001
90^{th} percentile of p-value prediction interval	0.95 [0.62, 0.99] (0.20-0.99)	0.99 [0.99, 0.99] (0.95-0.99)	0.67 [0.44, 0.73] (0.35-0.85)	0.20
Probability a replication study will have p ≤ 0.05	46.8 [40.8, 63.6] (37.6-87.1)	41.7 [39.3, 44.1] (37.6-46.8)	60.9 [57.9, 73.9] (51.4-78.5)	87.1

Values as: median [IQR] (range). BF: Bayes factor; FPR: false positive rate; Ho: null hypothesis; H1: alternative hypothesis; Pr: probability; PPV: positive predictive value. We report values to one decimal place for consistency of presentation; this is not intended to suggest the values can be known with such precision given the small sample sizes.

Table 5. Reverse Bayesian Implications of the Obtained p-Values ≤ 0.10 for Primary Outcomes in Consecutive RCTs Recently Published in the Field of Critical Care Research.

Study group	Overall (n = 90)	p > .05 to .10 (n = 7)	p = .05 to .0051 (n = 46)	p ≤ .005 (n = 37)
Reverse Bayesian Argument				
Likelihood Ratio: Pr(data | H1)/ Pr(data | Ho)	14.1 [5.0, 99.6] (1.1-419.6)	1.9 [1.3, 2.3] (1.1-2.6)	5.5 [3.7, 11.3] (2.8-24.4)	99.6 [85.2, 99.6] (28.6-419.6)
Prior Pr(H1) to have FPR 5%	57.5 [16.0, 79.2] (4.3-94.5)	90.9 [89.4, 93.6] (88.1-94.5)	77.6 [62.7, 83.8] (43.8-87.3)	16.0 [16.0, 18.6] (4.3-39.9)
Minimum FPR (using Prior Pr[H1] of 0.5)	6.7 [1.0, 16.7] (0.2-47.3)	34.4 [30.8, 43.7] (28.0-47.3)	15.4 [8.2, 21.4] (3.9-26.6)	1.0 [1.0, 1.2] (0.2-3.4)
Realistic FPR (using Prior Pr[H1] of 0.1)	39.1 [8.3, 64.3] (2.1-89.0)	82.5 [80.0, 87.5] (77.8-89.0)	62.1 [44.4, 71.0] (27.0-76.6)	8.3 [8.3, 9.8] (2.1-24.0)
Berger Bayesian Calculations				
BF Bound [upper bound for Pr(data | H1)/Pr(data | Ho)]	7.5 [3.5, 53.3] (1.6-399.4)	2.0 [1.7, 2.2] (1.6-2.4)	3.7 [2.9, 6.3] (2.5-12.0)	53.3 [44.5, 53.3] (13.9-399.4)
PPV: Highest Posterior Pr^U(H1 | data) using Prior Pr(H1) of 0.5.	88.2 [77.8, 98.2] (61.5-99.8)	67.0 [62.9, 68.8] (61.5-70.3)	78.8 [74.3, 86.2] (71.1-92.3)	98.2 [97.8, 98.2] (93.3-99.8)
PPV: Realistic Posterior Pr^R(H1 | data) using Prior Pr(H1) of 0.1	45.3 [28.0, 85.5] (15.1-97.8)	18.4 [15.9, 19.7] (15.1-20.8)	29.3 [24.3, 41.0] (21.4-57.1)	85.5 [82.7, 85.5] (60.7-97.8)
P-intervals implications				
10th percentile of p-value prediction interval	0.000013 [0.00000033, 0.000068]	0.00027 [0.00022, 0.00045]	0.000059 [0.000019, 0.00011]	0.00000033 [0.00000033, 0.00000050]
90th percentile of p-value prediction interval	0.47 [0.14, 0.72] (0.04-1.0)	0.98 [0.94, 1.0] (0.90-1.0)	0.70 [0.52, 0.80] (0.35-0.88)	0.14 [0.14, 0.16] (0.04-0.32)
Probability a replication study will have p ≤ 0.05	72.1 [58.3, 90.8] (37.6-97.3)	44.9 [39.6, 47.1] (37.6-49.0)	59.7 [54.1, 69.3] (50.0-78.5)	90.8 [89.6, 90.8] (80.2-97.3)

Values as: median [IQR] (range). BF: Bayes factor; FPR: false positive rate; Ho: null hypothesis; H1: alternative hypothesis; Pr: probability; PPV: positive predictive value. We report values to one decimal place for consistency of presentation; this is not intended to suggest the values can be known with such precision given the small sample sizes.

Reverse Bayesian implications according to Berger et al. were also similar for all cohorts of RCTs (Tables 3–6). As expected, the BF and PPVs (Pr^U[H1 | p] and Pr^R[H1 | p]) increased as the p-value category was more stringent. The BF was not large until the p-value category was ≤.005. The realistic Pr^R(H1 | p) for p-value category of .05–.0051 was 28.0% [24.1, 34.8], 30.6% [27.7, 48.5], 29.3% [24.3, 41.0], and 32.7% [24.1, 43.5] for the Adult-RCTs, Pediatric-RCTs, Consecutive-RCTs primary and secondary outcomes, respectively. The highest Pr^U(H1 | p) for p-value category .05–.0051 was a median of 77.8%, 79.8%, 78.8%, and 81.4% respectively. To have a Pr^U(H1 | p) >90% or Pr^R(H1 | p) >80%, RCTs needed to have obtained a p-value ≤.005.

Reverse Bayesian implications according to Matthews are given in Table 7. The credibility of significance (p ≤ .05) can be challenged by skeptics who believe prior evidence is likely not to exceed effect size SL.[15,16] The credibility of non-significance (p > .05) can be challenged by advocates who accept effect sizes are unlikely to exceed AL.[15,16] Results were similar in all cohorts of RCTs, even more concerning in Pediatric-RCTs and NHA-Consecutive-RCTs (Table 7). Except for p ≤ .005, the credibility of "statistically significant" findings were easy to challenge; for example, in Adult-RCTs with obtained p = .05–.0051 it was unlikely that prior evidence suggested the OR was likely to be >3.83 prior to the study. For p-values >.05 to .10, the credibility to rule out an effect was surprisingly low; for example, in Adult-RCTs if prior evidence could suggest an OR anywhere between 1 to >5000, the credibility was challenged. Once p-value was .11–.20 the AL was much lower, but still could be challenged for half the Adult-RCTs if prior evidence suggested an OR not outside of 1 to 38.

Outcome: Replication Study Implications

As expected, the 90th percentile of replication p-value decreased, and the probability a replication RCT would have p ≤ .05 increased as the p-value category became more stringent. When the obtained p-value was .05–.0051, for Adult-RCTs, Pediatric-RCTs, Consecutive-RCTs for primary and secondary outcome, 10% of replication p-values (ie, the 90th percentile of replication p-values) were expected to be p≥.72 [.60, .81], .67 [.44, .73], .70 [.52, .80], and .63 [.48, .81] respectively. When the obtained p-value was ≤.005, these values were p≥.13 [.13, .24], .20, .14 [.14, .16], and .14 [.14, .20] respectively. The probability that a replication study would obtain a p-value ≤.05 did not approach 90% unless the obtained p-value was ≤.005 (Tables 3–6).

Examples:

Some examples of applying these methods to high-profile individual trials (three with obtained p-value ≤.05, and

Table 6. Reverse Bayesian Implications of the Obtained p-Values for Secondary Outcomes in Consecutive RCTs Recently Published in the Field of Critical Care Research.

Study group	Overall (n = 86)	$p > 0.05$[a] (n = 17)	$p = 0.05$ to 0.0051 (n = 43)	$p \leq 0.005$ (n = 26)
Reverse Bayesian Argument				
Likelihood Ratio: Pr(data | H1)/ Pr(data | Ho)	7.7 [3.1, 47.7] (0.04-419.6)	0.50 [0.13, 1.93] (0.04-2.63)	7.0 [3.6, 12.9] (2.8-24.4)	99.6 [59.9, 99.6] (28.6, 4196)
Prior Pr(H1) to have FPR 5%	71.1 [28.8, 85.9] (4.3-99.8)	97.4 [90.8, 99.3] (87.9-99.8)	73.1 [59.6, 84.1] (43.8-87.1)	16.0 [16.0, 24.1] (4.3-39.9)
Minimum FPR (using Prior Pr[H1] of 0.5)	11.5 [2.1, 24.2] (0.2-96.4)	66.5 [34.6, 88.7] (27.6-96.4)	12.5 [7.2, 21.8] (3.9-26.1)	1.0 [1.0, 1.6] (0.2-3.4)
Realistic FPR (using Prior Pr[H1] of 0.1)	53.8 [16.1, 74.2] (2.1-99.6)	94.7 [82.5, 98.6] (77.4-99.6)	56.3 [41.1, 71.4] (27.0-76.1)	8.3 [8.2, 13.1] (2.1-24.0)
Berger Bayesian Calculations				
BF Bound [upper bound for Pr(data | H1)/Pr(data | Ho)]	4.7 [2.6, 23.2] (1.0-399.4)	1.2 [1.0, 2.0] (1.0-2.4)	4.4 [2.9, 6.9] (2.5-12.0)	53.3 [29.6, 53.3] (13.9-399.4)
PPV: Highest Posterior Pr^U(H1 | data) using Prior Pr(H1) of 0.5.	82.5 [72.5, 95.8] (50.0-99.8)	55.0 [50.1, 67.0] (50.0-70.5)	81.4 [74.1, 87.4] (71.3-92.3)	98.2 [96.7, 98.2] (93.3-99.8)
PPV: Realistic Posterior Pr^R(H1 | data) using Prior Pr(H1) of 0.1	34.3 [22.7, 71.8] (10.0-97.8)	11.9 [10.1, 18.5] (10.0-21.0)	32.7 [24.1, 43.5] (21.7-57.1)	85.5 [76.7, 85.5] (60.7-97.8)
P-intervals implications				
10^{th} percentile of p-value prediction interval	0.000035 [0.0000016, 0.00014]	0.0015 [0.00028, 0.0071]	0.000041 [0.000015, 0.00011]	0.00000033 [0.00000033, 0.00000095]
90^{th} percentile of p-value prediction interval	0.61 [0.24, 0.85] (0.04-1.0)	1.0 [0.98, 1.0] (0.90-1.0)	0.63 [0.48, 0.81] (0.35-0.88)	0.14 [0.14, 0.20] (0.04-0.32)
Probability a replication study will have $p \leq 0.05$	64.3 [51.8, 85.0] (7.0-97.3)	27.8 [14.4, 44.9] (7.0-14.4)	62.9 [53.7, 71.0] (50.3-78.5)	90.8 [87.1, 90.8] (80.2-97.3)

a.14 (16.3%) of p-values were >.10. BF: Bayes factor; FPR: false positive rate; Ho: null hypothesis; H1: alternative hypothesis; Pr: probability; PPV: positive predictive value. We report values to one decimal place for consistency of presentation; this is not intended to suggest the values can be known with such precision given the small sample sizes.

three with obtained p-value >.05) are shown in Supplemental Material 3.

Discussion

We examined three representative cohorts of critical care RCTs in order to demonstrate the meaning of obtained p-values. Our main findings include the following. First, most RCTs did not obtain a $p \leq .005$, not even most studies that obtained $p \leq .05$. Second, the obtained p-value category was not predicted by the field of study, study decade, number of centers, number of patients, or study ES (except for Pediatric-RCTs, which may have suffered from the "winner's curse,"[23] suggesting the p-value distributions were a general phenomenon in the field of critical care research. Third, unless obtained $p \leq .005$, the FPR of a finding was surprisingly high, and the PPV and probability of replication of "statistically significant" findings were surprisingly low. This was even more marked when a realistic prior Pr(H1) was assumed. As one example, in Adult-RCTs with an obtained $p = .05–.0051$, the median prior Pr(H1) necessary to obtain a FPR of 5% was 79.2%, and the minimum and realistic FPR were 16.7% and 64.3%; with an obtained $p \leq .005$ the median realistic FPR was 7.7%. In order to have an upper PPV >90% or realistic PPV >80%, RCTs needed to obtain a $p \leq .005$. Fourth, the Bayesian credibility of statistically significant and non-significant results were easy to challenge, but much less so with obtained $p \leq .005$; this was even more marked for Consecutive-NHA-RCTs. Overall, the findings suggest that most statistically significant and non-significant results from critical care RCTs should be reassessed as they often require implausible prior support and/or are not robustly credible.

Recent suggestions to improve reporting of p-values in published research include the methods we used here. [5] Colquhoun suggested reporting, given the obtained p-value, the prior Pr(H1) necessary to produce a specified desired FPR, and the minimum or realistic FPR; the discussion section of publications could then be used to argue whether the prior determined or used was plausible.[8,9] Berger et al. suggested reporting the answer to two main questions: The Strength of Evidence Question (how strongly does the evidence from the data favor H1 relative to Ho), answered by reporting the BFB; and The More Likely Hypothesis Question (how likely is it that there is truly an effect of the treatment as opposed to no effect), answered by reporting the final posterior Pr(H1 | p), which depends on defending the prior Pr(H1) used in that calculation.[10,11] Held developed a "nomogram for p values" based on Berger's method.[14] Matthews' suggested to report and defend the SL or AL in an Analysis of Credibility.[15,16] Many have suggested to consider $p \leq .005$ as "statistically

Table 7. Bayesian Analysis of Credibility Results for the Critical Care RCTs Cohorts.

Analysis of Credibility of Results	p = .05 to .0051	p ≤ .005	p > .05 to .10	p = .11 to .20
Adult RCTs				
Number of studies	n = 33	n = 11	n = 13	n = 13
Study ARD in mortality (%)	18.0 [9.8, 21.2]	15.0 [6.1, 25.2]	9.0 [8.0, 12.3]	-
Skepticism Limit for Odds or Hazard Ratios	3.83 [2.46, 10.59] (1.25, 44.57)	1.29 [1.22, 1.60] (1.09-5.92)	-	-
Advocacy Limit for Odds or Hazard Ratios	-	-	19 012 [5339, 5.7×10^9] ($21.2\text{-}1.4 \times 10^{16}$)	38.45 [2.02, 69.23] (1.20-37 167)
Pediatric RCTs				
Number of studies	n = 10	n = 2	n = 13	n = 13
Study ARD in mortality (%)	21.1 [12.7, 40.3]	27.4 and 33.3	10.0 [4.2, 23.1]	10.1 [5.4, 20.6]
Skepticism Limit for Odds or Hazard Ratios	7.28 [4.34, 20.34] (3.82-3578.23)	2.39 and 5.31	-	-
Advocacy Limit for Odds or Hazard Ratios	-	-	10^5 [1606.86, $>10^5$] ($2.55\text{->}10^5$)	7406 [1073, 90 725] (21.7-90 725)
Consecutive RCTs: Primary Outcomes[a]				
Number of studies	n = 44	n = 37	n = 7	-
Skepticism Limit for Odds or Hazard Ratios	3.40 [2.08, 12.21] ($1.09 - >10^5$) (n = 31)	1.50 [1.20, 2.38] (1.06-6.25) (n = 19)	-	-
Skepticism Limit for d	1.55 [.92, 2.58] (.20 to 5.09) (n = 13)	.71 [0.19, 1.81] (0.06-7.70) (n = 18)	-	-
Advocacy Limit for Odds or Hazard Ratios	-	-	11 111.0 [160.2, 16 424.2] (41.39-17 487.80) (n = 5)	-
Advocacy Limit for d	-	-	5.09 [4.57, -] (n = 2)	-
Consecutive RCTs: Secondary Outcomes[b]				
Number of studies	n = 43	n = 26	n = 17	-
Skepticism Limit for Odds or Hazard Ratios	2.73 [1.68, 8.56] (1.18-40.66) (n = 20)	1.64 [1.40, 1.92] (1.30-15.37) (n = 13)	-	-
Skepticism Limit for d	1.14 [0.58, 2.60] (0.07-6.74) (n = 23)	0.87 [0.39, 1.50] (0.11-4.68) (n = 13)	-	-
Advocacy Limit for Odds or Hazard Ratios	-	-	9.90 [3.19, 347910] ($1.43 - >10^5$) (n = 14)	-
Advocacy Limit for d	-	-	2.31 [0.60, -] (n = 3)	-

Values as: Median [IQR] (range). ARD: absolute risk difference; d: standardized mean difference.
The credibility of significance (a finding with p ≤ .05) can be challenged by skeptics who believe prior evidence is unlikely to exceed the Skepticism Limit effect size. The credibility of non-significance (a finding with p > .05) can be challenged by advocates who accept effect sizes are unlikely to exceed the Advocacy Limit.
a.For the SL in the Consecutive RCTs, the SL for OR/HR for the 45 human studies was 2.29 [1.55, 4.89] and for the 5 NHA studies was 16.95 [5.99, $>10^5$]; and for d for the 18 human studies was 0.66 [0.18, 1.92] and for the 13 NHA studies was 1.55 [1.09, 3.44]. For AL, the only NHA study had AL for OR 15360.
b.For the SL in the Consecutive RCTs, for the secondary outcome, the SL for OR/HR included only human studies (there were no NHA studies), and for d for the 21 human studies was 0.62 [0.33, 1.65], and for the 15 NHA studies was 1.53 [1.01, 2.60]. For AL, there was 1 NHA study for OR/HR with AL 1083, and 1 NHA study for d with AL 2.74.

significant", and p-values .05 to .0051 as "suggestive" (perhaps warranting further investigation).[19,20] We demonstrated that RCTs that obtained p ≤ .005 had much higher BF and PPV, and much lower FPR, better approaching what is generally considered acceptable evidence for an effect.

Our choice of Prior Pr(H1) may be criticized. The minimum Pr(H1) used was 50%, defensible as reflecting clinical equipoise that justifies blinding and randomization.[26] The realistic Pr(H1) used was 10%, defensible for several reasons. First, ≤10% has been suggested as a realistic estimate of the proportion of interventions tested in a clinical field that prove to be useful.[1,18] Second, systematic reviews of adult and pediatric critical care RCTs consistently find that ≤10% of tested interventions are useful.[17,18,27] Third, reviews of translation from

results in NHA to human RCTs consistently find that <10% of very promising interventions are found useful.[28,29] Fourth, interventions thought to be successful in human critical care RCTs often turn out to have been false positive findings.[17,27,30] Fifth, others calculating reverse Bayesian implications of obtained p-values have similarly used this as a reasonable estimate.[34] Sixth, the findings mentioned above are based on published RCTs, and due to publication bias, it is likely these overestimate Pr(H1). Finally, this was meant to be what we consider a plausible example, and regardless of this choice our main point remains - interpretation of RCT findings requires the attempt to defend a chosen prior Pr(H1) in order to produce a desired FPR or desired posterior Pr(H1 | data). In addition, incremental and replication research (as opposed to the search for novel findings) is a priority in order to provide the evidence for the choice of a higher Pr(H1).

Our results are compatible with studies in the field of critical care RCTs that reported a low fragility-index [defined as the minimum number of reversals in outcome that need to occur for the result to be no longer statistically significant] was common.[35] The fragility-index has an almost perfect negative correlation with the obtained p-value, and can be said to simply be "repackaging of the p-value".[39(p348)] Arguably, explicitly stating the reverse Bayesian implications of obtained p-values is a better way to demonstrate the fragility of findings in the research field.

This study has limitations. First, we did not consider study biases that can inflate the obtained p-value. Potential sources of bias include anything that produces greater flexibility in study designs, definitions, outcomes, and analytic modes, eg, data dredging (multiple testing, p-hacking), changed data preprocessing parameters, and changed statistical analytical methods or outcomes, sometimes influenced by financial and other interests and prejudices.[1,40] Biases may have been fewer in the multicenter mortality outcome Adult-RCTs that still demonstrated our main findings. Since bias would further reduce the Prior Pr(H1), our results can be considered a conservative minimum estimate. Second, we did not include non-RCTs, nor non-mortality outcomes for the Adult-RCTs and Pediatric-RCTs. Non-RCTs and often less objective non-mortality outcomes provide more flexibility for bias to influence results.[40,41,44,45] Non-mortality outcomes in the Consecutive-RCTs were included, with similar results. This suggests that our results can again be considered a conservative minimum estimate. Third, our results only reflect already described mathematical implications of p-values. This was our aim: we demonstrated the necessarily mathematical implications of p-values obtained in representative cohorts of RCTs, finding that $p \leq .05$ is a poor standard to adjudicate the "statistical significance" of findings. As many critical care clinicians may not be familiar with these reverse Bayesian implications of p-values and confidence intervals, we hope that demonstration using real-world RCTs can make the need for revised research standards clearer, leading to necessary change in research methods and reporting. In addition, the methods we used can easily be applied to any individual RCT of interest to a clinician (see detailed description in Supplemental Material 1) in order to determine the credibility of the findings. Specifically, the clinician should focus on whether the prior Pr(H1) necessary to produce a desired FPR is defensible [using Colquhoun's online calculator], whether the PPV of a finding is based on a realistic prior Pr(H1) [using our online calculator based on Berger's BFB], whether the final posterior Pr(H1 | p) that would convince is based on a defensible prior Pr(H1) [using Held's published nomogram], and/or whether the SL (or AL for non-significant findings) is credible [using our online calculator based on Matthews' formulas]. Based on our findings, we believe that results in most of these RCTs will be found to be less certain than clinicians believe. Examples of application to a few high-profile individual RCTs is provided in Supplemental Material 3.

This study has several strengths. First, we included three cohorts of RCTs in critical care to demonstrate generalizability. Second, we used a detailed case report form manual for data recording and calculations. Third, we used methods suggested to demonstrate the implications of obtained p-values. To our knowledge, this is the first study in the critical care research field to report detailed reverse Bayesian implications of p-values.

Conclusions

We examined three representative cohorts of critical care RCTs in order to demonstrate the often-misinterpreted meaning of obtained p-values. We suggest that to improve interpretation of obtained p-values in RCTs reverse Bayesian implications should be reported in the results section, argument to support the assumed prior Pr(H1) should be a focus of the discussion section, and an obtained p-value $\leq .005$ should be used to claim statistical significance.

Author Contributions

SN and ARJ contributed to conception and design of the work, acquisition, analysis and interpretation of the data, and substantial critical revisions of the manuscript for important intellectual content, have approved the submitted version, and have participated sufficiently in the work to take public responsibility for the content. ARJ wrote the first draft of the article.

Financial Disclosure

This work was supported by a Department of Pediatrics Resident Research Grant awarded to SN. The funding agency had no role in design and conduct of the study; collection, analysis or interpretation of the data; preparation, writing, review, or approval of the manuscript; or the decision to submit the manuscript for publication.

Availability of Data and Materials

The dataset used and/or analyzed during the current study is available at the following: Joffe, A. R. (2021, September 29). Reverse Bayesian Implications of p-values. Retrieved from osf.io/zmjya.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the University of Alberta, Department of Pediatrics (grant number Resident Research Grant).

Ethical Approval

Not applicable, because this article does not contain any studies with human or animal subjects.

ORCID iD

Ari R. Joffe https://orcid.org/0000-0002-4583-707X

Supplemental material

Supplemental material for this article is available online.

References

1. Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.
2. Goodman SN. Toward evidence-based medical statistics. 1: the p value fallacy. *Ann Intern Med*. 1999;130(12):995-1004.
3. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P-values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31:337-350.
4. American Statistical Association Board of Directors. ASA Statement on statistical significance and p—values. *Am Stat.* 2016;70(2):131-133.
5. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p<.05". *Am Stat.* 2019;73(Suppl1):1-19.
6. Price R, Bethune R, Massey L. Problem with p values: why p values do not tell you if your treatment is likely to work. *Postgrad Med J*. 2020;96(1131):1–3.
7. Goodman SN. Toward evidence-based medical statistics. 2: the Bayes factor. *Ann Intern Med*. 1999;130(12):1005-1013.
8. Colquhoun D. The false positive risk: a proposal concerning what to do about p-values. *Am Stat.* 2019;73(Suppl1):192-201.
9. Colquhoun D. The reproducibility of research and the misinterpretation of p-values. *R Soc Open Sci*. 2017;4:171085.
10. Benjamin DJ, Berger JO. Three recommendations for improving the use of p-values. *Am Stat.* 2019;73(S1):186-191.
11. Sellke T, Bayarri MJ, Berger JO. Calibration of p values for testing precise null hypotheses. *Am Stat*. 2001;55(1):62-71.
12. Held L, Matthews R, Ott M, Pawel S. Reverse-Bayes methods for evidence assessment and research synthesis. arXiv.org 2021 (preprint). DOI: 2102.13443.v2. Available at: https://arxiv.org/abs/2102.13443 (accessed July 30, 2021).
13. Allmark P. Bayes and health care research. *Med Health Care Philos*. 2004;7(3):321-332.
14. Held L. A nomogram for P values. *BMC Med Res Methodol*. 2010;10:21.
15. Matthews RAJ. Beyond 'significance': principles and practice of the analysis of credibility. *R Soc Open ci*. 2018;5:171047.
16. Matthews RAJ. Moving towards the post p<.05 era via the analysis of credibility. *Am Stat.* 2019;73(Suppl1):202-212.
17. Santacruz CA, Pereira AJ, Celis E, Vincent JL. Which multicenter randomized controlled trials in critical care medicine have shown reduced mortality? A systematic review. *Crit Care Med*. 2019;47(12):1680-1691.
18. Duffett M, Choong K, Hartling L, Menon K, Thabane L, Cook DJ. Randomized controlled trials in pediatric critical care: a scoping review. *Critical Care*. 2013;17(5):R256.
19. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. *Nat Hum Behav*. 2018;2(1):6-10.
20. Ioannidis JPA. The proposal to lower P-value thresholds to .005. *JAMA*. 2018;319(4):1429-1430.
21. Lazzeroni LC, Lu Y, Belitsakaya-Levy I. Solutions for quantifying p-value uncertainty and replication power. *Nat Methods*. 2016;13(2):107-108.
22. Cumming G. Replication and p intervals. P values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*. 2018;3(4):286-300.
23. Young NS, Ioannidis JPA, Al-Ubaydii O. Why current publication practices may distort science. *PLoS Med*. 2008;5(10):e201.
24. Altman N, Krzywinski M. Interpreting P values. *Nat Methods*. 2017;14(3):213-214.
25. Halsey LG. The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? *Biol Lett*. 2019;15:20190174.
26. Johnson N, Lilford RJ, Brazier W. At what level of collective equipoise does a clinical trial become ethical? *J Med Ethics*. 1991;17(1):30-34.
27. Abrams D, Montesi SB, Moore SKL, et al. Powering bias and clinically important treatment effects in randomized trials of critical illness. *Crit Care Med*. 2020;48(12):1710-1719.
28. Joffe AR, Bara M, Anton N, Nobis N. Expectations for the methodology and translation of animal research: a survey of the general public, medical students and animal researchers in North America. *Altern Lab Anim*. 2016;44(4):361-381.
29. Pippin JJ. Animal research in medical sciences: seeking a convergence of science, medicine, and animal law. *S Tex L Rev*. 2013;54:469.
30. Ranieri VM, Thompson BT, Barie PS, et al. Williams MD, for the PROWESS-SHOCK study group. *Drotrecogin alfa (activated) in adults with septic shock. NEJM*. 2012;366(22):2055-2064.
31. National Heart, Lung, and Blood Institute PETAL Clinical Trials Network; Moss M, Huang DT, Brower RG, et al. Early neuromuscular blockade in the acute respiratory distress syndrome. *NEJM*. 2019;380(21):1997-2008.
32. Mouncey PR, Osborn TM, Power GS, et al. for the ProMISe Trial Investigators. Trial of early, goal-directed resuscitation for septic shock. *NEJM*. 2015;372(14):1301-1311.
33. The NICE-SUGAR Study Investigators. Intensive versus conventional glucose control in critically ill patients. *NEJM*. 2009;360(13):1283-1297.
34. Held L. Reverse-Bayes analysis of two common misinterpretations of significance tests. *Clinical Trials*. 2013;10(2):236-242.
35. Ridgeon EE, Young PJ, Bellomo R, Muchetti M, Lembo R, Landoni G. The fragility index in multicenter randomized controlled critical care trials. *Crit Care Med*. 2016;44(7):1278-1284.
36. Grolleau F, Collins GS, Smarandache A, et al. The fragility and reliability of conclusions of anesthesia and critical care randomized trials with statistically significant findings: a systematic review. *Crit Care Med*. 2019;47(3):456-462.

37. Vargas M, Buonano P, Marra A, Iacovazzo C, Servillo G. Fragility index in multicenter randomized controlled trials in critical care medicine that have shown reduced mortality. *Crit Care Med*. 2020;48(3):e250-e251.
38. Matics TJ, Khan N, Jani P, Kane JM. The fragility of statistically significant findings in pediatric critical care randomized controlled trials. *Pediatr Crit Care Med*. 2019;20(6):e258-e262.
39. Carter RE, McKie PM, Storlie CB. The fragility index: a P-value in sheep's Clothing? *Eur Heart J*. 2017;38(5):346-348.
40. Forstmeier W, Wagenmakers E-J, Parker TH. Detecting and avoiding likely false-positive findings - a practical guide. *Biol Rev*. 2017;92(4):1941-1968.
41. Munafo MR, Nosek BA, Bishop DVM, et al. A manifesto for reproducible science. *Nat Hum Behav*. 2017;1:0021.
42. Higginson AD, Munafo MR. Current incentives for scientists lead to underpowered studies with erroneous conclusions. *PLoS Biol*. 2016;14(11):e2000995.
43. Smaldino PE, McElreath R. The natural selection of bad science. *R Soc Open Sci*. 2016;3:160384.
44. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psych Sci*. 2011;22(11):1359-1366.
45. Szucs D. A tutorial on hunting statistical significance by chasing n. *Front Psychol*. 2016;7:1444.