HHS Public Access

Author manuscript

Lancet. Author manuscript; available in PMC 2025 September 14.

Published in final edited form as:

Lancet. 2024 September 14; 404(10457): 1067–1076. doi:10.1016/S0140-6736(24)01295-9.

Bayesian Statistics for Clinical Research

Ewan C. Goligher, MD[1,2,3,4], **Anna Heath, PhD**[5,6,7], **Michael O. Harhay, PhD**[8,9]

[1]Interdepartmental Division of Critical Care Medicine, University of Toronto, Toronto, Canada

[2]Department of Medicine, Division of Respirology, University Health Network, Toronto, Canada

[3]Toronto General Hospital Research Institute, Toronto, Canada

[4]Department of Physiology, University of Toronto, Toronto, Canada

[5]Child Health Evaluative Sciences, The Hospital for Sick Children, Toronto, ON, Canada

[6]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

[7]Department of Statistical Science, University College London, London, UK

[8]Departments of Biostatistics, Epidemiology, and Informatics; and of Medicine, University of Pennsylvania, Philadelphia, PA

[9]MRC Clinical Trials Unit, University College London, London, UK

Abstract

Frequentist and Bayesian statistics represent two differing paradigms for analysis of data. Frequentist thinking became the dominant mode of statistical thinking in medical practice during the 20th century. The advent of modern computing has made Bayesian analysis techniques increasingly accessible, enabling growing use of Bayesian methods in a range of disciplines, including medical research. Rather than conceiving of probability as the expected frequency of an event (purported to be measurable and objective), Bayesian thinking conceives of probability as a measure of the strength of a belief (an explicitly subjective concept). Bayesian analysis combines prior information (represented by a mathematical probability distribution, the prior) with information from the study (likelihood) to generate an updated probability distribution (the posterior) representing the available information for clinical decision-making. Owing to its fundamentally different conception of probability, Bayesian statistics offers an intuitive, flexible, and informative approach that facilitates the design, analysis, and interpretation of clinical trials.

Address for Correspondence: Dr. Ewan Goligher, Toronto General Hospital, 585 University Ave., 9-MaRS-9024, Toronto ON M5G 2N2.

Contributors

E.C.G. developed the idea for this review. All authors contributed to the development of the proposed content. E.C.G. wrote the first draft of the review, and all authors contributed to and approved the final draft.

Search strategy and selection criteria

Data for this Review were identified by searches of MEDLINE, PubMed, and references from relevant articles using the search terms “Bayesian”, “probability theory”, and “frequentism”. Articles or books published in English between 1955 and 2023 were included.

Herein, we provide a brief account of the philosophical and methodlogical differences between Bayesian and frequentist approaches and survey the use of Bayesian methods for the design and analysis of clinical research.

> "They say that Understanding ought to work by the rules of right reason. These rules are, or ought to be, contained in Logic; but the actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the probability which is, or ought to be, in a reasonable man's mind."
>
> -James Clerk Maxwell[1]

> "Bayes' rule can be described in one sentence: by updating our initial beliefs with objective new information, we get a new and improved belief. To its adherents, it is an elegant statement about learning from experience. To its opponents, it is subjectivity run amok."
>
> -Sharon Bertsch McGrayne, *The Theory that Would Not Die*[2]

Introduction

In practicing medicine, the physician is often faced with considerable uncertainty, whether in making a diagnosis, offering a prognosis, or recommending a therapy. Where there is uncertainty, there is room for error. Skilled physicians are accustomed to dealing with uncertainty by reasoning in probabilistic terms, calibrating their strength of belief in a particular diagnosis or prognosis according to the available evidence (i.e., history, physical, laboratory, and imaging data). Diagnostic investigation is a process of progressively reducing uncertainty about the diagnosis by accumulating information to reach a conclusion with a sufficiently high probability to guide therapy. The Bayesian statistical framework conceives of clinical research as something closely analogous to diagnostic investigation—an effort to progressively reduce uncertainty about a hypothesis. In this review, we will consider the emerging role of Bayesian statistics in the design and conduct of clinical research.

Statistical science is integral to clinical investigation. Over the 20th century, medical statistics came to be dominated by what is called the "frequentist" mode.[3] Frequentism has bequeathed us the all-familiar p-value, the confidence interval, the notion of statistical significance, Type I and II errors, and the power calculation. Yet despite the fact that some modicum of exposure to conventional frequentist statistics is ubiquitous during scientific and clinical training, and although we are accustomed to interacting with frequentist statistical constructs when we read a scientific publication (p-values, etc.), comparatively few people can offer precise and accurate working definitions for these statistical constructs.[4] The abstract and counterintuitive nature of these constructs belies their utility. The Bayesian framework offers an alternate, more intuitive paradigm for the design and analysis of clinical research. Indeed, Bayesian analysis may sometimes suggest a markedly different conclusion from frequentist analysis of the same data (see Panel 1). With the advent of

modern statistical computing, the practical application of Bayesian methods has become increasingly feasible in clinical research.

What is the Bayesian statistical framework, and how does it work?

Bayesian statistics is based on Bayes' theorem, which describes the fundamental relationship between evidence (data) and explanation (hypothesis). The theorem is given by:

$$P(Hypothesis \mid Data) = \frac{P(Data \mid Hypothesis)}{P(Data)} P(Hypothesis)$$

Where "P" stands for "probability" and "|" stands for "given." This framework for evaluating probability is attributed to the Rev. Thomas Bayes, an 18th century Presbyterian minister and mathematician.[5,6] It was independently described a few decades later by the famous French mathematician, Pierre-Simon Laplace, and for this reason is sometimes referred to as the Bayes-Laplace theorem.[2]

As is evident from the relation above, there are three components to Bayesian statistical reasoning: the "prior" probability of a given explanation (denoted as P[Hypothesis]), the likelihood of obtaining the observed data given that explanation (denoted as P[Data|Hypothesis]), and the posterior probability of a given explanation after combining the prior with the observed likelihood (denoted as P[Hypothesis|Data]). This theorem was originally referred to as "inverse probability" in that it allows one to "invert" the probability of data under a given hypothesis (P[Data|Hypothesis])—what we can measure—to obtain the probability of a hypothesis given data (P[Hypothesis|Data]). Indeed, this was the significance of Bayes' discovery of the theorem: it provided a formal method for drawing inferences about hypotheses based on observed probabilities in data.[7] That is, one could make claims about hypotheses, not just claims about observed data.

It is helpful to appreciate that, mathematically, the prior, likelihood, and posterior are usually represented not as single numbers but rather as probability density functions (or distributions) (Figure 1).[3] Using probability distributions allows us to quantify the uncertainty about a parameter (e.g., mortality rate). Suppose we have followed 100 people with a given disease state for one year, and 30 have died at follow-up. Using a beta distribution, combined with binomial data, we can generate a distribution for the range of possible values for the one-year mortality rate in the population (Figure 1). With this distribution, we can, for example, estimate the probability that the true value for the mortality rate in the population lies between 0·25 and 0·35 (by computing the area under the curve in this region). The spread of the distribution tells us how much information or uncertainty is represented by the data: as information (e.g. sample size and number of events) increases, the distribution narrows (lower variance), signifying greater information and decreased uncertainty.

Bayesian statistics proceeds by combining a probability density distribution representing prior information (the "prior") with a probability density distribution representing the data (the "likelihood") to obtain an updated probability density distribution (the "posterior")

(Figure 2).[3,7] The likelihood function must be selected for both Bayesian and frequentist analyses and is computed directly from the available data. The major challenge in Bayesian analysis is to determine how the prior should be defined.[8] The role of priors is an important source of controversy and resistance to the Bayesian approach, and we will consider this further below.

Fundamental conceptual differences between Bayesian and frequentist statistics

To understand why the Bayesian approach might be preferred to frequentist statistics, we need to briefly explore a number of deeper philosophical issues in scientific epistemology (theory of knowledge) (summarized in Table 1).

Frequentist statistical inference focuses entirely on the probability of observing data (P[Data|Hypothesis]).[9] In evaluating a statistical association (quantified by, for example, an odds ratio), it asks the question, "Given a null hypothesis of no association (odds ratio = 1), what is the probability of obtaining data with an odds ratio as or more extreme as the point estimate observed in this sample of data?" The idea is that if this probability is very low, then the null hypothesis can be rejected (an inferential process referred to as "null hypothesis significance testing").[10] For frequentists, the concept of "probability" has a very specific meaning—it refers to the frequency of observing the results in a hypothetical infinite series of repeated trials.[11] The frequentist null hypothesis is defined in terms of a sampling distribution (Figure 2), which represents the range of frequencies at which data would be *expected* under the null hypothesis, if the study was repeated over and over *ad infinitum* in precisely the same manner. Although the study is conducted only once, the inference relies on the expectation that the study was hypothetically repeated many times in the same way, and the calculated probability equates to the frequency of the observed result in those hypothetical repeated trials.[12] For frequentists, probability statements are always and only statements about frequencies of data. Thus the "p-value" is the frequency at which data as or more extreme than that observed in the study would be expected if the null hypothesis were true and the study was repeated an infinite number of times.[12] The "confidence interval" is the interval generated from a sample of data such that in repeated samples, 95% of the time the estimated interval would contain the true value.[13] Crucially, one cannot infer from a confidence interval that there is a 95% probability that the true value lies within the reported interval (despite a widespread propensity to do so) because that's a statement about the strength of belief in a hypothesis, not about frequencies of obtaining data, and requires an implicit assumption of no prior information (see below).

The Bayesian approach is markedly different: it conceives of probability primarily as a "degree of belief."[14,15] To assert something with a high probability is to profess confident belief and low uncertainty. To suggest that something has a 50% probability is to express maximum uncertainty. This Bayesian conception of probability aligns with the usage of the term in everyday parlance. We may speak of the probability of rain this afternoon, the probability of winning in sports tonight, or the probability of electoral success this year (all one-off, non-repeatable events which can have no "frequency"). In clinical medicine,

we speak of pre- and post-test probability to refer to our strength of belief in a given diagnosis and to describe how that strength of belief is modified by diagnostic testing.[16] These are not statements about frequencies of outcomes in repeated events, but about belief in the outcomes of one-time occurrences or cases. Given this conception of probability as a "degree of belief", it is sensible to Bayesians to consider prior probability, because we approach studies with some degree of belief for or against a hypothesis before undertaking the study to test the hypothesis (even if we are maximally uncertain with a probability of 50%). Indeed, the exercise of expressing prior beliefs mathematically as a prior distribution forces us to think carefully about the pre-existing evidence and the appropriate level of uncertainty in the face of that prior information.

Frequentists have traditionally rejected this conception of probability on philosophical grounds.[9,11,12,14] They insist that scientific inference using probabilities should be based on objective ("world-based") statements about data, rather than subjective ("mind-based") statements about degrees of belief.[14] But there is a certain irony in such insistence. Frequentist inference is based on the expected frequencies of data that would be obtained if the study was hypothetically repeated many times exactly according to the intentions of the investigator.[17] So the interpretation of the data very much depends on the state of mind (intentions and expectations) of the investigator according to hypothetical (imaginary) repetitions of the study together with individual judgments about the exact analysis methods to be used.[18] It's not at all clear that we can separate statistical inference from human psychology, even under the frequentist paradigm.

Crucially, because frequentist inference is based on the expected frequency of observed data under many hypothetical repetitions of the study with the same sample size and primary outcome, frequentist study design requires that the study be conducted precisely according to the original intentions of the investigator: the sample size is fixed for the analysis, and the primary outcome (primary hypothesis test) must be pre-specified. If the actual study conduct deviates from those original intentions, the validity of statistical inference is seriously threatened.[17]

By contrast, in Bayesian inference, the "degree of belief" depends only on the available information from the study data and the prior. The available information from the study data is entirely encompassed by the likelihood function, a statistical axiom referred to as the "likelihood principle."[19–21] Under the likelihood principle, the intentions of the investigator are irrelevant to the information available in the study data. Hence, issues such as the original planned sample size and stopping rule, which outcome was deemed primary —features of the investigators' state of mind—are not relevant to assessing the hypothesis using the study data and do not affect the Bayesian posterior distribution. Provided the prior is appropriately specified, the computed posterior distribution is not affected by the reasons for stopping the study or whether the pre-specified sample size was obtained.

Specifying prior distributions

The incorporation of prior probabilities in Bayesian data analysis is a perpetually controversial issue. For some, using information from outside the study to evaluate the

study's meaning risks making science more subjective than objective, particularly since investigators may disagree about the most appropriate prior by virtue of their differing biases and perspectives.

Yet anyone familiar with scientific endeavour will understand that data are always interpreted in a broader context. Journals invite editorials from experts who can set the study in context, and presentations of important new clinical trial data at international meetings are often accompanied by editorials and panel discussions. The "Research in Context" panel in most Lancet family journals suggests a Bayesian approach–there is "Evidence before this study" (the prior), the "Added value of the study" (the likelihood), and the "Implications of all the available evidence" (the posterior). Priors seem to have a real place in the interpretation of data, the major question is whether they can be specified quantitatively in a valid and reproducible manner that appropriately reflects prior information.

In general, the goal of specifying a prior is to represent the plausible and defensible degree of belief in a hypothesis amongst the clinical community prior to the reporting of new study results. The most basic prior is a neutral prior centered on the null with a relatively wide variance reflecting the appropriate level of uncertainty in a hypothesis.[22,23] Such a prior is generally appropriate in the analysis of clinical trials since it reflects the clinical uncertainty and equipoise that motivated the trial. Alternately, non-informative ("flat") priors, which regard all possible values of the parameter as equally likely, can be used. In theory, these priors result in posterior distributions that are entirely dependent on the likelihood function computed from the study results. In practice, defining these non-informative priors can be challenging but guidance is available to select reasonable non-informative priors.[24] Because it is *prima facie* exceedingly unlikely that all possible values of the parameter are equally likely at the outset of a trial (e.g., an odds ratio of 0·1 seems much less likely than an odds ratio of 0·9 since very large effects are much less common than small treatment effects), the non-informative prior is generally not an appropriate prior for a primary analysis (though it may be useful as a sensitivity analysis to assess the posterior under the likelihood alone). Specifying the prior to represent the prior belief of a skeptic or pessimist about treatment effect can also be a powerful means of showing the evidence is persuasive if the posterior probability of benefit remains high.[25]

Priors can be specified directly based on previous data.[8,23] Data reported in previous clinical trials or meta-analyses can be used to construct the prior distribution. If the relevance of previous studies to the study under analysis is uncertain, the influence of the prior can be reduced (down-weighted) by inflating its variance.[8] One modified example of such a prior is to borrow historical information on control group event rates.[26,27] Priors can also be elicited empirically from stakeholders, including clinicians, researchers, and patient and caregiver representatives; a prior reflecting consensus can be generated by aggregating across these elicited distributions.[28,29] One common strategy for prior construction is to specify a range of priors representing varying degrees of enthusiasm or skepticism to assess the sensitivity of the conclusion (the posterior) to varying prior beliefs.[7,25,30] For example, the post hoc Bayesian re-analysis of EOLIA (see Panel 1) employed multiple strategies, including non-informative prior, priors reflecting varying degrees of a priori enthusiasm and skepticism about the benefit of early ECMO, and a prior derived from a meta-analysis of previous

studies. Finding a high posterior probability of benefit across all these priors strengthens one's confidence in the benefit of the intervention.

Prior specification requires a great deal of care and forethought, and each element of a prior distribution (its center, shape, and variance) should be carefully and explicitly justified.[25] Priors should be fully pre-specified in advance of unblinding the data to ensure that the prior is not influenced by knowledge of the study findings (i.e., it is a true prior).

Interpreting posterior distributions

Once the posterior distribution is estimated, a wide range of information can be gleaned from it. Typically, the "central tendency" of the posterior distribution is described using either the posterior mean or median value, depending on the shape of the distribution. The spread of the distribution is summarized by the credible interval. Unlike the frequentist confidence interval, the credible interval has a straightforward interpretation—the plausible range of values for the parameter given the available information.[9] For example, the 95% credible interval can be interpreted as communicating that there is a 95% probability that the true value lies within the interval (Table 1). The 95% interval is often reported (mainly because we are used to seeing 95% confidence intervals designed to correspond to a Type I error rate of 5%). However, 90% or 99% credible intervals are also frequently reported in Bayesian analyses.

The posterior distribution can also be used to compute the probability of benefit or harm (e.g., odds ratio <1 or >1, respectively in Figure 2) based on the area under the curve above or below the chosen threshold. Often, a region of practical equivalence (or zone of indifference) will be specified based on the minimum clinically relevant effect (Figure 2).[30] The posterior distribution can be used to estimate the probability that the effect exceeds the minimum clinically relevant effect or possibly even larger effects if of interest. Conversely, the probability that the effect does not exceed the minimum clinically relevant effect (sometimes called, 'futility') or the probability of harm can also be computed. All these computations are obtained from the same posterior distribution, emphasizing the interpretive value of this approach.[31]

Drawing conclusions from Bayesian analyses

Bayesian methods provide both a challenge and an opportunity when we are aiming to draw definitive conclusions from data. The standardized p-value threshold of 0·05 allows a declaration of a "significant" study result, without considering the nuances of the clinical question. However, the flexibility of Bayesian methods allows (and requires) researchers to incorporate the clinical context into the analysis before declaring a substantive conclusion from the data.[7,31]

A key question in interpreting the results of a clinical trial is what posterior probability of benefit should be sufficient to drive change in practice. There are no arbitrarily defined standard thresholds for posterior probability to declare a "practice-changing" result.[21] Rather, study investigators must provide a context-specific interpretation of these values. For example, in a clinical area with comparatively fewer effective interventions, clinicians

might accept a lower posterior probability of efficacy for a new and inexpensive therapy than when considering an expensive novel intervention in a clinical area with a range of currently available effective interventions. For example, such considerations are essential to interpreting the results of the EOLIA trial (Panel 1), where a complex, invasive, and costly intervention offers a high probability of reducing the risk of death. Is a 96% probability of any reduction in mortality sufficient to justify widespread implementation of routine early ECMO in very severe ARDS? And how might patients, caregivers, clinicians, health system administrators, and health economists all answer that question differently? Importantly, by reporting the prior and posterior probabilities directly, the analysis allows all stakeholders to come to their own conclusions about whether the prior is consistent with their understanding of prior evidence and whether the posterior probability is sufficient to change their practice or change the clinical guidelines.

An additional tool in the Bayesian toolkit, known as decision theory, can formalize this process of combining key external factors to determine the best decision based on the available data. The goal of Bayesian decision analysis is to combine information across different outcomes (including outcomes, costs, and patient priorities) to determine an overall measure of the “value” of each intervention. The intervention with the maximum “value” after accounting for uncertainty (maximum expected value) can be taken to be the optimal intervention for use in clinical practice.[32]

What difference does it make? Interpretation of clinical trials

Bayesian analysis can clarify the meaning of data from clinical trials (Table 2).[7,33,34] A frequentist analysis of data may fail to reject a null hypothesis of no effect for two reasons: because the treatment effect is absent or harmful, or because the information in the study (sample size) was insufficient. Computing the posterior probability of benefit would clarify whether benefit had been ruled out (low probability of benefit) or not (moderate-high probability of benefit).[34] In a systematic re-analysis of trials, Wijeysundera et al. demonstrated that in 49 studies with non-significant p-values, the posterior probability of benefit ranged between 2%−97%; the posterior probability of benefit was >80% in 15/49 studies.[33] In these cases, Bayesian analysis clarifies the meaning of a non-significant result. The case of EOLIA (Panel 1) serves as an especially striking example where the Bayesian perspective can alter the meaning of non-statistically significant findings.[358]

On the other hand, a “statistically significant” result may not equate to a high probability of a “clinically significant’ benefit. The ability to estimate the probability of benefit for clinically relevant effect sizes is especially important for costly, invasive, and burdensome interventions. Wijeysundera et al. reported that 9/39 “statistically significant” studies had a posterior probability of benefit <70% for large effect sizes (hazard ratio <0·8).[33] In these cases, certain treatments may not be of meaningful benefit even if the null hypothesis of no effect can be rejected (i.e., p<0·05). Consequently, Bayesian analysis often permits a more nuanced and informative interpretation of trial results.[34]

What difference does it make? Design of clinical trials

Since Bayesian analysis can clarify the information in clinical trials, using Bayesian statistics to design clinical trials can help to ensure that the information generated by clinical trials is sufficient to reach a definitive conclusion for or against an intervention. The Bayesian approach is especially well-suited to the design of adaptive trials that discontinue enrolment once a clear conclusion is reached (Table 2).[3,21] When designing clinical trials, we are generally accustomed to fixing sample size. Stopping a trial after an interim analysis before the pre-specified sample size was reached ("early stopping") has traditionally been regarded as a serious methodological error to be avoided if at all possible because the trial did not conform to the pre-specified intentions of the investigator.[36] Yet the claim that the information obtained from analysis depends on the stopping rule, and not just the data, merits scrutiny. Jerome Cornfield, an eminent NIH statistician involved in the development of the modern clinical trial, once wrote that "To most scientists without previous exposure to statistics, as well as to most intelligent laymen, any dependence on stopping rules… seems like a violation of common sense."[37] And if the likelihood principle underpinning Bayesian statistics can be embraced as common sense, then Cornfield's observation rings true. Bayesian statistics envisions information accumulating during a trial (Figure 3); since the information in the study depends only on the data (per the likelihood principle) and is independent of the investigators intentions vis-à-vis data analysis, study conclusions can be updated at any time until sufficient information is available to reach a conclusion, irrespective of the original intentions of the investigator.[38] It should be noted that trial adaptations can also be specified based on the frequentist paradigm, though this requires application of complex rules (e.g. "alpha spending functions") to account for inflated risk of a false positive conclusion (Type I error).[39]

Due to the dominance of the frequentist paradigm, particularly for regulatory bodies such as the FDA in the US, Bayesian trials (adaptive or otherwise) are often tested through simulation to ensure that they exhibit good control of frequentist error rates, i.e., Type I and II errors.[40,41] These are often known as hybrid trial designs as they preserve the Bayesian interpretation of the analysis but rely on predicted frequentist error rates (assessed through simulation) to appraise trial operating characteristics. Bayesian methods for sample size calculation have also been developed.[31,42] These do not consider the power that can be achieved by the study but aim to control the precision with which we can estimate the parameter of interest, i.e., the treatment effect.[43] These methods for determining the sample size of a study align closely with the Bayesian interpretation of research, which aims to gather information and build on their information with each subsequent study. By controlling the precision of the estimated treatment effect, we control the amount of information available about the parameter, rather than our ability to make decisions following the study.

What difference does it make? Bayesian hierarchical models

The Bayesian approach leverages information outside the data (the prior) to enhance the information generated by analysis (the posterior). Bayesian hierarchical models employ an analogous approach to combining information between clusters of patients (also referred to

as subgroups or subtypes). Effectively, information from all clusters is employed to generate a data-driven posterior distribution for each individual cluster.[44] "Pooling" or "borrowing" information between clusters (i.e. using information from one cluster to estimate treatment effect in another) in this way increases precision in estimated effects.[45] The degree of pooling between clusters can depend on the similarity between those clusters. For example, in the Bayesian adaptive multiplatform trial of therapeutic anticoagulation for patients hospitalized for Covid-19, patients with moderate Covid-19 and severe Covid-19 were analyzed in a single hierarchical model.[46–48] The patients with moderate Covid-19 were further subgrouped according to their baseline D-dimer (low, high, or unknown). The estimated effects for severe and moderate Covid-19 were very different, with minimal borrowing between them (Table 3). On the other hand, the estimated effects for subgroups within moderate Covid-19 were fairly similar, and an overall conclusion of benefit was reached in the pooled hierarchy of moderate Covid-19 (Table 3). This approach makes maximally efficient use of available information and can reduce sample size requirements (while maintaining a larger effective sample size as reflected by the precision of the credible intervals) when possible differences in treatment effect among groups are hypothesized but unknown.[49]

Conclusions

Joel Greenhouse records of the eminent statistician Jerome Cornfield that "[his] frustration with the prevailing frequentist methods of the time grew out of a need for a theory of statistics that would truly help advance scientific discovery and would provide meaningful measures of evidence."[19] As an enthusiast for Bayesian methods, Cornfield was ahead of his time. Subsequent technical developments in computing capacity, together with repeated historical lessons about the shortcomings of frequentist statistical inference, particularly for inference in clinical trials,[12] suggest that routine implementation of Bayesian methods is warranted. Bayesian and frequentist methods should ultimately be seen as complementary, rather than as rivals (Panel 2). Bayesian analysis can supplement interpretation of data in studies designed on a frequentist basis, and can put the design of clinical trials on a more informative footing. Utilization of Bayesian inference is unquestionably growing, and will continue to grow, as familiarity with this framework grows, and the intuitive nature of Bayesian inference is increasingly appreciated.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Thanks to Lindsay Berry, PhD, for advice on computing effective sample size from statistical borrowing in hierarchical models.

Funding

E.C.G. is supported by an Early Career Health Research Award from the National Sanitarium Association. M.O.H. is supported by grant number R01-HL168202 from NHLBI/NIH. A.H. is supported by a Canada Research Chair in Statistical Trial Design and the Discovery Grant Program of the Natural Sciences and Engineering Research Council of Canada (RGPIN-2021-03366)

Declaration of interests

E.C.G. receives fees for speaking or consulting from Vyaire, BioAge, Stimit LLC, Lungpacer Medical Inc., Getinge, Draeger, Heecap, and Zoll. He serves on the Clinical Advisory Board for Getinge and previously served on the Advisory Board for Lungpacer Medical Inc. He has received in-kind support for research from Timpel Medical Inc., Lungpacer Medical Inc., and Getinge. M.O.H received statistical consulting fees from Unlearn.AI, Guidepoint Global, and Berkeley Research Group, fees for editorial services from Elsevier and the American Thoracic Society, for serving on a data safety monitoring board from the University of California, San Francisco, and the University of Pittsburgh, and for pilot grant reviews from Brown University and New York University.

References

1. Jeffreys H Theory of probability. 3rd ed. Oxford Oxfordshire];: Clarendon Press; 1983.
2. McGrayne SB. The theory that would not die : how Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of C. New Haven, Connecticut: Yale University Press; 2011.
3. Bland JM, Altman DG. Statistics Notes: Bayesians and Frequentists. BMJ 1998; 317(7166): 1151. doi:10.1136/bmj.317.7166.1151. [PubMed: 9784463]
4. Hoekstra R, Morey RD, Rouder JN, Wagenmakers E Robust misinterpretation of confidence intervals. Psychon Bull Rev 2014; 21(5): 1157–64. doi:10.3758/s13423-013-0572-3. [PubMed: 24420726]
5. Stigler SM. Who Discovered Bayes's Theorem? The American statistician 1983; 37(4): 290–6. doi:10.1080/00031305.1983.10483122.
6. BAYES. AN ESSAY TOWARDS SOLVING A PROBLEM IN THE DOCTRINE OF CHANCES. Biometrika 1958; 45(3–4): 296–315. doi:10.1093/biomet/45.3-4.296.
7. 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. doi:10.1136/bmj.319.7208.508. [PubMed: 10454409]
8. Goligher EC, Tomlinson G, Hajage D, 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): 2251–9. doi:10.1001/jama.2018.14276 [doi]. [PubMed: 30347031]
9. Lewis RJ, Wears RL. An introduction to the Bayesian analysis of clinical trials. Annals of emergency medicine 1993; 22(8): 1328–36. doi:10.1016/S0196-0644(05)80119-2. [PubMed: 8333640]
10. Lehmann EL. The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two? Journal of the American Statistical Association 1993; 88(424): 1242–9. doi:10.1080/01621459.1993.10476404.
11. Neyman J Frequentist Probability and Frequentist Statistics. Synthese (Dordrecht) 1977; 36(1): 97–131. doi:10.1007/BF00485695.
12. Goodman SN. Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy. Annals of internal medicine 1999; 130(12): 995–1004. doi:10.7326/0003-4819-130-12-199906150-00008. [PubMed: 10383371]
13. Young KD, Lewis RJ. What Is Confidence? Part 2: Detailed Definition and Determination of Confidence Intervals. Annals of emergency medicine 1997; 30(3): 311–8. doi:10.1016/S0196-0644(97)70167-7. [PubMed: 9287893]
14. Rowbottom DP. Probability. Cambridge, UK;: Polity Press; 2015.
15. Lee PM. Bayesian statistics : an introduction. New York: Oxford University Press; 1989.
16. Browner WS, Newman TB. Are All Significant P Values Created Equal?: The Analogy Between Diagnostic Tests and Clinical Research. JAMA : the journal of the American Medical Association 1987; 257(18): 2459–63. doi:10.1001/jama.1987.03390180077027. [PubMed: 3573245]
17. Kruschke JK. Bayesian data analysis. Wiley interdisciplinary reviews.Cognitive science 2010; 1(5): 658–76. doi:10.1002/wcs.72. [PubMed: 26271651]
18. Gelman A, Loken E. The Statistical Crisis in Science. American scientist 2014; 102(6): 460–5. doi:10.1511/2014.111.460.

19. Greenhouse JB. On becoming a Bayesian: Early correspondences between J. Cornfield and L. J. Savage. Statistics in medicine 2012; 31(24): 2782–90. doi:10.1002/sim.5604. [PubMed: 22941781]
20. Berry DA. Interim Analysis in Clinical Trials: The Role of the Likelihood Principle. The American statistician 1987; 41(2): 117–22. doi:10.1080/00031305.1987.10475458.
21. Berry DA. Bayesian clinical trials. Nature reviews. Drug discovery 2006; 5(1): 27–36. doi:10.1038/nrd1927. [PubMed: 16485344]
22. Lemoine NP. Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. Oikos 2019; 128(7): 912–28. doi:10.1111/oik.05985.
23. Zampieri FG, Casey JD, Shankar-Hari M, Harrell J, Frank E, 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. American journal of respiratory and critical care medicine 2021; 203(5): 543–52. doi:10.1164/rccm.202006-2381CP. [PubMed: 33270526]
24. Gelman A, Jakulin A, Pittau MG, Su Y. A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models. The annals of applied statistics 2008; 2(4): 1360–83. doi:10.1214/08-AOAS191.
25. Goligher EC, Harhay MO. What Is the Point of Bayesian Analysis? American journal of respiratory and critical care medicine 2024; 209(5): 485–7. doi:10.1164/rccm.202310-1757VP. [PubMed: 37922491]
26. Pencina MJ, Thompson BT. Clinical Trials in the 21st Century — Promising Avenues for Better Studies. NEJM evidence 2022; 1(6). doi:10.1056/EVIDctw2200060.
27. Viele K, Berry S, Neuenschwander B, et al. Use of historical control data for assessing treatment effects in clinical trials. Pharmaceutical statistics : the journal of the pharmaceutical industry 2014; 13(1): 41–54. doi:10.1002/pst.1589.
28. Falconer JR. Methods for Eliciting Informative Prior Distributions: A Critical Review. Decision analysis 2022; 19(3): 189–204. doi:10.1287/deca.2022.0451.
29. Johnson SR, Tomlinson GA, Hawker GA, Granton JT, Feldman BM. Methods to elicit beliefs for Bayesian priors: a systematic review. Journal of clinical epidemiology 2010; 63(4): 355–69. doi:10.1016/j.jclinepi.2009.06.003. [PubMed: 19716263]
30. Spiegelhalter DJ, Freedman LS, Parmar Mahesh K. B.. Bayesian Approaches to Randomized Trials. Journal of the Royal Statistical Society. Series A, Statistics in society 1994; 157(3): 357–416. doi:10.2307/2983527.
31. Hobbs BP, Carlin BP. Practical Bayesian Design and Analysis for Drug and Device Clinical Trials. Journal of biopharmaceutical statistics 2008; 18(1): 54–80. doi:10.1080/10543400701668266. [PubMed: 18161542]
32. Fisher RA. Statistical methods and scientific inference. 2nd ed. Edinburgh: Oliver and Boyd; 1959.
33. Wijeysundera DN, Austin PC, Hux JE, Beattie WS, Laupacis A. Bayesian statistical inference enhances the interpretation of contemporary randomized controlled trials. Journal of clinical epidemiology 2009; 62(1): 13,21.e5. doi:10.1016/j.jclinepi.2008.07.006. [PubMed: 18947971]
34. Yarnell CJ, Abrams D, Baldwin MR, et al. Clinical trials in critical care: can a Bayesian approach enhance clinical and scientific decision making? The lancet respiratory medicine 2020;. doi:10.1016/S2213-2600(20)30471-9.
35. Combes A, Hajage D, Capellier G, et al. Extracorporeal Membrane Oxygenation for Severe Acute Respiratory Distress Syndrome. The New England journal of medicine 2018; 378(21): 1965–75. doi:10.1056/nejmoa1800385. [PubMed: 29791822]
36. Bassler D, Montori VM, Briel M, Glasziou P, Guyatt G. Early stopping of randomized clinical trials for overt efficacy is problematic. Journal of clinical epidemiology 2008; 61(3): 241–6. doi:10.1016/j.jclinepi.2007.07.016. [PubMed: 18226746]
37. Cornfield J Sequential Trials, Sequential Analysis and the Likelihood Principle. The American statistician 1966; 20(2): 18. doi:10.2307/2682711.
38. Jack Lee J, Chu CT. Bayesian clinical trials in action. Statistics in medicine 2012; 31(25): 2955–72. doi:10.1002/sim.5404. [PubMed: 22711340]

39. Stallard N, Todd S, Ryan EG, Gates S. Comparison of Bayesian and frequentist group-sequential clinical trial designs. BMC medical research methodology 2020; 20(1): 4. doi:10.1186/s12874-019-0892-8. [PubMed: 31910813]
40. Kalil A, Sun J. Bayesian Methodology for the Design and Interpretation of Clinical Trials in Critical Care Medicine: A Primer for Clinicians. Critical care medicine 2014; 42(10): 2267–77. doi:10.1097/CCM.0000000000000576. [PubMed: 25226118]
41. US Food and Drug Administration. Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. 2010; Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-use-bayesian-statistics-medical-device-clinical-trials. Accessed March 14, 2024.
42. Adcock CJ. Sample size determination: a review. Journal of the Royal Statistical Society. Series D (The Statistician) 1997; 46(2): 261–83. doi:10.1111/1467-9884.00082.
43. Cao J, Lee JJ, Alber S. Comparison of Bayesian sample size criteria: ACC, ALC, and WOC. Journal of statistical planning and inference 2009; 139(12): 4111–22. doi:10.1016/j.jspi.2009.05.041. [PubMed: 25554718]
44. McGlothlin AE, Viele K. Bayesian Hierarchical Models. JAMA : the journal of the American Medical Association 2018; 320(22): 2365–6. doi:10.1001/jama.2018.17977. [PubMed: 30535206]
45. Chen N, Lee JJ. Bayesian cluster hierarchical model for subgroup borrowing in the design and analysis of basket trials with binary endpoints. Statistical methods in medical research 2020; 29(9): 2717–32. doi:10.1177/0962280220910186. [PubMed: 32178585]
46. Lawler PR, Goligher EC, Berger JS, et al. Therapeutic Anticoagulation with Heparin in Noncritically Ill Patients with Covid-19. N Engl J Med 2021; 385(9): 790–802. doi:10.1056/NEJMoa2105911. [PubMed: 34351721]
47. Goligher EC, Bradbury CA, McVerry BJ, et al. Therapeutic Anticoagulation with Heparin in Critically Ill Patients with Covid-19. The New England journal of medicine 2021; 385: 777–89. [PubMed: 34351722]
48. Goligher EC, Lawler PR, Jensen TP, et al. Heterogeneous Treatment Effects of Therapeutic-Dose Heparin in Patients Hospitalized for COVID-19. JAMA : the journal of the American Medical Association 2023;.
49. Su L, Chen X, Zhang J, Yan F. Comparative Study of Bayesian Information Borrowing Methods in Oncology Clinical Trials. JCO precision oncology 2022; 6(6): e2100394. doi:10.1200/PO.21.00394. [PubMed: 35263169]
50. Lewis RJ, Angus DC. Time for Clinicians to Embrace Their Inner Bayesian?: Reanalysis of Results of a Clinical Trial of Extracorporeal Membrane Oxygenation. JAMA : the journal of the American Medical Association 2018; 320(21): 2208–10. doi:10.1001/jama.2018.16916. [PubMed: 30347047]

Panel 1.

Case study on the relevance of statistical paradigms to the meaning of data: extracorporeal membrane oxygenation in acute respiratory distress syndrome.

Frequentist and Bayesian statistics can sometimes entail nearly opposite conclusions from the same study. Extracorporeal membrane oxygenation (ECMO) for acute respiratory distress syndrome (ARDS) serves as a pertinent case. The ECMO to Rescue Lung Injury in Severe ARDS (EOLIA) trial compared a strategy of early ECMO to conventional rescue ECMO in patients with life-threatening hypoxemia or hypercapnia from severe ARDS. The original, planned sample size was 331 patients, but enrolment was discontinued after enrolling 249 patients because a planned interim analysis suggested that the trial was unlikely to reach a statistically significant conclusion of benefit.[35] The effect of treatment was numerically large: observed mortality was 46% in the control group and 35% in the intervention group (11% absolute risk difference) yet the results were not statistically significant (relative risk 0·76, 95% confidence interval, 0·55 to 1·04, p=0·09). The authors were forced to conclude that the trial did not support the hypothesis that early ECMO reduces mortality in very severe ARDS. By contrast, a post hoc Bayesian re-analysis of these very same data found a high posterior probability of mortality benefit ranging between 88–99% depending on the prior,[8] leading editorialists to conclude that ECMO should be regarded as efficacious.[50] In this case, the frequentist and Bayesian frameworks appear to give rise to opposing conclusions, raising fundamental questions about how we should approach statistical analysis and the interpretation of data.

Panel 2.

Foes or Friends? Reconsidering the relationship between frequentist and Bayesian statistics

This paper emphasizes the strengths of Bayesian statistics over the possible weaknesses of frequentist statistics. However, frequentist methods have salient strengths. For one, they are computationally much simpler than Bayesian methods, and conventional frequentist trial design is straightforward, with a wide array of methods for estimating sample size and statistical power for different types of outcomes and estimands. The p-value functions as a built-in decision analysis tool ("significant" versus "non-significant") and given its logic it is well-suited to analyses where the primary goal is to test a hypothesis rather than to estimate the magnitude of treatment effect (e.g. tests of interaction), though it can still be criticized for disregarding prior information and the optimal p-value threshold to declare significance is debated. Bayesian methods for hypothesis testing are well-described. In trials designed based on frequentist statistics, Bayesian analysis can function to complement the interpretation of data without necessarily being the primary basis for analyzing the trial. Spiegelhalter et al. suggested that Bayesian analysis results could be presented in trial reports in an Interpretation section located between Results and Discussion.[30] Used in this way, Bayesian analysis can help to establish whether meaningful benefit from an intervention has been ruled out when the trial results are not "significant". The primary analysis of a clinical trial should always be conducted according to the pre-specified design. When the sample size is adequately large, frequentist and Bayesian analyses will converge towards the same conclusion.

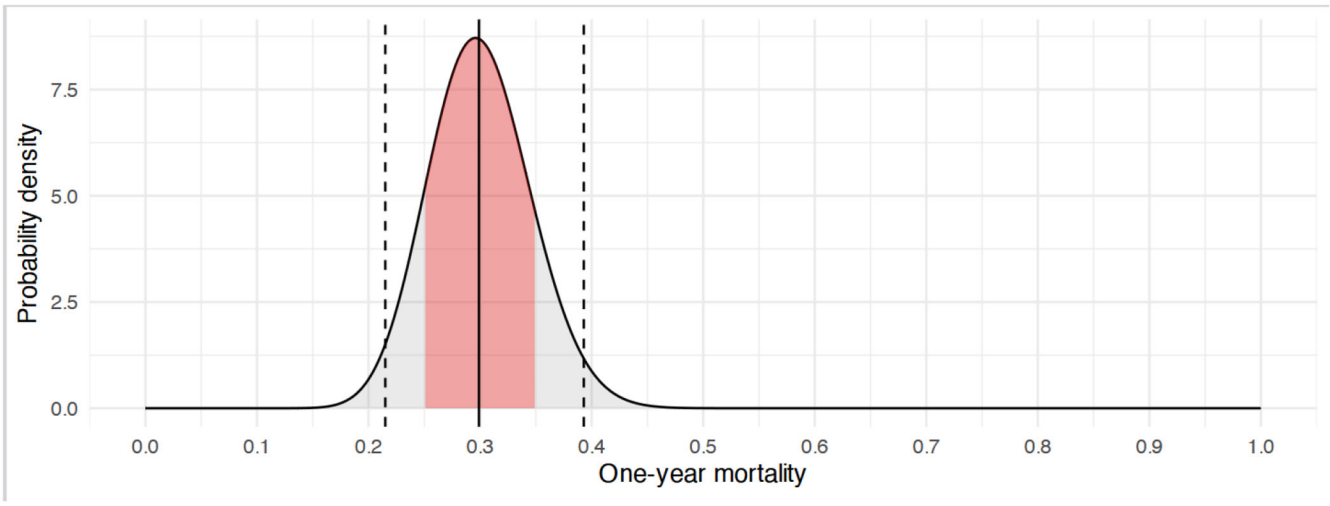

Figure 1.

Example of a probability density distribution obtained in a hypothetical observational cohort study of mortality at one-year. In a sample of 100 patients, 30 were found to have died at one year of follow-up. These data are assumed to have a binomial likelihood. The probability density distribution computed from the likelihood function for these data, assuming a non-informative prior (constructed using a beta distribution), is shown in the plot. This data-based probability distribution represents the posterior distribution for this study. The distribution can be described in terms of its median value (solid vertical line) and 2·5th and 97·5th percentiles (dashed vertical lines) representing the 95% credible interval. The area under the curve in the shaded red region gives the probability that the mortality rate lies between 0·25 and 0·35.

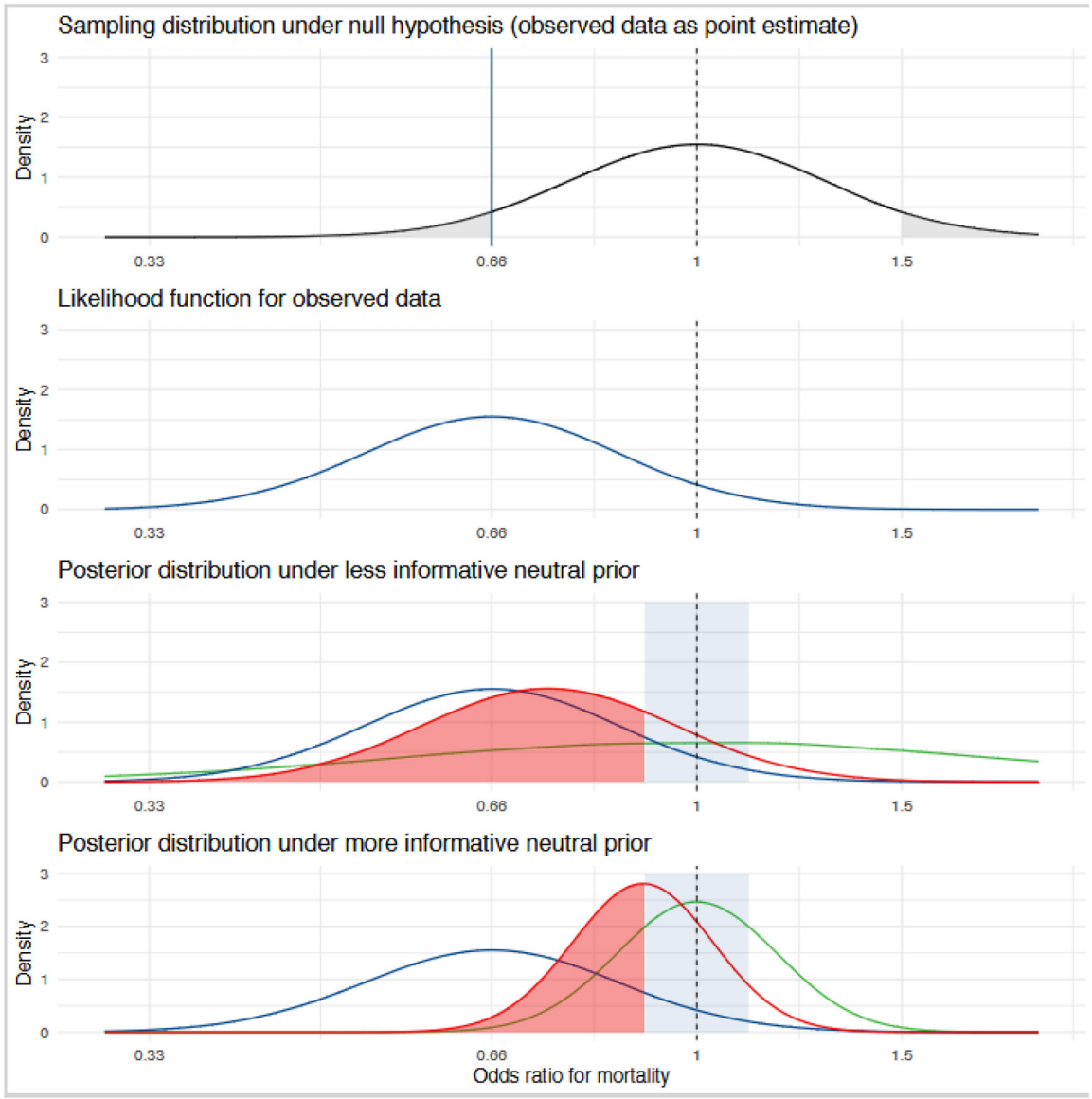

Figure 2.

Comparison of frequentist and Bayesian statistical inference. In a hypothetical study evaluating an association between intervention and mortality, 280 patients were randomized 1:1 to intervention or control. Mortality in the control group was 38% and in the intervention group was 29%, yielding a point estimate for the odds ratio of 0·66. The sampling distribution for the null hypothesis in this study (top panel), defined based on the primary outcome and planned sample size (stopping rule), gives the expected frequency of obtaining data with a point estimate (vertical blue line) as or more extreme than the observed point estimate if the study was hypothetically repeated in precisely the same way many times (grey-shaded area). If this expected frequency is less than 5 out of 100 ("p<0·05"), the null hypothesis of no effect is generally rejected. In frequentist inference, the hypothesis is fixed (effect=null, vertical dashed line), and the *expected* data under the hypothesis are treated as a random variable with (hypothetical) repeated sampling (represented by the sampling

distribution). In Bayesian inference, the *observed* data are treated as a fixed distribution (the likelihood function, middle panel) rather than as a point estimate and the hypothesis is treated as a random variable with a probability distribution (not as a point estimate). Information on the hypothesis prior to the study is represented by the prior distribution (green curve, bottom panels). Combining the prior with the likelihood yields a new random probability distribution for the updated hypothesis, the posterior distribution (red curve, bottom panels); the posterior distribution will vary according to the prior. The area under the posterior distribution may be used to estimate the probability of treatment effect below the region of practical equivalence which defines the minimum clinically relevant effect (blue-shaded region, bottom panels). The posterior distribution depends only on the prior and the likelihood; it is not determined by the analysis plan and stopping rule for the study (unlike the frequentist sampling distribution).

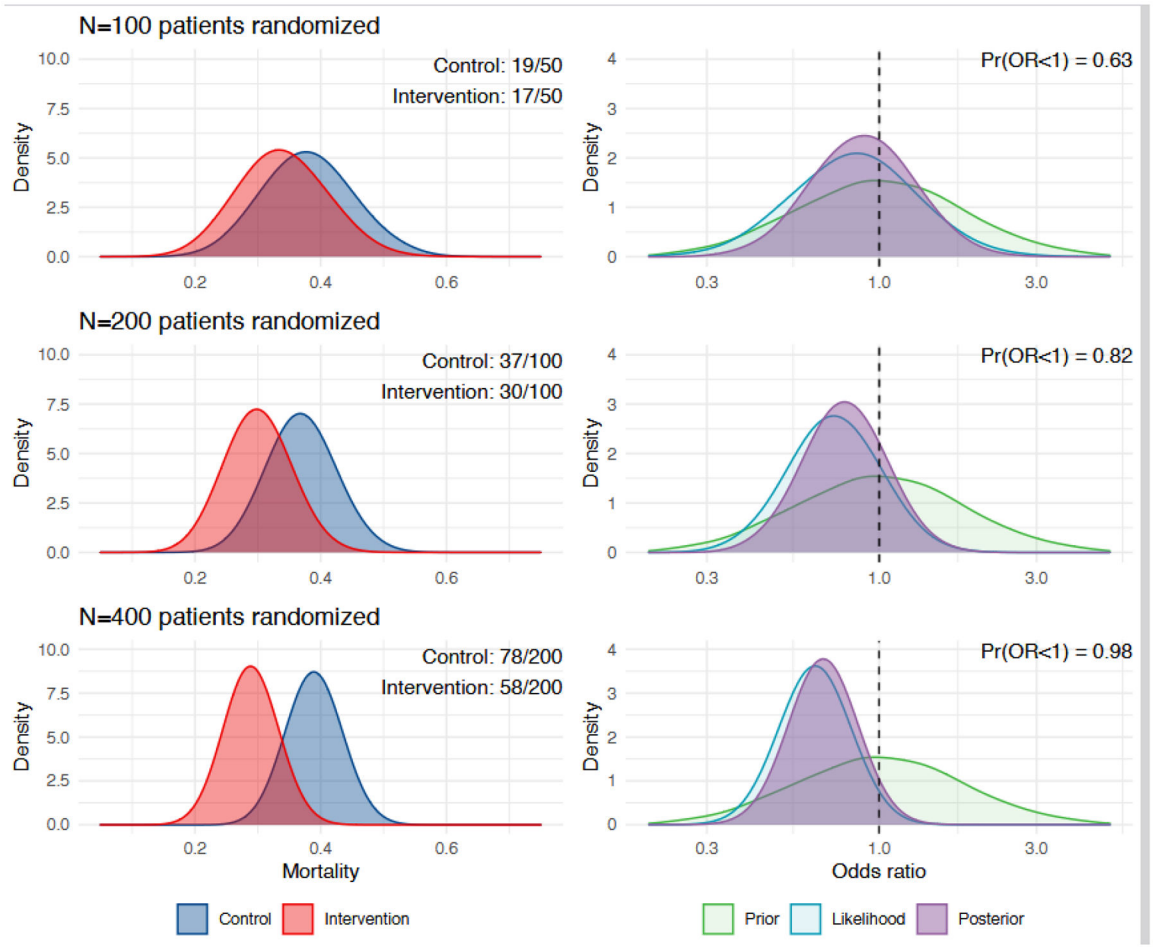

Figure 3.

Bayesian sequential analysis of a clinical trial. A clinical trial was simulated by randomly sampling from a population where the true mortality rate at 28 days after randomization in the Intervention and Control Groups is 0·3 and 0·4, respectively. The prior distribution for the odds ratio is neutral, with information equivalent to a sample size of 50 patients. As increasing numbers of patients are enrolled and randomized, sequential analyses reveal accumulating information about mortality in the Intervention and Control groups (left panels). This increase in information is represented by a progressive decrease in the variance of the probability distributions for mortality in each group (left panels). The increasing information about mortality allows more precise estimates of the difference in mortality between groups (quantified by the posterior odds ratio, right panels), yielding a progressively higher strength of belief that the intervention lowers mortality (posterior probability of superiority, Pr(OR<1)). Under the likelihood principle, the information in the posterior distribution is not affected by the number of interim analyses, facilitating adaptive decision-making in trial design (e.g., stop the trial once the posterior probability of superiority exceeds a pre-specified threshold).

Table 1.

Conceptual differences between frequentist and Bayesian statistical inference

Concept	Frequentist paradigm	Bayesian paradigm
Meaning of "probability"	Expected frequency of events	Degree of belief in a hypothesis
Basic approach	Focus on the expected frequency of observing data under the null hypothesis to base scientific inference on "objective" data rather than "subjective" degree of belief	Represent available information and uncertainty ("degree of belief") about a hypothesis quantitatively using probability distributions
Model of science	Subject hypotheses to rigorous testing one-at-a-time, prioritizing avoidance of Type I error (false positive conclusions)	Progressive accumulation of information to reduce uncertainty
Determinants of statistical information about the hypothesis in a study	Statistical information in a study depends on both the study results and the investigator's intentions represented in the pre-specified statistical design (stopping rule, primary comparison, etc.)	Statistical information deriving from a study depends only on the study results (likelihood principle)
Primary question for statistical inference	Probability of observing data given the null hypothesis	Probability of a hypothesis given the observed data
	"Should the null hypothesis be rejected?"	"How strongly should we believe the hypothesis?"
Inferential probability	P-value: given the null hypothesis, expected frequency of observing data as or more extreme than that observed in the study if the exact same study is hypothetically repeated over and over in accordance with the original pre-specified intentions of the investigator	Posterior probability: probability statements about the parameter of interest, given the available information from the study data and prior
Inferential interval	95% confidence interval: an interval generated from a sample of data such that in repetitions of the same study, 95% of such intervals would contain the true value of the parameter	95% credible interval: the plausible range of values for the parameter of interest given the available information from the study and the prior
Pre-defined probability threshold for inference	Generally, $p<0.05$	No single value of posterior probability defines "belief" in a hypothesis
Considers prior information	No, insists that inference should be based on study data alone	Yes, requires specification of prior probability to estimate posterior probability

Table 2.

Practical application of frequentist and Bayesian statistical outputs.

Paradigm	Output	Application to data analysis	Application to trial design
Frequentist	p-value	Determine whether evidence against the null hypothesis is sufficient to reject the null hypothesis and conclude "statistical significance"	Used in sample size calculations to determine that statistical power is sufficient
		Generally fixed at <5%, i.e., $p < 0.05$	In adaptive designs, can be used to determine whether to stop the study
			The p-value for stopping depends on the pre-specified study design (e.g., number of interim analyses, etc.)
	Confidence interval	Determine whether the range of values for treatment effect compatible with the data includes or excludes the null	n/a
		Does not account for prior information	
		Generally applied using 95% interval range	
Bayesian	Posterior probability	Quantify the appropriate strength of belief in favour of a given treatment effect based on the available information	In adaptive designs, can be used to determine whether to stop the study
		No single fixed posterior probability used to declare "significance", although fixed thresholds to conclude superiority are pre-specified in Bayesian trial design	The posterior probability criterion for stopping enrolment is fixed
	Credible interval	Represent the plausible range of values for treatment effect given available information	Can be used in sample size determination to control precision of treatment effect estimate
		Accounts for prior information	
		Multiple possible interval ranges can be reported to characterize the posterior distribution, depending on clinical context	

Table 3.

Borrowing between groups in a Bayesian hierarchical model of the multiplatform randomized clinical trial of therapeutic anticoagulation for patients hospitalized for Covid-19

Severity group	D-dimer group	With borrowing (primary analysis)		Without borrowing (sensitivity analysis)		Increase in effective sample size with borrowing
		Adjusted odds ratio (95% CrI)	Effective sample size*	Adjusted odds ratio (95% CrI)	Actual sample size	
Moderate	Low	1·22 (0·93–1·57)	1516	1·12 (0·82–1·54)	1075	441
	High	1·31 (1·00–1·76)	1272	1·39 (0·95–2·01)	630	642
	Unknown	1·32 (1·00–1·86)	1228	1·47 (0·96–2·28)	514	714
Severe	n/a	0·83 (0·67–1·03)	1098	0·82 (0·66-1·03)	1098	0

*Effective sample size is the sample size required to generate the level of precision (variance) in the estimated effect (adjusted odds ratio). Information in this Table is derived from references 46 and 47.