Using Bayesian Methods to Augment the Interpretation of Critical Care Trials

An Overview of Theory and Example Reanalysis of the Alveolar Recruitment for Acute Respiratory Distress Syndrome Trial

Fernando G. Zampieri[1,2*], Jonathan D. Casey[3*], Manu Shankar-Hari[4,5], Frank E. Harrell, Jr.[6], and Michael O. Harhay[7,8,9]

[1]Research Institute, HCor-Hospital do Coração, São Paulo, Brazil; [2]Center for Epidemiological Research, Southern Denmark University, Odense, Denmark; [3]Division of Allergy, Pulmonary, and Critical Care Medicine and [6]Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, Tennessee; [4]Guy's and St. Thomas' NHS Foundation Trust, ICU Support Offices, St. Thomas' Hospital, London, United Kingdom; [5]School of Immunology & Microbial Sciences, King's College London, London, United Kingdom; and [7]PAIR (Palliative and Advanced Illness Research) Center Clinical Trials Methods and Outcomes Lab, [8]Department of Biostatistics, Epidemiology, and Informatics, and [9]Division of Pulmonary and Critical Care, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania

ORCID IDs: 0000-0001-9315-6386 (F.G.Z.); 0000-0002-0977-290X (J.D.C.); 0000-0002-5338-2538 (M.S.-H.); 0000-0002-0553-674X (M.O.H.).

Abstract

Most randomized trials are designed and analyzed using frequentist statistical approaches such as null hypothesis testing and *P* values. Conceptually, *P* values are cumbersome to understand, as they provide evidence of data incompatibility with a null hypothesis (e.g., no clinical benefit) and not direct evidence of the alternative hypothesis (e.g., clinical benefit). This counterintuitive framework may contribute to the misinterpretation that the absence of evidence is equal to evidence of absence and may cause the discounting of potentially informative data. Bayesian methods provide an alternative, probabilistic interpretation of data. The reanalysis of completed trials using Bayesian methods is becoming increasingly common, particularly for trials with effect estimates that appear clinically significant despite *P* values above the traditional threshold of 0.05. Statistical inference using Bayesian methods produces a distribution of effect sizes that would be compatible with observed trial data, interpreted in the context of prior assumptions about an intervention (called "priors"). These priors are chosen by investigators to reflect existing beliefs and past empirical evidence regarding the effect of an intervention. By calculating the likelihood of clinical benefit, a Bayesian reanalysis can augment the interpretation of a trial. However, if priors are not defined *a priori*, there is a legitimate concern that priors could be constructed in a manner that produces biased results. Therefore, some standardization of priors for Bayesian reanalysis of clinical trials may be desirable for the critical care community. In this Critical Care Perspective, we discuss both frequentist and Bayesian approaches to clinical trial analysis, introduce a framework that researchers can use to select priors for a Bayesian reanalysis, and demonstrate how to apply our proposal by conducting a novel Bayesian trial reanalysis.

Keywords: Bayesian; randomized trials; critical care; statistical significance; *P* value

(Received in original form June 16, 2020; accepted in final form December 3, 2020)

*Co–first authors.

M.O.H. was supported by the NHLBI of the NIH (R00HL141678). J.D.C. was supported by the NIH/NHLBI (K12HL133117 and K23HL153584). M.S.-H. is funded by the National Institute for Health Research Clinician Scientist Award (CS-2016-16-011). F.E.H. was supported by a Clinical and Translational Science Award (UL1 TR002243) from the National Center for Advancing Translational Sciences. The views expressed in this publication are those of the authors and not necessarily those of the National Health Service, the National Institute for Health Research, the NIH, the National Center for Advancing Translational Sciences, or the Department of Health and Social Care.

Author Contributions: Conception and design: All authors. Analysis and interpretation: All authors. Drafting the manuscript for important intellectual content: All authors.

Correspondence and requests for reprints should be addressed to Michael O. Harhay, Ph.D., Perelman School of Medicine, University of Pennsylvania, 304 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021. E-mail: mharhay@pennmedicine.upenn.edu.

This article has an online supplement, which is accessible from this issue's table of contents at www.atsjournals.org.

Am J Respir Crit Care Med Vol 203, Iss 5, pp 543–552, Mar 1, 2021

Originally Published in Press as DOI: 10.1164/rccm.202006-2381CP on December 3, 2020
Internet address: www.atsjournals.org

Randomized clinical trials (RCTs) are designed to estimate the impact of an intervention on selected outcomes. The design, analysis, and interpretation of RCTs have increasingly come under scrutiny, both within the critical care community and in medicine more broadly. The primary controversy stems from confusion and frustration regarding the meaning (and thus the interpretation) of *P* values and null hypothesis testing, which are statistical concepts employed to design and analyze trials in an approach to statistics known as frequentism (1). One consequence of this debate regarding RCTs is a surging interest in Bayesian statistical thinking, which can be used to design, analyze, and interpret new clinical trials or to reanalyze trials reported using frequentist methods in an attempt to contextualize the results (2–8).

This Critical Care Perspective focuses on the use of Bayesian methods to augment the interpretation of critical care trial results. Bayesian methods provide a probabilistic interpretation of data that encompasses both the trial data and preexisting knowledge (or beliefs) about the effect of an intervention under study. However, a Bayesian trial analysis requires a subtle set of analytic decisions that differ from the decisions made with frequentist methods. These decisions, which may be unfamiliar to many clinicians, have the potential to affect the interpretation of results. Accordingly, standardizing Bayesian reanalysis of critical care trials could promote transparency and increase the rigor and reproducibility of the results of these analyses. Thus, the goal of this Critical Care Perspective is to introduce readers to the key differences between Bayesian and frequentist trial interpretation and to provide a generalizable framework for authors to apply Bayesian assessments to future studies. To support these aims, we provide guidance on design decisions, an illustrative example of a Bayesian reanalysis of a completed trial, and adaptable statistical code for researchers to apply to future work (provided in the online supplement).

The Traditional Approach to Trial Design and Analysis in Critical Care

To design a trial of a binary outcome using frequentist methods, investigators must estimate the event rate in the target population on the basis of existing evidence. Next, investigators must define the size of the treatment effect that the trial will be able to detect (9). Then, the investigators must select the probability of rejecting a true null hypothesis, which is usually defined as the chance that the *P* value will be less than 0.05, given that no effect exists. Finally, investigators must select the probability of failing to reject a false null hypothesis (i.e., the chance that the *P* value will be greater than 0.05 despite a true treatment effect equal to the size of the treatment effect the trial is designed to detect). Usually, this chance is assumed to be between 10% and 20%. All of these assumptions will impact the interpretation of the resulting trial.

Clinical trials designed using frequentist methods rely on null hypothesis testing using *P* values for trial interpretation. Assuming the usual null hypothesis of no treatment effect, *P* values provide evidence of how incompatible trial results are with the assumption that the treatment has no effect. Thus, *P* values provide no information regarding the size or clinical significance of the measured treatment effect. Confidence intervals for treatment effects improve interpretation by demonstrating the range of effect sizes that are compatible with the observed data, but they do not answer the clinical question of the likelihood of a clinically meaningful treatment effect.

Potential Challenges with Interpretation of Frequentist Clinical Trials

There is nothing wrong with the frequentist approach to trials (10, 11). The challenge with the frequentist approach largely arises with how results are interpreted. Specifically, the reporting of trial results using null hypothesis testing often leads to the misconception that trials that fail to reach arbitrary *P* value thresholds are "negative." Though such trials are more accurately termed "indeterminate" (12), such precision in language and interpretation is unfortunately rare. Consequently, researchers, readers, and editors often equate trials with a *P* value greater than 0.05 as providing no evidence regarding the true treatment effect or, worse, suggesting that it provides evidence for the inefficacy of the intervention (13). For example, frequentist analyses of the EOLIA (Extracorporeal Membrane Oxygenation to rescue Lung Injury in Severe Acute Respiratory Distress Syndrome) (14) and ANDROMEDA-SHOCK (15) trials, both with *P* values greater than 0.05, were simply reported as failing to improve their primary outcomes while simultaneously observing 11% and 8.5% intervention-associated absolute reductions in mortality, respectively. Many clinicians would deem the mortality differences observed in both EOLIA and ANDROMEDA-SHOCK as clinically important. This highlights the fundamental tension about what the *P* value does not tell us, which is the probability of meaningful treatment effect sizes given the observed data, and how these results should be interpreted within the context of existing knowledge and beliefs.

An Introduction to Bayesian Methods

Bayesian methods provide a different approach to data interpretation (5–7, 16). In simple terms, Bayesian methods focus on providing plausible values for the treatment effect that are compatible with both the observed data and prior knowledge or beliefs. This admixture is created by combining two distributions (one distribution of potential effect sizes from prior knowledge or beliefs [called the "prior" probability distribution] and a second distribution representing the new trial results [called the "likelihood"]) to produce a new distribution of effect sizes, termed the "posterior" (after data) probability distribution. The posterior (i.e., after trial) distribution of an effect size can be summarized and presented either graphically or numerically. It is not necessary to compute point estimates with Bayesian methods, but it is commonplace to report the posterior mean, median, and numerical ranges to summarize the likelihood of the true effect (e.g., 0.95 credible interval), all of which are straightforwardly obtainable from the posterior distribution. Credible intervals represent the interval that contains the true value of the effect size within a given probability (commonly set to 0.95, similar to frequentist confidence coverage). More clinically useful is the ability of Bayesian reanalysis to calculate the probability that an intervention causes a specific treatment effect (e.g., a clinically important benefit).

Furthermore, one can provide several empirical summaries and effect size interpretations that demonstrate how results would be influenced by the use of a range of prior distributions that incorporate prior data and hypothetical examples. For example, the reanalyses of the previously mentioned EOLIA (17) and ANDROMEDA-SHOCK (18) trials demonstrated a high likelihood that the interventions were effective across a broad range of assumptions. Thus, to summarize, the three major components of a Bayesian analysis are the prior (which, in the context of randomized trials, reflects an *a priori* belief regarding the possible effect of an intervention [e.g., general skepticism]), the likelihood (the new trial data), and the posterior probability distribution (the distribution of possible effects derived from combining the likelihood with the prior).

Priors: The Key Aspect of a Bayesian Analysis

The key difference between Bayesian analysis and frequentism is the use of a prior, and the process of defining a prior is the aspect of a Bayesian analysis that is most foreign to those accustomed to frequentist methods.

For the following examples, we will describe a hypothetical study that uses a binary outcome and is analyzed by logistic regression. For our example, we will assume that the outcome represents a poor outcome (e.g., mortality), so positive effects mean harm. The range of possible treatment effect sizes for a given intervention can be expressed as a distribution, which is commonly assumed to be a function with the values that mostly likely represent the magnitude of the treatment effect in the middle. The range of possible prestudy treatment effects encapsulated by the prior can be described using a mean (μ) and SD (σ) if the assumed prior is a normal distribution. The mean of the distribution represents the average expected treatment effect (i.e., beneficial, no effect, or harmful), whereas the SD represents the spread of the distribution (i.e., our confidence in our belief regarding the treatment effect). For binary, ordinal, or time-to-event outcomes, the prior is usually described using an effect ratio, such as the odds ratio (OR), and is conveniently specified by the log of this ratio [e.g., log(OR)], in which a μ of greater than 0 represents higher odds for the event.

It is common to first consider a "neutral" prior in Bayesian analyses. These neutral priors are centered at the absence of effect (symmetric); that is, they consider that benefit and harm are equally possible (i.e., the probability of an OR greater than or less than 1 is 0.50). Assuming a normal distribution for a neutral prior, the mean of this distribution is 0, which is equivalent to an OR of 1 (the log of 1 equals 0). In contrast, an "optimistic" prior is one that represents the belief that benefit is more likely than harm and is therefore centered at values less than 0, equivalent to an OR of less than 1 (the log of a number between 0 and 1 results in a negative number), with more of the probability distribution falling below an OR of 1 than above an OR of 1. A "pessimistic" prior represents the belief that harm is more likely than benefit by having a mean greater than 0, the equivalent of an OR greater 1 (the log of a number greater than 1 results in a positive number), with more of the probability distribution falling above an OR of 1 than below an OR of 1.

In the specific case of normally distributed priors, the strength of a prior belief (i.e., our confidence that the treatment effect is close to the mean of our prior) is captured by the SD of the prior distribution. The stronger the belief, the smaller the SD, the narrower the prior probability distribution, and the bigger effect the prior will have on the posterior probability distribution. The weaker the belief, the larger the SD, the broader the prior probability distribution, and the less effect the prior will have on the posterior probability distribution. The strength of the prior can be defined mathematically to create a distribution that allows for a specific probability of benefit or harm or it can be created on the basis of a summary of results from a previous trial, an approach not considered here.

It is also common to see other terms being applied to nominate priors. Two examples include "skeptical" priors and "flat" priors. In this analysis, we will refer to a skeptical prior as one that assumes the most likely treatment effect is zero and places a small probability on a large benefit (hence its description as skeptical) or a large harm. The term "skeptical" can also describe any prior that has a narrow deviation around its central measurement and that is either centered at the absence of effect (or "neutral") or at a harmful effect, which is slightly different from the terminology we use in this paper. A "flat" prior is one with an infinite SD, meaning that every treatment effect is equally likely. A "flat" prior, therefore, contains almost no information at all. Thus, not surprisingly, the results of a Bayesian analysis using a flat prior will be similar to the results obtained from a traditional frequentist analysis when there is only one look at the data. For example, in the ANDROMEDA-SHOCK trial, a frequentist analysis calculated that the OR for the intervention was 0.61 (95% confidence interval, 0.38–0.92; $P = 0.022$), whereas the subsequent Bayesian reanalysis based on a flat prior estimated an OR of 0.62 (95% credible interval, 0.38–0.92) (15, 18).

How to Create a Prior?

Priors are generated to reflect the beliefs that existed regarding an intervention before a trial was conducted. If data from previous clinical trials are available, that data may be used as a prior in a Bayesian analysis. Most priors, however, are mathematically derived to generate distributions that reflect expert opinion regarding the effectiveness of an intervention.

In the suggestions for Bayesian reanalysis detailed in this paper, we propose that researchers use normally distributed priors (using a symmetric scale such as the log OR or log hazard ratio). Our selection is principally motivated by an effort to balance simplicity in generation and ease of interpretation by a clinical audience, as evidenced in the Bayesian reanalysis of the EOLIA and ANDROMEDA-SHOCK trials (17, 18).

There are, however, many acceptable approaches to defining priors. Additional discussion on the derivation of priors is included in Appendix E1 in the online supplement, including Figures E1–E3, and is available in other texts (19).

Why Standardize the Bayesian Reanalysis Process?

One may argue that many of the challenges with frequentist analysis, such as the dichotomous interpretation, result from oversimplifications of interpretation.

Frequentist methods can be interpreted in a continuous fashion, for example, using *P* value functions (20) or *S* values (11). By extension, readers may be concerned that standardizing Bayesian trial analysis could similarly oversimplify analyses and prevent thoughtful and careful interpretation of trial data. A Bayesian reanalysis of a completed trial, however, will face unique challenges that could be minimized by a more standardized approach. To conduct a Bayesian reanalysis, investigators must choose the number of priors, the prior beliefs, and the strengths of those prior beliefs. For a Bayesian reanalysis, all of these decisions are, by definition, *post hoc* and therefore susceptible to the possibility that knowledge of trial results could contribute to bias, skewing the posterior probability distributions. Therefore, some basic tenets to guide the selection of priors *a posteriori* may be useful to both homogenize future Bayesian reanalysis and avoid the reporting (and perception by readers) of biased results. The core of our proposal is the presentation of a minimum set and range of theoretical priors that should be included in any critical care trial reanalysis that uses Bayesian methodology. Using a range of standardized priors minimizes the risk that investigators would be able to skew study results toward desired outcomes. This approach is related to the "community of priors" approach of Spiegelhalter and colleagues (19).

Proposal for a Minimum Set of Priors for a Bayesian Reanalysis

We propose that Bayesian reanalysis should typically consider the full range of possible beliefs through the use of optimistic, skeptical, and pessimistic priors. However, we do not believe that each prior should be given equal interpretive weight. To conduct a clinical trial, investigators must establish that equipoise exists regarding which treatment arm is expected to prove superior. We, therefore, suggest that most of the emphasis during a Bayesian reanalysis should be given to skeptical priors that are symmetric around 0, which are the statistical equivalent of the clinical concept of equipoise. It has been argued that pessimistic priors are rarely necessary, as investigators are motivated to conduct a clinical trial by a belief that the studied intervention will prove beneficial. We consider this to be a flawed argument. Consider, for example, prior beliefs regarding the use of corticosteroids for coronavirus disease (COVID-19). Previous evidence from trials of viral pneumonia and acute respiratory distress syndrome suggested that corticosteroids were ineffective or potentially harmful (21, 22), leading many societies to argue against their use for COVID-19 (23, 24). The results of the RECOVERY (Randomized Evaluation of COVID-19 Therapy) trial, however, demonstrated sufficient benefit from corticosteroids for COVID-19 to overcome prior pessimistic beliefs (25). Similarly, it would be inappropriate to consider only an optimistic prior. If every clinician truly believed that an intervention was beneficial, there would not be sufficient equipoise to conduct a trial.

Proposed Guidance for Conducting and Reporting a Bayesian Reanalysis of a Trial

As noted in Priors: The Key Aspect of a Bayesian Analysis and Appendix E1, priors can be defined in numerous ways. We believe that normally distributed priors (using a symmetric scale such as the log OR or log hazard ratio) are appropriate in most cases (17, 18). Accordingly, building on prior suggestions by Sung and colleagues (26), we propose the following four principles for designing and reporting a Bayesian reanalysis of a critical care trial to the community.

1. *Use at least one skeptical, one pessimistic, and one optimistic prior (*Table 1*).* Pessimistic, skeptical, and optimistic refer to where most of the probability mass is located for the prior and where is it centered. Pessimistic and optimistic priors should be mirrored at the same magnitude of effect size, although differences in SDs may be justifiable (Table 2). A visual distribution of all possible priors is shown in Figure E4, Appendix E2.
2. *Justify selection of each "knowledge or belief" strength (*Table 1*).* Each prior includes both a belief and the strength of that belief (which is a surrogate for the variance of the expected effect size). Belief strength could be weak (high uncertainty on the effect size), moderate, or strong (little uncertainty). Not all possible combinations of belief and belief strength need to be reported, but a rationale should be provided for each chosen belief strength. If little prior data is available, a weak strength for each prior may be the most appropriate. In other scenarios, different combinations may be acceptable. Some hypothetical scenarios are shown in Table 2 (14, 25, 27–29).
3. *Consider previous trials and or other valid external evidence.* When available, existing trial data should be used in the creation of priors in addition to the hypothetical priors described above. Any properly justified prior can be added to the analysis, and priors derived from metanalysis of prior evidence can be useful and may help contextualizing research.
4. *Provide sufficient detail to interpret results by providing*:
   a. A plot of the posterior distribution.
   b. Mean and credible intervals for the intervention. A 0.95 credible interval is suggested for comparison with traditional frequentist analysis, but any credible interval can be used. After the posterior distribution for the effect size is obtained, any summary method can be applied, such as the probability that the intervention is associated with any benefit.
   c. Probability that the intervention is similar to the control, reported as the probability mass that falls within the range of practical equivalence (ROPE). The ORs used to define the boundaries of ROPE will depend on the studied intervention and outcome and are somewhat subjective. The aim of this analysis is to provide a measurement of how much of the probability mass is centered around values close to the absence of an effect. The ROPE range is specific to the studied intervention, the studied outcome, and the baseline event rate. For example, an OR of 1.05 may be considered clinically important for a trial studying a commonly used intervention (such as intravenous fluids or oxygen).
   d. Probability of *severe* harm or *outstanding* benefit extracted from the posterior distribution. This is also context sensitive but may aid the reader to visualize the probability that the intervention has an extreme effect on the outcome (30).
   e. Summary of the impact of different prior selections on the interpretation of the reanalysis. This could be done by comparing differences between effect

Table 1. Recommended Guidance for the Selection and Application of a Minimum Set of Priors and Analyses to Be Used in a Bayesian Reanalysis of a Completed Trial

	Defining Priors		
Prior Belief	**Belief Strength**		
	Weak	**Moderate**	**Strong**
Neutral	"I know almost nothing about the intervention and cannot rule out extreme effect sizes."	"I have no reason to believe the intervention is good or bad, but I am mostly sure I can rule out large effect sizes."	"I strongly believe the intervention has no effect or a very small effect."
	Bayesian analysis will not provide additional information, as the results will converge with results from frequentist approaches	Consider a normal prior centered at an OR of 1 that allows a 0.95 probability that the OR is between 2 and 0.5; that is, Pr(OR $<$ 0.5) = 0.025 and Pr(OR $>$ 2) = 0.025	Consider a normal prior centered at an OR of 1 that allows a 0.95 probability that the OR is between 1.5 and 1/1.5; that is, Pr(OR $<$ 0.66) = 0.025 and Pr(OR $>$ 1.5) = 0.025
	Example prior distribution: N(0, 5)*	Example prior distribution: N(0, 0.355)*	Example prior distribution: N(0, 0.205)*
Optimistic	"I believe the intervention is good, but there are few data, and I cannot rule out harm."	"I believe the intervention is good, but I acknowledge there is a nonnegligible chance it may be harmful."	"I strongly believe the intervention is good and that there is a very low chance that it is harmful."
	Consider a normal prior centered at the log of the expected OR for the intervention with variance set to allow at least 0.30 probability of Pr(OR $>$ 1)	Consider a normal prior centered at the log of the expected OR for the intervention with variance set to allow at least a 0.15 probability of Pr(OR $>$ 1)	Only useful in special cases Consider a normal prior centered at the log of the expected OR for the intervention with variance set to allow at least 0.05 probability of Pr(OR $>$ 1)
Pessimistic	"I believe the intervention is harmful, but there are few data, and I cannot rule out eventual benefit."	"I believe the intervention is harmful, but I acknowledge there is a nonnegligible chance it may be beneficial."	"I strongly believe the intervention is harmful and that there is a very low chance that it is beneficial."
	Consider a normal prior centered at the log of the expected OR for the intervention with the variance set to allow at least 0.30 probability of Pr(OR $<$ 1)	Consider a normal prior centered at log of the expected OR for the intervention with the variance set to allow at least 0.15 probability of Pr(OR $<$ 1)	Only useful in special cases Consider a normal prior centered at log of the expected OR for the intervention with the variance set to allow at least 0.05 probability of Pr(OR $<$ 1)
	Summarizing Results		
Key points	• Include at least one skeptical, one pessimistic, and one optimistic prior. • Justify the use of prior belief strengths. • Provide a graphical representation of priors and posteriors. • For each prior, provide the posterior distribution and provide the probability of obtaining benefit/harm for the intervention. • Provide the probability of obtaining relevant effect sizes, including the region of practical equivalence and the chance of significant benefit or harm. Justify choices of the cutoffs used. • Summarize the impact of different prior selections on the interpretation of the reanalysis. This could be done by comparing differences between effect estimates for each prior or by applying a Bayesian metanalysis considering the results of each prior simulation as a different study.† • Discuss the results with a focus on the priors that were used with individual results for each prior.		

Definition of abbreviations: OR = odds ratio; Pr = probability.
For this example, the primary outcome is mortality, so the proportion of the distribution with an OR less than 1.0 [Pr(OR $<$ 1)] is the probability of benefit. Quotes represent a nontechnical statement on what priors mean for clarity.
*N means the prior follows a normal distribution with two parameters (mean and SD). Creating a prior requires the selection of the mean of prior distribution (μ, reflecting the prior belief of the intervention as providing benefit, no effect, or harm) and the SDs (σ, the spread of the possible effect sizes around the mean, which is a reflection of the "strength" of that belief). A description of the prior can be summarized as N(μ, σ), which indicates a normal distribution with mean = μ and SD = σ. The prior is for the log(OR) of the intervention.
†*See* Appendix E3 for details.

estimates for each prior or by applying a Bayesian metanalysis considering the results of each prior simulation as a different study. Methods to complete a Bayesian metanalysis are discussed in Appendix E3 of the online supplement.

Bayesian Reanalysis of the ART Trial Using Our Proposed Framework

To illustrate how a Bayesian reanalysis should be conducted using this framework, we include below a reanalysis of ART (Alveolar Recruitment for Acute Respiratory Distress Syndrome Trial) (27). The statistical code and dataset to reproduce the analysis and generate all nine of the priors in Table 1 is provided in the online supplement.

Table 2. Suggestions for Selecting Prior Belief Strengths Given Hypothetical Scenarios and Examples from the Critical Care Literature

Scenario	Neutral Prior Strength Suggestion	Optimistic Prior Strength Suggestion	Pessimistic Trial Strength Suggestion
Little to no information previously available Example: most trials run in the COVID-19 pandemic	Weak	Weak	Weak
Conflicting evidence, with some trials showing benefit and others pointing toward harm Example: EOLIA trial (14)	Moderate ("skeptical" prior)	Moderate	Moderate
Evidence pointing toward benefit (for example, positive previous metanalysis). No outliers in previous literature. Usually occurs for trials designed to confirm benefit Example: ART trial (27)	Moderate ("skeptical" prior)	Moderate	Weak
Evidence pointing toward benefit (for example, previous metanalysis). Presence of outliers (one or few studies) pointing toward an opposite direction	Moderate ("skeptical" prior)	Moderate	Moderate
Consecrated intervention deemed to be beneficial above reasonable doubt inside the medical community Example: assessing the effects of proton-pump inhibitors to avoid gastric bleeding using data from the SUP-ICU trial (28)	Moderate ("skeptical" prior)	Strong	Weak
Interventions with a very low rationale of exerting a direct effect on a given outcome, but data is available Example: mortality outcome in the PEPTIC trial (29) and/or SUP-ICU trial (28)	Strong ("very skeptical")	Weak	Weak
Several previous trials reporting neutral results, sometimes reaching futility thresholds on trial sequential analysis	Strong ("very skeptical")	Weak	Weak

Definition of abbreviations: ART = Alveolar Recruitment for Acute Respiratory Stress Syndrome; COVID-19 = coronavirus disease; EOLIA = Extracorporeal Membrane Oxygenation to Rescue Lung Injury in Severe Acute Respiratory Distress Syndrome; PEPTIC = Proton Pump Inhibitors versus Histamine-2 Receptor Blockers for Ulcer Prophylaxis Treatment in the ICU; SUP-ICU = Stress Ulcer Prophylaxis in the ICU.

ART compared a strategy of open-lung mechanical ventilation (using lung recruitment maneuvers and positive end-expiratory pressure [PEEP] titration according to the best respiratory system compliance; $n = 501$; experimental group) to a control strategy of mechanical ventilation with low PEEP ($n = 509$). After adjustment, the trial showed that the open-lung strategy was associated with an increase in 28-day mortality (hazard ratio of 1.20; 95% confidence interval, 1.01–1.42; $P = 0.041$). For simplicity, we will consider 28-day mortality as a binary endpoint in this example and will not consider other model adjustments. In this scenario, the trial suggested harm in the experimental group with an OR of 1.27 (95% confidence interval, 0.99–1.63; $P = 0.057$). Like the ANDROMEDA-SHOCK reanalysis, a Bayesian regression with flat priors results in an essentially identical OR of 1.28 and similar 0.95 credible intervals ranging from 1.00 to 1.63 (Figures 1 and E5).

Implementation of Principles 1–3 of Our Proposal to Develop Our Priors

Following the principles in Table 1, we included at least one skeptical, one pessimistic, and one optimistic prior (i.e., principle 1). Now, we will provide a justification for the strength of each prior we used (i.e., principle 2) in the context of existing knowledge or beliefs (i.e., principle 3). At the time ART was conducted, results of previous trials were already available (31, 32). These results suggested an intervention-associated benefit, but many studies were neutral or indeterminate. Accordingly, it is sensible to consider a moderate belief strength for both the optimistic and neutral prior and to consider a weak pessimistic prior (Table 2). This a common situation for multicenter trials and will likely apply to many reanalyses, as the time, cost, and complexity of large multicenter trials necessitate that only interventions with promising data from pilot trials or observational studies are selected. Furthermore, a moderate-belief optimistic prior represents

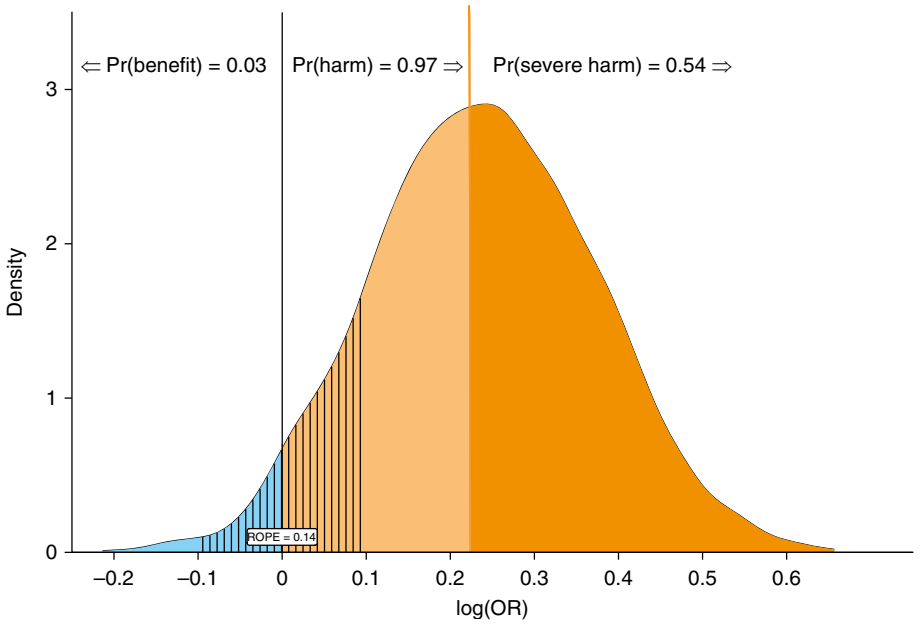

Figure 1. Posterior distribution of the log odds ratio (OR) in ART (Alveolar Recruitment for Acute Respiratory Distress Syndrome Trial) using a "flat" prior. The distribution represents 100,000 draws from the posterior, which approximates to a normal distribution with a mean of 0.24 and an SD of 0.13. The vertical line at 0 represents the point at which the OR is equal to 1 [i.e., log(OR) = 0]. The area to the right (in orange) represents the probability that the intervention is harmful (0.97 probability). The probability of severe harm [Pr(OR > 1.25)] is shown in dark orange and is equal to 0.54. Values <0 mean the intervention is beneficial [Pr(log(OR) < 0); Pr(OR < 1.0)] and are shown in light blue (which equals 0.03). The ROPE is defined as the OR between 1/1.1 and 1.1 (vertically hatched area) and is 0.14. A similar figure with the OR on the *x*-axis is shown in Figure E5 for comparison. All these findings provide compelling evidence against the experimental treatment even in the context of a flat prior. Pr = probability; ROPE = region of practical equivalence.

the type of equipoise that is common for many trials, namely, that there is belief that the intervention could help patient outcomes, but the possibility of harm has not been ruled out.

The neutral prior of moderate strength ("skeptical" prior) is centered at the absence of effect [OR = 1; log(OR) = 0] with an SD of 0.355, such that 0.95 of the probability falls in the range 0.5 < OR < 2, the range of treatment effects that might be reasonably expected. Therefore, our skeptical prior will follow a normal distribution with a mean of 0 and an SD of 0.355 [N(0, 0.355)]. The next step is to define the mean and SD for the optimistic and pessimistic priors. Similar to other trials in acute respiratory distress syndrome, ART was designed to have enough power to detect a reduction in mortality, with an OR close to 0.66 [log(OR) = −0.41]. Therefore, we set the optimistic prior with a mean of −0.41 (i.e., OR = 0.66), and pessimistic prior with a mean of 0.41 (i.e., OR = 1.5). Guided by the framework in Table 1, the SD of the optimistic prior is defined to retain a 0.15 probability of harm [Pr(OR > 1)], and the pessimistic prior is chosen to retain a 0.30 probability of benefit [Pr(OR < 1)]. We derived that an SD of 0.40 for the optimistic prior and of 0.8 for the pessimistic priors would provide such probabilities (additional details in the online supplement).

Implementation of Principle 4 of Our Proposal to Report Our Results

Now that we have fully defined our family of three priors for this reanalysis, we are ready to proceed with the analysis. For the purpose of this reanalysis, we will consider the ROPE for the intervention to be between 1/1.1 and 1.1 and define severe harm as an OR of greater than 1.25 and outstanding benefit as an OR of less than 1/1.25. The definition of the ROPE and thresholds for benefit and harm will vary by trial based on the studied intervention, the chosen outcome, and baseline event rates in the study population. If the minimum clinically important treatment effect has been established in the study setting, the ROPE may be set to cover the range below this threshold. The results of this ART reanalysis are shown in Table 3 and Figure 2. In all scenarios, the probability of harm was high [Pr(OR > 1) > 0.9]. The probability of a significant benefit was roughly 0 regardless of the prior, and there was a low probability mass located in the previously defined ROPE between the intervention and the control arm (as low as 0.127 for the pessimistic prior). There was a significant probability that the intervention could cause severe harm, defined as an OR of greater than 1.25, with even an optimistic prior resulting in a posterior probability of severe harm greater than 0.30.

As suggested, we can demonstrate the influence of the priors on the posterior probability distribution by estimating the heterogeneity in the results induced by the priors in a metanalytic context. This approach is discussed in Appendix E3. In the specific case of this ART reanalysis, the estimated heterogeneity (I^2) was close to 0.11, which can be interpreted as a low degree of heterogeneity. That is, the results of the Bayesian reanalysis were not sensitive to which prior was used, as approximately 0.11 of all variance was caused by the priors (Figures E6 and E7). These results are also shown from a different perspective in Table 3 by reporting the differences in the posterior OR between the used priors. Those differences were small, suggesting that the priors had a minor effect on conclusions. Taken together, these results point toward a significant probability that the intervention used in ART was hazardous with regard to the outcome of 28-day mortality.

Opportunities for Continued Innovation and Consensus Building

Our hope with this article was to provide both an introduction to a Bayesian reanalysis for clinicians and propose a foundation for additional development and discussion. Accordingly, our proposal focused on a suggested minimum set of analyses and data presentations. The suggestions presented in the manuscript represent the opinions of the authors and are not based on a consensus-generating process. We make a compelling argument for using a Bayesian reanalysis to contextualize the results of completed clinical trials, and we believe that the field would benefit from more thorough, consensus-based guidelines on the optimal methods for analyzing treatment effects (i.e., absolute risk, relative risk, or OR),

Table 3. Results of a Bayesian Reanalysis of the ART trial (27) Using the Reanalysis Framework Developed in This Manuscript

Prior: Mean (SD)	OR (95% CrI)	Difference in OR vs. Skeptical Prior (95% CrI)*	Difference in OR vs. Optimistic Prior (95% CrI)*	Difference in OR vs. Pessimistic Prior (95% CrI)*	Probability of Harm; Pr(OR > 1)	Probability of Important Benefit; Pr(OR < 1/1.25)	Probability of Important Harm;† Pr(OR > 1.25)	ROPE;‡ Pr(OR > 1/1.1, OR < 1.1)	Interpretation
Skeptical: 0 (0.355)	1.24 (0.98 to 1.55)	—	0.03 (0.02 to 0.04)	−0.05 (−0.06 to −0.03)	0.956	0.00	0.465	0.168	When assuming a moderate strength neutral prior ("skeptical" prior), the probability of harm of the intervention is more than 0.95. There is also an important probability that the intervention is very harmful, following the chosen definition of severe harm (an OR >1.25).
Optimistic: −0.41 (0.40)	1.19 (0.95 to 1.51)	−0.03 (−0.04 to −0.02)	—	−0.08 (−0.10 to −0.06)	0.936	0.00	0.348	0.255	Even when assuming a moderate strength optimistic prior, the probability of harm of the intervention is greater than 0.90. The probability of severe harm remains clinically relevant at 0.35, and there is only a probability of 1 in 4 that the intervention is within the defined limits of equivalence.
Pessimistic: 0.41 (0.80)	1.28 (1.01 to 1.62)	0.05 (0.03 to 0.06)	0.08 (0.06 to 0.10)	—	0.971	0.00	0.563	0.127	When assuming a weak strength pessimistic prior, the probability of harm of the intervention is very high. Not only is the intervention probably harmful under these assumptions, but the probability of severe harm is greater than 0.50.

Definition of abbreviations: ART = Alveolar Recruitment for Acute Respiratory Distress Syndrome Trial; CrI = credible interval; OR = odds ratio; Pr = probability; ROPE = region of practical equivalence.

Data are shown on the log scale, with negative values meaning OR <1 and positive values meaning OR >1.

*Difference obtained by sampling the posterior OR distribution.

†Please note that this is a suggestion. "Significant harm" is subjective and should be tailored to the scenario.

‡Please note that this is a suggestion. "Equivalence" is subjective and should be tailored to the scenario.

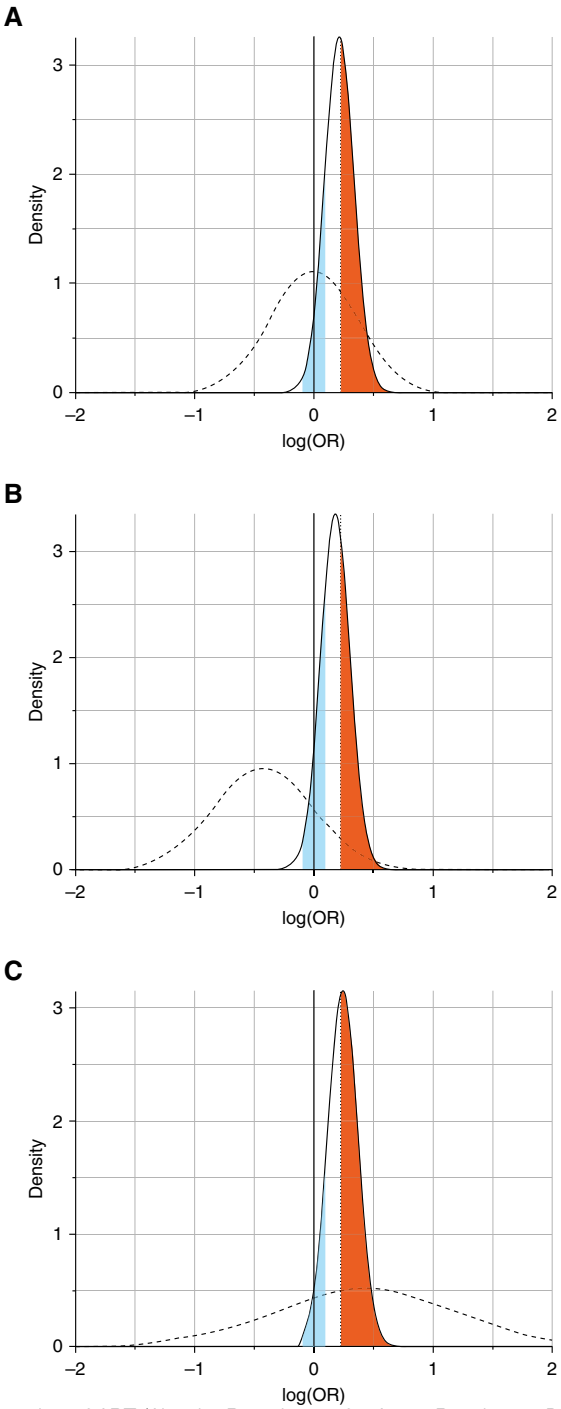

Figure 2. Reinterpretation of ART (Alveolar Recruitment for Acute Respiratory Distress Syndrome Trial). Priors were set following the suggested principles outlined in the main manuscript using optimistic, skeptical, and pessimistic priors of moderate strength at (*A*) N(0, 0.355), (*B*) N(−0.44, 0.40), and (*C*) N(0.44, 0.80). Priors are shown in dashed lines. For each selected prior, the black line shows the posterior distribution of the odds ratio (OR). The probability of significant harm [Pr(OR > 1.25)] is filled in red (values in Table 3). The ROPE, defined as an OR between 1/1.1 and 1.1, is filled in blue. ROPE = region of practical equivalence.

defining equivalence (i.e., ROPE) boundaries, quantifying the strength of a prior belief, and communicating the impact of chosen priors on the posterior probability distribution. In the last sense, for example, the use of I^2 as a measurement of impact of priors on results can be useful but may also be seen as too technical for readers.

Although not covered in this manuscript, future work is also needed to standardize the identification of trial-specific minimal clinically important treatment effects and to determine how these calculations should be incorporated into trial design and interpretation. We hope that by introducing the reader to Bayesian methods in the specific context of trial reanalysis we succeed in starting a detailed discussion on all the aforementioned points.

Discussion

A Bayesian reanalysis can be a helpful tool to augment the interpretation of critical care trials (8, 17, 18, 33–37). In this review, we have highlighted potential challenges with the interpretation of results from a trial conducted using frequentist methods. We have described how Bayesian reanalysis can be used to provide important clinical insights, including the clinically relevant probabilities that trial interventions are associated with benefit or harm in contrast to the more indirect frequentist approach. We have also provided a framework for how a Bayesian reanalysis of a frequentist trial can be conducted, including suggestions regarding the selection of priors (Table 1). Finally, using the ART trial, we have provided an example of how these suggestions can be applied to conduct and report a Bayesian reanalysis, finding that ART suggests a high probability of harm, regardless of prior beliefs. For simplicity, we focused our conceptual framework on Bayesian principles for binary outcomes. Although they are out of the scope of this manuscript, the same principles could also be applied to continuous, count, or time-to-event endpoints. We hope that the suggestions included here may form the basis of future consensus-based guidelines, which we believe would improve the reporting and reproducibility of Bayesian reanalyses and support across study comparisons. Finally, we hope that this discussion enhances physicians' understanding of Bayesian methods and further improves the critical appraisal of RCTs. ■

Author disclosures are available with the text of this article at www.atsjournals.org.

Acknowledgment: The authors thank Dr. Alexandre B. Cavalcanti for providing raw anonymized data from ART for the example in this manuscript and for providing feedback on prior drafts of the manuscript.

References

1. Neyman J. Outline of a theory of statistical estimation based on the classical theory of probability. *Philos Trans R Soc Lond A* 1937;236:333–380.
2. Berry DA. Bayesian clinical trials. *Nat Rev Drug Discov* 2006;5:27–36.
3. Angus DC, Berry S, Lewis RJ, Al-Beidh F, Arabi Y, van Bentum-Puijk W, *et al*. The REMAP-CAP (randomized embedded multifactorial adaptive platform for community-acquired pneumonia) study: rationale and design. *Ann Am Thorac Soc* 2020;17:879–891.
4. Laterre PF, Berry SM, Blemings A, Carlsen JE, Francois B, Graves T, *et al*.; SEPSIS-ACT investigators.Effect of selepressin vs placebo on ventilator- and vasopressor-free days in patients with septic shock: the SEPSIS-ACT randomized clinical trial. *JAMA* 2019;322: 1476–1485. [Published erratum appears in *JAMA* 15:e1917553.]
5. Bittl JA, He Y. Bayesian analysis: a practical approach to interpret clinical trials and create clinical practice guidelines. *Circ Cardiovasc Qual Outcomes* 2017;10:e003563.
6. Wijeysundera DN, Austin PC, Hux JE, Beattie WS, Laupacis A. Bayesian statistical inference enhances the interpretation of contemporary randomized controlled trials. *J Clin Epidemiol* 2009;62:13–21, e5.
7. Harhay MO, Casey JD, Clement M, Collins SP, Gayat É, Gong MN, *et al*. Contemporary strategies to improve clinical trial design for critical care research: insights from the First Critical Care Clinical Trialists Workshop. *Intensive Care Med* 2020;46:930–942.
8. 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* [online ahead of print] 20 Nov 2020; DOI: 10.1016/S2213-2600(20)30471-9.
9. Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. *Lancet* 2005;365:1348–1353.
10. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. *Am Stat* 2016;70:129–133.
11. Greenland S. Valid p-values behave exactly as they should: some misleading criticisms of p-values and their resolution with s-values. *Am Stat* 2019;73:106–114.
12. Sackett DL. Superiority trials, noninferiority trials, and prisoners of the 2-sided null hypothesis. *ACP J Club* 2004;140:A11.
13. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995;311:485.
14. Combes A, Hajage D, Capellier G, Demoule A, Lavoué S, Guervilly C, *et al*.; EOLIA Trial Group, REVA, and ECMONet. Extracorporeal membrane oxygenation for severe acute respiratory distress syndrome. *N Engl J Med* 2018;378:1965–1975.
15. Hernández G, Ospina-Tascón GA, Damiani LP, Estenssoro E, Dubin A, Hurtado J, *et al*.; The ANDROMEDA SHOCK Investigators and the Latin America Intensive Care Network (LIVEN). Effect of a resuscitation strategy targeting peripheral perfusion status vs serum lactate levels on 28-day mortality among patients with septic shock: the ANDROMEDA-SHOCK randomized clinical trial. *JAMA* 2019;321:654–664.
16. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. Boca Raton, FL: CRC Press; 2013.
17. 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;320:2251–2259.
18. Zampieri FG, Damiani LP, Bakker J, Ospina-Tascon 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: 423–429.
19. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health-care evaluation. Chichester: John Wiley & Sons; 2004.
20. Infanger D, Schmidt-Trucksäss A. P value functions: an underused method to present research results and to promote quantitative reasoning. *Stat Med* 2019;38:4189–4197.
21. Arabi YM, Mandourah Y, Al-Hameed F, Sindi AA, Almekhlafi GA, Hussein MA, *et al*.; Saudi Critical Care Trial Group. Corticosteroid therapy for critically ill patients with Middle East respiratory syndrome. *Am J Respir Crit Care Med* 2018;197:757–767.
22. Ni YN, Chen G, Sun J, Liang BM, Liang ZA. The effect of corticosteroids on mortality of patients with influenza pneumonia: a systematic review and meta-analysis. *Crit Care* 2019;23:99.
23. Alhazzani W, Møller MH, Arabi YM, Loeb M, Gong MN, Fan E, *et al*. Surviving sepsis campaign: guidelines on the management of critically ill adults with Coronavirus Disease 2019 (COVID-19). *Crit Care Med* 2020;48:e440–e469.
24. Jamil S, Mark N, Carlos G, Cruz CSD, Gross JE, Pasnick S. Diagnosis and management of COVID-19 disease. *Am J Respir Crit Care Med* 2020;201:P19–P20.
25. Horby P, Lim WS, Emberson JR, Mafham M, Bell JL, Linsell L, *et al*.; RECOVERY Collaborative Group. Dexamethasone in hospitalized patients with covid-19 - preliminary report. *N Engl J Med* [online ahead of print] 17 Jul 2020; DOI: 10.1056/NEJMoa2021436.
26. Sung L, Hayden J, Greenberg ML, Koren G, Feldman BM, Tomlinson GA. Seven items were identified for inclusion when reporting a Bayesian analysis of a clinical study. *J Clin Epidemiol* 2005;58: 261–268.
27. Cavalcanti AB, Suzumura EA, Laranjeira LN, Paisani DM, Damiani LP, Guimaraes HP, *et al*.; Writing Group for the Alveolar Recruitment for Acute Respiratory Distress Syndrome Trial (ART) Investigators. Effect of lung recruitment and titrated positive end-expiratory pressure (PEEP) vs low PEEP on mortality in patients with acute respiratory distress syndrome: a randomized clinical trial. *JAMA* 2017;318: 1335–1345.
28. Krag M, Marker S, Perner A, Wetterslev J, Wise MP, Schefold JC, *et al*.; SUP-ICU trial group. Pantoprazole in patients at risk for gastrointestinal bleeding in the ICU. *N Engl J Med* 2018;379: 2199–2208.
29. Young PJ, Bagshaw SM, Forbes AB, Nichol AD, Wright SE, Bailey M, *et al*.; PEPTIC Investigators for the Australian and New Zealand Intensive Care Society Clinical Trials Group, Alberta Health Services Critical Care Strategic Clinical Network, and the Irish Critical Care Trials Group. Effect of stress ulcer prophylaxis with proton pump inhibitors vs histamine-2 receptor blockers on in-hospital mortality among ICU patients receiving invasive mechanical ventilation: the PEPTIC randomized clinical trial. *JAMA* 2020;323:616–626.
30. Brown SM, Peltan ID, Webb B, Kumar N, Starr N, Grissom C, *et al*. Hydroxychloroquine versus Azithromycin for Hospitalized Patients with Suspected or Confirmed COVID-19 (HAHPS): protocol for a pragmatic, open-label, active comparator trial. *Ann Am Thorac Soc* 2020;17:1008–1015.
31. Suzumura EA, Figueiró M, Normilio-Silva K, Laranjeira L, Oliveira C, Buehler AM, *et al*. Effects of alveolar recruitment maneuvers on clinical outcomes in patients with acute respiratory distress syndrome: a systematic review and meta-analysis. *Intensive Care Med* 2014;40:1227–1240.
32. Villar J, Kacmarek RM, Pérez-Méndez L, Aguirre-Jaime A. A high positive end-expiratory pressure, low tidal volume ventilatory strategy improves outcome in persistent acute respiratory distress syndrome: a randomized, controlled trial. *Crit Care Med* 2006;34: 1311–1318.
33. Zampieri FG, Costa EL, Iwashyna TJ, Carvalho CRR, Damiani LP, Taniguchi LU, *et al*.; Alveolar Recruitment for Acute Respiratory Distress Syndrome Trial Investigators. Heterogeneous effects of alveolar recruitment in acute respiratory distress syndrome: a machine learning reanalysis of the Alveolar Recruitment for Acute Respiratory Distress Syndrome Trial. *Br J Anaesth* 2019;123:88–95.
34. Granholm A, Marker S, Krag M, Zampieri FG, Thorsen-Meyer HC, Kaas-Hansen BS, *et al*. Heterogeneity of treatment effect of prophylactic pantoprazole in adult ICU patients: a post hoc analysis of the SUP-ICU trial. *Intensive Care Med* 2020;46:717–726.
35. Brophy JM. Bayesian interpretation of the EXCEL trial and other randomized clinical trials of left main coronary artery revascularization. *JAMA Intern Med* 2020;180:986–992.
36. Wagenmakers E-J, Gronau QF. Absence of evidence and evidence of absence in the FLASH trial: a Bayesian reanalysis [preprint]. PsyArXiv; 2020 [2020 Oct 30]. Available from: https://psyarxiv.com/4pf9j.
37. Field SM, Hoek JM, de Vries YA, Linde M, Pittelkow M-M, Muradchanian J, *et al*. Rethinking remdesivir for COVID-19: a Bayesian reanalysis of trial findings [preprint]. MetaArXiv; 2020 [2020 Oct 30]. Available from: https://osf.io/preprints/metaarxiv/2kam7/.