Mazzinari *et al. Perioperative Medicine* (2026) 15:28
https://doi.org/10.1186/s13741-026-00659-4

Perioperative Medicine

METHODOLOGY Open Access

Bayesian statistics: a primer for perioperative medicine clinicians

Guido Mazzinari[1,2,3,4*], Fernando G. Zampieri[5], Michael O. Harhay[6], Marcus J. Schultz[4,7,8,9,10], David M. van Meenen[4,11] and Ary Serpa-Neto[4,12,13,14]

Abstract

Bayesian methods offer an intuitive and coherent statistical framework for updating probabilistic beliefs by integrating prior knowledge—whether from existing data or expert consensus—with new evidence via likelihood functions to generate posterior probability distributions. This approach yields clinically meaningful outputs, such as credible intervals and probabilities of treatment benefit, and can incorporate thresholds relevant to practice, like the region of practical equivalence (ROPE). Recent advances in computation—including Markov chain Monte Carlo (MCMC) sampling, Hamiltonian Monte Carlo algorithms, and probabilistic programming languages like Stan and JAGS— have made Bayesian approaches feasible even for complex hierarchical models. In perioperative medicine, these methods are particularly valuable for (1) complementing trial results by quantifying clinically important effects in the context of statistically nonsignificant findings or modest probabilities of benefit despite statistical significance, (2) enhancing meta-analyses through coherent integration of heterogeneous studies and sparse data, and (3) enabling adaptive and platform trial designs through continuous evidence synthesis. The ability to incorporate informative priors can complement existing knowledge, especially in small–sample studies, which are common in perioperative medicine, where traditional approaches provide insufficient precision. Although concerns remain regarding subjectivity in prior specification, these are increasingly addressed through structured guidelines, benchmark priors, and comprehensive sensitivity analyses. Altogether, Bayesian methods provide a flexible and powerful alternative for generating actionable insights in complex clinical settings, including in perioperative care.

*Correspondence:
Guido Mazzinari
gmazzinari@gmail.com
[1]Department of Anesthesiology and Pain Medicine, Hospital Universitario y Politécnico La Fe, Avenida Fernando Abril Martorell 106, Valencia 46026, Spain
[2]Perioperative Medicine Research Group, Instituto de Investigación Sanitaria La Fe, Avenida Fernando Abril Martorell 106, Valencia 46026, Spain
[3]Department of Statistics and Operational Research, Universidad de Valencia, Calle Dr. Moliner, 50., Burjassot, Valencia 46100, Spain
[4]PROtective VEntilation Network, https://www.provenetwork.org/home
[5]Department of Critical Care Medicine, Faculty of Medicine, and Dentistry, University of Alberta and Alberta Health Services, Edmonton, AB, Canada
[6]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
[7]Department of Intensive Care, Amsterdam University Medical Centers, Location 'AMC', Amsterdam, the Netherlands
[8]Mahidol–Oxford Tropical Medicine Research Unit (MORU), Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand
[9]Nuffield Department of Medicine, University of Oxford, Oxford, UK
[10]Department of Anesthesia, General Intensive Care and Pain Management, Division of Cardiothoracic and Vascular Anesthesia & Critical Care Medicine, Medical University Vienna, Vienna, Austria
[11]Department of Anesthesiology, Amsterdam University Medical Centers, Location 'AMC', Amsterdam, the Netherlands
[12]Department of Critical Care Medicine, Hospital Israelita Albert Einstein, Av. Albert Einstein, 627/701 - Morumbi, São Paulo - SP 05652-900, Brasil
[13]Australian and New Zealand Intensive Care Research Centre (ANZIC-RC), Monash University, Melbourne, Australia
[14]Department of Intensive Care, Melbourne Medical School, University of Melbourne, Austin Hospital, Melbourne, Australia

Introduction

Bayesian methods, first proposed in the eighteenth century as a way to update beliefs in light of new evidence, have always promised a more intuitive framework for medical reasoning. This approach assesses the probability of the effect of a studied intervention given the observed data. It provides an interpretable probabilistic framework and allows for the inclusion of preexisting knowledge into the analysis. It is somewhat analogous to interpreting a diagnostic test, a familiar scenario for doctors. To decide whether a patient has a suspected condition, we base our judgment on prior information, which in this case is represented by the prevalence of the specific condition and the patient's clinical signs and symptoms, and then we modify our belief after seeing the test results (Spiegelhalter et al., 1999). In short, Bayesian analysis is how we naturally think—just adding numbers and probabilities (Goligher and Harhay 2024). Yet, despite its logical appeal, this method has often stood at the fringe of mainstream statistics; critics question its practicality and call it subjective (Efron 1986).

We will outline in this paper why there is renewed interest in Bayesian methods in perioperative medicine, and their advantages and limitations, to provide the reader with a clear foundation for interpreting studies that use this approach. There are many resources to dive deeper and have a more detailed understanding (Spiegelhalter 2004; Gelman et al., 2013; Kruschke 2014; McElreath 2020).

Frequentist statistics *status quo*

Frequentist statistics, the dominant statistical approach, was developed in the early twentieth century and derives its name from its reliance on the long-run frequency behaviour of random variables. Under this framework, if an experiment were repeated infinitely, we could—given certain assumptions—determine how often a particular outcome occurs., frequentist analysis works as follows: It assumes a null hypothesis, e.g., no difference between treatments, model the expected distribution of the data if the null were true and calculate the probability of observing the actual, or more extreme data, under this assumption, the *P*-value. In simple terms, it tries to answer the question: how likely is it to see current or more extreme data if the null hypothesis were correct? We can then reject or fail to reject this null hypothesis. Then, when a test returns a *P*-value that exceeds the prespecified threshold for rejection, also known as the significance level, the correct interpretation is not that the studies' results are negative; it is simply that the available data are insufficient to reject the null hypothesis. Furthermore, the threshold for rejecting or accepting the null hypothesis is often considered arbitrary, and a traditional 0.05 value has become the norm (Goodman 1999a).

Confidence Intervals (CIs) have been promoted as a remedy to the perceived problems of *P*-values. Advocates argue that CIs offer insight into the possible range and size of that effect. As such, they are often seen as more informative and less susceptible to binary thinking. Unfortunately, their definition is no more intuitive than that of P-values. A frequentist confidence interval (CI) is the range of values derived from sample data that would contain the true population parameter in a specified proportion of repeated experiments, e.g., 95%. It does not mean there's a 95% probability that the true value lies within this specific interval—instead, it reflects how often such intervals would capture the parameter over infinite repetitions. Like *P*-values, CIs rely on long-run frequency properties and share the same limitations (Feinstein 1998; Gelman et al., 2019).

In brief, the main issue is the misplaced belief that frequentist operating characteristics simultaneously reflect both the long-run probability of error and the evidential strength of a particular result. This erroneous conflation is particularly striking for fundamental frequentist concepts—such as power and type I (α) and type II (β) errors. Power is the probability of detecting a true effect of a specified size—that is, the probability of rejecting the null hypothesis when a particular alternative hypothesis is true. It is typically calculated before a study begins, based on assumptions about the expected effect size, variability, and significance level (α, often set at 0.05). The complement of power is the type II error rate (β), which is the probability of failing to reject the null when the alternative is true. These concepts are meant to describe performance across many repetitions of the same study design and are not applicable to interpreting the results of a single study in isolation (Zwet et al., 2024). Moreover, testing multiple hypotheses or selectively reporting favourable results inflates the type I error rate, while unplanned interim analyses or data-driven stopping rules distort the nominal P-value, making it no longer a valid measure of compatibility with the null (Stefan and Schönbrodt 2023). In the last decade, statisticians have raised concerns over the misuse of frequentist methods (Greenland et al., 2016), and formal statements on how *P*-values should be implemented and interpreted have also been published (Wasserstein and Lazar 2016). Also, considerable efforts have been made to improve frequentist methods, (Wasserstein et al., 2019; Greenland 2019) although the traditional way to make statistics in practice is still entrenched (Goodman 2019; Matthews 2021). Another potential solution is to use different methods, such as Bayesian statistics, that provide direct probabilistic statements about treatment effects, including whether they exceed clinically meaningful thresholds, without relying on dichotomous significance testing. An overview of key features and differences between frequentist and

Bayesian methods is shown in Table 1, and the strengths and disadvantages of Bayesian methods are summarised in Table 2.

Bayesian methods

Bayesian inference centres on Bayes' theorem, a mathematical rule for updating the probability of a hypothesis as new evidence emerges. It is defined formally in Table 3. This framework establishes a direct relationship between three key components: the *prior*, the *likelihood*, which, when combined, produce the *posterior* (Spiegelhalter et al., 1994).

The *prior* quantifies existing knowledge—whether from past data, theory, expert judgment or acknowledging no or limited insight (i.e., a flat or noninformative prior)—before analysing new evidence. Selecting a prior distribution requires specifying its location and scale, or whichever parameter defines the chosen distribution; these parameters define our best guess on the actual value and how confident we are in it. For instance, to define a prior distribution for age, the mean and variance of a normal distribution could be specified. Priors can be categorised as either noninformative or informative based on the amount of information they convey. This flexibility constitutes both Bayes' power and its perennial critique: priors necessarily incorporate judgment (De Grooth and Cremer 2024; De Grooth and Elbers 2022). A sensible way to think about how to set priors is that even when precise effect estimates are unavailable or there are conflicting beliefs, there are at least biologically plausible bounds on possible effects. For instance, no drug or treatment produces infinite benefit, and no one is taller than four meters; thus, there is no reason to place any prior probabilities on such unrealistic values (Prior Distributions for rstanarm Models n.d.; Harrell 2024). On top of this scientifically sound way of proceeding, in recent years, guidelines have been published in various fields on how priors should be set, and some benchmarks have been proposed (Zampieri et al., 2021; Heuts et al., 2025). Also, computation-based sensitivity analyses to assess the potential perturbations on the posterior by prior and likelihood have been developed (Kallioinen et al., 2024). A more detailed description of how readers could construct priors is provided in Table 4, and some examples of prior distributions are shown in Fig. 1.

The *likelihood* calculates the probability of observing the data given a specific hypothesis or set of parameter values. It is the mathematical representation of the statistical model that describes the data-generating process. In Bayesian analysis, the statistical model used to construct the likelihood is just as critical as in traditional frequentist methods. When these models are simple, the posterior distribution can be derived exactly with a closed-form solution for the mathematical formulas.

Table 1 The frequentist and the Bayesian methodology compared

Design feature	Frequentist	Bayesian
Prior information besides the data	It is informally introduced in the design by setting type I (α) and II (β) errors and difference or precision threshold	Formally specified either by expert opinion or previous studies posterior estimations
Parameters estimation	It is assumed to be fixed and can be hypothetically estimated with infinite repeated sampling	It has an unknown quantity that can be represented by a probability distribution
Starting question	It starts making a guess, called the hypothesis; one common guess is that there is no difference or effect for a parameter of interest, the null–hypothesis It estimates the probability of the observed data given that the prespecified hypothesis is true	Bayesian methodology starts with initial beliefs, the priors, about how likely different options are distributed after observing the data It estimates the probability of a hypothesis given the observed data
Results	Point estimates based on likelihood estimators and P-values	Posterior probability distributions that can return actual probabilities for any given hypothesis of interest
Uncertainty presentation	Confidence intervals. It is a range constructed such that, if the same experiment were repeated many times with new data, 1 – α, usually 95%, of those intervals would include the true value of the parameter of interest. It does not state the probability that the true parameter lies within a specific interval from a single study	Posterior probability can be used to make informed estimates or compute ranges of credibility on a predefined percentage of probability, such as 95% highest density (HDI) or Credibility Intervals (CrI); these can coincide in the case of symmetric distributions
Sequential analyses	Estimates should be adjusted for the number of interim analyses	Interim analysis flows naturally, and the number of repeated looks does not affect the posterior estimation
Sensitivity analyses	Multiplicity correction such as Bonferroni or Holm	Different stakeholders' opinions can be tested by defining different priors

However, real-world scenarios quickly become complex. The posterior becomes a function that cannot be integrated analytically because it requires evaluating high-dimensional integrals and intense calculations. This was a hurdle for Bayesian methods in the past, as high-dimensional integrals in multiparameter models proved intractable for all but small-scale problems. However, advances in computing power, the rise of Monte Carlo (MC) techniques (Hastings 1970; Casella and George 1992; Brooks et al., 2011; Betancourt 2018; Gómez-Rubio 2020),

Table 2 Strengths and weaknesses of Bayesian methods

Preexisting knowledge or data can be implemented into the analysis: This allows sequential learning, in which posterior distributions from a study can be used as a new prior for future research. This is also useful when data are scarce or when models are complex. This also provides transparency if priors are adequately reported and justified.
We can figure out how likely the effect estimates are by using probabilities: Unlike frequentist confidence intervals, which only give a range of possible values for the true parameter without saying how likely they are, Bayesian credibility intervals can quantify such probability.
It allows us to test specific hypotheses, like whether the estimate is smaller or larger than a certain threshold we are interested in: For instance, we can estimate the probability of having an effect as the posterior probability density that is over an odds ratio (OR) of 1 and make statements such as, there is a 70% probability of an effect. Basically, Bayesian methods provide what clinicians want to know: the probability of benefit and harm, given new data.
It is more robust compared to the standard meta-analytic methods and models' misspecifications:Standard meta-analytical methods can have issues in accurately estimating confidence intervals and have rigid assumptions about the studies' heterogeneity. Bayesian estimation yields more precise estimates of pooled effects, especially when the number of studies included in the analysis is small.
It eliminates the need for statistical penalties with repeated data analyses
All these advantages have some costs since Bayesian methods: *Require priors*: Results can depend on the choice of prior, especially with limited data *Require more computation*: Bayesian models often take longer to run and need more processing power *Are harder to explain*: Some clinicians and researchers find the concepts (priors, posteriors, probabilities) unfamiliar *May have reproducibility concerns*: Different reasonable priors can yield slightly different results, which can raise questions during peer review

Table 3 Bayes theorem

It is formally defined as: $p\left(\theta \mid d\right) = \frac{p(d \mid \theta) \bullet p(\theta)}{p(d)}$
$p\left(\theta \mid d\right)$: the posterior probability of the parameter given the data. This is what we ultimately want to estimate
$p\left(d \mid \theta\right)$: is the likelihood. It is the probability of observing the data given a specific configuration of parameters
$p(\theta)$: is the prior probability. It is what is believed to be true before observing the data and quantifies existing external information or expert knowledge
$p(d)$: it is the marginal likelihood, the probability of observing the data under all possible hypotheses. This acts as a normalising constant to ensure the posterior probability is a valid probability

Table 4 How a prior is set

Priors represent the assumptions or guess about something before it is tested; past data can be used as a prior––otherwise, mathematical methods can establish them, e.g., based on expert opinions or general principles, or previous results can inform a subsequent study; priors can also be set to reflect no initial knowledge, or assume that all possibilities are equally likely. Whatever the choice, priors should be carefully and clearly formulated
Parameters represent the unknown quantities in a statistical model that we estimate from data. When specifying a probability distribution for a parameter, two key features define its space: location, e.g., the mean of a normal distribution represents our prior belief about where the true value lies, scale, e.g., variance, captures the expected dispersion around this central value, reflecting our uncertainty. Together, these determine the shape and range of plausible values before observing data. For instance, a prior for a normal distribution that can be used on a log-odds scale for a treatment effect measured in Odds Ratio (OR) would a have a μ and a σ term representing mean and standard deviation, respectively, it can be then written as N[μ,σ] and we can assign values to these terms according to our beliefs
When setting our prior belief about the true effect of a parameter, i.e., what we guess the benefit of a treatment or the mean of a certain quantitative variable could be, guidelines recommend using different values to reflect potential discrepancies among stakeholders. Minimal clinically important differences (MCID), if available, can inform this choice. A common framework reported by various authors is to set a range of sceptical, i.e., no effect is expected, optimistic, i.e., a benefit is expected and pessimistic, i.e., harm is expected
When setting our prior belief uncertainty according to how strong our beliefs are, we can use different priors: non-informative (or vague/flat or improper), – – used when there is little to no prior information available. They are designed to have minimal influence on the posterior distribution, allowing the data to primarily drive the inference; weakly informative – – they span a realistic range of parameter values but are not completely flat and are often used to prevent unrealistic parameter values from having undue influence while still allowing the data to dominate the inference; informative priors – – they incorporate substantial external information or expert knowledge. They can be based on previous studies, historical data, or expert elicitation. Informative priors can lead to more precise estimates, particularly with small sample sizes

dedicated probabilistic programming languages (Stan n.d.; JAGS n.d.) and specialised packages for popular statistical software such as R or Stata have transformed this limitation into a key strength for posterior estimation. In Table 5, we provide an overview of the most used algorithms and software for implementing Bayesian analysis. Briefly, Monte Carlo methods are a family of computational techniques that use repeated random sampling to approximate quantities that would be difficult or impossible to calculate directly. In a Bayesian context, we can explore the range of plausible parameter values by simulating many scenarios, yielding practical estimates of posterior distributions without requiring complex mathematics.

The *posterior* represents an updated belief about the parameters after considering both prior knowledge and observed data. Such posterior is a probability distribution from which we can obtain point estimates, e.g., mean, median, and Credible Intervals (CrIs), and we can even use it as the new prior for the next trial. It allows us, for instance, to assess the probability that one treatment is superior to another and the likelihood of achieving a clinically meaningful difference. it may be interesting to determine whether the effect of an intervention is significant in practical terms rather than statistically significant. Instead of relying solely on confidence/credible intervals, we can define a range of practical equivalence (ROPE)

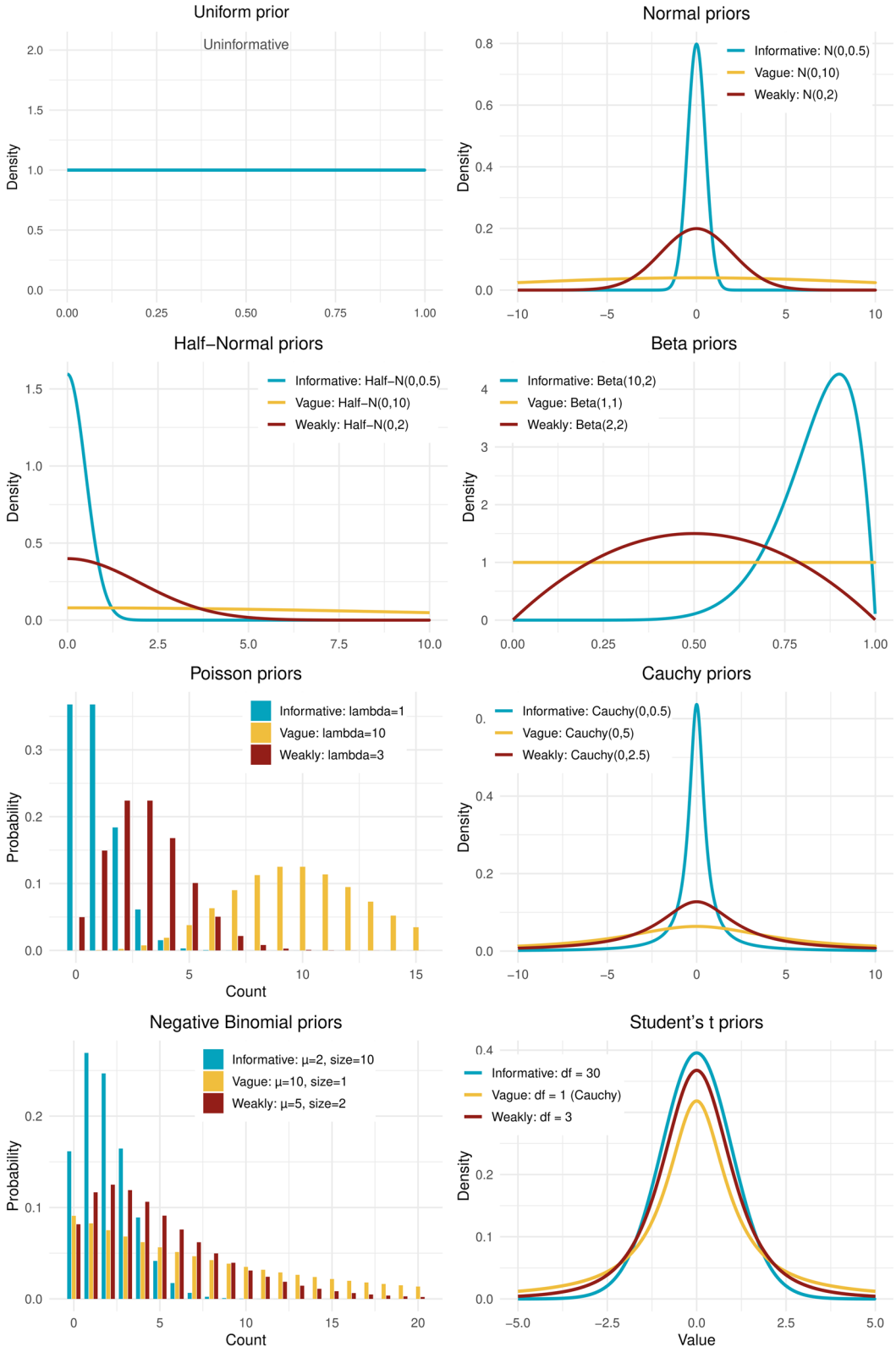

Fig. 1 (See legend on next page.)

(See figure on previous page.)

Fig. 1 Graphical overview of common priors showing how uninformative and informative priors are constructed. Uniform prior: Assumes all parameter values are equally likely ("flat" or "uninformative"). Used when no prior knowledge exists; results align with frequentist methods. Normal prior: its location parameter is the average, i.e. the highest point of the curve; the scale parameter is the standard deviation, which tells how much the data varies from the average; a *half–normal* distribution resulting from cutting a normal bell curve in half, keeping only the part that is above or below the mean, is commonly used to define the heterogeneity parameters in a random effect model. Beta prior: Models probabilities between 0 and 1 (e.g., bias of a coin). Useful for bounded parameters like proportions. Poisson prior: Counts events in fixed time–space, shaped by the average occurrence rate. Cauchy prior: Symmetric like the normal but with heavier tails, accommodating outliers. A half-Cauchy distribution, like a half-normal distribution, can define heterogeneity. Negative binomial prior: Models overdispersed count data in which the variance exceeds the mean, offering flexibility for rare but variable events via dispersion and success-probability parameters. Student t prior: Heavy-tailed alternative to the normal, robust to outliers. Degrees of freedom (df) control tail thickness: low df = vague priors, high df ≈ normal. Used for location/regression coefficients with rare extremes

Table 5 Common methods for approximate Bayesian inference

In many realistic settings, Bayesian inference cannot be performed analytically. Several approximate computational methods have been developed to estimate the posterior distribution of parameters of interest:
1. Markov Chain Monte Carlo (MCMC): The workhorse of Bayesian computation. MCMC generates a sequence of dependent samples from the posterior distribution using algorithms such as *Metropolis–Hastings*, which propose new values and accept them based on the posterior density ratio, or Gibbs Sampling, which samples each parameter conditionally, one at a time. *Pros*: General-purpose and widely applicable. *Cons*: Computationally intensive, slow convergence in high dimensions, autocorrelated samples, requires careful convergence diagnostics, e.g., trace plots, R-hat.
2. Hamiltonian Monte Carlo (HMC)An advanced MCMC method that uses gradient information of the log-posterior to make informed proposals and reduce random walk behaviour. Implemented in software like Stan using the No-U-Turn Sampler (NUTS). *Pros*: Much more efficient than standard MCMC, especially in high-dimensional models. *Cons*: Requires gradient evaluations; tuning parameters are critical.
3. Variational Inference (VI)Reframes Bayesian inference as an optimisation problem. VI posits a simpler family of distributions (e.g., multivariate normal) and minimises a divergence metric, often Kullback–Leibler divergence, between this family and the true posterior. *Pros*: Faster than MCMC; scales well to large datasets. *Cons*: Can underestimate uncertainty and perform poorly when posteriors are multimodal or skewed. Less robust in clinical contexts where decision-making under uncertainty is critical.
4. Integrated Nested Laplace Approximation (INLA)A deterministic alternative to MCMC, developed for latent Gaussian models (e.g., spatial, temporal models). INLA approximates marginal posteriors using nested Laplace approximations. *Pros*: Highly efficient and accurate for specific model classes, e.g., spatial or time-series. *Cons*: Not suitable for arbitrary Bayesian models; limited flexibility beyond latent Gaussian frameworks.

that encompasses values so similar in effect that they're functionally equivalent. For instance, when comparing an intervention to standard care, if the effect falls within the ROPE, the two strategies are considered equally effective. The ROPE is study-specific, informed by prior evidence, expert judgment, or clinical relevance. Effects inside this range are deemed trivial, even if statistically detectable. By examining how much the 95% highest density interval (HDI) of a posterior distribution overlaps with the ROPE, we can better gauge whether an intervention is likely beneficial, harmful, or inconsequential in practice (Kruschke 2018). To illustrate these principles, we report a posterior distribution from simulated data simulate data in Fig. 2.

Bayesian methods applied

Bayesian methods were first applied in critical care literature to reanalyse results from RCTs (Zampieri et al., 2021; Goligher et al., 2018; Granholm et al., 2020; Zampieri et al., 2020; Schouteden et al., 2025). These methods are particularly valuable in critical care, where small sample sizes often fail to exclude clinically important mortality differences or when frequentist interim analysis could potentially lead to untimely recruitment stopping (Ferreira and Meyer 2019). A recent review of intensive care trials (Yarnell et al., 2021) found that while frequentist and Bayesian methods are generally in agreement, some trials showed a high likelihood of clinical benefit despite nonsignificant frequentist results, while others exhibited low probabilities of benefit despite statistically significant effects. Also, certain trials demonstrated substantial sensitivity to the choice of prior distribution, indicating insufficient data to conclusively determine treatment effects. These cases illustrate how integrating Bayesian analyses into trial reporting could improve interpretation and reduce ambiguity, and have been applied recently in perioperative medicine as well (Mazzinari et al., 2025a, 2025b).

A Bayesian framework enhances meta-analysis. As we outlined, Bayesian models are built explicitly on the idea that the effect of interest is itself uncertain and should be treated as a random quantity. In other words, the true effect size is not an unknown but fixed number; rather, it is something we model probabilistically. Therefore, all uncertainty, including that arising from small study sizes, heterogeneity between studies, and missing or sparse data, can be integrated coherently into a single model (Friede et al., 2017; Rhodes et al., 2016). Also, Bayesian models supply a robust framework compared to the standard meta-analytic methods, which can have issues in accurately estimating confidence intervals and have rigid assumptions about the studies' heterogeneity (Cornell et al., 2014). Network meta-analysis (NMA) extends pairwise meta-analysis by comparing multiple treatments

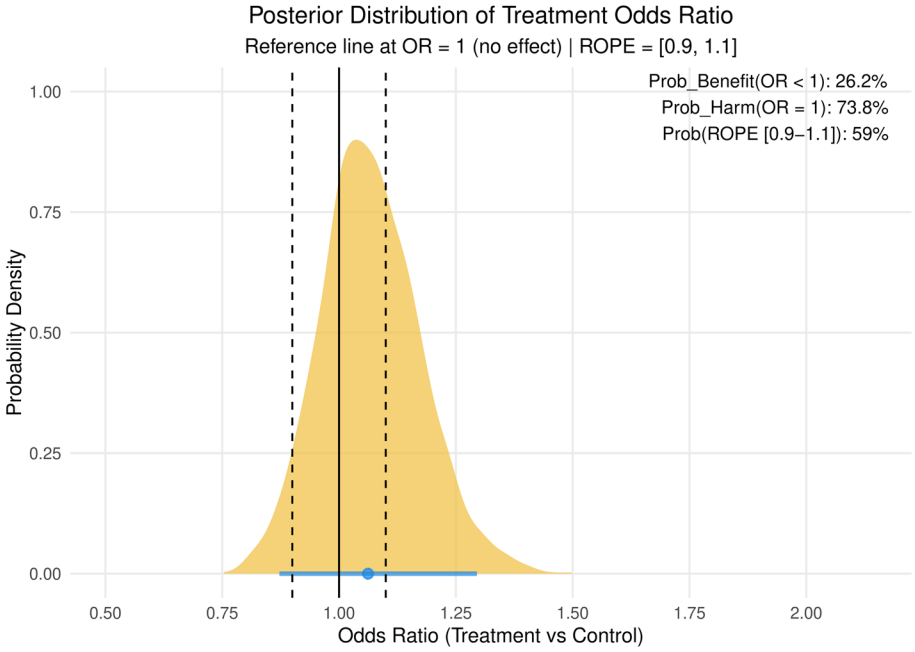

Fig. 2 Posterior distribution of the odds ratio (OR) fitting a logistic model using a sceptical prior: The outcome variable is a hypothetical binary variable for complications, and the dependent variable is a binary variable for a hypothetical beneficial treatment; thus OR > 1 represents a harmful effect for the treatment. We simulated 2000 patients assigned 1:1 to either the control or treatment group, and the true effect was set at OR = 1.1. The ROPE is defined as the OR between 1/1.1 and 1.1. The 95% High-Density Interval (HDI) is reported as a blue line in each panel

simultaneously, including treatments that have never been compared head-to-head in any study. Bayesian hierarchical modelling in this context provides a clean way to account simultaneously for: the correlations induced by multi-arm trials, the consistency assumption, i.e., the idea that indirect and direct evidence about the same treatment comparison should agree, the estimation of relative treatment effects for all treatment pairs and the ranking of treatments. Bayesian NMA is an active field in perioperative medicine with several publications in recent years (Jivraj et al., 2025; Schorer et al., 2025; Li et al., 2025; Zhou et al., 2025).

The Bayesian framework can also be used to directly inform the design of RCTs. Again, all unknowns, i.e., treatment effects, future outcomes, and even trial-design adaptations, are modelled by probability distributions that are updated continuously as data accrue, and the probability that an experimental therapy is superior (or inferior) given the observed data is computed. This allows interim and even continuous monitoring without penalty, using predictive probabilities to guide decisions such as early stopping for futility or efficacy, adaptive randomization to favour better-performing arms, and modifications to accrual or patient-subgroup focus based on emerging biomarker signals Hierarchical Bayesian models further extend this flexibility by borrowing "strength" from historical or parallel trials—down-weighting disparate data but pooling when concordance suggests exchangeability—while formally adjusting for between-trial heterogeneity. Such models can integrate patient-level covariates and early auxiliary endpoints to sharpen inference on long-term outcomes, improving both ethical treatment of participants and trial efficiency. Bayesian design has been used in intensive care (Francio et al., 2025; Perkins et al., 2018) and cardiovascular medicine trials (Holmes et al., 2009; Reardon et al., 2017; Popma et al., 2019; Foley et al., 2024; Mistry et al., 2023), but it remains relatively uncommon in perioperative medicine (Fields et al., 2025; Reitz et al., 2025).

Bayesian pitfalls and scientific methodology

Bayesian methods in clinical trials have several recognised limitations. They often involve intensive computation, commonly relying on MCMC algorithms, which can be burdensome for large hierarchical models and typically require careful convergence monitoring. Results also depend sensitively on the prior specification: subjective or poorly justified priors can substantially influence conclusions, and there is no universally objective rule for choosing noninformative priors. These factors add methodological complexity and raise a learning barrier for clinicians: indeed, a lack of expertise has been identified as a challenge, since unfamiliarity may increase the risk of misapplication. Bayesian approaches remain a powerful option, though their assumptions and limitations should be explicitly acknowledged in trial design and analysis. Several guidelines on how to carry out an analysis workflow (Gabry et al., 2019; Gelman 2020) and on reporting

results are available (Kruschke 2021; Sung et al., 2005; Spiegelhalter et al., 2000).

Some critics of Bayes frame the debate into a forced choice between evidence and belief. Science aims to shape well-justified beliefs from evidence rather than to remove them entirely. Bayesian analysis helps physicians assess how trial results should update their confidence in a treatment's benefit and guide action. Also, judgment is unavoidable in scientific inference. Data interpretation demands expert reasoning, not just automatic reliance on an arbitrary threshold. In this regard, there is also a debate on whether the Bayes factor should be routinely used, with some authors in favour (Goodman 1999b) and others against (Gelman 2008; Robert 2015), and there are even other schools of thought apart from frequentist or Bayesian being proposed as a solution, like the evidential approach (Zampieri et al., 2025). In the end, no statistical method is flawless. Researchers should prioritise approaches that solve the most real-world problems with the fewest drawbacks.

Looking ahead, Bayesian methods offer real promise for medical research by providing clearer, more intuitive ways to quantify uncertainty and update evidence as new data emerge. They can support more flexible study designs, make better use of small or evolving datasets, and communicate results in terms that align with clinical decision-making. Nevertheless, to get the most out of them, we need to use them carefully, be open about how we choose our priors, explain results in simple and clear language, and avoid the kind of oversimplification that led to widespread confusion with frequentist statistics. If we do that, Bayesian methods can become a powerful and trustworthy tool for the field.

Acknowledgements
Not applicable.

Authors' contributions
GM: Concept, design, writing first draft of manuscript, study guarantor, critical revision of manuscript for important intellectual content, final version approval. FGZ: Design, literature review, writing first draft of manuscript, critical revision of manuscript for important intellectual content, final version approval. MOH: Critical revision of manuscript for important intellectual content, and final version approval. MJS: Critical revision of manuscript for important intellectual content, and final version approval. DVM: Critical revision of manuscript for important intellectual content, and final version approval. ASN: Critical revision of manuscript for important intellectual content, and final version approval.

Funding
MH partially supported by NIH/NHLBI R01-HL168202. No others are declared.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 25 July 2025 / Accepted: 13 February 2026
Published online: 24 February 2026

References

Betancourt M. A Conceptual Introduction to Hamiltonian Monte Carlo. arXiv; 2018. Available from: http://arxiv.org/abs/1701.02434. Accessed 16 Jul 2025.
Brooks S, Gelman A, Jones G, Meng XL. Handbook of Markov Chain Monte Carlo. 1st ed. New York: Chapman and Hall/CRC; 2011.
Casella G, George EI. Explaining the Gibbs sampler. Am Stat. 1992;46(3):167–74.
Cornell JE, Mulrow CD, Localio R, Stack CB, Meibohm AR, Guallar E, et al. Random-effects meta-analysis of inconsistent effects: a time for change. Ann Intern Med. 2014;160(4):267–70.
De Grooth HJ, Cremer OL. Bayes and the evidence base: reanalyzing trials using many priors does not contribute to consensus. Am J Respir Crit Care Med. 2024;209(5):483–4.
De Grooth HJ, Elbers P. Pick your prior: scepticism about sceptical prior beliefs. Intensive Care Med. 2022;48(3):374–5.
Efron B. Why isn't everyone a Bayesian? Am Stat. 1986;40:1–5.
Feinstein AR. *P*-Values and confidence intervals: two sides of the same unsatisfactory coin. J Clin Epidemiol. 1998;51(4):355–60.
Ferreira D, Meyer N. Post hoc Bayesian analyses. JAMA. 2019;321(16):1632.
Fields BC, Soliz JM, Speer BB, Hancher-Hodges S, Popat KU, Ghebremichael SJ, et al. Repeat versus single quadratus lumborum block to reduce opioids after open pancreatectomy (RESQU-BLOCK): a randomized clinical trial. Ann Surg. 2025;283(2):212–8. https://doi.org/10.1097/SLA.0000000000006767.
Foley MJ, Rajkumar CA, Ahmed-Jushuf F, Simader FA, Chotai S, Pathimagaraj RH, et al. Coronary sinus reducer for the treatment of refractory angina (ORBITA-COSMIC): a randomised, placebo-controlled trial. The Lancet. 2024;403(10436):1543–53.
Friede T, Röver C, Wandel S, Neuenschwander B. Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases. Biom J. 2017;59(4):658–71.
Gabry J, Simpson D, Vehtari A, Betancourt M, Gelman A. Visualization in Bayesian workflow. J R Stat Soc Ser A Stat Soc. 2019;182(2):389–402.
Gelman A. Objections to Bayesian statistics. Bayesian Anal. 2008;3(3):445–9.
Gelman A, Greenland S. Are confidence intervals better termed "uncertainty intervals"? BMJ. 2019;366(10):l5381.
Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. 3rd edition. CRC press; 2013.
Gelman A, Vehtari A, Simpson D, Margossian CC, Carpenter B, Yao Y, et al. Bayesian Workflow. arXiv. 2020. Available from: http://arxiv.org/abs/2011.01808. Accessed 18 Jul 2025.
Goligher EC, Harhay MO. What is the point of Bayesian analysis? Am J Respir Crit Care Med. 2024;209(5):485–7.
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.
Gómez-Rubio V. Bayesian inference with INLA. 1st ed. Chapman and Hall/CRC press; 2020.
Goodman SN. Toward evidence-based medical statistics. 1: The *P* value fallacy. Ann Intern Med. 1999a;130:995–1004.
Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med. 1999b;130(12):1005–13.
Goodman SN. Why is Getting Rid of *P*-Values So Hard? Musings on Science and Statistics. Am Stat. 2019;73(sup1):26–30.
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(4):717–26.
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(sup1):106–14.

Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31(4):337–50.
Harrell F. Statistical Thinking. 2024. Traditional Frequentist Inference Uses Unrealistic Priors. Available from: https://www.fharrell.com/post/uprior/. Accessed 16 Jul 2025.
Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970;57(1):97.
Heuts S, Kawczynski MJ, Sayed A, Urbut SM, Albuquerque AM, Mandrola JM, et al. Bayesian Analytical Methods in Cardiovascular Clinical Trials: Why, When, and How. Can J Cardiol. 2025;41(1):30–44.
Holmes DR, Reddy VY, Turi ZG, Doshi SK, Sievert H, Buchbinder M, et al. Percutaneous closure of the left atrial appendage versus warfarin therapy for prevention of stroke in patients with atrial fibrillation: a randomised non-inferiority trial. Lancet. 2009;374(9689):534–42.
JAGS - Just Another Gibbs Sampler. n.d. Available from: https://mcmc-jags.sourceforge.io/. Accessed 16 Jul 2025.
Jivraj NK, Lakbar I, Sadeghirad B, Müller MM, Sohn SY, Peel JK, et al. Intra-operative ventilation strategies and their impact on clinical outcomes: a systematic review and network meta-analysis of randomised trials. Anaesthesia. 2025;80(8):973–87.
Kallioinen N, Paananen T, Bürkner PC, Vehtari A. Detecting and diagnosing prior and likelihood sensitivity with power-scaling. Stat Comput. 2024;34:57.
Kruschke JK. Doing bayesian data analysis: a tutorial with R, JAGS, and Stan. 2nd ed. Academic Press; 2014.
Kruschke JK. Rejecting or accepting parameter values in Bayesian estimation. Adv Methods Pract Psychol Sci. 2018;1(2):270–80.
Kruschke JK. Bayesian analysis reporting guidelines. Nat Hum Behav. 2021;5(10):1282–91.
Li H, Wang H, Wang C, Huang Y, Yuan R, Zhao X, et al. Comparison of clinical characteristics of different ventilation devices for one-lung ventilation in adults: a network meta-analysis. Int J Surg. 2025;111(6):3989–4001.
Matthews R. The *p*-value statement, five years on. Significance. 2021;18(2):16–9.
Mazzinari G, Díaz-Cambronero O, Garutti I, Errando CL, Ferrando C, Spadaro S, et al. Impact of neuromuscular block monitoring and reversal on postoperative pulmonary complications in thoracic surgery: a Bayesian analysis of the iPROVE-OLV trial. Br J Anaesth. 2025a;135(5):1428–40 Online ahead of print.
Mazzinari G, Zampieri FG, Ball L, Campos NS, Bluth T, Hemmes SNT, et al. High Positive End-expiratory Pressure (PEEP) with Recruitment maneuvers versus low PEEP during general anesthesia for surgery: a Bayesian individual patient data meta-analysis of three randomized clinical trials. Anesthesiology. 2025b;142(1):72–97.
McElreath R. Statistical rethinking. A bayesian course with examples in R and Stan. 2nd Edition. CRC press; 2020.
Mistry EA, Hart KW, Davis LT, Gao Y, Prestigiacomo CJ, Mittal S, et al. Blood Pressure Management After Endovascular Therapy for Acute Ischemic Stroke: The BEST-II Randomized Clinical Trial. JAMA. 2023;330(9):821.
Perkins GD, Ji C, Deakin CD, Quinn T, Nolan JP, Scomparin C, et al. A Randomized Trial of Epinephrine in Out-of-Hospital Cardiac Arrest. N Engl J Med. 2018;379(8):711–21.
Popma JJ, Deeb GM, Yakubov SJ, Mumtaz M, Gada H, O'Hair D, et al. Transcatheter Aortic-Valve Replacement with a Self-Expanding Valve in Low-Risk Patients. N Engl J Med. 2019;380(18):1706–15.
Prior Distributions for rstanarm Models. n.d. Available from: https://mc-stan.org/rstanarm/articles/priors.html#how-to-specify-flat-priors-and-why-you-typically-shouldn-t-. Accesed 16 Jul 2025.
Reardon MJ, Van Mieghem NM, Popma JJ, Kleiman NS, Søndergaard L, Mumtaz M, et al. Surgical or Transcatheter Aortic-Valve Replacement in Intermediate-Risk Patients. N Engl J Med. 2017;376(14):1321–31.
Reitz KM, Nassereldine H, Kennedy J, Zeh R, Khandwala F, Seymour CW, et al. Strategies to promote resiliency: a randomized embedded multifactorial adaptative platform (REMAP) clinical trial to study interventions to improve recovery after surgery in high-risk patients. Ann Surg Open. 2025;6(2):e566.
RENOVATE Investigators and the BRICNet Authors, Francio F, Weigert RM, Mattei EDB, Grion CMC, Festti J, et al. High-flow nasal oxygen vs noninvasive ventilation in patients with acute respiratory failure: the RENOVATE randomized clinical trial. JAMA. 2025;333(10):875.
Rhodes KM, Turner RM, White IR, Jackson D, Spiegelhalter DJ, Higgins JPT. Implementing informative priors for heterogeneity in meta-analysis using meta-regression and pseudo data. Stat Med. 2016;35(29):5495–511.
Robert CP. The expected demise of the Bayes factor. arXiv. 2015. Available from: http://arxiv.org/abs/1506.08292. Accessed 18 Jul 2025.
Schorer R, Ibsen A, Hagerman A, Ellenberger C, Putzu A. Diagnostic accuracy of vascular ultrasonography for postanesthesia induction hypotension: a systematic review and network meta-analysis. Anesth Analg. 2025;141(1):26–37.
Schouteden E, Heuts S, Bels JLM, Thiesen S, Van Gassel RJJ, Lee ZY, et al. The impact of high versus standard enteral protein provision on functional recovery following intensive care admission: a pre-planned Bayesian analysis of the PRECISe trial. Clin Nutr. 2025;48:153–60.
Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomized trials. J R Stat Soc Ser A Stat Soc. 1994;157(3):357.
Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Methods in health service research: an introduction to Bayesian methods in health technology assessment. BMJ. 1999;319(7208):508–12.
Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Bayesian methods in health technology assessment: a review. Health Technol Assess Winch Engl. 2000;4(38):1–130.
Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health care evaluation. Chichester; Hoboken, NJ: Wiley; 2004.
Stan: software for Bayesian analysis. n.d. Available from: https://mc-stan.org/. Accessed 16 Jul 2025.
Stefan AM, Schönbrodt FD. Big little lies: a compendium and simulation of*p*-hacking strategies. R Soc Open Sci. 2023;10(2):220346.
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(3):261–8.
Van Zwet E, Gelman A, Greenland S, Imbens G, Schwab S, Goodman SN. A new look at *P* values for randomized clinical trials. NEJM Evid. 2024;3(1):EVIDoa2300003 Epub 2023 Dec 22.
Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "$p < 0.05$." Am Stat. 2019;73(sup1):1–19.
Wasserstein RL, Lazar NA. The ASA statement on *p*-Values: context, process, and purpose. Am Stat. 2016;70(2):129–33.
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):207–16.
Zampieri FG, Casey DJ, Shankar-Hari M, Harrell FE, Harhay MO. 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. Am J Respir Crit Care Med. 2021;203(5):543–52.
Zampieri FG, Cahusac PMB, Maia IS, Yehya N, Meyer NJ, Li F, et al. Trial analysis and interpretation in critical care using the evidential (likelihood) approach: rationale and practical considerations. Am J Respir Crit Care Med. 2025;211(9):1610–21. https://doi.org/10.1164/rccm.202504-0809TR.
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):423–9.
Zhou S, Yu S, Bi Y, Tian Z, Pan R, Yan T, et al. The safety and efficacy of remimazolam, ciprofol, and propofol anesthesia in endoscopy: a systematic review and network meta-analysis. BMC Anesthesiol. 2025;25(1):230.

Publisher's Note