VIEWPOINT

Using Bayesian Hypothesis-testing to Reanalyze Randomized Controlled Trials: Does it Always Tell the Truth, the Whole Truth and Nothing but the Truth?

Kwok Ming Ho[1], Anna Lee[2]

Abstract

Adequately powered randomized controlled trials (RCTs) are considered the highest level of evidence in guiding clinical practice. Reports using Bayesian hypothesis-testing to reanalyze RCTs are increasing. One distinct advantage of Bayesian analysis is that we can obtain a range of numerical probabilities that reflect how likely a study intervention is more effective than the alternative after considering both pre-existing available evidence and the alternate hypotheses. A recent analysis of critical care trials showed that some trials with an indeterminate result according to the frequentist analysis could have a high probability of being effective when reinterpreted by Bayesian analysis. In this perspective article, we will discuss the caveats in interpreting the results of Bayesian reanalysis of RCTs before we change clinical practice. When overoptimistic hypothesis prior probabilities are used, it carries a risk to translate noises into false signals. Using Bayes factors (BFs) to quantify evidence contained in data (by the ratio of the probability of data under each hypothesis) is thus more preferable than using a single prior probability, such that the BF approach becomes the mainstream in Bayesian hypothesis-testing. Still, BFs are dependent on the prior parameter distributions; comparing different hypotheses would invariably result in different results.

Keywords: Bayes factor, effectiveness, evidence-based medicine, likelihood ratio, trials.

Indian Journal of Critical Care Medicine (2024): 10.5005/jp-journals-10071-24833

Highlights

- Reports using Bayesian hypothesis-testing to reanalyze randomized controlled trials (RCTs) are increasing.
- The Bayes factor (BF) approach is the mainstream when using Bayesian hypothesis-testing to reanalyze results of RCTs.
- Bayes factors are dependent on the prior parameter distributions and the associated assigned weight or variance, such that comparing different hypotheses would invariably result in different results that may confound the interpretation of the results.

Introduction

Randomized controlled trials have been considered the gold standard in guiding clinical practice in the paradigm of evidence-based medicine. In superiority RCTs, confirmation of effectiveness using a frequentist approach relies on rejecting the null hypothesis. The *p*-value generated from a clinical trial—defined as assuming the null thesis is true, what's the probability of obtaining the trial's data as extreme or more extreme than observed (in a similar trial in the future)—is counterintuitive; and as such, dichotomizing the p-value into significant or insignificant becomes a default position in handling the results of RCTs.[1]

Reports using Bayesian analysis to reinterpret RCTs, in particular those with results that were considered indeterminate according to the frequentist analysis, are growing.[2–5] This isn't surprising because humans frequently don't make intuitive decisions following the frequentist approach when we face uncertainties.[6] Indeed, educationists have long supported that using a Bayesian approach will optimize clinicians' clinical logic when they face uncertainties.[7,8] One distinct advantage of Bayesian analysis is that we can gather a range of numerical chances that reflect how likely the study intervention is more effective than null at a population level after taking into account what pre-existing information we have and what the alternate hypotheses are.

A recent analysis of critical care trials showed that some trials with an indeterminate result could have a high probability of being effective when reinterpreted by Bayesian reanalysis, and the conversion rate was heavily dependent on whether previous belief about the study intervention's effectiveness was skeptical, uninformative, or enthusiastic.[9] As some clinicians and researchers aren't familiar with Bayesian analysis, there's an implicit risk for some to believe that Bayesian analysis could reveal the fact that an intervention is truly effective beyond reasonable doubt that has not been uncovered by a frequentist analysis. While we agree that Bayesian analysis is an extremely useful tool in decision-making it

[1,2]Department of Anaesthesia and Intensive Care, The Chinese University of Hong Kong, Hong Kong SAR, Hong Kong

Corresponding Author: Kwok Ming Ho, Department of Anaesthesia and Intensive Care, The Chinese University of Hong Kong, Hong Kong SAR, Hong Kong, Phone: +852 64631926, e-mail: kmho@cuhk.edu.hk

How to cite this article: Ho KM, Lee A. Using Bayesian Hypothesis-testing to Reanalyze Randomized Controlled Trials: Does it Always Tell the Truth, the Whole Truth and Nothing but the Truth? Indian J Crit Care Med 2024;28(11):1005–1008.

Source of support: Nil

Conflict of interest: KMH is an Intensive Care Editor for anaesthesia and intensive care. AL is a Content Editor for cochrane anaesthesia and cochrane emergency and critical care review groups.

has its own limitations. In this perspective article, we will discuss some caveats that should be thoroughly considered before we use the results of Bayesian reanalysis of RCTs as a foundation to change clinical practice.

Assuming a trial is impeccably designed and has included the use of a clinically plausible effect size to adequately power the study,[10] a frequentist analysis should be instrumental in assessing whether the study intervention is effective or not at a population leve. Of course, we admit that the chances of type I and II errors are frequently set at 5 and 20%, respectively,[11] and that we need to appreciate the fact that absence of substantiation of efficacy is not the same as substantiation of absence. Thus, there's an argument to use Bayesian approac to reanalyze the data because, by doing so, we can put together current with former information, which would improve precision of the posterior distribution.[12] Note that Bayesian procedures can either strengthen or weaken a null hypothesis,[13] not dissimilar to doing a meta-analysis by combining the results of a new trial with all previous similar trials. There are two caveats in conducting Bayesian reanalysis of RCTs, however.

Firstly, both the magnitude and validity of the results of Bayesian reanalysis can be heavily influenced by the previous information (or the prior) in a multiplicative way. In Equation 1, for ease of demonstration, we've employed the likelihood rate (LR) approach to explain this issue. LR is an estimate grounded on comparing two hypotheses – the null and alternate thesis. Mathematically, LR is the ratio of the chances of the observed result under the distinct alternate and null hypotheses.[14,15] It's easy to interpret because RCTs would have generated an alternate hypothesis prospectively in their power computation. A more complicated but superior approach than the LR approach is to use BF. Bayes factor involves integrating the probability of the data across possible priors after specifying their weights (and a smaller variance is associated with a greater weight).[8]

Materials and Methods

When we use LR to estimate post-trial probability – as suggested by some authors[14] – a second caveat is that *how the alternate hypothesis is formulated* can affect the LR of the new trial. Extending Equations 1 and 2 mathematically will help us to calculate LR in illustrating how the alternate hypothesis can affect post-trial odds and post-trial probability (assuming relative risk or hazard ratio outcome data and alternate hypothesis of the trial are available).[14] Note that post-trial probability is equal to post-trial odds/(1 + post-trial odds).

$$\boxed{\begin{array}{c}\textbf{Post-trial odds}\\\text{or updated}\\\text{information}\\\text{(or the}\\\text{posterior}\\\text{probability}\\\text{distribution)}\end{array}} = \boxed{\begin{array}{c}\textbf{Pretrial odds}\\\text{or prior}\\\text{information/}\\\text{belief (or}\\\text{prior}\\\text{probability}\\\text{distribution)}\end{array}} \times \boxed{\begin{array}{c}\textbf{Likelihood}\\\textbf{function}\\\text{(based on new}\\\text{trial's results}\\\text{and}\\\text{its alternative}\\\text{hypothesis)}\end{array}}$$

Equation 1: Post-trial odds is related to pre-trial odds and likelihood ratio of the results of the new trial in a multiplicative fashion

$$\boxed{\text{Post-trial odds}} = \boxed{\text{Pretrial odds}} \times \boxed{e^{\wedge\,[Z*A\,-\,(A2/2)]}}$$

Equation 2: Likelihood ratio can be mathematically calculated

NB: CI, confidence interval. LR, likelihood ratio. Ln, natural logarithm function.

LR is estimated as follows:

Z = Ln (Relative risk $_{\text{observed}}$)/[(width of the natural logarithm of 95% CI bound of the observed results)/(2 × 1.96)]

A = Ln (Relative risk $_{\text{proposed in the alternate hypothesis in sample size calculation}}$/[(width of the natural logarithm of the 95%CI bound of the observed results)/(2 × 1.96)]

$\text{Ln (LR)} = Z*A - (A^2/2) \rightarrow \text{LR} = e^{\wedge\, Z*A - (A2/2)}$

Results

Let's illustrate this by using data from a hypothetical superiority RCT with borderline significance according to the frequentist analysis (75 events/200 patients in the intervention group vs 91 events/200 patients in the placebo group, relative risk 0.82 (95% confidence interval: 0.65–1.04; $p = 0.106$) (Table 1).

In scenarios 1–3 when the prior is uninformative (e.g. with insufficient prior information and hence no better than tossing the coin in determining whether the study intervention is better

Table 1: Hypothetical examples illustrating how the pre-trial odds as well as the alternate hypothesis (used for power calculation) can substantially affect post-trial odds and post-trial probabilities, exemplified by the vastly different post-trial probabilities in scenarios 6 and 7 that the intervention is effective, despite having the same observed new trial results (with a relative risk of 0.82, 95% confidence interval of 0.65–1.04; $p = 0.106$), based on the likelihood ratio equation used in reference 14

Scenario	*Pre-trial probability of being effective*	*Pre-trial odds*	*Relative risk for the alternate hypothesis*	*Likelihood ratio of the new trial*	*Post-trial odds*	*Bayesian post-trial probability of being effective*
1	50%	1 (uninformative prior)	0.80	3.854	3.854	79.4%
2	50%	1 (uninformative prior)	0.70	1.646	1.646	62.2%
3	50%	1 (uninformative prior)	0.65	0.535	0.535	34.9%
4	40%	0.67 (pessimistic prior)	0.80	3.854	2.582	72.1%
5	40%	0.67 (pessimistic prior)	0.70	1.646	1.103	52.5%
6	40%	0.67 (pessimistic prior)	0.65	0.535	0.358	26.4%
7	80%	4 (optimistic prior)	0.80	3.854	15.416	93.9%
8	80%	4 (optimistic prior)	0.70	1.646	6.584	86.8%
9	80%	4 (optimistic prior)	0.65	0.535	2.140	68.2%

An online calculator at https://cjdbarlow.au/pages/bf.html can also be used to generate the Bayes factor and assess how different pre-trial probabilities affect the post-trial probabilities by entering the trial's raw outcome data. NB, pre-trial odds = pre-trial probability/(1-pre-trial probability); post-trial probability, post-trial odd/(1 + post-trial odds) than the placebo), the post-trial probability from the Bayesian analysis will theoretically not be different from or no better than a frequentist analysis; the post-trial probabilities do vary substantially from 34.9 to 79.4% dependent on what alternate hypothesis is used to power the study. Having a pessimistic prior (as in scenarios 4–6) decreases the post-trial probabilities (to between 26.4 and 72.1%), and the extent of this effect will again depend on what alternate hypothesis is used to power the study. Conversely, having an optimistic prior (as in scenarios 7–9) would substantially increase the post-trial probabilities (to between 68.2 and 93.9%). The alternate hypothesis used to power the study will still affect the post-trial probabilities, but only to a smaller extent compared to when uninformative and pessimistic priors are used. Having an unrealistic alternate hypothesis (of large effect size) would generate a lower post-trial probability in a Bayesian analysis using a LR approach.

Discussion

The two caveats in interpreting Bayesian reanalysis of a RCT are quite obvious in this hypothetical exercise. Higher prior odds lead to higher posterior odds. The other is related to the intrinsic nature of Bayesian hypothesis testing: comparing different hypotheses leads to different "evidence". Using BFs to incorporate a model of the prior distributions with Markov chain Monte Carlo (MCMC) convergence, we can compare the null with the range of possible alternate hypotheses (instead of a single LR or alternate hypothesis) to generate full posterior probabilities.[2–4] This can partially bypass the limitation of choosing an inappropriate (single) alternate hypothesis when the LR approach is used. Still, the BF is affected by the weighting (or variances) assigned to different prior parameter distributions. As such, sensitivity analyses to assess the robustness of the results will be appropriate.[8,15] Credible interval is an important estimate of Bayesian analysis and is generated from the posterior distribution, which can similarly be affected by the accuracy and magnitude of the prior[16] and, despite its name, should also be interpreted with caution.

Finally, determining when a posterior probability is sufficiently high to call an intervention as "effective" (or credible, or outside the Region of Practical Equivalence) in a Bayesian analysis requires careful consideration of the nature of the intervention itself, beyond the numerical probability of the intervention being more effective than placebo.[13,17] As a minimum, some patient-centered implications, such as patient preferences, whether a study intervention is expensive, and its effectiveness exceeding the minimal clinically important (or meaningful) difference without potential serious harms, should all be thoroughly evaluated.

In conclusion, generating a numerical probability to reflect the chance a study intervention is more effective than its alternative in a RCT after taking pre-existing evidence into account is appealing and may help us overcome the temptation to interpret the *p* value as a dichotomized outcome. However, both the validity and optimism of the chosen priors have a direct multiplicative impact on the accuracy and magnitude of the posterior (or post-trial) probabilities, respectively. Using BFs has some distinct advantages over the LR approach and should be used whenever possible.

In our conservative view, a reasonable high post-trial probability (e.g., >90%) in the Bayesian reanalysis of a RCT with indeterminate frequentist analysis results should, at best, be considered hypothesis-generating similar to a meta-analysis, instead of interpreted as being conclusive to change clinical practice. This is particularly important if the study interventions are expensive and/or potentially harmful, such as Extracorporeal Membrane Oxygenation.[4] Clinical decisions should not be solely determined by posterior distributions but also, importantly, loss functions, utilities, and thresholds. Designing a study *up front* under a Bayesian framework by pre-specifying all analytic parameters is more preferable than conducting *post-hoc* Bayesian reanalysis, particularly for phase I, II, and dose optimization trials (https://www.trialdesign.org/#newsSection). With the scarcity of health resources, 'Choosing Wisely' is important, and this requires critical thinking and careful consideration of research data in a holistic fashion.

As in most detective stories, "the true culprit is usually someone who has a very small prior probability of being guilty".[17] Perhaps this may also apply to major scientific discoveries. As always, the best way to find the culprit in a crime (or discover effective interventions in medical science) is to conduct high-quality investigations (or well-designed, adequately powered trials). A carefully generated low BF (or LR) will help to eliminate the impossible in favor of a null hypothesis. Conversely, a high BF (or LR) will help to reallocate the suspicion to the true culprit in a crime (or confirm a truly effective intervention). For forensic evidence alone to be conclusive in a criminal trial, a BF of at least 1000 is recommended.[18] In this sense, the quality of our investigation—in either medical research or criminal investigation—remains paramount in determining whether we can find the truth beyond doubt. Blaming either the frequentist or Bayesian would appear to have missed the point.[19]

Contributions by the Authors

All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by KMH. The first draft of the manuscript was written by KMH, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Orcid

Kwok Ming Ho https://orcid.org/0000-0002-6705-6004
Anna Lee https://orcid.org/0000-0003-2864-0045

References

1. Pocock SJ, Ware JH. Translating statistical findings into plain English. Lancet 2009;373(9679):1926–1928. DOI: 10.1016/S0140-6736(09)60499-2.
2. Harhay MO, Blette BS, Granholm A, Moler FW, Zampieri FG, Goligher EC, et al. A Bayesian interpretation of a pediatric cardiac arrest trial (THAPCA-OH). NEJM Evid 2023;2(1):EVIDoa2200196. DOI: 10.1056/EVIDoa2200196.
3. Zampieri FG, Damiani LP, Bakker J, Ospina-Tascón GA, Castro R, Cavalcanti AB, et al. Effects of a resuscitation strategy targeting peripheral perfusion status versus serum lactate levels among patients with septic shock. A Bayesian reanalysis of the ANDROMEDA-SHOCK trial. Am J Respir Crit Care Med 2020;201(4):423–429. DOI: 10.1164/rccm.201905-0968OC.
4. Goligher EC, Tomlinson G, Hajage D, Wijeysundera DN, Fan E, Jüni P, et al. Extracorporeal membrane oxygenation for severe acute respiratory distress syndrome and posterior probability of mortality benefit in a post hoc Bayesian analysis of a randomized clinical trial. JAMA 2018;320(21):2251–2259. DOI: 10.1001/jama.2018.14276.
5. Lammers D, Richman J, Holcomb JB, Jansen JO. Use of Bayesian statistics to reanalyze data from the pragmatic randomized optimal platelet and plasma ratios trial. JAMA Netw Open 2023;6(2):e230421. DOI: 10.1001/jamanetworkopen.2023.0421.

6. Ma WJ, Kording KP, Goldreich D. Bayesian Models of Perception and Action: An Introduction. MIT press. 2023; chapter 2 Pages: 43–56. Available from: https://mitpress.mit.edu/9780262047593/bayesian-models-of-perception-and-action/. ISBN: 9780262047593.
7. Bours MJ. Bayes' rule in diagnosis. J Clin Epidemiol 2021;131:158–160. DOI: 10.1016/j.jclinepi.2020.12.021.
8. Sidebotham D, Barlow CJ, Martin J, Jones PM. Interpreting frequentist hypothesis tests: Insights from Bayesian inference. Can J Anaesth 2023;70(10):1560–1575. DOI: 10.1007/s12630-023-02557-5.
9. Yarnell CJ, Abrams D, Baldwin MR, Brodie D, Fan E, Ferguson ND, et al. Clinical trials in critical care: Can a Bayesian approach enhance clinical and scientific decision making? Lancet Respir Med 2021;9(2):207–216. DOI: 10.1016/S2213-2600(20)30471-9.
10. Gibbs NM, Weightman WM. Beta errors in anaesthesia randomized controlled trials in which no statistical significance is found: Is there an elephant in the room? Anaesth Intensive Care 2022;50(3):153–158. DOI: 10.1177/0310057X221086590.
11. Ranganathan P, Cs P. An introduction to statistics: Understanding hypothesis testing and statistical errors. Indian J Crit Care Med 2019;23(Suppl 3):S230–S231. DOI: 10.5005/jp-journals-10071-23259.
12. Edwards W, Lindman H, Savage LJ. Bayesian Statistical Inference for Psychological Research. 1963;70(3):531-578; In Samuel Kotz & Norman L. Johnson (Eds.), Breakthroughs in Statistics, Volume 1 Foundations and Basic Theory. Springer Series in Statistics. Available from: https://psycnet.apa.org./record/1964-00040-001.
13. Ferguson J. Bayesian interpretation of p values in clinical trials. BMJ Evid Based Med 2022;27(5):313–316. DOI: 10.1136/bmjebm-2020-111603.
14. Perneger TV. How to use likelihood ratios to interpret evidence from randomized trials. J Clin Epidemiol 2021;136:235–242. DOI: 10.1016/j.jclinepi.2021.04.010.
15. Johnson VE, Pramanik S, Shudde R. Bayes factor functions for reporting outcomes of hypothesis tests. Proc Natl Acad Sci U S A 2023;120(8):e2217331120. DOI: 10.1073/pnas.2217331120.
16. Kruschke JK. Bayesian analysis reporting guidelines. Nat Hum Behav 2021;5(10):1282–1291. DOI: 10.1038/s41562-021-01177-7.
17. Kruschke JK, Liddell TM. Bayesian data analysis for newcomers. Psychon Bull Rev 2018;25(1):155–177. DOI: 10.3758/s13423-017-1272-1.
18. Wei Z, Yang A, Rocha L, Miranda MF, Nathoo FS. A review of Bayesian hypothesis testing and its practical implementations. Entropy (Basel) 2022;24(2):161. DOI: 10.3390/e24020161.
19. Udani S. A good workman never blames his tools: Appropriate use of severity of illness scoring systems determines their utility! Indian J Crit Care Med 2020;24(8):628–629. DOI: 10.5005/jp-journals-10071-23545.