Using Bayesian statistics in confirmatory clinical trials in the regulatory setting: a tutorial review

Se Yoon Lee[1*]

Abstract

Bayesian statistics plays a pivotal role in advancing medical science by enabling healthcare companies, regulators, and stakeholders to assess the safety and efficacy of new treatments, interventions, and medical procedures. The Bayesian framework offers a unique advantage over the classical framework, especially when incorporating prior information into a new trial with quality external data, such as historical data or another source of co-data. In recent years, there has been a significant increase in regulatory submissions using Bayesian statistics due to its flexibility and ability to provide valuable insights for decision-making, addressing the modern complexity of clinical trials where frequentist trials are inadequate. For regulatory submissions, companies often need to consider the frequentist operating characteristics of the Bayesian analysis strategy, regardless of the design complexity. In particular, the focus is on the frequentist type I error rate and power for all realistic alternatives. This tutorial review aims to provide a comprehensive overview of the use of Bayesian statistics in sample size determination, control of type I error rate, multiplicity adjustments, external data borrowing, etc., in the regulatory environment of clinical trials. Fundamental concepts of Bayesian sample size determination and illustrative examples are provided to serve as a valuable resource for researchers, clinicians, and statisticians seeking to develop more complex and innovative designs.

Keywords Bayesian hypothesis testing, Sample size determination, Regulatory environment, Frequentist operating characteristics

Background

Clinical trials are a critical cornerstone of modern healthcare, serving as the crucible in which medical innovations are tested, validated, and ultimately brought to patients [1]. Traditionally, since the 1940s, these trials have adhered to frequentist statistical methods, offering valuable insights into decision-making to demonstrate treatment effects. However, they may fall short in addressing the increasing complexity of modern clinical

trials, such as personalized medicine [2, 3], innovative study designs [4, 5], and the integration of real-world data into randomized controlled trials [6–8], among many other challenges [9–11].

These new challenges commonly necessitate innovative solutions. The US 21st Century Cures Act and the US Prescription Drug User Fee Act VI include provisions to advance the use of complex innovative trial designs [12]. Generally, complex innovative trial designs have been considered to refer to complex adaptive, Bayesian, and other novel clinical trial designs, but there is no fixed definition because what is considered innovative or novel can change over time [12–15]. A common feature of many of these designs is the need for simulations rather than mathematical formulae to estimate trial

*Correspondence:
Se Yoon Lee
seyoonlee.stat.math@gmail.com
[1] Department of Statistics, Texas A &M University, 3143 TAMU, College Station, TX 77843, USA

operating characteristics. This highlights the growing embrace of complex innovative trial designs in regulatory submissions.

In this paper, our particular focus is on Bayesian methods. Guidance from the U.S. Food and Drug Administration (FDA) [16] defines Bayesian statistics as an approach for learning from evidence as it accumulates. Bayesian methods offer a robust and coherent probabilistic framework for incorporating prior knowledge, continuously updating beliefs as new data emerge, and quantifying uncertainty in the parameters of interest or outcomes for future patients [17]. The Bayesian approach aligns well with the iterative and adaptive nature of clinical decision-making, offering opportunities to maximize clinical trial efficiency, especially in cases where data are sparse or costly to collect.

The past two decades have seen notable demonstrations of Bayesian statistics addressing various types of modern complexities in clinical trial designs. For example, Bayesian group sequential designs are increasingly used for seamless modifications in trial design and sample size to expedite the development process of drugs or medical devices, while potentially leveraging external resources [18–22]. One recent example is the COVID-19 vaccine trial, which includes four Bayesian interim analyses with the option for early stopping to declare vaccine efficacy before the planned trial end [23]. Other instances where Bayesian approaches have demonstrated their promise are umbrella, basket, or platform trials under master protocols [24]. In these cases, Bayesian adaptive approaches facilitate the evaluation of multiple therapies in a single disease, a single therapy in multiple diseases, or multiple therapies in multiple diseases [25–32]. Moreover, Bayesian approaches provide an effective means to integrate multiple sources of evidence, a particularly valuable aspect in the development of pediatric drugs or medical devices where small sample sizes can impede traditional frequentist approaches [33–35]. In such cases, Bayesian borrowing techniques enable the integration of historical data from previously completed trials, real-world data from registries, and expert opinion from published resources. This integration provides a more comprehensive and probabilistic framework for information borrowing across different sub-populations [36–39].

It is important to note that the basic tenets of good trial design are consistent for both Bayesian and frequentist trials. Sponsors using the Bayesian approach for sizing a trial should adhere to the principles of good clinical trial design and execution, including minimizing bias, as outlined in regulatory guidance [16, 40, 41], following almost the same standards as those given to frequentist approaches. For example, regulators often recommend that sponsors submit a Bayesian design that effectively maintains the frequentist type I and type II error rates (or some analog of it) at the nominal levels for all realistic scenarios by carefully calibrating design parameters.

In the literature, numerous articles [13, 42–47] and textbooks [17, 48] extensively cover both basic and advanced concepts of Bayesian designs. While several works focus on regulatory issues in developing Bayesian designs [49–51], there seems to be a lack of tutorial-type review papers explaining how to develop Bayesian designs for regulatory submissions within the evolving regulatory environment, along with providing tutorial-type examples. Such papers are crucial for sponsors, typically pharmaceutical or medical device companies, preparing to use Bayesian designs to gain insight and build more complex Bayesian designs.

In this paper, we provide a pedagogical understanding of Bayesian designs by elucidating key concepts and methodologies through illustrative examples and address the existing gaps in the literature. For the simplicity of explanation, we apply Bayesian methods to construct single-stage designs, two-stage designs, and parallel designs for single-arm trials, but the illustrated key design principles can be generalized to multiple-arm trials. Specifically, our focus in this tutorial is on Bayesian sample size determination, which is most useful in confirmatory clinical trials, including late-phase II or III trials in the drug development process or pivotal trials in the medical device development process. We highlight the advantages of Bayesian designs, address potential challenges, examine their alignment with evolving regulatory science, and ultimately provide insights into the use of Bayesian statistics for regulatory submissions.

This tutorial paper is organized as follows. Figure 1 displays the diagram of the paper organization. We begin by explaining a simulation-based approach to determine the sample size of a Bayesian design in Sizing a Bayesian trial section, which is consistently used throughout the paper as the building blocks to develop many kinds of Bayesian designs. Next, the specification of the prior distribution for Bayesian submission is discussed in Specification of prior distributions section, and two important Bayesian decision rules, namely, the posterior probability approach and the predictive probability approach, are illustrated in Decision rule - posterior probability approach and Decision rule - predictive probability approach sections, respectively. These are essential in the development of Bayesian designs for regulatory submissions. Advanced design techniques for multiplicity adjustment using Bayesian hierarchical modeling are illustrated in Multiplicity adjustments section, and incorporating external data using power prior modeling is explained in External data borrowing section. We conclude the paper with a discussion in Conclusions section.
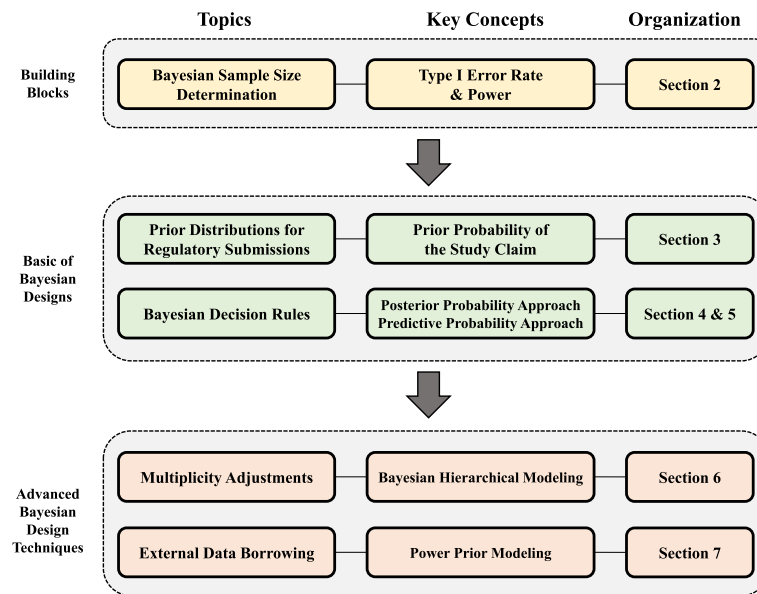
Fig. 1 Topics, key concepts, and organization of paper

Sizing a Bayesian trial

A simulation principle of Bayesian sample size determination

Although practical and ethical issues need to be considered, one's initial reasoning when determining the trial size should focus on the scientific requirements [52]. Scientific requirements refer to the specific criteria, conditions, and standards that must be met in the design, conduct, and reporting of scientific research to ensure the validity, reliability, and integrity of the findings. Much like frequentist approaches for determining the sample size of the study [53], its Bayesian counterpart also proceeds by first defining a success criterion to align with the primary objective of the trial. Subsequently, the number of subjects is determined to provide a reliable answer to the questions addressed within regulatory settings.

In the literature, various studies have explored the sizing of Bayesian trials [54–60]. Among these, the simulation-based method proposed by [60] stands out as popular, and it was further explored by [61, 62] for practical applications. This method is widely used by many healthcare practitioners, including design statisticians at companies or universities, for its practical applicability in a broad range of Bayesian designs. Furthermore, this method, with a particular prior setting, is well-suited for the regulatory submission, where the evaluation of the frequentist operating characteristics of the Bayesian design is critical. This will be discussed in Calibration of Bayesian trial design to assess frequentist operating characteristics section.

In this section, we outline the framework of the authors' work [60]. Similar to the notation in Reference [63] assume that the endpoint has probability density function $f(y|\theta)$, where the $\theta \in \Theta$ represents the parameter of main interest. The hypotheses to be investigated are the null and alternative hypotheses,

$$\mathcal{H}_0 : \theta \in \Theta_0 \text{ versus } \mathcal{H}_a : \theta \in \Theta_a, \quad (1)$$

where Θ_0 and Θ_a represent the disjoint parameter spaces for the null and alternative hypotheses, respectively. $\Theta = \Theta_0 \cup \Theta_a$ denotes the entire parameter space. Suppose that the objective of the study is to evaluate the efficacy of a new drug, achieved by rejecting the null hypothesis. Let $\mathbf{y}_N = (y_1, \cdots, y_N)^\top$ denotes a set of N outcomes such that y_i $(i = 1, \cdots, N)$ is identically and independently distributed according a distribution $f(y|\theta)$.

Throughout the paper, we assume that the parameter space Θ is a subset of real numbers. The range of the parameter space Θ is determined by the type of outcomes. For example, for continuous outcomes y, the distribution $f(y|\theta)$ may be a normal distribution, where the parameter space is the set of real numbers, $\Theta = \mathbb{R}$; and for binary outcomes, the distribution $f(y|\theta)$ is the Bernoulli distribution, where the parameter space is the set of fractional numbers, $\Theta = [0, 1]$. In this formulation, typically, the hypotheses (1) are one-sided; for example, $\mathcal{H}_0 : \theta \leq \theta_0$ versus $\mathcal{H}_a : \theta > \theta_0$ or $\mathcal{H}_0 : \theta \geq \theta_0$ versus $\mathcal{H}_a : \theta < \theta_0$. Throughout the paper, when we denote hypotheses in the abstract form (1), it is considered a one-sided superiority test for the coherency of the paper. The logic explained in this paper can be

generalized to a form of a two-sided test, non-inferiority test, or equivalence test in a similar manner, but discussion on these forms is out of scope for this paper.

The simulation-based approach incorporates two essential components: the 'sampling prior' $\pi_s(\theta)$ and the 'fitting prior' $\pi_f(\theta)$. The sampling prior is utilized to generate observations $\mathbf{y}_N$ by considering the scenario of 'what if the parameter θ is likely to be within a specified portion of the parameter space?' The fitting prior is employed to fit the model once the data $\mathbf{y}_N$ has been obtained upon completion of the study. We note that the sampling prior should be a proper distribution, while the fitting prior does

$$T(\mathbf{y}_N) = \mathbb{P}_f[\theta \in \Theta_a|\mathbf{y}_N] = \int \mathbf{1}\{\theta \in \Theta_a\}\pi_f(\theta|\mathbf{y}_N)d\theta,$$

where the indicator function $\mathbf{1}\{A\}$ is 1 if A is true and 0 otherwise. A typical success criterion takes the form of

$$\text{Study Sucess} = \mathbf{1}\{T(\mathbf{y}_N) > \lambda\} = \mathbf{1}\{\mathbb{P}_f[\theta \in \Theta_a|\mathbf{y}_N] > \lambda\}, \quad (4)$$

where $\lambda \in [0, 1]$ is a pre-specified threshold value.

At this point, we introduce a key quantity to measure the expected behavior of the Bayesian test statistics $T(\mathbf{y}_N)$ – the probability of study success based on the Bayesian testing procedure – by considering the idea of repeated sampling of the outcomes $\mathbf{y}_N \sim f_s(\mathbf{y}_N)$:

$$\beta_\Theta^{(N)} = \mathbb{P}_s[T(\mathbf{y}_N) > \lambda|\mathbf{y}_N \sim f_s(\mathbf{y}_N)] = \int \mathbf{1}\{\mathbb{P}_f[\theta \in \Theta_a|\mathbf{y}_N] > \lambda\}f_s(\mathbf{y}_N)d\mathbf{y}_N. \quad (5)$$

not need to be proper as long as the resulting posterior, $\pi(\theta|\mathbf{y}_N) \propto f(\mathbf{y}_N|\theta) \cdot \pi_f(\theta)$, is proper. We also note that the sampling prior is a unique Bayesian concept adopted in the simulation-based approach, whereas the fitting prior refers to the prior distributions used in the daily work of Bayesian data analyses [64], not confined to the context of sample size determination.

In the following, we illustrate how to calculate the Bayesian test statistic, denoted as $T(\mathbf{y}_N)$, using the posterior probability approach by using a sampling prior and a fitting prior. (Details of the posterior probability approach will be explained in Decision rule - posterior probability approach section). First, one generates a value of parameter of interest θ from the sampling prior $\pi_s(\theta)$, and then generates the outcome vector $\mathbf{y}_N = (y_1, \cdots, y_N)^\top$ based on that θ.

In the notation $\beta_\Theta^{(N)}$ (5), the superscript 'N' indicates the dependence on the sample size N, and the subscript 'Θ' represents the support of the sampling prior $\pi_s(\theta)$. Note that in the Eq. (5), the probability inside of $\mathbf{1}\{A\}$ (that is, $\mathbb{P}_f[\cdot]$ is computed with respect to the posterior distribution $\pi_f(\theta|\mathbf{y}_N)$ (3) under the fitting prior, while the probability outside (that is, $\mathbb{P}_s[\cdot]$) are taken with respect to the marginal distribution $f_s(\mathbf{y}_N)$ (2) under the sampling prior. Note that the value $\beta_\Theta^{(N)}$ (5) also depends on the choice of the threshold (λ), the parameter spaces corresponding to the null and alternative hypothesis (Θ_0 and Θ_a), and the sampling and fitting priors ($\pi_s(\theta)$ and $\pi_f(\theta)$).

Monte Carlo simulation is employed to approximate the value of $\beta_\Theta^{(N)}$ (Eq. 5) in cases where it is not expressed as a closed-form formula:

$$\hat{\beta}_\Theta^{(N)} \approx \frac{1}{R}\sum_{r=1}^R \mathbf{1}\{\mathbb{P}_f[\theta \in \Theta_a|\mathbf{y}_N^{(r)}] > \lambda\}, \quad \mathbf{y}_N^{(r)} \sim f_s(\mathbf{y}_N), \quad (r = 1, \cdots, R),$$

This process produces N outcomes $\mathbf{y}_N$ from its prior predictive distribution (also called, marginal likelihood function)

$$\mathbf{y}_N \sim f_s(\mathbf{y}_N) = \int f(\mathbf{y}_N|\theta)\pi_s(\theta)d\theta. \quad (2)$$

After that, one calculates the posterior distribution of θ given the data $\mathbf{y}_N$, which is

$$\pi_f(\theta|\mathbf{y}_N) = \frac{f(\mathbf{y}_N|\theta)\pi_f(\theta)}{\int f(\mathbf{y}_N|\theta)\pi_f(\theta)d\theta}. \quad (3)$$

Eventually, a measure of evidence to reject the null hypothesis is summarized by the Bayesian test statistics, the posterior probability of the alternative hypothesis being true given the observations $\mathbf{y}_N$, which is

where R is the number of simulated datasets. When Monte Carlo simulation is used for regulatory submission in a Bayesian design to estimate the expected behavior of the Bayesian test statistics $T(\mathbf{y}_N)$, typically, one uses $R = 10,000$ or $100,000$ and also reports a 95% confidence interval for $\beta_\Theta^{(N)}$ to describe the precision of the approximation. Often, for complex designs, computing the Bayesian test statistic $T(\mathbf{y}_N) = \mathbb{P}_f[\theta \in \Theta_a|\mathbf{y}_N]$ itself requires the use of Markov Chain Monte Carlo (MCMC) sampling techniques, such as the Gibbs sampler or Metropolis-Hastings algorithm [65–67]. In such cases, a nested simulation technique is employed to approximate $\beta_\Theta^{(N)}$ (5) (Algorithm 1 in Supplemental material). It is important to note that when MCMC techniques are used, regulators recommend sponsors check the

convergence of the Markov chain to the posterior distribution [16], using various techniques to diagnose non-convergence [64, 65].

Now, we are ready to apply the above concept to Bayesian sample size determination. We consider two different populations from which the random sample of N observations $\mathbf{y}_N$ may have been drawn, with one population corresponding to the null parameter space Θ_0 and another population corresponding to the alternative parameter space Θ_a – similar to Neyman & Pearson's approach (based on hypothesis testing and type I and II error rates) [68].

This can be achieved by separately considering two scenarios: 'what if the parameter θ is likely to be within a specified portion of the null parameter space?' and 'what if the parameter θ is likely to be within a specified portion of the alternative parameter space?' Following notations from [62], let $\bar{\Theta}_0$ and $\bar{\Theta}_a$ denote the closures of Θ_0 and Θ_a, respectively. In this formulation, the null sampling prior $\pi_{s0}(\theta)$ is the distribution supported on the boundary $\Theta_B = \bar{\Theta}_0 \cap \bar{\Theta}_a$, and the alternative sampling prior $\pi_{s1}(\theta)$ is the distribution supported on the set $\Theta_a^* \subset \Theta_a$. For a one-sided test, such as $\mathcal{H}_0 : \theta \leq \theta_0$ versus $\mathcal{H}_a : \theta > \theta_0$, one may choose the null sampling prior $\pi_{s0}(\theta)$ as a point-mass distribution at θ_0, and the alternative sampling prior $\pi_{s1}(\theta)$ as a distribution supported on $\Theta_a^* \subset (\theta_0, \infty)$.

Eventually, for a given $\alpha > 0$ and $\beta > 0$, the Bayesian sample size is the value

$$N = \max\left(\min\{N : \beta_{\Theta_B}^{(N)} \leq \alpha\}, \min\{N : \beta_{\Theta_a^*}^{(N)} \geq 1 - \beta\}\right), \qquad (6)$$

where $\beta_{\Theta_B}^{(N)}$ and $\beta_{\Theta_a^*}^{(N)}$ are given in (5) corresponding to $\pi_s(\theta) = \pi_{s0}(\theta)$ and $\pi_s(\theta) = \pi_{s1}(\theta)$, respectively. The values of $\beta_{\Theta_B}^{(N)}$ and $\beta_{\Theta_a^*}^{(N)}$ are referred to as the Bayesian type I error and power, while $1 - \beta_{\Theta_a^*}^{(N)}$ is referred to as the Bayesian type II error. The sample size N satisfying the condition $\beta_{\Theta_B}^{(N)} \leq \alpha$ meets the Bayesian type I error requirement. Similarly, the sample size N satisfying the condition $\beta_{\Theta_a^*}^{(N)} \geq 1 - \beta$ meets the Bayesian Power requirement. Eventually, the selected sample size N (6) is the minimum value that simultaneously satisfies the

Bayesian type I error and power requirement. Typical values for α are 0.025 for a one-sided test and 0.05 for a two-sided test, and β is typically set to 0.1 or 0.2 regardless of the direction of the alternative hypothesis [16].

Figure 2 provides a flowchart illustrating the process of Bayesian sample size determination. We explain the practical algorithm for selecting an optimal Bayesian sample size N (6), subject to the maximum sample size $N_{\max}$ – typically chosen under budgetary limits. To begin, we consider a set of K candidate sample sizes, denoted as $\mathcal{N} = \{N_k; k = 1, \ldots, K, N_k < N_{k+1}, N_K = N_{\max}\}$. Often, one may include the frequentist sample size as a reference.

The process commences with the evaluation of the smallest sample size, N_1, checking whether it meets the Bayesian type I error and power requirements, i.e., $\beta_{\Theta_B}^{(N_1)} \leq \alpha$ and $\beta_{\Theta_a^*}^{(N_1)} \geq 1 - \beta$. To that end, we independently generate N_1 outcomes, $\mathbf{y}_{N_1}$, from the marginal distributions $f_{s0}(\mathbf{y}_{N_1})$ and $f_{s1}(\mathbf{y}_{N_1})$, which are based on the null and alternative sampling priors $\pi_{s0}(\theta)$ and $\pi_{s1}(\theta)$, respectively. The data drawn in this manner corresponds to the type I error and power scenarios, respectively. Subsequently, we independently compute the Bayesian test statistics, $T(\mathbf{y}_{N_1})$, using the common fitting prior $\pi_f(\theta)$, and record the testing results, whether it rejects the null hypothesis or not, $\mathbf{1}\{T(\mathbf{y}_{N_1}) > \lambda\}$ (4) for each scenario. By repeating this procedure R times (for example, $R = 10,000$), we can estimate the expected behaviors of the Bayesian test statistics $\beta_{\Theta_B}^{(N_1)}$ and $\beta_{\Theta_a^*}^{(N_1)}$ through Monte-Carlo approximation and evaluate whether the size N_1 meets both Bayesian type I error and power requirements. If these requirements are met, then N_1 is deemed the Bayesian sample size for the study. If not, we evaluate the next sample size, N_2, and reassess its suitability for meeting the requirements. This process continues until we identify the Bayesian sample size meeting the requirements within the set $\mathcal{N}$. If it cannot be found within this set $\mathcal{N}$, it may be necessary to explore a broader range of candidate sizes, adjust the values of α and β under regulatory consideration, modify the threshold λ, or consider other potential modifications such as changing the hyper-parameters of the fitting prior.
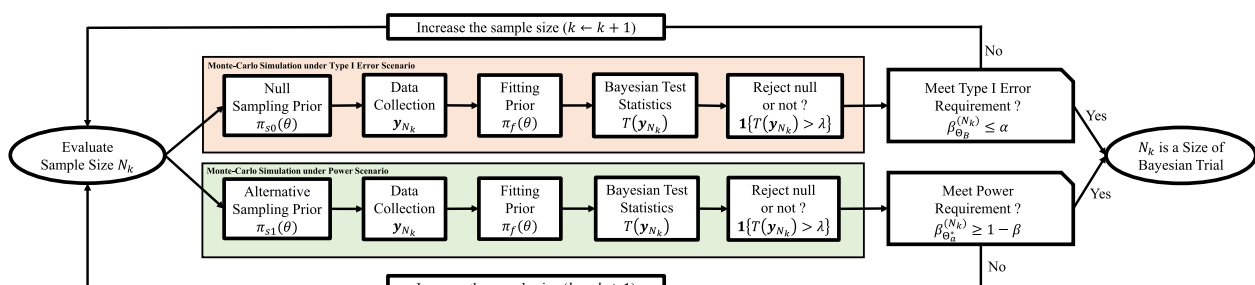

Fig. 2 Flow chart of Bayesian sample size determination within the collection of possible sizes of Bayesian trial $\mathcal{N} = \{N_k; k = 1, \cdots, K, N_k < N_{k+1}\}$

It is evident that Bayesian sample size determination is computationally intensive. It becomes even more intense when the complexity of the design increases. For instance, one needs to consider factors like the number and timing of interim analyses for Bayesian group sequential design, as well as the number of sub-groups and ratios in Bayesian platform design. Moreover, the computational complexity increases when the Bayesian test statistic requires MCMC sampling, as the convergence of the Markov chain should be diagnosed for each iteration within the Monte Carlo

$\pi_{s0}(\theta) = \delta(\theta_0)$ for some $\theta_0 \in \Theta_B = \bar{\Theta}_0 \cap \bar{\Theta}_a$ where $\delta(x)$ is the Direc-Delta function), and the alternative sampling prior is specified by a Dirac measure with the point-mass at the value inducing the minimally detectable treatment effect, representing the smallest effect size (i.e., $\pi_{s1}(\theta) = \delta(\theta_a)$ for some $\theta_a \in \Theta_a^* \subset \Theta_a$).

In this calibration, the expected behavior of the Bayesian test statistics $T(\mathbf{y}_N) = \mathbb{P}_f[\theta \in \Theta_a | \mathbf{y}_N]$ can be represented as the frequentist type I error and power of the design as follow:

$$\text{Type I error} : \beta_{\theta_0}^{(N)} = \mathbb{P}[T(\mathbf{y}_N) > \lambda | \mathbf{y}_N \sim f(\mathbf{y}_N|\theta_0)] = \mathbb{P}_{\theta_0}[T(\mathbf{y}_N) > \lambda], \tag{7}$$

simulation. In such scenarios, the use of parallel computation techniques or modern sampling schemes can significantly reduce computation time [69, 70].

Calibration of Bayesian trial design to assess frequentist operating characteristics

Scientifically sound clinical trial planning and rigorous trial conduct are important, regardless of whether trial sponsors use a Bayesian or frequentist design. Maintaining some degree of objectivity in the interpretation of testing results is key to achieving scientific soundness. The central question here is how much we can trust a testing result based on a Bayesian hypothesis testing procedure, which is driven by the Bayesian type I error and power in the planning phase. More specifically, suppose that such a Bayesian test, where the threshold of the decision rule was chosen to meet the Bayesian type I error rate of less than 0.025 and power greater than 0.8, yielded the rejection of the null hypothesis, while a frequentist test did not upon completion of the study. Then, can we still use the result of the Bayesian test for registration purposes? Perhaps, this can be best addressed by calculating the frequentist type I error and power of the Bayesian test during the planning phase so that the Bayesian test can be compared with some corresponding frequentist test in an apple-to-apple comparison, or as close as possible.

In most regulatory submissions, Bayesian trial designs are 'calibrated' to possess good frequentist properties. In this spirit, and in adherence to regulatory practice, regulatory agencies typically recommend that sponsors provide the frequentist type I and II error rates for the sponsor's proposed Bayesian analysis plan [16, 71].

The simulation-based approach for Bayesian sample size determination [60], as illustrated in A simulation principle of Bayesian sample size determination section, is calibrated to measure the frequentist operating characteristics of a Bayesian trial design if the null sampling prior is specified by a Dirac measure with the point-mass at the boundary value of the null parameter space Θ_0 (i.e.,

$$\text{Power} : \beta_{\theta_a}^{(N)} = \mathbb{P}[T(\mathbf{y}_N) > \lambda | \mathbf{y}_N \sim f(\mathbf{y}_N|\theta_a)] = \mathbb{P}_{\theta_a}[T(\mathbf{y}_N) > \lambda]. \tag{8}$$

Throughout the paper, we interchangeably use the notations $\mathbb{P}_\theta[\cdot]$ and $\mathbb{P}[\cdot | \mathbf{y}_N \sim f(\mathbf{y}_N|\theta)]$. The former notation is simpler, yet it omits specifying which values are being treated as random and which are not; hence, the latter notation is sometimes more convenient for Bayesian computation.

With the aforementioned calibration, the prior specification problem of the Bayesian design essentially boils down to the choice of the fitting prior $\pi_f(\theta)$. This is because the selection of the null and alternative sampling prior is essentially determined by the formulation of the null and alternative hypotheses, aligning with the frequentist framework. In other words, the fitting prior provides the unique advantage of Bayesian design by incorporating prior information about the parameter θ, which is then updated by Bayes' theorem, leading to the posterior distribution. The choice of the fitting prior will be discussed in Specification of prior distributions section. In what follows, to avoid notation clutter, we omit the subscript 'f' in the notation of the fitting prior $\pi_f(\theta)$.

Example - standard single-stage design based on beta-binomial model

Suppose a medical device company aims to evaluate the primary safety endpoint of a new device in a pivotal trial. The safety endpoint is the primary adverse event rate through 30 days after a surgical procedure involving the device. The sponsor plans to conduct a single-arm study design in which patient data is accumulated throughout the trial. Only once the trial is complete, the data will be unblinded, and the pre-planned statistical analyses will be executed. Suppose that the null and alternative hypotheses are: $\mathcal{H}_0 : \theta \geq \theta_0$ versus $\mathcal{H}_a : \theta < \theta_0$. Here, θ_0 represents the performance goal of the new device, a numerical value (point estimate) that is considered sufficient by a regulator for use as a comparison for the safety endpoint. It is recommended that the performance goal

not originate from a particular sponsor or regulator. It is often helpful if it is recommended by a scientific or medical society [72].

A fundamental regulatory question is "when a device passes a safety performance goal, does that provide evidence that the device is safe?". To answer this question, the sponsor sets a performance goal by $\theta_0 = 0.12$, and anticipates that the safety rate of the new device is $\theta_a = 0.05$. The objective of the study is, therefore, to detect a minimum treatment effect of $7\% = 12\% - 5\%$ in reducing the adverse event rate of patients treated with the new medical device compared to the performance goal. The sponsor targeted to achieve a statistical power of $1 - \beta = 0.8$ with the one-sided level $\alpha = 0.025$ test of a proposed design. The trial is successful if the null hypothesis $\mathcal{H}_0$ is rejected after observing the outcomes from N patients upon completion of the study.

The following Bayesian design is considered:

- One-sided significance level: $\alpha = 0.025$,
- Power: $1 - \beta = 0.8$,
- Null sampling prior: $\pi_{s0}(\theta) = \delta(\theta_0)$, where $\theta_0 = 0.12$,
- Alternative sampling prior: $\pi_{s1}(\theta) = \delta(\theta_a)$, where $\theta_a = 0.05$,
- Prior: $\theta \sim \pi(\theta) = \mathcal{B}eta(\theta|a, b)$,
- Hyper-parameters: $a > 0$ and $b > 0$,
- Likelihood: $y_i \sim f(y|\theta) = \mathcal{B}ernoulli(y|\theta)$, $(i = 1, \cdots, N)$,
- Decision rule: Reject null hypothesis if $T(\mathbf{y}_N) = \mathbb{P}[\theta < \theta_0|\mathbf{y}_N] > 0.975$.

Under the setting, (frequentist) type I error and power of the Bayesian design can be expressed as:

Here, the integral expression $(\int \cdot)$ can be further simplified to summation expression $(\sum \cdot)$ by using a binomial distribution, similar to [73].

The Bayesian sample size satisfying the type I & II error requirements are then

$$N = \max(\min\{N : \beta_{\theta_0}^{(N)} \le 0.025\}, \min\{N : \beta_{\theta_a}^{(N)} \ge 0.8\}).$$

Due the conjugate relationship between the binomial distribution and beta prior, the posterior distribution is the beta distribution, $\pi(\theta|\mathbf{y}_N) = \mathcal{B}eta(x + a, N - x + b)$ such that $x = \sum_{i=1}^{N} y_i$. Therefore, the Bayesian test statistics $T(\mathbf{y}_N) = \mathbb{P}[\theta < \theta_0|\mathbf{y}_N]$ can be represented as a closed-form in this case.

We consider $N = 100, 150$, and 200 as the possible sizes for the Bayesian trial. We evaluate three prior options: (1) a non-informative prior with $a = b = 1$ (prior mean is 50%), (2) an optimistic prior with $a = 0.8$ and $b = 16$ (prior mean is 4.76%), and (3) a pessimistic prior with $a = 3.5$ and $b = 20$ (prior mean is 14.89%). An optimistic prior assigns a probability mass that is favorable for rejecting the null hypothesis before observing any new outcomes, while a pessimistic prior assigns a probability mass that is favorable for accepting the null hypothesis before observing any new outcomes. As a reference, we consider a frequentist design in which the decision criterion is determined by the p-value associated with the z-test statistic, $Z = (x/N - \theta_0)/\sqrt{\theta_0(1 - \theta_0)/N}$, being less than the one-sided significance level of $\alpha = 0.025$ to reject the null hypothesis.

Table 1 shows the results of the power analysis obtained by simulation. Designs satisfying the requirement of type

$$\beta_{\theta_0}^{(N)} = \mathbb{P}_{\theta_0}[\mathbb{P}[\theta < \theta_0|\mathbf{y}_N] > 0.975] = \int \mathbf{1}(\mathbb{P}[\theta < \theta_0|\mathbf{y}_N] > 0.975) \cdot \prod_{i=1}^{N} \theta_0^{y_i}(1 - \theta_0)^{1-y_i} d\mathbf{y}_N,$$

$$\beta_{\theta_a}^{(N)} = \mathbb{P}_{\theta_a}[\mathbb{P}[\theta < \theta_0|\mathbf{y}_N] > 0.975] = \int \mathbf{1}(\mathbb{P}[\theta < \theta_0|\mathbf{y}_N] > 0.975) \cdot \prod_{i=1}^{N} \theta_a^{y_i}(1 - \theta_a)^{1-y_i} d\mathbf{y}_N.$$

Table 1 Frequentist operating characteristics of Bayesian designs with different prior options

Sample Size (*N*)	Bayesian Design (Non-informative prior)		Bayesian Design (Optimistic prior)		Bayesian Design (Pessimistic prior)		Frequentist Design (Z-test statistics)	
	Type I Error	Power	Type I Error	Power	Type I Error	Power	Type I Error	Power
100	0.0148	0.6181	0.0755	0.8767	0.0148	0.6181	0.0155	0.6214
150	**0.0231**	**0.8690**	0.0448	0.9268	0.0114	0.7838	**0.0242**	**0.8690**
200	**0.0164**	**0.9184**	0.0467	0.9767	**0.0164**	**0.9184**	0.0158	0.9231

Note: Bayesian designs are based on the beta-binomial models with prior options: (1) a non-informative prior with $a = b = 1$, (2) an optimistic prior with $a = 0.8$ and $b = 16$, and (3) a pessimistic prior with $a = 3.5$ and $b = 20$

I error $\leq 2.5\%$ and power $\geq 80\%$, are highlighted in bold in the table. The results indicate that the operating characteristics of the Bayesian design based on a non-informative prior are very similar to those obtained using the frequentist design. This similarity is typically expected because a non-informative prior has minimal impact on the posterior distribution, allowing the data to play a significant role in determining the results.

The results show that the Bayesian design based on an optimistic prior tends to increase power at the expense of inflating the type I error. Technically, the inflation is expected because, by definition, the type I error is evaluated by assuming the true treatment effect is null (i.e. $\theta = \theta_0$), then it is calculated under a scenario where the prior is in conflict with the null treatment effect, resulting in the inflation of the type I error. In contrast, the Bayesian design based on a pessimistic prior tends to decrease the type I error at the cost of deflating the power. The deflation is expected because, by definition, the power is evaluated by assuming the true treatment effect is alternative (i.e. $\theta = \theta_a$), then it is calculated under a scenario where the prior is in conflict with the alternative treatment effect, resulting in the deflation of the power.

Considering the trade-off between power and type I error, which is primarily influenced by the prior specification, thorough pre-planning is essential for selecting the most suitable Bayesian design on a case-by-case basis for regulatory submission. Particularly, when historical data is incorporated into the hyper-parameter of the prior as an optimistic prior, there may be inflation of the type I error rate, even after appropriately discounting the historical data [74]. In such cases, it may be appropriate to relax the type I error control to a less stringent level compared to situations where no prior information is used. This is because the power gains from using external prior information in clinical trials are typically not achievable when strict type I error control is required [75, 76]. Refer to Section 2.4.3 in [77] for relevant discussion. The extent to which type I error control can be relaxed is a case-by-case decision for regulators, depending on various factors, primarily the confidence in the prior information [16]. We discuss this in more detail by taking the Bayesian borrowing design based on a power prior [36] as an example in External data borrowing section.

Numerical approximation of power function

In this subsection, we illustrate a numerical method to approximate the power function of a Bayesian hypothesis testing procedure. The power function of a test procedure is the probability of rejecting the null hypothesis, with the true parameter value as the input. The power function plays a crucial role in assessing the ability of a statistical test to detect a true effect or relationship between the design parameters. Visualizing the power function over the parameter space, as provided by many statistical software (SAS, PASS, etc), is helpful for trial sizing because it displays the full spectrum of the behavior of the testing procedure. Understanding such behaviors is crucial for regulatory submission, as regulators often recommend simulating several likely scenarios and providing the expected sample size and estimated type I error for each case.

Consider the null and alternative hypotheses, $\mathcal{H}_0 : \theta \in \Theta_0$ versus $\mathcal{H}_a : \theta \in \Theta_a$, where $\Theta = \Theta_0 \cup \Theta_a$, and Θ_0 and Θ_a are disjoint. Let outcomes y_i ($i = 1, \cdots, N$) be identically and independently distributed according to a density $f(y|\theta)$. Given a Bayesian test statistics $T(\mathbf{y}_N)$, suppose that a higher value of $T(\mathbf{y}_N)$ raises more doubt about the null hypothesis being true. We reject the null hypothesis if $T(\mathbf{y}_N) > \lambda$, where λ is a pre-specified threshold. Then, the power function $\psi : \Theta \to [0, 1]$ is defined as follows:

$$
\begin{aligned}
\psi(\theta) &= \mathbb{P}_\theta[T(\mathbf{y}_N) > \lambda] = \mathbb{P}[T(\mathbf{y}_N) > \lambda | \mathbf{y}_N \sim f(y|\theta)] \\
&= \int \mathbf{1}\{T(\mathbf{y}_N) > \lambda\} \prod_{i=1}^{n} f(y_i|\theta) d\mathbf{y}_N.
\end{aligned}
\tag{9}
$$

Eventually, one needs to calculate $\psi(\theta)$ over the entire parameter space Θ to explore the behavior of the testing procedure. However, the value of $\psi(\theta)$ is often not expressed as a closed-form formula, mainly due to two reasons: no explicit formula for the outside integral $\mathbb{P}_\theta[\cdot]$ or the Bayesian test statistics $T(\mathbf{y}_N)$. Thus, it is often usual that the value of $\psi(\theta)$ is approximated through a nested simulation strategy. See Algorithm 1 in Supplemental material. The idea of the Algorithm 1 is that the outside integral in (9) is approximated by a Monte-Carlo simulation (with R number of replicated studies), and the test statistics is approximated by Monte-Carlo or Markov Chain Monte-Carlo simulation (with S number of posterior samples) when the test statistics are not expressed in closed form. It is important to note that this approximation is exact in the sense that if R and S go to infinity, then $\widetilde{\psi}(\theta)$ converges to the truth $\psi(\theta)$. This contrasts with the formulation of the power functions of many frequentist tests, which are derived based on some large sample theory [78], to induce a closed-form formula.

Specification of prior distributions
Classes of prior distributions

The prior distributions for regulatory submissions can be broadly classified into non-informative priors and informative priors. A non-informative prior is a prior distribution with no preference for any specific parameter value. A Bayesian design based on a

non-informative prior leads to objective statistical inference, resembling frequentist inference, and is therefore the least controversial. It is important to note that choosing a non-informative prior distribution can sometimes be challenging, either because there may be more than one way to parameterize the problem or because there is no clear mathematical justification for defining non-informativeness. [79] reviews the relevant literature but emphasizes the continuing difficulties in defining what is meant by 'non-informative' and the lack of agreed reference priors in all but simple situations.

For example, in the case of a beta-binomial model (as illustrated in Calibration of Bayesian trial design to assess frequentist operating characteristics section), choices such as $\mathcal{B}eta(1,1)$, $\mathcal{B}eta(0.5,0.5)$, $\mathcal{B}eta(1/3,1/3)$, or $\mathcal{B}eta(0.001,0.001)$ could all be used as non-informative priors. Refer to Subsection 5.5.1 of [17] and the paper by [80] for a relevant discussion. In Bayesian hierarchical models, the mathematical meaning of a non-informative prior distribution is not obvious due to the complexity of the model. In those cases, we typically set the relevant hyper-parameters to diffuse the prior evenly over the parameter space and minimize the prior information as much as possible, leading to a nearly non-informative prior.

On the other hand, an informative prior is a prior distribution that expresses a preference for a particular parameter value, enabling the incorporation of prior information. Informative priors can be further categorized into two types: prior distributions based on empirical evidence from previous trials and prior distributions based on personal opinions, often obtained through expert elicitation. The former class of informative priors is less controversial when the current and previous trials are similar to each other. Possible sources of prior information include: clinical trials conducted overseas, patient registries, clinical data on very similar products, and pilot studies. Recently, there has been breakthrough development of informative prior distribution that enables incorporating the information from previous trials, and eventually reducing sample size of a new trial, while providing appropriate mechanism of discounting [81–84]. We provide details on the formulation of an informative prior and relevant regulatory considerations in External data borrowing section. Typically, informative prior distribution based on personal opinions is not recommended for Bayesian submissions due to subjectivity and controversy [85].

Incorporating prior information formally into the statistical analysis is a unique feature of the Bayesian approach but is also often criticized by non-Bayesians. To mitigate any conflict and skepticism regarding prior information, it is crucial that sponsors and regulators meet early in the process to discuss and agree upon the prior information to be used for Bayesian clinical trials.

Prior probability of the study claim

The prior predictive distribution plays a key role in pre-planning a Bayesian trial to measure the prior probability of the study claim – the probability of the study claim before observing any new data. Regulators recommend that this probability should not be excessively high, and what constitutes 'too high' is a case-by-case decision [16]. Measuring this probability is typically recommended when an informative prior distribution is used for the Bayesian submission. Regulatory agencies make this recommendation to ensure that prior information does not overwhelm the data of a new trial, potentially creating a situation where unfavorable results from the proposed study get masked by a favorable prior distribution. In an evaluation of the prior probability of the claim, regulators will balance the informativeness of the prior against the efficiency gain from using prior information, as opposed to using noninformative priors.

To calculate the prior probability of the study claim, we simulate multiple hypothetical trial data using the prior predictive distribution (2) by setting the sampling prior as the fitting prior, and then calculate the probability of rejecting the null hypothesis based on the simulated data. We illustrate the procedure for calculating this probability using the beta-binomial model illustrated in Calibration of Bayesian trial design to assess frequentist operating characteristics section as an example. First, we generate the data $(\mathbf{y}_N)^{(r)} \sim f(\mathbf{y}_N) = \int f(\mathbf{y}_N|\theta)\pi(\theta)d\theta$ ($r = 1, \cdots, R$), where R represents the number of simulations. Here, f is the Bernoulli likelihood, and π is the beta prior with hyper-parameters a and b. In this particular example, a and b represent the number of hypothetical patients showing adverse events and not showing adverse events *a priori*, hence $a + b$ is the prior effective sample size. The number of patients showing adverse events out of N patients, $X^{(r)} = \sum_{i=1}^{N} y_i^{(r)}$, is distributed according to a beta-binomial distribution [86], denoted as $X^{(r)} \sim \mathcal{B}eta\text{-}\mathcal{B}inom(N, a, b)$. One can use a built-in function rbetabinom.ab($\cdot$) within the R package VGAM to generate the r-th outcome $X^{(r)}$. Second, we compute the posterior probability and make a decision whether to reject the null or not, i.e., $d(r) = \mathbf{1}\{\mathbb{P}[\theta < \theta_0|\mathbf{y}_N^{(r)}] > 0.975\} = 1$ if $\mathcal{H}_0$ is rejected and 0 otherwise. Finally, the value of $\sum_{r=1}^{R} d(r)/R$ is the prior probability of the study claim based on the prior choice of $\theta \sim \pi(\theta) = \mathcal{B}eta(\theta|a, b)$.

We consider four prior options where the hyper-parameters have been set to induce progressively stronger prior information to reject the null *a priori*. Table 2 shows the results of the calculations of this probability. For the

Table 2 Prior probability of the study claim based on beta-binomial model

Prior Distribution	Number of hypothetical patients showing adverse events	Number of hypothetical patients not showing adverse events	Prior mean (standard deviation)	Prior probability of a study claim
$\mathcal{B}eta(1,1)$	1	1	50% (5.8%)	5.8%
$\mathcal{B}eta(1,9)$	1	9	10% (9%)	47.1%
$\mathcal{B}eta(1,19)$	1	19	5% (4.9%)	77.3%
$\mathcal{B}eta(1,49)$	1	49	2% (2%)	99.1%

non-informative prior, the prior probability of the study claim is only 5.8%, implying that the outcome from a new trial will most likely dominate the final decision. However, the third and fourth options provide probabilities greater than 50%, indicating overly strong prior information; hence, appropriate discounting on the prior effective sample size is recommended.

Decision rule - posterior probability approach
Posterior probability approach
The central motivation for utilizing the posterior probability approach in decision-making is to quantify the evidence to address the question, "Does the current data provide convincing evidence in favor of the alternative hypothesis?" The key quantity here is the posterior probability of the alternative hypothesis being true based on the data observed up to the point of analysis. This Bayesian tail probability can be used as the test statistic in a single-stage Bayesian design upon completion of the study, similar to the role of the p-value in a single-stage frequentist design [77]. Furthermore, one can measure it in both interim and final analyses within the context of Bayesian group sequential designs [19, 46], akin to a z-score in a frequentist group sequential design [87, 88].

It is important to note that if the posterior probability approach is used in decision-making at the interim analysis, it does not involve predicting outcomes of the future remaining patients. This distinguishes it from the predictive probability approach, where the remaining time and statistical information to be gathered play a crucial role in decision-making at the interim analysis (as discussed in Decision rule - predictive probability approach section). Consequently, the posterior probability approach is considered conservative, as it may prohibit imputation for incomplete data or partial outcomes. For this reason, the posterior probability approach is standardly employed in interim analyses to declare early success or in the final analysis to declare the trial's success to support marketing approval of medical devices or drugs in the regulatory submissions [23, 89].

Suppose that $\mathbf{y}$ denotes an analysis dataset, and θ is the parameter of main interest. A sponsor wants to test $\mathcal{H}_0 : \theta \in \Theta_0$ versus $\mathcal{H}_a : \theta \in \Theta_a$, where $\Theta = \Theta_0 \cup \Theta_a$,

and Θ_0 and Θ_a are disjoint. Bayesian test statistics following the posterior probability approach can be represented as a functional $\mathcal{F}\{\cdot\} : \mathcal{Q}_{\theta|\mathbf{y}} \to [0,1]$, such that:

$$\mathcal{F}\{\pi(\theta|\mathbf{y})\} = T(\mathbf{y}) = \mathbb{P}[\theta \in \Theta_a|\mathbf{y}] = \int \mathbf{1}(\theta \in \Theta_a) \cdot \pi(\theta|\mathbf{y})d\theta, \tag{10}$$

where $\mathcal{Q}_{\theta|\mathbf{y}}$ represents the collection of posterior distributions. Finally, to induce a dichotomous decision, we need to pre-specify the threshold $\lambda \in [0,1]$. By introducing an indicator function φ (referred as a 'critical function' in [63]), the testing result is determined as follow:

$$\varphi(\mathbf{y}) = \begin{cases} 1 & \text{if } \mathcal{F}\{\pi(\theta|\mathbf{y})\} = \mathbb{P}[\theta \in \Theta_a|\mathbf{y}] > \lambda \\ 0 & \text{if } \mathcal{F}\{\pi(\theta|\mathbf{y})\} = \mathbb{P}[\theta \in \Theta_a|\mathbf{y}] \leq \lambda, \end{cases}$$

where 1 and 0 indicate the rejection and acceptance of the null hypothesis, respectively.

In the interim analysis, rejecting the null can be interpreted as claiming the early success of the trial, and in the final analysis, rejecting the null can be interpreted as claiming the final success of the trial. Figure 3 displays a pictorial description of the decision procedure.

The formulation of Bayesian test statistics is universal regardless of the hypothesis being tested (e.g., mean comparison, proportion comparison, association), and it does not rely on asymptotic theory. The derivation procedure for Bayesian test statistics based on the posterior probability approach is intuitive, considering the backward process of the Bayesian theorem. A higher value of $T(\mathbf{y}) = \mathbb{P}[\theta \in \Theta_a|\mathbf{y}]$ implies that more mass has been concentrated on the alternative parameter space Θ_a *a posteriori*. Consequently, there is a higher probability that the data were originally generated from the density indexed with parameters belonging to Θ_a, that is, $\mathbf{y} \sim f(\mathbf{y}|\theta)$, $\theta \in \Theta_a$. The prior distribution in this backward process acts as a moderator by appropriately allocating even more or less mass on the parameter space Θ before seeing any data $\mathbf{y}$. If there is no prior information, the prior distribution plays a minimal role in this process.

This contrasts with the derivation procedure for frequentist test statistics, which involves formulating a point estimator such as sufficient statistics from
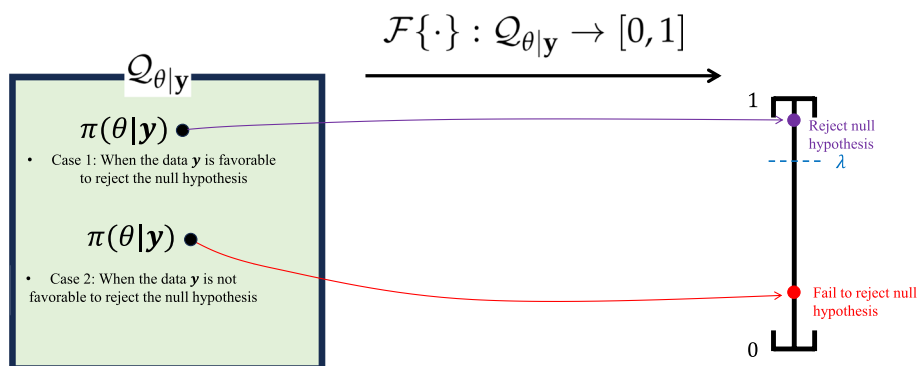
Fig. 3 Pictorial illustration of the decision rule based on the posterior probability approach: If the data **y** were generated from the alternative (or null) density $f(\mathbf{y}|\theta)$ where $\theta \in \Theta_a$ (or $\theta \in \Theta_0$), then the posterior distribution would be more concentrated on the alternative space Θ_a (or null parameter Θ_0), resulting in a higher (or lower) value of the test statistic $\mathcal{F}\{\pi(\theta|\mathbf{y})\} = \mathbb{P}[\theta \in \Theta_a|\mathbf{y}]$. The pre-specified threshold λ is used to make the dichotomous decision based on the test statistic

the sample data to make a decision about a specific hypothesis. The derivation may vary depending on the type of test (e.g., t-test, chi-squared test, z-test) and the hypothesis being tested. Furthermore, asymptotic theory is often used if the test statistics based on exact calculation are difficult to obtain [53].

For a single-stage design with the targeted one-sided significance level of α, the threshold λ is normally set to $1 - \alpha$, provided that the test is a one-sided test and the prior distribution is a non-informative prior. This setting is frequently chosen, particularly when there is no past historical data to be incorporated into the prior; see the example of the beta-binomial model in Calibration of Bayesian trial design to assess frequentist operating characteristics section. If an informative prior is used, this convention (that is, $\lambda = 1 - \alpha$) should be carefully used because the type I error rate can be inflated or deflated based on the direction of the informativeness of prior distribution (see Table 1).

Asymptotic property of posterior probability approach
Bernstein-Von Mises theorem [90, 91], also called Bayesian central limit theorem, states that if the sample size N is sufficiently large, the influence of the prior $\pi(\theta)$ diminishes, and the posterior distribution $\pi(\theta|\mathbf{y}_N)$ closely resembles the likelihood $f(\mathbf{y}_N|\theta)$ under suitable regularity conditions (for e.g., conditions stated in [91] or Section 4.1.2 of [92]). Consequently, it simplifies the complex posterior distribution into a more manageable normal distribution, independent of the form of prior, as long as the prior distribution is continuous and positive on the parameter space.

By using Bernstein-Von Mises theorem, we can show that if the sample size N is sufficiently large, the posterior probability approach asymptotically behaves similarly to the frequentist testing procedure based on the *p*-value approach [93] under the regularity conditions. For the ease of exposition, we consider a one-sided testing problem. In this specific case, we further establish an asymptotic equation between the Bayesian tail probability (10) and *p*-value.

Theorem 1 *Let a random sample of size N, y_i, $(i = 1, \ldots, N)$, be independently and identically taken from a distribution $f(y|\theta)$ depending on the real parameter $\theta \in \Theta \subset \mathbb{R}$. Consider a one-sided testing problem $\mathcal{H}_0 : \theta \leq \theta_0$ versus $\mathcal{H}_A : \theta > \theta_0$, where θ_0 denotes the performance goal. Consider testing procedures with two paradigms:*

Frequentist testing procedure : $T_1(\mathbf{y}_N) > \lambda_1 \iff$ Reject $\mathcal{H}_0$;

Bayesian testing procedure : $T_2(\mathbf{y}_N) > \lambda_2 \iff$ Reject $\mathcal{H}_0$,

where $T_1(\mathbf{y}_N)$ is the maximum likelihood estimator and $T_2(\mathbf{y}_N)$ is the Bayesian test statistics based on posterior probability approach, that is, $T_2(\mathbf{y}_N) = \mathbb{P}[\theta > \theta_0|\mathbf{y}_N]$. λ_1 and λ_2 denote threshold values for the testing procedures. For frequentist testing procedure, we assume that $T_1(\mathbf{y}_N)$ itself serves as the frequentist test statistics of which higher values cast doubt against the null hypothesis $\mathcal{H}_0$, and $p(\mathbf{y}_N)$ denotes the p-value. For Bayesian testing procedure, assume that the prior density $\pi(\theta)$ is continuous and positive on the parameter space Θ.

Under the regularity conditions necessary for the validity of normal asymptotic theory of the maximum likelihood estimator and posterior distribution, and assuming the null hypothesis to be true, it holds that

$$\mathbb{P}[\theta > \theta_0|\mathbf{y}_N] \approx 1 - p(\mathbf{y}_N) \quad \text{for large } N, \qquad (11)$$

independently of the form of $\pi(\theta)$.

The proof can be found in Supplemental material.

Typically, for regulatory submissions, the significance level of the one-sided superiority test (e.g., $\mathcal{H}_0 : \theta \leq \theta_0$ versus $\mathcal{H}_A : \theta > \theta_0$, with the performance goal θ_0) is 2.5%. To achieve a one-sided significance level of $\alpha = 0.025$ for a frequentist design, one would use the decision rule $p(\mathbf{y}_N) < 0.025$ to reject the null hypothesis, where $p(\mathbf{y}_N)$ denotes the *p*-value. The *p*-value is often called the 'observed significance level' because the value by itself represents the evidence against a null hypothesis based on the observed data $\mathbf{y}_N$ [94].

Theorem 1 states that the value of the Bayesian tail probability (10) itself also serves as the evidence for the statistical significance. Furthermore, a Bayesian decision rule of $\mathbb{P}[\theta > \theta_0 | \mathbf{y}_N] > 0.975$ will lead to the one-sided significance level of 0.025, regardless of the choice of prior, whether it is informative or non-informative, under regularity conditions, if the sample size N is sufficiently large.

We illustrate Theorem 1 by using the beta-binomial model described in Calibration of Bayesian trial design to assess frequentist operating characteristics section as an example. Recall that, under sample sizes of $N = 100$, $N = 150$, and $N = 250$, Bayesian designs with non-informative priors meet the type I error requirement, while Bayesian designs with optimistic and pessimistic priors inflate and deflate the type I error, respectively (see Table 1). Under the same settings (that is, Bayesian threshold $\lambda = 0.975$), we now increase the sample size N up to 100,000 to explore the asymptotic behavior of the Bayesian designs. Figure 4 shows the results, where the inflation and deflation induced by the choice of the prior are getting washed out as N increases. When N is as large as 25,000 or more, the type I errors of all the Bayesian designs approximately achieve the type I error rate of 2.5%, implying that the asymptotic Eq. (10) holds.

In practice, the sample size (N) for pivotal trials in medical device development and phase II trials in drug development often leads to a modest sample size, and there are practical challenges limiting the feasibility of conducting larger studies [95]. Consequently, the asymptotic Eq. (10) may not hold in such limited sample sizes. Therefore, sponsors need to conduct extensive simulation experiments in the pre-planning of Bayesian clinical trials to best leverage existing prior information while controlling the type I error rate.

Bayesian group sequential design

An adaptive design is defined as a clinical study design that allows for prospectively planned modifications based on accumulating study data without undermining the study's integrity and validity [16, 40, 41]. In nearly all situations, to preserve the integrity and validity of a study, modifications should be prospectively planned and described in the clinical study protocol prior to initiation of the study [16]. Particularly, for Bayesian adaptive designs, including Bayesian group sequential designs, clinical trial simulation is a fundamental tool to explore, compare, and understand the operating characteristics, statistical properties, and adaptive decisions to answer the given research questions [96].

Posterior probability approach is widely adopted as a decision rule for complex innovative designs. In such designs, the choice of the threshold value(s) often depends on several factors, including the complexity of trial design, specific objectives, the presence of interim analyses, ethical considerations, statistical methodology, prior information, and type I & II error requirements.
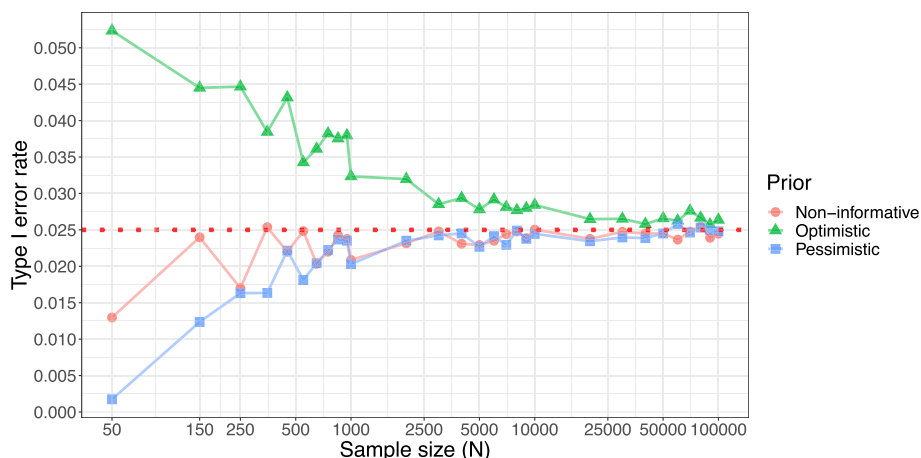

Fig. 4 Type I error rates of Bayesian designs based on the beta-binomial model with three prior options for testing $\mathcal{H}_0 : \theta \geq \theta_0$ versus $\mathcal{H}_a : \theta < \theta_0$, where $\theta_0 = 0.12$. Prior options are (1) a non-informative prior with $a = b = 1$, (2) an optimistic prior with $a = 0.8$ and $b = 16$, and (3) a pessimistic prior with $a = 3.5$ and $b = 20$

Consider a multi-stage design where the sponsor wants to use the posterior probability approach as an early stopping option for the trial success at interim analyses as well as the success at the final analysis. Let $\mathbf{y}^{(k)}$ $(k = 1, \ldots, K)$ denote the analysis dataset at the k-th interim analysis (thus, the K-th interim analysis is the final analysis), and θ is the parameter of main interest. The sponsor wants to test $\mathcal{H}_0 : \theta \in \Theta_0$ versus $\mathcal{H}_a : \theta \in \Theta_a$, where $\Theta = \Theta_0 \cup \Theta_a$, and Θ_0 and Θ_a are disjoint. One can use the following sequential decision criterion:

to declare the trial a success. In contrast, sequential design offers K chances to declare success at interim analyses and the final analysis. However, having K opportunities to declare success implies that there are K ways the trial can be falsely considered successful when it is not truly successful. These are the K false positive scenarios, and controlling the overall type I error rate is crucial to maintain scientific integrity for regulatory submission [16].

Similar to frequentist group sequential designs, our primary concern here is to control the overall type I

$$\text{1-st interim analysis} : T(\mathbf{y}^{(1)}) = \mathbb{P}[\theta \in \Theta_a | \mathbf{y}^{(1)}] > \lambda_1 \iff \text{Reject } \mathcal{H}_0,$$

$$\text{2-nd interim analysis} : T(\mathbf{y}^{(2)}) = \mathbb{P}[\theta \in \Theta_a | \mathbf{y}^{(2)}] > \lambda_2 \iff \text{Reject } \mathcal{H}_0,$$

$$\vdots$$

$$\text{K-1-th interim analysis} : T(\mathbf{y}^{(K-1)}) = \mathbb{P}[\theta \in \Theta_a | \mathbf{y}^{(K-1)}] > \lambda_{K-1} \iff \text{Reject } \mathcal{H}_0,$$

$$\text{K-th interim analysis} : T(\mathbf{y}^{(K)}) = \mathbb{P}[\theta \in \Theta_a | \mathbf{y}^{(K)}] > \lambda_K \iff \text{Reject } \mathcal{H}_0.$$

Figure 5 displays the processes of decision rules based on single-stage design and K-stage group sequential design. In practice, a general rule suggests that planning for a maximum of five interim analyses ($K = 5$) is often sufficient [52]. In single-stage design, there is only one opportunity

error rate of the sequential testing procedure. The overall type I error rate refers to the probability of falsely rejecting the null hypothesis $\mathcal{H}_0$ at any analysis, given that $\mathcal{H}_0$ is true. In this example, the overall type I error rate is given by:

$$
\begin{aligned}
&\mathbb{P}[T(\mathbf{y}^{(1)}) > \lambda_1 \text{ or } \cdots \text{ or } T(\mathbf{y}^{(K)}) > \lambda_K | \mathbf{y}^{(l)} \sim f(y|\theta_0), (l = 1, \cdots, K)] \\
&= \mathbb{P}[T(\mathbf{y}^{(1)}) > \lambda_1 | \mathbf{y}^{(1)} \sim f(y|\theta_0)] \\
&+ \mathbb{P}[T(\mathbf{y}^{(1)}) \leq \lambda_1 \text{ and } T(\mathbf{y}^{(2)}) > \lambda_2 | \mathbf{y}^{(l)} \sim f(y|\theta_0), (l = 1, 2)] \\
&+ \mathbb{P}[T(\mathbf{y}^{(1)}) \leq \lambda_1 \text{ and } T(\mathbf{y}^{(2)}) \leq \lambda_2 \text{ and } T(\mathbf{y}^{(3)}) > \lambda_3 | \mathbf{y}^{(l)} \sim f(y|\theta_0), (l = 1, 2, 3)] \\
&+ \cdots \\
&+ \mathbb{P}[T(\mathbf{y}^{(l)}) \leq \lambda_l, (l = 1, \cdots, K-1) \text{ and } T(\mathbf{y}^{(K)}) > \lambda_K | \mathbf{y}^{(l)} \sim f(y|\theta_0), (l = 1, \cdots, K)],
\end{aligned}
\tag{12}
$$

(a) Single-stage Bayesian Design
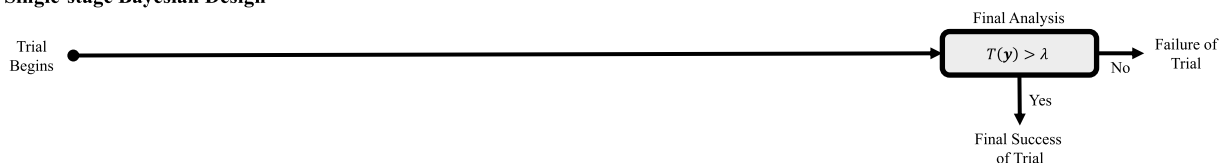
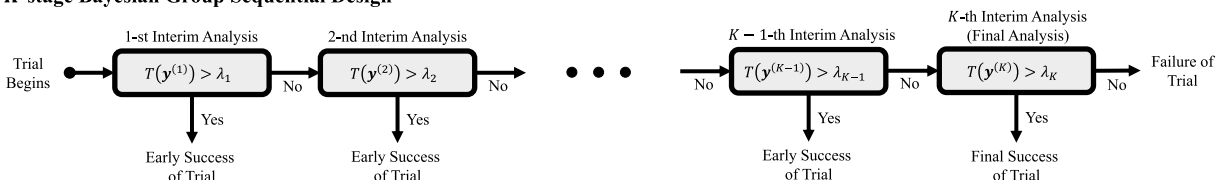**(b) *K*-stage Bayesian Group Sequential Design**

Fig. 5 Processes of fixed design (**a**) and sequential design (**b**). The former allows only a single chance to declare success for the trial, while the latter allows K chances to declare success. The test statistic for the former design is denoted as $T(\mathbf{y}) = \mathbb{P}[\theta \in \Theta_a | \mathbf{y}]$, and for the latter design, they are $T(\mathbf{y}^{(k)}) = \mathbb{P}[\theta \in \Theta_a | \mathbf{y}^{(k)}]$, where $(k = 1, \cdots, K)$. In both designs, threshold values (λ and $\lambda_k, k = 1, \cdots, K$) should be pre-specified before the trial begins to control the type I error rate

where $\theta_0 \in \Theta_0$ denotes the null value which leads to the maximum type I error rate (for e.g., θ_0 is the performance goal for a single-arm superiority design). Noting from Eq. (12), the overall type I error rate is a summation of the error rates at each interim analysis. For the relevant calculations corresponding to the frequentist group sequential design, refer to page 10 of [97], where Bayesian test statistics $T(\mathbf{y}^{(l)})$ and thresholds λ_l $(l = 1, \dots, K)$ are replaced by Z-test statistics based on interim data $\mathbf{y}^{(k)}$ and pre-specified critical values, respectively.

The crucial design objective in the development of a Bayesian group sequential design is to control the overall type I error rate to be less than a significance level of α (typically, 0.025 for a one-sided test and 0.05 for a two-sided test). This objective is similar to what is typically achieved in its frequentist counterparts, such as O'Brien-Fleming [98] or Pocock plans [99], or through the alpha-spending approach [100]. To achieve this objective, adjustments to the Bayesian thresholds $(\lambda_1, \dots, \lambda_K)$ are important, and this adjustment necessitates extensive simulation work. Failing to make these adjustments may result in an inflation of the overall type I error. For example, if one were to use the same thresholds of $\lambda_l = 1 - \alpha$ $(l = 1, \dots, K)$ for all the interim analyses, then the overall type I error would lead to the value greater than α regardless of the maximum number of interim analyses. Furthermore, the overall type I error may eventually converge to 1 as the number of interim analyses K goes to infinity, similar to the behavior observed in a frequentist group sequential design [101]. Additionally, compared to single stage designs, group sequential designs may require a larger sample size to achieve the same power all else being equal, as there is an inevitable statistical cost for repeated analyses.

Example - two-stage group sequential design based on beta-binomial model

We illustrate the advantage of using a Bayesian group sequential design compared to the single-stage Bayesian design described in Calibration of Bayesian trial design to assess frequentist operating characteristics section. Similar research using frequentist designs can be found in [102]. Recall that the previous fixed design based on a non-informative prior led to a power of 86.90% and a type I error rate of 2.31% with a sample size of 150 and a threshold of $\lambda = 0.975$ (Table 1). Our goal here is to convert the fixed design into a two-stage design that is more powerful, while controlling the overall the type I error rate ≤ 0.025. For fair comparison, we aim for the expected sample size $E(N)$ of the two-stage design to be as close to 150 as possible. Having a smaller value of $E(N)$ than 150 is even more desirable in our setting because it means that two-stage design can shorten the length of the trial of the fixed design. To compensate for the inevitable statistical cost of repeated analyses, the total sample size of the two-stage design is set to $N = 162$, representing an 8% increase in the final sample size of the single-stage design. The stage 1 sample size N_1 and stage 2 sample size N_2 are divided in the ratios of $3 : 7$, $5 : 5$, or $7 : 3$ to see the pattern of probability of early termination with different timing of interim analysis. Finally, we choose $\lambda_1 = 0.996$ and $\lambda_2 = 0.978$ as the thresholds for the interim analysis and the final analysis, respectively. Note that a more stringent stopping rule has been applied for early interim analyses than for the final analysis, similar to the proposed design of O'Brien and Fleming [98]. The same adaptation procedure will be taken to the single-stage designs with final sample sizes of 100 and 200 as reference.

Table 3 shows the results of the power analysis. It is observed that the overall type I error rates have

Table 3 Operating characteristics of two-stage designs based on beta-binomial model

Total Sample Size (N)	Stage 1 Sample Size (N_1)	Stage 2 Sample Size (N_2)	Expected Sample Size($E(N)$)	Probability of Early Termination (*PET*)	Type I Error (α)	Power ($1 - \beta$)	% Change in Power Compared with Single-stage Design
108	32	76	108	0.0000	0.0199	0.7053	+14.10
	54	54	105	0.0603	0.0220	0.6945	+12.36
	76	32	100	0.2632	0.0219	0.7094	+14.77
162	49	113	153	0.0819	0.0200	0.8865	+2.01
	81	81	145	0.2202	0.0228	0.8862	+1.97
	113	49	146	0.3348	0.0208	0.8860	+1.95
216	65	151	191	0.1659	0.0219	0.9598	+4.50
	108	108	177	0.3642	0.0205	0.9570	+4.20
	151	65	183	0.5120	0.0197	0.9568	+4.18

Note: All two-stage designs are based on the beta-binomial model with a non-informative prior. The expected sample size ($E(N)$) and the probability of early termination (*PET*) have been calculated under alternative, $\mathcal{H}_a$. Formula of $E(N)$ is given by $E(N) = N_1 + (1 - PET) \cdot N_2$, where N_1 and N_2 denote the sample sizes for stages 1 and 2, respectively. Thresholds for stage 1 and stage 2 are $\lambda_1 = 0.996$ and $\lambda_2 = 0.978$, respectively, for all designs. The percentage change in the last column has been calculated by comparing the powers between the two-stage design and the single-stage design (non-informative) in Table 1

been protected at 2.5% for all the considered designs. The expected sample sizes of the two-stage designs using a total sample size of $N = 162$ are $E(N) = 153$ $(N_1 : N_2 = 3 : 7)$, $E(N) = 145$ $(N_1 : N_2 = 5 : 5)$, and $E(N) = 146$ $(N_1 : N_2 = 7 : 3)$, with the power improved from 86.9% (single-stage design, see Table 1) to approximately 88.6% for all three cases. The power gain is even greater for the two-stage designs using a total sample size of $N = 216$, where the expected sample sizes are smaller than $N = 200$, which is advantageous for using a group-sequential design. Power gains occur for the two-stage designs using a total sample size of $N = 108$ as well, but the expected sample sizes are larger than $N = 100$; therefore, the single-stage design would be preferable in terms of expected sample sizes.

To summarize, the results show that, with an 8% increase in the final sample size of the single-stage design, we can construct a two-stage design in which the expected sample size is smaller or equal to the final sample size of the single-stage design. This is while still protecting the type I error rate below 2.5% and benefiting from an increase in the overall power of the designs by as much as 14% ($N = 108$), 2% ($N = 162$), and 4% ($N = 216$), assuming the alternative hypothesis is true. In other words, a Bayesian group sequential design allowing the claim of early success at interim analysis can help save costs by possibly reducing length of a trial when there is strong evidence of a treatment effect for the new medical device. Even if the evidence turns out to be not as strong as expected upon completion of the study (the null hypothesis seems more likely to be true in the observed final results), the potential risk for the sponsor would be the additional cost spent on enrolling 8% more patients than with the single-stage design.

Decision rule - predictive probability approach
Predictive probability approach
The primary motivation for employing the predictive probability approach in decision-making is to answer the question at an interim analysis: "Is the trial likely to present compelling evidence in favor of the alternative hypothesis if we gather additional data, potentially up to the maximum sample size?" This question fundamentally involves predicting the future behavior of patients in the remainder of the study, where the prediction is based on the interim data observed thus far. Consequently, its idea is akin to measuring conditional power given interim data in the stochastic curtailment method [103, 104]. The key quantity here is the predictive probability of observing a statistically significant treatment effect if the trial were to proceed to its predefined maximum sample size, calculated in a fully Bayesian way.

One of the most standard applications of predictive probability approach for regulatory submission is the interim analysis for futility stopping (i.e., early stopping the trial in favor of the null hypothesis) [23, 105–107]. This is motivated primarily by an ethical imperative; the goal here is to assess whether the trial, based on interim data, is unlikely to demonstrate a significant treatment effect even if it continues to its planned completion. This information can then be utilized by the monitoring committee to assess whether the trial is still viable midway through the trial [108]. The study will stop for lack of benefit if the predictive probability of success at the final analysis is too small. Other areas where this approach are useful include the early termination for success with consideration of the current sample size (i.e., early stopping the trial in favor of the alternative hypothesis) [18, 109, 110], or sample size re-estimation to evaluate whether the planned sample size is sufficiently large to detect the true treatment effect [111].

We focus on illustrating the use of the predictive probability approach for futility interim analysis. To simplify the discussion, we consider the two-stage futility design where only one interim futility analysis exists. The idea illustrated here can be extended to a multi-stage design by implementing the following testing procedure at each of the interim analyses in the multi-stage design. The logic explained here can be extended to the applications of early success claims and sample size re-estimation after a few modifications.

Suppose that $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ denote the datasets at the interim and final analyses, respectively, and θ is the main parameter of interest. We distinguish all incremental quantities from cumulative ones using the notation "tilde". Therefore, $\tilde{\mathbf{y}}^{(2)}$ and $\mathbf{y}^{(2)} = \{\mathbf{y}^{(1)}, \tilde{\mathbf{y}}^{(2)}\}$ represent the incremental stage 2 data and the final data, respectively.

At the final analysis, a sponsor plans to test the null hypothesis $\mathcal{H}_0 : \theta \in \Theta_0$ versus the alternative hypothesis $\mathcal{H}_a : \theta \in \Theta_a$, where $\Theta = \Theta_0 \cup \Theta_a$, and Θ_0 and Θ_a are disjoint sets. Suppose that $H(\mathbf{y}^{(2)})$ is the final test statistic to be used, and a higher value casts doubt that the null hypothesis is true. Therefore, the sponsor will claim the success of the trial if it is demonstrated that $H(\mathbf{y}^{(2)}) > \lambda_2$ with a predetermined threshold λ_2, where the threshold is chosen to satisfy the type I & II error requirement of the futility design. It is at the sponsor's discretion whether to use frequentist or Bayesian statistics to construct the final test statistic $H(\mathbf{y}^{(2)})$. This is because the purpose of using the predictive probability approach is to make a decision at the interim analysis, not at the final analysis.

At the interim analysis, the outcomes from stage 1 patients $\mathbf{y}^{(1)}$ are observed. We measure the predictive probability of

success at the final analysis, which is the Bayesian test statistics of the predictive probability approach represented as a functional $\mathcal{G}(\cdot) : \mathcal{Q}_{\tilde{\mathbf{y}}^{(2)}|\mathbf{y}^{(1)}} \to [0, 1]$, such that:

$$\mathcal{G}\{f(\tilde{\mathbf{y}}^{(2)}|\mathbf{y}^{(1)})\} = T(\mathbf{y}^{(1)}) = \mathbb{P}[H(\mathbf{y}^{(1)}, \tilde{\mathbf{y}}^{(2)}) > \lambda_2|\mathbf{y}^{(1)}]$$
$$= \int \mathbf{1}(H(\mathbf{y}^{(1)}, \tilde{\mathbf{y}}^{(2)}) > \lambda_2) \cdot f(\tilde{\mathbf{y}}^{(2)}|\mathbf{y}^{(1)})d\tilde{\mathbf{y}}^{(2)},$$
(13)

where $\mathcal{Q}_{\tilde{\mathbf{y}}^{(2)}|\mathbf{y}^{(1)}}$ represents the collection of posterior predictive distributions of stage 2 patient outcome $\tilde{\mathbf{y}}^{(2)}$ given the interim data $\mathbf{y}^{(1)}$. As seen from the integral (13), the fully Bayesian nature of the predictive probability approach is characterized by its integration of final decision results $\mathbf{1}(H(\mathbf{y}^{(1)}, \tilde{\mathbf{y}}^{(2)}) > \lambda_2)$ over the data space of all possible scenarios of future patients' outcome $\tilde{\mathbf{y}}^{(2)}$, with the weight of the integral respecting the posterior predictive distribution $f(\tilde{\mathbf{y}}^{(2)}|\mathbf{y}^{(1)})$. Note that the posterior predictive distribution is again a mixture distribution of the likelihood function of the future outcome $\tilde{\mathbf{y}}^{(2)}$ and the posterior distribution given the interim data:

$$f(\tilde{\mathbf{y}}^{(2)}|\mathbf{y}^{(1)}) = \int f(\tilde{\mathbf{y}}^{(2)}|\theta) \cdot \pi(\theta|\mathbf{y}^{(1)})d\theta.$$

It is important to note that the predictive probability (13) differs from the predictive power [112, 113], which represents a weighted average of the conditional power, given by $\int \mathbb{P}[H(\mathbf{y}^{(1)}, \tilde{\mathbf{y}}^{(2)}) > \lambda_2|\theta] \cdot \pi(\theta|\mathbf{y}^{(1)})d\theta$. The calculation of the predictive probability (13) follows the fully Bayesian paradigm. However, the predictive power is a mix of both frequentist and Bayesian paradigms, constructed based on the conditional power (frequentist statistics) and posterior distribution (Bayesian statistics). Both can be used as the metric of a Bayesian stochastic curtailment method [114], but the recent trend seems to be that the predictive probability is more prevalently used for regulatory submissions than predictive power [23, 115].

Finally, to induce a dichotomous decision at the interim analysis, we need to pre-specify the futility threshold $\gamma_1 \in [0, 1]$. By introducing an indicator function ψ, the testing result for the futility analysis is determined as follow:

$$\psi(\mathbf{y}^{(1)}) = \begin{cases} 1 & \text{if } \mathcal{G}\{f(\tilde{\mathbf{y}}^{(2)}|\mathbf{y}^{(1)})\} = \mathbb{P}[H(\mathbf{y}^{(1)}, \tilde{\mathbf{y}}^{(2)}) > \lambda_2|\mathbf{y}^{(1)}] \geq \gamma_1 \\ 0 & \text{if } \mathcal{G}\{f(\tilde{\mathbf{y}}^{(2)}|\mathbf{y}^{(1)})\} = \mathbb{P}[H(\mathbf{y}^{(1)}, \tilde{\mathbf{y}}^{(2)}) > \lambda_2|\mathbf{y}^{(1)}] < \gamma_1, \end{cases}$$

where 1 and 0 indicate the rejection and acceptance of the null hypothesis, respectively. Figure 6 displays a pictorial description of the decision procedure.

Theoretically, it is important to note that allowing early termination of a trial for futility tends to reduce both the trial's power and the type I error rate [107]. To explain this, suppose that one uses the identical final threshold λ_2 in both of the two-stage futility design, as explained above, and the fixed design. Then, the following inequality holds:

$$\mathbb{P}_\theta[H(\mathbf{y}^{(2)}) > \lambda_2] \geq \mathbb{P}_\theta[T(\mathbf{y}^{(1)}) \geq \gamma_1 \text{ and } H(\mathbf{y}^{(2)}) > \lambda_2], \quad \text{for all } \theta \in \Theta,$$
(14)

which means that the power function of the fixed design is uniformly greater or equal to the power function of the two-stage futility design over the entire parameter space Θ. This implies that equipping a futility rule to a fixed design leads to a reduction of both the type I error rate and power compared to the fixed design.

We briefly discuss the choice of the futility threshold γ_1 and the final threshold λ_2 in the two-stage futility
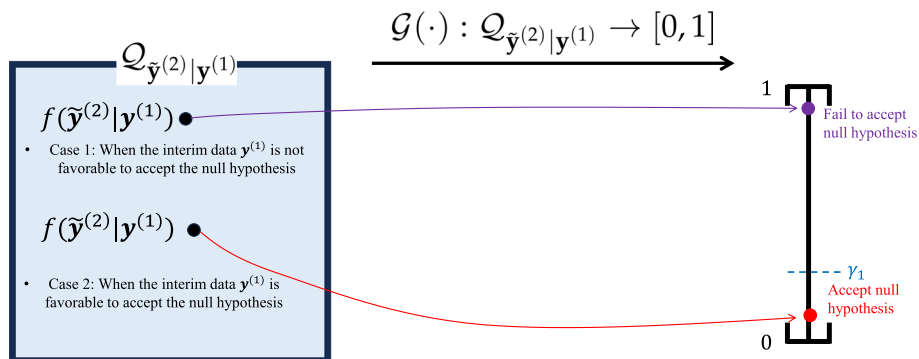

Fig. 6 Pictorial illustration of the decision rule based on the predictive probability approach for futility analysis. If the interim data $\mathbf{y}^{(1)}$ favors accepting the null hypothesis (Case 2 in the figure), it is also likely that the future remaining patients' outcomes $\tilde{\mathbf{y}}^{(2)}$ would be predicted to be more favorable for accepting the null hypothesis. This prediction results in a lower value of the test statistic $\mathcal{G}\{f(\tilde{\mathbf{y}}^{(2)}|\mathbf{y}^{(1)})\} = \mathbb{P}[H(\tilde{\mathbf{y}}^{(2)}, \mathbf{y}^{(1)}) > \lambda_2|\mathbf{y}^{(1)}]$ (13). The pre-specified threshold γ_1 is then used to make the dichotomous decision based on the test statistic

design. Futility threshold γ_1 is typically chosen within the range of 1% to 20% in many problems. Having fixed the threshold λ_2, a higher threshold for γ_1 increases the likelihood of discontinuing a trial involving an ineffective treatment, which is desirable because it shortens the trial length when there is a true negative effect. However, it may reduce both the type I error rate and power compared to a lower threshold for γ_1. On the other hand, the final threshold λ_2 of the futility design is typically chosen to align with the nominal significance level of the corresponding fixed design. This is mainly due to the relevant operational risk of inflating the

$$\text{Study Sucess} = \mathbf{1}\left\{ LB(\mathbf{y}) = \hat{S}(\text{52-week}) - 1.96 \cdot \sqrt{\hat{Var}[\hat{S}(\text{52-week})]} > 0.5 \right\}. \tag{15}$$

type I error rate if futility stopping were not executed as planned, even after the final threshold λ_2 has been chosen to make rejection easier to reclaim the lost type I error rate [107, 116]. In summary, when constructing a futility design, the sponsor needs to choose the futility threshold that does not substantially affect the operating characteristics of the original fixed-sample size, while also curtailing the trial length when there is a negative effect.

Example - two-stage futility design with Greenwood test
Suppose that a sponsor considers a single-arm design for a phase II trial to assess the efficacy of a new anti-arrhythmic drug in treating patients with a mild atrial fibrillation [117]. The primary efficacy endpoint is the freedom from recurrence of the indication at 52 weeks (1 year) after the intervention. The sponsor sets the null and alternative hypotheses by $\mathcal{H}_0 : \theta \le 0.5$ versus $\mathcal{H}_a : \theta > 0.5$, where θ denotes the probability of freedom from recurrence at 52 weeks. Let $S(t)$ represent the survival function; then the main parameter of interest is $\theta = S(\text{52-week})$. At the planning stage, regulator agreed on the proposal of sponsor that the time to recurrence follows a three-piece exponential model, with a hazard function given as $h(t) = 0.1 \cdot \xi$ if $t \in [0, \text{8-week}]$, $h(t) = 0.05 \cdot \xi$ if $t \in (\text{8-week}, \text{24-week}]$, and $h(t) = 0.01 \cdot \xi$ if $t \in (\text{24-week}, \text{52-week}]$, where ξ is a positive number. In order to simulate the survival data in the power calculation, the value of ξ will be derived to set the true data-generating parameter

to be $\theta = S(\text{52-week}) = 0.50, 0.55, 0.60, 0.65$, and 0.7. Note that $\theta = 0.50$ corresponds to the type I error scenario, and the rest of the settings correspond to power scenarios.

We first construct a single-stage design with the final sample size of $N = 100$ patients. The final analysis is conducted by a frequentist hypothesis testing based on the one-sided level-0.025 Greenwood test using a confidence interval approach [118]. More specifically, the testing procedure is that the null hypothesis is rejected if the lower bound of the 95% two-sided confidence interval evaluated at $t = \text{52-week}$ is greater than 0.5, that is,

Here, the mean estimate $\hat{S}(t)$ is the Kaplan-Meier estimate of $S(t)$ [119], and its variance estimate $\hat{Var}[\hat{S}(t)]$ is based on the Greenwood formula [120], and notation $\mathbf{y}$ represents the final data from $N = 100$ patients. The results of the power analysis obtained by simulation indicate that the probabilities of rejecting the null hypothesis are 0.0185, 0.1344, 0.461, 0.8332, and 0.9793 when the effectiveness success rates (θ) are 0.5, 0.55, 0.60, 0.65, and 0.7, respectively. Note that the type I error rate is 0.0185 less than the 0.025.

Next, we construct a two-stage futility design by equipping the above single-stage design with a non-binding futility stopping option based on the predictive probability approach. Non-binding means that the investigators can freely decide whether they really want to stop or not. This is more common in practice because a stopping decision is typically influenced not only by interim data but also by new external data or safety information [121]. The final sample size of the futility design is again $N = 100$, and we keep the decision criterion for the study success of the final test the same as that of the single-stage design (15). This means that there are no adjustments to the final threshold to reclaim a loss of type I error rate. The futility analysis will be performed when $N_1 = 30$ patients have completed the 52 weeks of follow-up (30% of participants). A non-informative Gamma prior $\mathcal{G}a(0.1, 0.1)$ will be used for each of the hazard rate parameters of the three-piece exponential model. Futility stopping (i.e., accepting the null hypothesis) is triggered if the predictive probability of trial success at the maximum sample size is less than the pre-specified futility threshold $\gamma_1 = 0.05$. Technically, the predictive probability is

$$T(\mathbf{y}^{(1)}) = \mathbb{P}[LB(\mathbf{y}^{(1)}, \tilde{\mathbf{y}}^{(2)}) > 0.5 | \mathbf{y}^{(1)}] = \int \mathbf{1}(LB(\mathbf{y}^{(1)}, \tilde{\mathbf{y}}^{(2)}) > 0.5) \cdot f(\tilde{\mathbf{y}}^{(2)} | \mathbf{y}^{(1)}) d\tilde{\mathbf{y}}^{(2)},$$

where $\mathbf{y}^{(1)}$ and $\tilde{\mathbf{y}}^{(2)}$ denote the time-to-event outcomes from $N_1 = 30$ patients and $\tilde{N}_2 = N - N_1 = 70$ patients, respectively, and $f(\tilde{\mathbf{y}}^{(2)}|\mathbf{y}^{(1)})$ denotes the posterior predictive distribution of outcomes of the future remaining patients $\tilde{\mathbf{y}}^{(2)}$.

In the power analysis, we vary the number of stage 1 patients, N_1, to 50 and 70 and set the futility threshold, γ_1, to 0.1 and 0.15 to explore the operating characteristics of the futility design. Figure 7 illustrates the testing procedures of the single-stage design and the two-stage futility design. In this setting, the only difference between the futility and single-stage designs is that the former has the option to stop the trial due to futility when N_1 patients had completed the follow-up of 52 weeks, while the latter does not. Table 4 shows the power analysis results of the two-stage futility designs.

The results demonstrate that the probability of rejecting the null hypothesis in the futility design is consistently lower than that in the single-stage design across various effectiveness success rates ($\theta = 0.5, 0.55, 0.6, 0.65,$ and 0.7). This finding aligns with the theoretical result (refer to inequality (14)). For example, in the case where the futility threshold $\gamma_1 = 0.05$ with a stage 1 sample size of $N_1 = 30$, the percentage change in the probability of rejecting the null hypothesis compared to a single-stage design is $-8.82\%, -15.86\%, -4.53\%, -1.858\%,$ and -0.441% when the true effectiveness success rate (θ) is 0.5, 0.55, 0.6, 0.65, and 0.7, respectively.

We examine the general pattern of the reduction in the type I error rate and power of the futility design compared to the single-stage design as the futility threshold γ_1 changes. Note that the average of type I error rates across three different stage 1 sample size for the futility design are 0.0173, 0.0160, and 0.0156 when the futility thresholds γ_1 are set at 0.05, 0.10, and 0.15, respectively. These results reflect reductions of 6.4%, 13.5%, and 15.6% in the type I error rate compared to the single-stage design. (Recall that the type I error rate of the single-stage design is 0.0185.) This implies that a higher value for the futility threshold γ_1 leads to a more substantial reduction in the type I error rate compared to the single-stage design. A similar pattern of reduction is observed in the power scenarios when $\theta = 0.55, 0.6, 0.65,$ and 0.7.

Notably, the probability of early termination tends to increase as the stage 1 sample size grows from $N_1 = 30$ to $N_3 = 70$. This increase is particularly significant in the type I error scenario when $\theta = 0.5$. Across all the scenarios examined, the expected sample size consistently stays below $N = 100$. This indicates that the futility design outperforms the single-stage design in terms of expected sample size as a performance criterion. Furthermore, this reduction in expected sample size is even more pronounced in the type I error scenarios. In conclusion, it is evident that for long-term survival endpoints, like the example discussed here, the futility design can lead to substantial resource savings by allowing the trial to be terminated midway when the lack of clinical benefit becomes clear.

Multiplicity adjustments
Multiplicity problem - primary endpoint family

Efficacy endpoints are measures designed to reflect the intended effects of a drug or medical device. Clinical trials are often conducted to evaluate the relative efficacy of two or more modes of treatment. For instance, consider a new drug developed for the treatment of heart failure [122]. In this case, it may be unclear whether the heart failure drug primarily promotes a decrease in mortality, a reduction in heart failure hospitalization, or an improvement in quality of life (such as Kansas City Cardiomyopathy Questionnaire score overall summary score [123]). However, demonstrating any
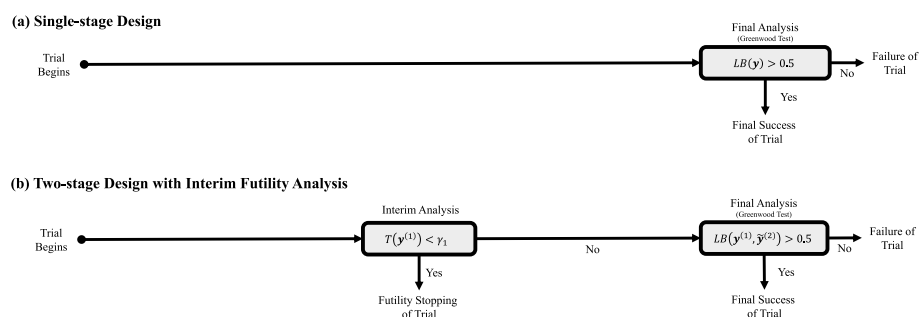

Fig. 7 Testing procedures of the single-stage design and the two-stage futility design are as follows: at the final analysis, both designs employ the one-sided level-0.025 Greenwood test with a final sample size of $N = 100$. Only the futility design has the option to stop the trial due to futility when N_1 patients had completed 52 weeks of follow-up. In the power analysis, we use $N_1 = 30, 50,$ and 70, along with $\gamma_1 = 0.05, 0.1,$ and 0.15 to assess the operating characteristics of the design

Table 4 Operating characteristics of two-stage futility designs with the final sample size $N = 100$

Futility Threshold	Effectiveness Success Rate	Stage 1 Sample Size $N_1 = 30$			Stage 1 Sample Size $N_1 = 50$			Stage 1 Sample Size $N_1 = 70$		
		Probability of Rejecting Null Hypothesis	Expected Sample Size ($E(N)$)	Probability of Early Termination (PET)	Probability of Rejecting Null Hypothesis	Expected Sample Size ($E(N)$)	Probability of Early Termination (PET)	Probability of Rejecting Null Hypothesis	Expected Sample Size ($E(N)$)	Probability of Early Termination (PET)
0.05	0.5	0.017	70.11	0.427	0.017	66.80	0.664	0.018	76.12	0.796
	0.55	0.116	84.18	0.226	0.117	80.75	0.385	0.12	85.06	0.498
	0.6	0.441	93.98	0.086	0.439	92.85	0.143	0.446	93.97	0.201
	0.65	0.818	98.32	0.024	0.817	97.95	0.041	0.829	98.74	0.042
	0.7	0.975	99.58	0.006	0.973	99.50	0.01	0.976	99.88	0.004
0.10	0.5	0.016	59.68	0.576	0.016	63.05	0.739	0.016	74.11	0.863
	0.55	0.114	76.48	0.336	0.115	76.45	0.471	0.117	82.00	0.600
	0.6	0.428	89.01	0.157	0.427	89.65	0.207	0.44	92.08	0.264
	0.65	0.799	95.31	0.067	0.810	96.95	0.061	0.826	97.93	0.069
	0.7	0.966	98.74	0.018	0.969	99.15	0.017	0.971	99.55	0.015
0.15	0.5	0.016	57.93	0.601	0.016	61.20	0.776	0.015	73.15	0.895
	0.55	0.110	73.61	0.377	0.114	74.40	0.512	0.115	80.56	0.648
	0.6	0.419	87.12	0.184	0.423	88.20	0.236	0.429	90.67	0.311
	0.65	0.795	94.47	0.079	0.806	96.60	0.068	0.817	97.27	0.091
	0.7	0.964	98.6	0.020	0.968	99.05	0.019	0.969	99.31	0.023

Note: Probabilities of rejecting null hypothesis of single-stage design with the final sample size of $N = 100$ are 0.0185, 0.1344, 0.461, 0.8332, and 0.9793 when the effectiveness success rates are 0.5, 0.55, 0.60, 0.65, and 0.7, respectively

of these effects individually would hold clinical significance; there are multiple chances to 'win.' Consequently, all three endpoints – mortality rate, number of heart failure hospitalizations, and an index for quality of life – might be designated as separate primary endpoints. This is an illustrative example of a primary endpoint family, and failure to adjust for multiplicity can lead to a false conclusion that the heart failure drug is effective. Here, multiplicity refers to the presence of numerous comparisons within a clinical trial [124–127]. See Section III of the FDA guidance document for the multiple endpoints for more details on the primary endpoint family [128].

In the following, we formulate the multiplicity problem of the primary endpoint family. We consider a family of K primary endpoints, any one of which could support the conclusion that a new treatment has a beneficial effect. For simplicity, we assume that the outcomes of the patients are binary responses, where a response of 1 (yes) indicates that the patient shows a treatment effect. Using the example of a heart failure drug, the first efficacy endpoint measures mortality: whether a patient has survived (yes/no), the second endpoint measures morbidity: whether a patient experienced heart failure hospitalization (no/yes), and the third endpoint measures the quality of life: whether the Kansas City Cardiomyopathy Questionnaire overall summary score

effective). To explain this, suppose that at the final analysis upon completion of the study, the rejection of any one of the null hypotheses among K null hypotheses will lead to marketing approval for a new drug. If there are $K = 2$ independent endpoints, each tested at $\alpha = 0.025$, and success on either endpoint by itself would lead to a conclusion of a drug effect, the type I error rate is approximately $5 \approx 1 - (1 - 0.025)^2$ percent. With $K = 4$ endpoints, the type I error rate increases to about $10 \approx 1 - (1 - 0.025)^4$ percent. When there are $K = 10$ endpoints, the type I error rate escalates to about $22 \approx 1 - (1 - 0.025)^{10}$ percent. The problem becomes more severe as the number of endpoints (K) increases.

Familywise type I error rate and power

It is important to ensure that the evaluation of multiple hypotheses will not lead to inflation of the study's overall type I error probability relative to the planned significance level. This is the primary regulatory concern, and it is required to minimize the chances of a false positive conclusion for any of the endpoints, regardless of which and how many endpoints in the study have no effect [128]. This probability of incorrect conclusions is known as the familywise type I error rate [129]. Technically, it is the probability of erroneously rejecting at least one null hypothesis under the global null, and can be written as,

$$
\begin{aligned}
\alpha^{family} &= \mathbb{P}[\text{Reject at least one null hypothesis}|\text{All null hypotheses are true}] \\
&= \mathbb{P}[\text{Reject the collection } \{\mathcal{H}_{0,i}\}_{i \in A} \text{ for all } A \in \mathcal{K}|\{\mathcal{H}_{0,i}\}_{i=1}^{K} \text{ are true}] \\
&= \mathbb{P}[V \geq 1|\{\mathcal{H}_{0,i}\}_{i=1}^{K} \text{ are true}],
\end{aligned}
\tag{17}
$$

has improved by more than 15 points (yes/no) during a defined period after the treatment. The logic explained in the following can be applied to various types of outcomes, including continuous outcomes and time-to-event outcomes.

We consider a form of parallel group trial design, each associated with hypotheses given by:

$$
\mathcal{H}_{0,i} : \theta_i \leq \theta_{0,i} \quad \text{versus} \quad \mathcal{H}_{a,i} : \theta_i > \theta_{0,i}, \quad (i = 1, \cdots, K),
\tag{16}
$$

where θ_i denotes the response rate for the i-th endpoint (where a higher rate indicates a better treatment effect), and $\theta_{0,i}$ represents the performance goal associated with the i-th endpoint.

In a clinical trial with a single endpoint ($K = 1$) tested at $\alpha = 0.025$, the probability of finding a treatment effect by chance alone is at most 0.025. However, multiple testing ($K > 1$) can increase the likelihood of type I error (a false conclusion that a new drug is

where $\mathcal{K} = P(\{1, 2, \cdots, K\}) - \emptyset$. Here, $P(A)$ and $\emptyset$ denote the power set of set A and the empty set, respectively. If there are $K = 4$ endpoints, one needs to consider $15 = 2^4 - 1$ false positive scenarios, each of which contributes to an increase in α^{family}. When $K = 10$ endpoints are examined in a study, the number of false positive scenarios increases to $1023 = 2^{10} - 1$ scenarios. V denotes the number of hypotheses rejected among the K hypotheses, taking an integer value from 0 to K.

Another regulatory concern for a primary endpoint family is to maximize the chances of a true positive conclusion. The desired power is an important factor in determining the sample size. Unlike the type I error scenario where α^{family} is standardly used in most cases, the concept of power can be generalized in various ways when multiple hypotheses are considered (see Chapter 2 in [129] for more details). The following two types of power are frequently used under the global alternative

$$\pi^{dis} = \mathbb{P}[\text{Reject at least one null hypothesis}|\text{All alternative hypotheses are true}]$$
$$= \mathbb{P}[\text{Reject the collection } \{\mathcal{H}_{0,i}\}_{i \in A} \text{ for all } A \in \mathcal{K}|\{\mathcal{H}_{a,i}\}_{i=1}^{K} \text{ are true}] \qquad (18)$$
$$= \mathbb{P}[V \geq 1|\{\mathcal{H}_{a,i}\}_{i=1}^{K} \text{ are true}],$$

$$\pi^{con} = \mathbb{P}[\text{Reject all null hypotheses}|\text{All alternative hypotheses are true}]$$
$$= \mathbb{P}[\text{Reject the collection } \{\mathcal{H}_{0,i}\}_{i=1}^{K}|\{\mathcal{H}_{a,i}\}_{i=1}^{K} \text{ are true}] \qquad (19)$$
$$= \mathbb{P}[V = K|\{\mathcal{H}_{a,i}\}_{i=1}^{K} \text{ are true}].$$

The former π^{dis} (18) and latter π^{con} (19) are referred to as disjunctive power and conjunctive power, respectively [130]. By definition, the disjunctive power is greater than the conjunctive power if the number of endpoints is more than one ($K = 2, 3, \cdots$), and both are equal when $K = 1$.

Typically, regulators require the study design to have $\alpha^{family} \leq \alpha$ with a target level $\alpha = 0.025$ for a one-sided test and $\alpha = 0.05$ for a two-sided test for a primary endpoint family. On the other hand, study specific discussion is necessary to determine which power (disjunctive power, conjunctive power, or another type) should be used for a given study. For example, if the study's objective is to detect all existing treatment effects, then one may argue that conjunctive power π^{con} should be used. However, if the objective is to detect at least one true effect, then disjunctive power π^{dis} is recommended [128].

Frequentist method - *p*-value based procedures

Much has been written and published on the mathematical aspects of frequentist adjustment procedures for multiple comparisons, and we refer the reader elsewhere for the details [131–133]. Here, we briefly explain three popular *p*-value based multiplicity adjustment procedures: the Bonferroni, Holm, and Hochberg methods [134, 135]. These methods utilize the *p*-values from individual tests and can be applied to a wide range of test situations [136]. The fundamental difference is that the Bonferroni method uses non-ordered *p*-values, while the Holm and Hochberg methods use ordered *p*-values. Refer to Section 18 from [137] for excellent summary of these methods.

● *Bonferroni Method*   The Bonferroni method is a single-step procedure that is commonly used, perhaps because of its simplicity and broad applicability. It is known that Bonferroni method provides the most conservative multiplicity adjustment [126]. Here, we use the most common form of the Bonferroni method which divides the overall significance level of α (typically 0.025 for the one-sided test) equally among the K endpoints for testing K hypotheses (16). The method then concludes that a treatment effect is significant at the α level for each one of the K endpoints for which the endpoint's *p*-value is less than α/K.

● *Holm Method*   The Holm procedure is a multi-step step-down procedure. It is less conservative than the Bonferroni method because a success with the smallest *p*-value allows other endpoints to be tested at larger endpoint-specific alpha levels than does the Bonferroni method. The endpoint *p*-values resulting from the final analysis are ordered from the smallest to the largest (or equivalently, the most significant to the least significant), denoted as $p_{(1)} \leq \cdots \leq p_{(K)}$.

We take the following stepwise procedure: (Step 1) the test begins by comparing the smallest *p*-value, $p_{(1)}$, to α/K, the same threshold used in the equally-weighted Bonferroni correction. If this $p_{(1)}$ is less than α/K, the treatment effect for the endpoint associated with this *p*-value is considered significant; (Step 2) the test then compares the next-smallest *p*-value, $p_{(2)}$, to an endpoint-specific alpha of the total alpha divided by the number of yet-untested endpoints. If $p_{(2)} < \alpha/(K-1)$, then the treatment effect for the endpoint associated with this $p_{(2)}$ is also considered significant; (Step 3) The test then compares the next ordered *p*-value, $p_{(3)}$, to $\alpha/(K-2)$, and so on until the last *p*-value (the largest *p*-value) is compared to α; (Step 4) The procedure stops, however, whenever a step yields a non-significant result. Once an ordered *p*-value is not significant, the remaining larger *p*-values are not evaluated and it cannot be concluded that a treatment effect is shown for those remaining endpoints.

● *Hochberg Method*   The Hochberg procedure is a multi-step step-up testing procedure. It compares the *p*-values to the same alpha critical values of $\alpha/K, \alpha/(K-1), \cdots, \alpha/2, \alpha$, as the Holm procedure. However, instead of starting with the smallest *p*-value as performed in Holm procedure, Hochberg procedure

starts with the largest *p*-value (or equivalently, the least significant *p*-value), which is compared to the largest endpoint-specific critical value α. If the first test of hypothesis does not show statistical significance, testing proceeds to compare the second-largest *p*-value to the second-largest adjusted alpha value, $\alpha/2$. Sequential testing continues in this manner until a *p*-value for an endpoint is statistically significant, whereupon the Hochberg procedure provides a conclusion of statistically-significant treatment effects for that endpoint and all endpoints with smaller *p*-values.

● *Examples*   For illustration, suppose that a trial with four endpoints ($K = 4$) yielded one-sided *p*-values of $p_1 = 0.006$ (1-st endpoint), $p_2 = 0.013$ (2-nd endpoint), $p_3 = 0.008$ (3-rd endpoint), and $p_4 = 0.0255$ (4-th endpoint) at the final analysis.

The Bonferroni method compares each of these *p*-values to $0.00625 = 0.025/4$, resulting in a significant treatment effect at the 0.025 level for only the 1-st endpoint because only the 1st endpoint has a *p*-value less than 0.00625.

The Holm method considers the successive endpoint-specific alphas, $0.00625 = 0.025/4$, $0.00833 = 0.025/(4 - 1)$, $0.0125 = 0.025/(4 - 2)$, and $0.025 = 0.025/(4 - 3)$. We start by comparing the smallest *p*-value $p_1 = 0.006$ with 0.00625. The treatment effect for the 1-st endpoint is thus successfully demonstrated, and the test continues to the second step. In the second step, the second smallest *p*-value is $p_3 = 0.008$, which is compared to 0.00833. The 3-rd endpoint has, therefore, also successfully demonstrated a treatment effect, as 0.008 is less than 0.00833. Testing can now proceed to the third step, in which the next ordered *p*-value of $p_2 = 0.013$ is compared to 0.0125. In this comparison, as 0.013 is greater than 0.0125, the test is not statistically significant. This non significant result stops further tests. Therefore, in this example, the Holm procedure concludes that treatment effects have been shown for the 1st and 3rd endpoints.

The Hochberg method considers the same successive endpoint-specific alphas as the Holm method. In the first step, the largest *p*-value of $p_4 = 0.0255$ is compared to its alpha critical value of $\alpha = 0.025$. Because this *p*-value of 0.0255 is greater than 0.025, the treatment effect for the 4-th endpoint is considered not significant. The procedure continues to the second step. In the second step, the second largest *p*-value, $p_2 = 0.013$, is compared to $\alpha/2 = 0.0125$. Because p_2 is greater than the allocated alpha, and the 2-nd endpoint is also not statistically significant, the test continues to the third step. In the third

step, the next largest *p*-value, $p_3 = 0.008$, is compared to its alpha critical value of $\alpha/3 = 0.00833$, and the 3-rd endpoint shows a significant treatment effect. This result automatically causes the treatment effect for all remaining untested endpoints, which have smaller *p*-values than 0.008, to be significant as well. Therefore, the 1-st endpoint also shows a significant treatment effect.

Bayesian multiplicity adjustment methods

Bayesian adjustments for multiplicity [138–141] can be acceptable for regulatory submissions, provided the analysis plan is pre-specified and the operating characteristics of the analysis are adequate [16]. It is advisable to consult regulators early on with regard to a Statistical Analysis Plan that includes Bayesian adjustment for multiplicity.

Generally, the development of Bayesian multiplicity adjustment involves three steps:

- Step 1: Statistical modeling for the outcomes of endpoints,
- Step 2: Performing the test for individual hypotheses (16) with pre-specified thresholds,
- Step 3: Interpreting the results of Step 2 in terms of the familywise error rate (17).

One of the unique advantages of Bayesian multiplicity adjustment is the flexibility of statistical modeling in the planning phase of Step 1, tailored to the study's objectives, the characteristics of the sub-population, and other relevant factors. For example, if a certain hierarchical or multilevel structure exists among sub-populations (such as, center - doctor - patients as discussed in [142]), then one would use a Bayesian hierarchical model to account for the heterogeneity between sub-populations and patient-to-patient variability simultaneously [26, 143, 144]. Furthermore, adaptive feature can be also incorporated to the Bayesian multiplicity adjustment [145, 146]. This stands in contrast to traditional frequentist approaches, which evaluate the outcomes from each sub-population independently or simply combine data from all sub-populations through a pooled analysis [147].

In Step 2, sponsors need to provide detailed descriptions of the decision rules that will be used to reject the *i*-th null hypothesis $\mathcal{H}_{0i}$ ($i = 1, \cdots, K$) in the Statistical Analysis Plan. The sponsor can choose either the posterior probability approach (Decision rule - posterior probability approach section) or the predictive probability approach (Decision rule - predictive probability approach section) as the decision rules. Most importantly, the threshold value for rejecting each null hypothesis should be pre-specified in the Statistical Analysis Plan, which often requires extensive simulations across all plausible scenarios (such as global null ("0

success" or "K failures"), the global alternative ("K successes"), and the mixed alternative scenarios).

Finally, in Step 3, the results of the K individual tests are interpreted to ensure that the frequentist familywise type I error rate α^{family} (17) is lower than or equal to the overall significance level α. Additionally, power specific to the study objective (disjunctive power, conjunctive power, or another type) may be measured to estimate the sample size of the study.

Bayesian multiplicity adjustment using Bayesian hierarchical modeling

Here, we illustrate the simplest form of the Bayesian multiplicity adjustment method using Bayesian hierarchical modeling. [83, 84, 146, 148]. Bayesian hierarchical modeling is a specific Bayesian methodology that combines results from multiple arms or studies to obtain estimates of safety and effectiveness parameters [149]. This approach is particularly appealing in the regulatory setting when there is an association between the outcomes of K endpoints so that exchangeability of patients' outcomes across K endpoints can be assumed [140]. Figure 8 outlines the three steps of the multiplicity control procedure using a Bayesian hierarchical model.

Let N_i be the number of patients to be enrolled in the i-th arm associated with the i-th endpoint for testing the null and alternative hypotheses, $\mathcal{H}_{0,i} : \theta_i \leq \theta_{0,i}$ versus $\mathcal{H}_{a,i} : \theta_i > \theta_{0,i}$, $(i = 1, \cdots, K)$ (16). The total sample size of the study is therefore $N = \sum_{i=1}^{K} N_i$. Let y_i denote the number of responders to a treatment, where a higher number indicates better efficacy. Then, the number of responders associated with the i-th endpoint is distributed according to a binomial distribution:

$$y_i | \theta_i \sim \mathcal{BN}(N_i, \theta_i), \quad (i = 1, \cdots, K). \tag{20}$$

Note that the parameters of main interest are $(\theta_1, \cdots, \theta_K) \in [0,1]^K$. Suppose that there is an

association between the outcomes of the K endpoints, and K sub-populations are exchangeable, a priori. We assume the most basic formulation of hierarchical prior on the $(\theta_1, \cdots, \theta_K)$ given by:

$$\phi_i | \mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2), \quad (i = 1, \cdots, K), \tag{21}$$

$$(\mu, \sigma^2) \sim \mathcal{NIG}(\nu, \omega, a, b), \tag{22}$$

where the parameter θ_i is logit-transformed to ϕ_i (i.e., $\theta_i = \exp(\phi_i)/\{1 + \exp(\phi_i)\}$, or equivalently, $\phi_i = \log(\theta_i/(1 - \theta_i)))$. The normal-inverse-gamma prior, denoted as $(\mu, \sigma^2) \sim \mathcal{NIG}(\nu, \omega, a, b)$, is equivalent to a mixture of normal and inverse gamma priors: $\mu | \sigma^2 \sim \mathcal{N}(\nu, \sigma^2/\omega)$ and $\sigma^2 \sim \mathcal{IG}(a, b)$. $(\nu, \omega, \alpha, \beta)$ represent the hyper-parameters, which we set as (0, 1/100, 0.001, 0.001). This choice ensures that the normal-inverse-gamma prior is diffused over the parameter space, and the prior information is almost vague (essentially, nearly non-informative), similar to the choice made by [146].

The hierarchical formulation (20)–(22) is designed to induce a shrinkage effect [150, 151]. Under this formulation, the Bayesian estimators of the parameters $\phi_i, (i = 1, \cdots, K)$ (or equivalently, $\theta_i, (i = 1, \cdots, K)$) will be pulled toward the global mean μ (or equivalently, $\exp(\mu)/1 + \exp(\mu)$), leading to a reduction in the width of the interval estimates of the parameters, a posteriori, similar to the James-Stein shrinkage estimator [152]. This shrinkage effect is also referred to as "borrowing strength", recognized in numerous regulatory guidance documents related to clinical trials for medical devices and small populations [16, 153].

To test the null and alternative hypotheses associated with the i-th endpoint (16), we use the posterior probability approach for decision-making as follow. Upon completion of the study, for each i $(i = 1, \cdots, K)$, we reject the i-th null hypothesis, $\mathcal{H}_{0,i} : \theta_i \leq \theta_{0,i}$, if the
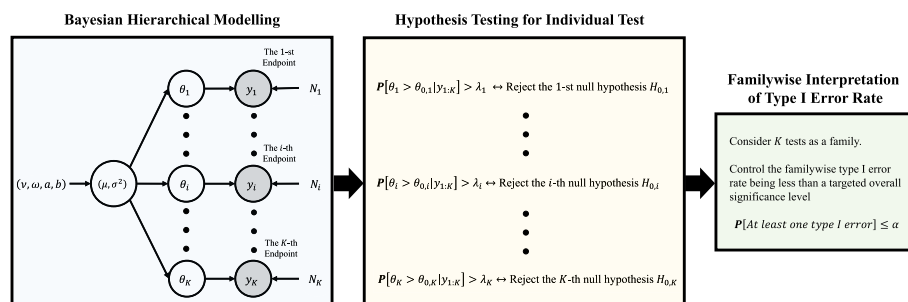

Fig. 8 Three steps to control the familywise type I error rate through Bayesian hierarchical modeling. The first step involves specifying a Bayesian hierarchical model, which depends on the context of the problem. In the second step, the decision rule for each individual test is specified. The third step involves interpreting the combination of individual type I error rates in terms of the familywise type I error rate, which is restricted by the overall significance level

posterior probability of the i-th alternative hypothesis, $\mathcal{H}_{a,i} : \theta_i > \theta_{0,i}$, being true is greater than a pre-specified threshold $\lambda_i \in [0, 1]$. That is, the decision criterion for the i-th endpoint is as follow:

$$\text{Sucess for the } i\text{-th endpoint} = \mathbf{1}\{\mathbb{P}[\theta_i > \theta_{0,i}|y_{1:K}] > \lambda_i\}, \quad (i = 1, \cdots, K), \tag{23}$$

where $y_{1:K}$ denotes the numbers of responses from the K endpoints. A higher value of λ_i leads to a more conservative testing for the i-th endpoint, resulting in a lower type I error rate and a lower power, given a fixed sample size N_i. The posterior probability in (23) is typically stochastically approximated by an analogous form of Algorithm 1 in Supplemental material, based on an MCMC method because the posterior distribution, $\pi(\theta_{1:K}, \mu, \sigma^2|y_{1:K})$, is not represented as a closed-form distribution.

Suppose that the i-th null hypothesis has been rejected at the final analysis. In this case, the drug is considered to have demonstrated effects for the i-th endpoint. The K threshold values $(\lambda_1, \ldots, \lambda_K)$ in the decision criteria (23) should be pre-specified during the design stage and chosen through simulation to ensure that the frequentist familywise type I error α^{family} (17) is less than the overall significance level α.

Simulation experiment

We evaluate the performance of Bayesian hierarchical modeling and frequentist methods (specifically, Bonferroni, Holm, and Hochberg procedures) as described in Frequentist method - p-value based procedures section under varying assumptions of the number of endpoints (K) from 1 to 10. Regarding the threshold for the decision rule (23) of Bayesian hierarchical modeling, we use the same value, $\lambda_i = 0.985$, for all endpoints $i = 1, \cdots, K$, irrespective of the number of endpoints K. In other words, there is no specific threshold adjustment concerning the number of endpoints (K).

The thresholds (adjusted alphas) for the Bonferroni, Holm, and Hochberg procedures are described in Frequentist method - p-value based procedures section. Note that the thresholds for the three procedures are set to be increasingly stringent as the number of endpoints (K) increases, aiming to keep the familywise type I error α^{family} less than α.

The sample size for each sub-population, N_i ($i = 1, \cdots, K$), is set to 85 or 100. For a single endpoint ($K = 1$), these sample sizes lead to a power of approximately 80% ($N_i = 85$) and 86% ($N_i = 100$) based on the Z-test for one proportion at the one-sided significance level $\alpha = 0.025$.

The followings are summary of the simulation setting:

- Number of endpoints: $K = 1, 2, \cdots, 10$,
- One-sided significance level: $\alpha = 0.025$,

- Number of patients: $N_i = 85$ or 100, $(i = 1, \cdots, K)$,
- Performance goals: $\theta_{0,i} = 0.35$, $(i = 1, \cdots, K)$,
- Anticipated rates: $\theta_{a,i} = 0.5$, $(i = 1, \cdots, K)$,
- Multiplicity adjustment methods:

1. Bayesian hierarchical modeling (Bayesian method),
2. Bonferroni, Holm, and Hochberg procedures (Frequentist methods),

- Decision rule:

1. Bayesian hierarchical modeling: Posterior probability approach (23) with the threshold $\lambda_i = 0.985$, $(i = 1, \cdots, K)$ across all settings,
2. Bonferroni, Holm, and Hochberg procedures: Use the adjusted p-value as described in Frequentist method - p-value based procedures section such that the unadjusted p-value are obtained by the exact binomial test [154].

Figure 9 displays the results of simulation experiments. Panels (a) and (b) demonstrate that all the considered adjustment methods successfully control the familywise type I error rate, α^{family}, at the one-sided significance level of $\alpha = 0.025$ across the number of endpoints K. Notably, these two panels show that the familywise type I error rate, α^{family}, based on Bayesian method decreases as K increases, even when the same thresholds $\lambda_i = 0.985$ are universally used across all settings. This result implies that there is no need for adjustments of the Bayesian threshold [140]. Essentially, this nice property is due to the shrinkage effect: borrowing strength across sub-populations automatically adjusts the familywise type I error rate α^{family} to be less than $\alpha = 0.025$.

Panels (c) and (d) demonstrate that the disjunctive powers π^{dis} (18) of all the considered adjustment methods increase as K increases. The Bayesian method is the most powerful, while the Bonferroni method is the least powerful among the four methods. The Hochberg method is marginally more powerful than the Holm method. Panels (e) and (f) show that only the Bayesian method leads to an increase in the conjunctive power π^{conj} (19) as K increases. These results indicate that the shrinkage effect of Bayesian hierarchical modeling is beneficial under the two power scenarios. In contrast, p-value-based multiplicity adjustment procedures are only appropriate to use under the disjunctive power scenario. This implies that the total sample size $N = \sum_{i=1}^{K} N_i$ required for the study can be significantly reduced if the Bayesian hierarchical
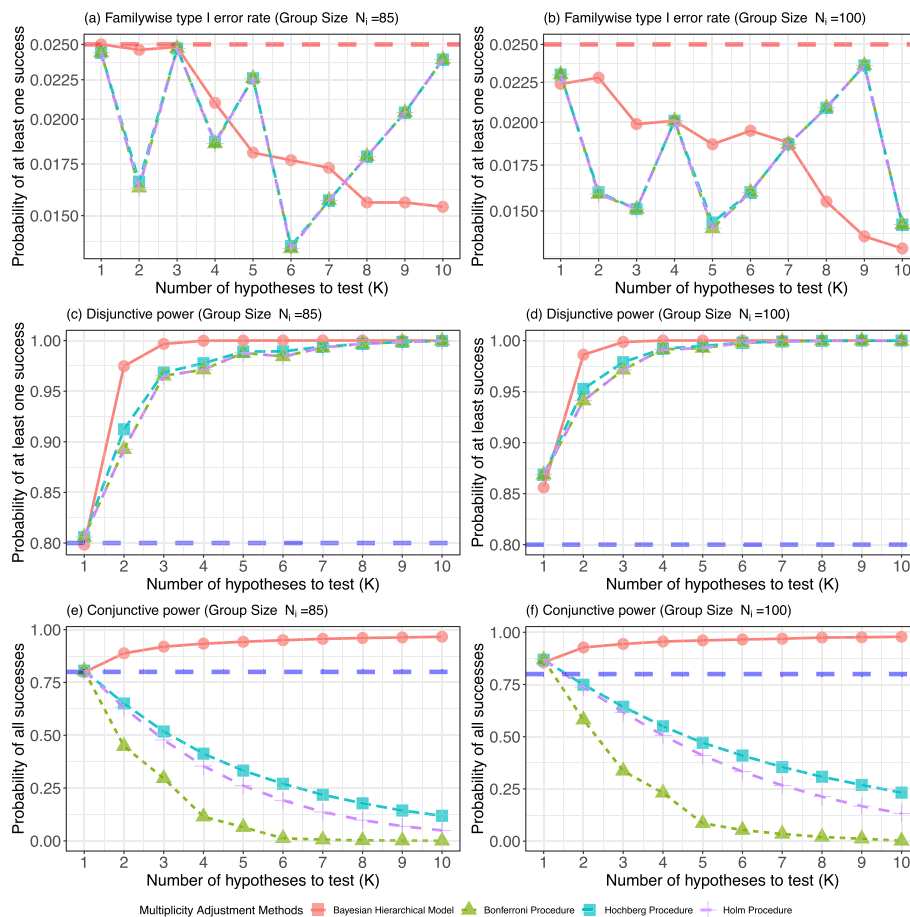
Fig. 9 Results of simulation experiment with different number of endpoints ($K = 1, \cdots, 10$) and group size ($N_i = 85, 100$)

model is used, compared to the frequentist methods. Particularly for the conjunctive power scenario, only the Bayesian hierarchical model possesses this unique advantage.

To summarize, the simulation experiment implies that the mechanism of multiplicity adjustment (shrinkage effect or borrowing strength) is automatically embedded in Bayesian hierarchical modeling. This controls the familywise type I error rate to be less than the significance level and improves both disjunctive and conjunctive powers as the number of hypotheses increases. This contrasts with the *p*-value-based procedures, which are criticized by their overconservatism, which becomes acute when the number of hypotheses is large [134, 136, 155, 156].

External data borrowing

Bayesian information borrowing for regulatory submission

There is a growing interest in Bayesian clinical trial designs with informative prior distributions, allowing the borrowing of information from an external source. Borrowing information from previously completed trials

is used extensively in medical device trials [16, 20, 157] and is increasingly seen in drug trials for extrapolation of adult data to pediatrics [35] or leveraging historical datasets for rare diseases [158–160]. In general, sponsors benefit in multiple ways by using Bayesian borrowing designs, including reductions in sample size, time, expense, and increased statistical power.

In practice, the key difficulty facing stakeholders hoping to design a trial using Bayesian borrowing methods is understanding the similarity of previous studies to the current study, including factors such as enrollment and treatment criteria, and achieving exchangeability between the studies in discussions with regulators. For example, outcomes of medical device trials for a device can vary substantially due to the device evolvement from the previous to the next generation, or by site influenced by differences such as physician training, technique, experience with the device, patient management, and patient population, among many other factors. Regulatory agencies recognize that two studies are never exactly alike; nonetheless, it is recommended

that the studies used to construct the informative prior be similar to the current study in terms of the protocol (endpoints, target population, etc.) and the time frame of the data collection to ensure that the practice of medicine and the study populations are comparable [16]. It is crucial that companies and regulators reach an agreement regarding the prior information and the Bayesian design before enrolling any patients in the new study [161].

One perceptible trend in the Bayesian regulatory environment is that the strict control of the type I error rate in the frequentist framework may need to be relaxed to a less stringent level for Bayesian submissions using information borrowed from external evidence, due to the unavoidable inflation of the type I error rate in certain scenarios [61, 75, 76, 162]. Such an inflation scenario typically occurs when the external data is more favorable for rejecting the null hypothesis of the current trial. Regulators are also increasingly aware of the substantial limitations that stringent control of the frequentist type I error may entail. For example, an FDA guidance [16] states that, '*If the FDA considers the type I error rate of a Bayesian experimental design to be too large, we recommend modifying the design or the model to reduce that rate. Determination of "too large" is specific to a submission because some*

putting restrictions on the amount of borrowing allowed from previous studies, one of the standard ways is to control the weight parameter on the external study data, which is typically a fractional real number [81, 163–166], and calibrate it to satisfy the requirement of the agreed maximally allowable type I error rate. In the next section, we illustrate the use of a power prior model to leverage historical data from a pilot study and explore the influence of the weight parameter on the frequentist operating characteristics of the Bayesian design.

Example - Bayesian borrowing design based on power prior

We illustrate a Bayesian borrowing design based on a power prior [36, 81] by taking the primary safety endpoint discussed in Example - standard single-stage design based on beta-binomial model section as an example. Suppose that a single-arm pilot trial with the number of patients $N_0 = 100$ is done under similar enrollment and treatment criteria as a new pivotal trial. The pilot study provides binary outcome data $\mathbf{y}_{N0} = (y_{10}, \cdots, y_{i0}, \cdots, y_{N_00})^\top$ for the informative prior in the Bayesian power prior method. The power prior raises the likelihood of the pilot data to the power parameter a_0, which quantifies the discounting of the pilot data due to heterogeneity between pilot and pivotal trials:

$$\pi(\theta|\mathbf{y}_{N_0}, a_0) \propto f(\mathbf{y}_{N_0}|\theta)^{a_0} \cdot \pi_0(\theta) \propto \left\{ \prod_{i=1}^{N_0} \theta^{y_{i0}}(1-\theta)^{1-y_{i0}} \right\}^{a_0} \cdot \mathcal{B}\text{eta}(\theta|0.01, 0.01) \qquad (24)$$
$$\propto \mathcal{B}\text{eta}(\theta|a_0 x_0 + 0.01, a_0(N_0 - x_0) + 0.01),$$

sources of type I error inflation (e.g., large amounts of valid prior information) may be more acceptable than others (e.g., inappropriate choice of studies for construction of the prior, inappropriate statistical model, or inappropriate criteria for study success). The seriousness (cost) of a Type I error is also a consideration.' Several approvals were granted both in the US and in Europe based on non-randomized studies using external controls [160]. Even though these approvals were typically for rare diseases, they signal the increasing willingness of regulators to review applications for Bayesian borrowing designs.

In order to control the type I error rate at a reasonable level with which stakeholders agree, one of the key aspects of Bayesian borrowing designs is to appropriately discount historical/prior information if the prior distribution is too informative relative to the current study [16]. Although such discounting can be achieved by directly changing the hyper-parameters of the prior, as exemplified by a beta-binomial model seen in Table 1, or by

where $x_0 = \sum_{i=1}^{N_0} y_{i0}$ represents the number of patients who experienced a primary adverse event within 30 days after a surgical procedure involving the device in the pilot trial.

In the power prior formulation (24), $\pi_0(\theta)$ denotes the prior distribution for θ before observing the pilot study data $\mathbf{y}_0$; this is referred to as the initial prior. The initial prior is often chosen to be noninformative, and in this example, we use $\pi_0(\theta) = \mathcal{B}\text{eta}(\theta|0.01, 0.01)$.

The power parameter $a_0 \in [0, 1]$ weighs the pilot data relative to the likelihood of the pivotal trial. The special cases of using the pilot data fully or not at all are covered by $a_0 = 1$ and $a_0 = 0$, respectively, while values of a_0 between 0 and 1 allow for differential weighting of the pilot data. The value $a_0 N_0$ can be interpreted as the prior effective sample size, the number of patients to be borrowed from the pilot study. The parameter a_0 can be estimated by using the normalized power prior formulation [163, 167]. However, in this paper, we fix a_0 since our purpose is to explore the influence of the power parameter a_0 on

the frequentist operating characteristics of the Bayesian design.

Finally, the posterior distribution, given the outcomes from patients in pivotal and pilot trials, is once again the beta distribution due to the conjugation relationship:

$$\pi(\theta|\mathbf{y}_N, \mathbf{y}_{N_0}, a_0) \propto f(\mathbf{y}_N|\theta) \cdot \pi(\theta|\mathbf{y}_{N_0}, a_0)$$
$$\propto \mathcal{B}\text{eta}(\theta|x + a_0 x_0 + 0.01, N - x + a_0(N_0 - x_0) + 0.01). \quad (25)$$

and $\beta_{\theta_a}^{(N)}(\mathbf{y}_{N_0}, a_0)$ (27) reduce to the values of $\beta_{\theta_0}^{(N)}$ (7) and $\beta_{\theta_a}^{(N)}$ (8), respectively. Otherwise ($0 < a_0 \leq 1$), the former values could be significantly different from the latter values.

Building upon the scenario presented in Example - standard single-stage design based on beta-binomial model section, the sponsor, during the planning stage of the pivotal trial, anticipated a safety rate of $\theta_a = 0.05$ with a performance goal set at $\theta_0 = 0.12$. At this stage, $\mathbf{y}_N$ is a random quantity, while $\mathbf{y}_{N_0}$ is observed, and a_0 is fixed at a specific value to control the influence of $\mathbf{y}_{N0}$ in the decision-making process. The decision rule states that if $T(\mathbf{y}_N, \mathbf{y}_{N_0}, a_0) = \mathbb{P}[\theta < 0.12|\mathbf{y}_N, \mathbf{y}_{N_0}, a_0] > 0.975$, then the null hypothesis $\mathcal{H}_0 : \theta \geq 0.12$ is rejected, implying the success of the study in ensuring the safety of the device.

Frequentist operating characteristics of this Bayesian borrowing design can be summarized by two following quantities:

In the following, we explore the operating characteristics of this Bayesian borrowing design under the two different scenarios regarding the direction of the pilot study data, whether it is favorable or unfavorable to reject the null hypothesis. In the optimistic external scenario, $x_0 = 5$ out of $N_0 = 100$ patients experienced the adverse event, resulting in a historical event rate of 0.05, which is lower than the performance goal of $\theta_0 = 0.12$. In contrast, the pessimistic external scenario is where $x_0 = 15$ out of $N_0 = 100$ patients experienced the adverse event, leading to a historical event rate of 0.15, which is higher than the performance goal.

Figure 10 displays the probability of rejecting the null hypothesis versus the power parameter a_0 for the two
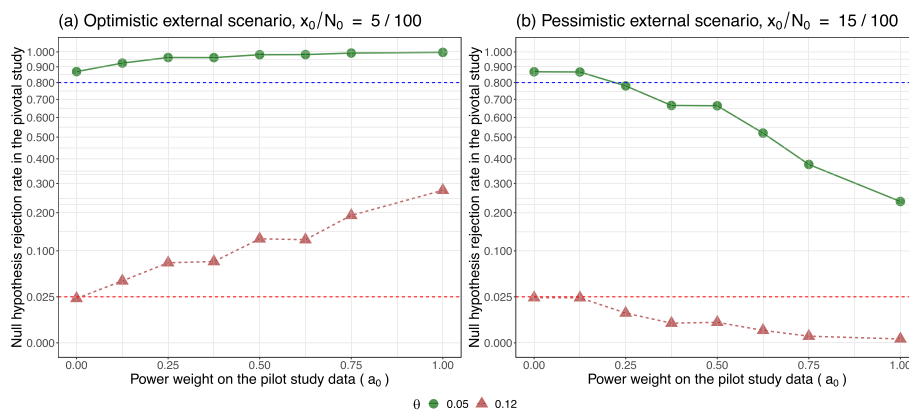
$$\text{Type I error} : \beta_{\theta_0}^{(N)}(\mathbf{y}_{N_0}, a_0) = \mathbb{P}[T(\mathbf{y}_N, \mathbf{y}_{N_0}, a_0) > 0.975|\mathbf{y}_N \sim f(\mathbf{y}_N|\theta_0), \mathbf{y}_{N_0}, a_0], \quad (26)$$

$$\text{Power} : \beta_{\theta_a}^{(N)}(\mathbf{y}_{N_0}, a_0) = \mathbb{P}[T(\mathbf{y}_N, \mathbf{y}_{N_0}, a_0) > 0.975|\mathbf{y}_N \sim f(\mathbf{y}_N|\theta_a), \mathbf{y}_{N_0}, a_0]. \quad (27)$$

It is important to note that the type I error rate and power of Bayesian borrowing designs depend on the pilot study data $\mathbf{y}_{N_0}$ and the power parameter a_0. In the case of no borrowing ($a_0 = 0$), the values of $\beta_{\theta_0}^{(N)}(\mathbf{y}_{N_0}, a_0)$ (26)

scenarios, provided that the sample size for the pivotal trial is $N = 150$. The true safety rate θ is set to be either $\theta_a = 0.05$ or $\theta_0 = 0.12$, corresponding to the power and type I error scenarios, respectively. In the case of no

Fig. 10 Null hypothesis rejection rate $\beta_\theta^{(N)} = \mathbb{P}[T(\mathbf{y}_N, \mathbf{y}_{N_0}, a_0) > 0.975|\mathbf{y}_N \sim f(\mathbf{y}_N|\theta)]$, ($\theta = \theta_a = 0.05$: alternative parameter, $\theta = \theta_0 = 0.12$: null parameter) versus power parameter a_0 under the optimistic external scenario (Panel **a**, $x_0/N_0 = 5/100$) and pessimistic external scenario (Panel **b**, $x_0/N_0 = 15/100$). Sample sizes of the pivotal and pilot trials are $N = 150$ and $N_0 = 100$, respectively

borrowing (that is, $a_0 = 0$), the type I error rate is 0.0225, and power is 0.8681, which is almost identical to those obtained from the Bayesian design with a non-informative beta prior and the frequentist design based on z-test statistics seen in Table 1.

Panel (a) in Fig. 10 demonstrates that, in the optimistic external scenario, the type I error rate (26) and power (27) simultaneously increase as the power parameter a_0 increases. Conversely, in the pessimistic external scenario (Panel (b)), the type I error rate (26) and power (27) simultaneously decrease as the power parameter a_0 increases. It is important to note that the inflation of the type I error in panel (a) and the deflation of the power in panel (b) are expected (see Example - standard single-stage design based on beta-binomial model section for relevant discussion).

The central question at this point is, 'Is the inflation of the type I error rate (26) under the optimistic scenario scientifically sound for the regulatory submission?' To answer this question, let us assume that the pilot and pivotal studies are very similar and that the pilot study data provide high quality so that the two studies are essentially exchangeable (refer to Subsection 3.7 in [16] for the concept of exchangeability). Under this idealistic assumption, this inflation is a mathematical result due to the opposite direction of pilot study data $\mathbf{y}_{N_0}$ (favoring the alternative hypothesis) and pivotal study data $\mathbf{y}_N$ (generated under the null hypothesis), not due to the incorrect use of the Bayesian borrowing design. Therefore, the inflation of the type I error rate under the optimistic scenario is scientifically sound for the regulatory submission only when the two studies are exchangeable.

In practice, establishing exchangeability between the two studies poses a somewhat intricate challenge, and regulatory bodies acknowledge that no two studies are entirely identical [16]. Therefore, the key to the successful submission of a Bayesian borrowing design is to mitigate any potential systematic biases (and consequently the risk of incorrect conclusions) when the pivotal study data $\mathbf{y}_N$ appears to be inconsistent with the pilot study data $\mathbf{y}_{N_0}$. This ultimately involves finding an appropriate degree of down-weighting for the pilot study data when such a prior-data conflict is present [168]. However, this is again a challenging task because, from an operational viewpoint, the pivotal study data $\mathbf{y}_N$ will be observed upon completion of the study, while the pilot study data $\mathbf{y}_{N_0}$ has already been observed during the planning phase. The key difficulty here is that the power parameter a_0 should be determined in the planning phase specified in the protocol or Statistical Analysis Plan before seeing any pivotal study data $\mathbf{y}_N$. One can estimate the power parameter a_0 through dynamic borrowing techniques [169], but such methods may have their own tuning parameters to

control the power parameter a_0 so the central issue does not completely disappear.

For this reason, thorough pre-planning is essential when employing Bayesian borrowing designs. This necessitates subject matter expertise, interactions, and a consensus among all stakeholders. It is crucial to establish an agreement on analysis and design priors, with the latter being utilized to assess the operating characteristics of the trial design under all conceivable scenarios. In this regard, a graphical approach can be used to help select design parameters, including the degree of discounting for the pilot study data [170].

Figure 11 presents heatmaps for the type I error rate (left heatmaps) and power (right heatmaps) to explore how changing the power parameter (a_0) and sample size in the pivotal study (N) impacts the type I error and power. As seen from panels (a) and (d), the inflation of the type I error under the optimistic external scenario and the deflation of power under the pessimistic external scenario are evident across the sample size of the pivotal trial (N). Another interesting phenomenon is that, as N increases, the tendencies of inflation/deflation diminish across the parameter (a_0), showcasing the Bernstein-Von Mises phenomenon [90, 91] as discussed in Asymptotic property of posterior probability approach section. This suggests that sponsors can benefit from Bayesian borrowing designs in reducing the sample size N only when the pilot study data favorably support rejecting the null hypothesis and N is not excessively large. The acceptable amount of pilot study data to be borrowed should be agreed upon in discussions with regulators because inflation of the type I error rate is expected in this scenario.

Conclusions

There have been many Bayesian clinical studies conducted and published in top-tier journals [18, 20, 23, 37, 171]. Nevertheless, the adoption of Bayesian statistics for the registration of new drugs and medical devices requires a significant advancement in regulatory science, presenting a range of potential benefits and challenges. In this section, we discuss key aspects of this transformation.

Complex innovative trial designs: The Bayesian framework provides a promising method to address a variety of modern design complexities as part of complex innovative trial designs. For example, it enables real-time adjustments to trial design, sample size, and patient allocation based on accumulating data from subjects in the trial. These adaptive features can expedite the development of medical products, reduce costs, and enhance patient safety: as exemplified in Example - two-stage group sequential design based on beta-binomial model
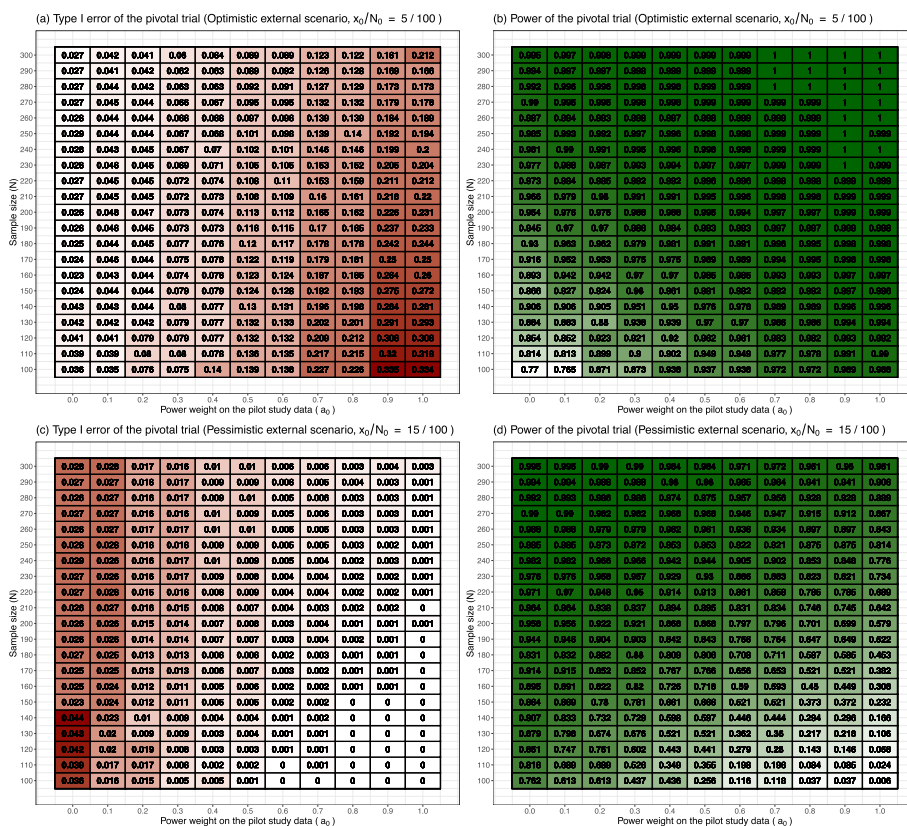
Fig. 11 Heatmaps to illustrate the frequentist operating characteristics of the Bayesian borrowing design. The y-axis and x-axis represent the sample size of the pivotal trial (N) and the power parameter (a_0), respectively. The contents in the heatmaps are the null hypothesis rejection rates $\beta_\theta^{(N)} = \mathbb{P}[T(\mathbf{y}_N, \mathbf{y}_{N_0}, a_0) > 0.975 | \mathbf{y}_N \sim f(\mathbf{y}_N | \theta)]$, where the type I error rate and power are obtained by setting $\theta = \theta_0 = 0.12$ and $\theta = \theta_a = 0.05$, respectively

and Example - two-stage futility design with Greenwood test sections. More recently, platform clinical trials have offered a flexible, efficient, and patient-centered approach to drug development and evaluation, with the potential to improve outcomes for patients and streamline the drug development process [9, 172, 173]. While adaptive features provide the design with great flexibility, it is important to note that such trial adaptations are scientifically valid only when prospectively planned and specified in the protocol or Statistical Analysis Plan, considering all alternative scenarios, and when conducted according to the pre-specified decision rules [174, 175]. Therefore, it is advisable for sponsors to seek early interaction with regulators regarding the details of their plans for using Bayesian methods [12].

Incorporating prior information: One defining feature of Bayesian statistics is the ability to incorporate prior information into the analysis. This contrasts with classical frequentist statistics, which may use information from previous studies only at the design stage. This feature is invaluable when designing clinical trials, especially

in situations where historical or more generally study-external data are available. The utilization of informative priors can improve statistical efficiency and enhance the precision of treatment effect estimates. However, it is essential to carefully consider the source and relevance of prior information to ensure the validity and integrity of the trial. Furthermore, as discussed in External data borrowing section, type I error inflation is expected to occur in certain situations. More theoretical work needs to be done in this area to clarify that the stringent control of the type I error probability when there is prior information is not an appropriate way to think about this problem. See Subsection 2.4.3 from [77] for relevant discussion.

Rare diseases and small sample sizes: In the context of rare diseases, where limited patient populations hinder traditional frequentist approaches, Bayesian methods are useful. They allow for the integration of diverse data sources, such as historical data or data from similar diseases, to provide robust evidence with a possibly smaller sample size than traditional frequentist approaches.

Obtaining ethical and institutional approval is easier in small studies compared with large multicentre studies [176]. However, as discussed in Asymptotic property of posterior probability approach section, the operating characteristics of clinical trial designs with a small sample size are more sensitive to the choice of the prior than those with a moderate or large sample size. This implies that smaller clinical trials are more vulnerable to the conflict between the trial data and prior evidence than larger clinical trials. More research is needed in both regulatory science and methodology in this area to mitigate such a conflict and ensure a safe path to regulatory submission, minimizing potential systemic bias.

Regulatory considerations: The integration of Bayesian statistics into the regulatory setting requires adherence to established guidelines and frameworks. In the past decade, the FDA has recognized the potential of Bayesian approaches and has provided guidance on their use [16, 40, 41, 72]. However, the adoption of Bayesian statistics is not without challenges and debates. Some statisticians and stakeholders remain cautious about the subjective nature of prior elicitation, potential biases, and the interpretation of Bayesian results. The ongoing debate surrounding the calibration of Bayesian methods, particularly in the context of decision-making, underscores the need for further research and consensus in the field.

Software implementation   For simple Bayesian designs, using built-in R functions or specialized tools like STAN [177] and JAGS [178] facilitates power analysis without requiring the user to construct an MCMC sampler. Parallel computation may not be necessary in these cases. However, for complex designs involving multiple arms, statistical modeling for enrollment, or multiple interim analyses, computational times increase significantly. Parallel computing becomes essential, often requiring high-performance computing resources. Specific expertise in Bayesian computation tailored for regulatory submission is crucial. Thus, having a skilled Bayesian statistician, either as an employee or consultant, is highly beneficial for guiding statistical aspects and developing customized Bayesian software in R, SAS [179, 180], or similar tools.

In conclusion, the use of Bayesian statistics in clinical trials within the regulatory setting is a promising evolution that can enhance the efficiency and effectiveness of the development process for new drugs or medical devices. However, successful implementation requires rigorous prior specification, careful consideration of decision rules to achieve the study objective, and adherence to regulatory guidelines. The Bayesian paradigm has demonstrated its potential in addressing the complexities of modern clinical trials, offering a versatile tool for researchers and regulators alike. As researchers, clinicians, and regulatory agencies continue to explore the benefits of Bayesian statistics, it is essential to foster collaboration, transparency, and ongoing dialogue to refine and harmonize the use of Bayesian approaches in clinical trials.

Abbreviations

FDA   Food and Drug Administration
MCMC   Markov Chain Monte Carlo

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02235-0.

> Supplementary Material 1.

References
1. Friedman LM, Furberg CD, DeMets DL, Reboussin DM, Granger CB. Fundamentals of clinical trials. Boston: Springer; 2015.
2. Zhou X, Liu S, Kim ES, Herbst RS, Lee JJ. Bayesian adaptive design for targeted therapy development in lung cancer-a step toward personalized medicine. Clin Trials. 2008;5(3):181–93.
3. Fountzilas E, Tsimberidou AM, Vo HH, Kurzrock R. Clinical trial design in the era of precision medicine. Genome Med. 2022;14(1):1–27.
4. Carlin BP, Nollevaux F. Bayesian complex innovative trial designs (CIDs) and their use in drug development for rare disease. J Clin Pharmacol. 2022;62:S56–71.
5. Wilson DT, Wason JM, Brown J, Farrin AJ, Walwyn RE. Bayesian design and analysis of external pilot trials for complex interventions. Stat Med. 2021;40(12):2877–92.

6. Yue LQ. Leveraging Real-World Evidence Derived from Patient Registries for Premarket Medical Device Regulatory Decision-Making. Stat Biopharm Res. 2018;10(2):98–103.

7. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-world evidence-what is it and what can it tell us. N Engl J Med. 2016;375(23):2293–7.

8. Wang C, Li H, Chen WC, Lu N, Tiwari R, Xu Y, et al. Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies. J Biopharm Stat. 2019;29(5):731–48.

9. Woodcock J, LaVange LM. Master protocols to study multiple therapies, multiple diseases, or both. N Engl J Med. 2017;377(1):62–70.

10. Moscicki RA, Tandon P. Drug-development challenges for small biopharmaceutical companies. N Engl J Med. 2017;376(5):469–74.

11. Bhatt DL, Mehta C. Adaptive designs for clinical trials. N Engl J Med. 2016;375(1):65–74.

12. U.S. Food and Drug Administration. Interacting with the FDA on Complex Innovative Trial Designs for Drugs and Biological Products. 2020. www.fda.gov/regulatory-information/search-fda-guidance-documents/interacting-fda-complex-innovative-trial-designs-drugs-and-biological-products. Accessed 23 Nov 2023.

13. Berry DA. Bayesian clinical trials. Nat Rev Drug Discov. 2006;5(1):27–36.

14. Jack Lee J, Chu CT. Bayesian clinical trials in action. Stat Med. 2012;31(25):2955–72.

15. Landau S, Stahl D. Sample size and power calculations for medical studies by simulation when closed form expressions are not available. Stat Methods Med Res. 2013;22(3):324–45.

16. U.S. Food and Drug Administration. Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. 2010. www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-use-bayesian-statistics-medical-device-clinical-trials. Accessed 23 Nov 2023.

17. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health-care evaluation, vol. 13. London: Wiley; 2004.

18. Wilber DJ, Pappone C, Neuzil P, De Paola A, Marchlinski F, Natale A, et al. Comparison of antiarrhythmic drug therapy and radiofrequency catheter ablation in patients with paroxysmal atrial fibrillation: a randomized controlled trial. Jama. 2010;303(4):333–40.

19. Gsponer T, Gerber F, Bornkamp B, Ohlssen D, Vandemeulebroecke M, Schmidli H. A practical guide to Bayesian group sequential designs. Pharm Stat. 2014;13(1):71–80.

20. Böhm M, Kario K, Kandzari DE, Mahfoud F, Weber MA, Schmieder RE, et al. Efficacy of catheter-based renal denervation in the absence of antihypertensive medications (SPYRAL HTN-OFF MED Pivotal): a multicentre, randomised, sham-controlled trial. Lancet. 2020;395(10234):1444–51.

21. Schmidli H, Häring DA, Thomas M, Cassidy A, Weber S, Bretz F. Beyond randomized clinical trials: use of external controls. Clin Pharmacol Ther. 2020;107(4):806–16.

22. Schmidli H, Bretz F, Racine-Poon A. Bayesian predictive power for interim adaptation in seamless phase II/III trials where the endpoint is survival up to some specified timepoint. Stat Med. 2007;26(27):4925–38.

23. Polack FP, Thomas SJ, Kitchin N, Absalon J, Gurtman A, Lockhart S, et al. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. N Engl J Med. 2020;383(27):2603–15.

24. U.S. Food and Drug Administration. Master Protocols: Efficient Clinical Trial Design Strategies to Expedite Development of Oncology Drugs and Biologics Guidance for Industry. 2022. www.fda.gov/regulatory-information/search-fda-guidance-documents/master-protocols-efficient-clinical-trial-design-strategies-expedite-development-oncology-drugs-and. Accessed 23 Nov 2023.

25. Berry SM, Petzold EA, Dull P, Thielman NM, Cunningham CK, Corey GR, et al. A response adaptive randomization platform trial for efficient evaluation of Ebola virus treatments: a model for pandemic response. Clin Trials. 2016;13(1):22–30.

26. Chu Y, Yuan Y. A Bayesian basket trial design using a calibrated Bayesian hierarchical model. Clin Trials. 2018;15(2):149–58.

27. Hirakawa A, Asano J, Sato H, Teramukai S. Master protocol trials in oncology: review and new trial designs. Contemp Clin Trials Commun. 2018;12:1–8.

28. Hobbs BP, Landin R. Bayesian basket trial design with exchangeability monitoring. Stat Med. 2018;37(25):3557–72.

29. Dodd LE, Proschan MA, Neuhaus J, Koopmeiners JS, Neaton J, Beigel JD, et al. Design of a randomized controlled trial for Ebola virus disease medical countermeasures: PREVAIL II, the Ebola MCM Study. J Infect Dis. 2016;213(12):1906–13.

30. Quintana M, Saville BR, Vestrucci M, Detry MA, Chibnik L, Shefner J, et al. Design and Statistical Innovations in a Platform Trial for ALS. Ann Neurol. 2023;94(3):417–609.

31. Alexander BM, Ba S, Berger MS, Berry DA, Cavenee WK, Chang SM, et al. Adaptive global innovative learning environment for glioblastoma: GBM AGILE. Clin Cancer Res. 2018;24(4):737–43.

32. I-SPY COVID Consortium. Clinical trial design during and beyond the pandemic: the I-SPY COVID trial. Nat Med. 2022;28(1):9–11.

33. Wang Y, Travis J, Gajewski B. Bayesian adaptive design for pediatric clinical trials incorporating a community of prior beliefs. BMC Med Res Methodol. 2022;22(1):118.

34. Psioda MA, Xue X. A Bayesian adaptive two-stage design for pediatric clinical trials. J Biopharm Stat. 2020;30(6):1091–108.

35. Gamalo-Siebers M, Savic J, Basu C, Zhao X, Gopalakrishnan M, Gao A, et al. Statistical modeling for Bayesian extrapolation of adult clinical trial information in pediatric drug evaluation. Pharm Stat. 2017;16(4):232–49.

36. Ibrahim JG, Chen MH. Power prior distributions for regression models. Stat Sci. 2000;15:46–60.

37. Richeldi L, Azuma A, Cottin V, Hesslinger C, Stowasser S, Valenzuela C, et al. Trial of a preferential phosphodiesterase 4B inhibitor for idiopathic pulmonary fibrosis. N Engl J Med. 2022;386(23):2178–87.

38. Müller P, Chandra N, Sarkar A. Bayesian approaches to include real-world data in clinical studies. Phil Trans R Soc A. 2023;381(2247):20220158.

39. U.S. Food and Drug Administration. Leveraging Existing Clinical Data for Extrapolation to Pediatric Uses of Medical Devices. 2016. www.fda.gov/regulatory-information/search-fda-guidance-documents/leveraging-existing-clinical-data-extrapolation-pediatric-uses-medical-devices. Accessed 23 Nov 2023.

40. U.S. Food and Drug Administration. Adaptive Designs for Medical Device Clinical Studies. 2016. www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-designs-medical-device-clinical-studies. Accessed 23 Nov 2023.

41. U.S. Food and Drug Administration. Adaptive Design Clinical Trials for Drugs and Biologics Guidance for Industry. 2019. www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry. Accessed 23 Nov 2023.

42. Spiegelhalter DJ, Freedman LS, Parmar MK. Bayesian approaches to randomized trials. J R Stat Soc Ser A (Stat Soc). 1994;157(3):357–87.

43. Zhou T, Ji Y. On Bayesian Sequential Clinical Trial Designs. New England J Stat Data Sci. 2023;2(1):136–51.

44. Bittl JA, He Y. Bayesian analysis: a practical approach to interpret clinical trials and create clinical practice guidelines. Circ Cardiovasc Qual Outcome. 2017;10(8):e003563.

45. Hobbs BP, Carlin BP. Practical Bayesian design and analysis for drug and device clinical trials. J Biopharm Stat. 2007;18(1):54–80.

46. Stallard N, Todd S, Ryan EG, Gates S. Comparison of Bayesian and frequentist group-sequential clinical trial designs. BMC Med Res Methodol. 2020;20:1–14.

47. Yin G, Lam CK, Shi H. Bayesian randomized clinical trials: From fixed to adaptive design. Contemp Clin Trials. 2017;59:77–86.

48. Berry SM, Carlin BP, Lee JJ, Muller P. Bayesian adaptive methods for clinical trials. Boca Raton: CRC Press; 2010.

49. Hirakawa A, Sato H, Igeta M, Fujikawa K, Daimon T, Teramukai S. Regulatory issues and the potential use of Bayesian approaches for early drug approval systems in Japan. Pharm Stat. 2022;21(3):691–5.

50. Rosner GL. Bayesian methods in regulatory science. Stat Biopharm Res. 2020;12(2):130–6.

51. Campbell G, Irony T, Pennello G, Thompson L. Bayesian Statistics for Medical Devices: Progress Since 2010. Ther Innov Regul Sci. 2023;57(3):453–63.

52. Pocock SJ. Clinical trials: a practical approach. London: Wiley; 2013.

53. Chow SC, Shao J, Wang H, Lokhnygina Y. Sample size calculations in clinical research. Boca Raton: CRC Press; 2017.

54. Inoue LY, Berry DA, Parmigiani G. Relationship between Bayesian and frequentist sample size determination. Am Stat. 2005;59(1):79–87.

55. Katsis A, Toman B. Bayesian sample size calculations for binomial experiments. J Stat Plan Infer. 1999;81(2):349–62.

56. Joseph L, Wolfson DB, Berger RD. Sample size calculations for binomial proportions via highest posterior density intervals. J R Stat Soc Ser D Stat. 1995;44(2):143–54.

57. Rubin DB, Stern HS. Sample size determination using posterior predictive distributions. Sankhyā Indian J Stat Ser B. 1998;60:161–75.

58. Joseph L, Wolfson DB, Berger RD. Some comments on Bayesian sample size determination. J R Stat Soc Ser D (Stat). 1995;44(2):167–71.

59. Lindley DV. The choice of sample size. J R Stat Soc Ser D (Stat). 1997;46(2):129–38.

60. Wang F, Gelfand AE. A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. Stat Sci. 2002;17:193–208.

61. Psioda MA, Ibrahim JG. Bayesian design of a survival trial with a cured fraction using historical data. Stat Med. 2018;37(26):3814–31.

62. Chen MH, Ibrahim JG, Lam P, Yu A, Zhang Y. Bayesian design of non-inferiority trials for medical devices using historical data. Biometrics. 2011;67(3):1163–70.

63. Lehmann EL, Romano JP, Casella G. Testing statistical hypotheses, vol. 3. New York: Springer; 1986.

64. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. New York: Chapman and Hall/CRC; 1995.

65. Gamerman D, Lopes HF. Markov chain Monte Carlo: stochastic simulation for Bayesian inference. New York: CRC Press; 2006.

66. Andrieu C, De Freitas N, Doucet A, Jordan MI. An introduction to MCMC for machine learning. Mach Learn. 2003;50:5–43.

67. Lee SY. Gibbs sampler and coordinate ascent variational inference: A set-theoretical review. Commun Stat-Theory Methods. 2022;51(6):1549–68.

68. Neyman J, Pearson ESIX. On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc Lond Ser A Containing Pap Math Phys Character. 1933;231(694–706):289–337.

69. Ma YA, Chen Y, Jin C, Flammarion N, Jordan MI. Sampling can be faster than optimization. Proc Natl Acad Sci. 2019;116(42):20881–5.

70. Hoffman MD, Gelman A, et al. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J Mach Learn Res. 2014;15(1):1593–623.

71. Grieve AP. Idle thoughts of a 'well-calibrated' Bayesian in clinical drug development. Pharm Stat. 2016;15(2):96–108.

72. U.S. Food and Drug Administration. Design Considerations for Pivotal Clinical Investigations for Medical Devices. 2013. www.fda. gov/regulatory-information/search-fda-guidance-documents/ design-considerations-pivotal-clinical-investigations-medical-devic es. Accessed 23 Nov 2023.

73. Storer BE, Kim C. Exact properties of some exact test statistics for comparing two binomial proportions. J Am Stat Assoc. 1990;85(409):146–55.

74. Burger HU, Gerlinger C, Harbron C, Koch A, Posch M, Rochon J, et al. The use of external controls: To what extent can it currently be recommended? Pharm Stat. 2021;20(6):1002–16.

75. Best N, Ajimi M, Neuenschwander B, Saint-Hilary G, Wandel S. Beyond the classical type I error: Bayesian metrics for Bayesian designs using informative priors. Stat Biopharm Res. 2024;0:1–37.

76. Kopp-Schneider A, Calderazzo S, Wiesenfarth M. Power gains by using external information in clinical trials are typically not possible when requiring strict type I error control. Biom J. 2020;62(2):361–74.

77. Lesaffre E, Baio G, Boulanger B. Bayesian methods in pharmaceutical research. Boca Raton: CRC Press; 2020.

78. Hall W, Mathiason DJ. On large-sample estimation and testing in parametric models. Int Stat Rev/Rev Int Stat. 1990;77–97.

79. Kass RE, Wasserman L. The selection of prior distributions by formal rules. J Am Stat Assoc. 1996;91(435):1343–70.

80. Kerman J. Neutral noninformative and informative conjugate beta and gamma prior distributions. Electron J Stat. 2011;5(none):1450–1470.

81. Ibrahim JG, Chen MH, Gwon Y, Chen F. The power prior: theory and applications. Stat Med. 2015;34(28):3724–49.

82. Ibrahim JG, Chen MH, Sinha D. On optimality properties of the power prior. J Am Stat Assoc. 2003;98(461):204–13.

83. Thall PF, Wathen JK, Bekele BN, Champlin RE, Baker LH, Benjamin RS. Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. Stat Med. 2003;22(5):763–80.

84. Lee SY. Bayesian Nonlinear Models for Repeated Measurement Data: An Overview, Implementation, and Applications. Mathematics. 2022;10(6):898.

85. Irony TZ, Pennello GA. Choosing an appropriate prior for Bayesian medical device trials in the regulatory setting. Am Stat Assoc 2001 Proc Biopharm Sect. 2001;1000:85.

86. Griffiths D. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. Biometrics. 1973;29:637–48.

87. Fleming TR, Harrington DP, O'Brien PC. Designs for group sequential tests. Control Clin Trials. 1984;5(4):348–61.

88. Jennison C, Turnbull BW. Group sequential methods with applications to clinical trials. New York: CRC Press; 1999.

89. Böhm M, Fahy M, Hickey GL, Pocock S, Brar S, DeBruin V, et al. A re-examination of the SPYRAL HTN-OFF MED Pivotal trial with respect to the underlying model assumptions. Contemp Clin Trials Commun. 2021;23:100818.

90. Johnstone IM. High dimensional Bernstein-von Mises: simple examples. Inst Math Stat Collect. 2010;6:87.

91. Walker AM. On the asymptotic behaviour of posterior distributions. J R Stat Soc Ser B Stat Methodol. 1969;31(1):80–8.

92. Ghosh JK, Delampady M, Samanta T. An introduction to Bayesian analysis: theory and methods, vol. 725. New York: Springer; 2006.

93. Fisher RA. Design of experiments. Br Med J. 1936;1(3923):554.

94. Cox DR. Statistical significance. Ann Rev Stat Appl. 2020;7:1–10.

95. Faris O, Shuren J. An FDA viewpoint on unique considerations for medical-device clinical trials. N Engl J Med. 2017;376(14):1350–7.

96. Mayer C, Perevozskaya I, Leonov S, Dragalin V, Pritchett Y, Bedding A, et al. Simulation practices for adaptive trial designs in drug and device development. Stat Biopharm Res. 2019;11(4):325–35.

97. Wassmer G, Brannath W. Group sequential and confirmatory adaptive designs in clinical trials, vol. 301. London: Springer; 2016.

98. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics. 1979;50:549–56.

99. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika. 1977;64(2):191–9.

100. Demets DL, Lan KG. Interim analysis: the alpha spending function approach. Stat Med. 1994;13(13–14):1341–52.

101. Armitage P, McPherson C, Rowe B. Repeated significance tests on accumulating data. J R Stat Soc Ser A (Gen). 1969;132(2):235–44.

102. Pocock SJ. Interim analyses for randomized clinical trials: the group sequential approach. Biometrics. 1982;38:153–62.

103. Lachin JM. A review of methods for futility stopping based on conditional power. Stat Med. 2005;24(18):2747–64.

104. Gordon Lan K, Simon R, Halperin M. Stochastically curtailed tests in long-term clinical trials. Seq Anal. 1982;1(3):207–19.

105. Freidlin B, Korn EL. A comment on futility monitoring. Control Clin Trials. 2002;23(4):355–66.

106. Saville BR, Connor JT, Ayers GD, Alvarez J. The utility of Bayesian predictive probabilities for interim monitoring of clinical trials. Clin Trials. 2014;11(4):485–93.

107. Snapinn S, Chen MG, Jiang Q, Koutsoukos T. Assessment of futility in clinical trials. Pharm Stat J Appl Stat Pharm Ind. 2006;5(4):273–81.

108. DeMets DL, Ellenberg SS. Data monitoring committees-expect the unexpected. N Engl J Med. 2016;375(14):1365–71.

109. Lee JJ, Liu DD. A predictive probability design for phase II cancer clinical trials. Clin Trials. 2008;5(2):93–106.

110. Herson J. Predictive probability early termination plans for phase II clinical trials. Biometrics. 1979;24:775–83.

111. Broglio KR, Connor JT, Berry SM. Not too big, not too small: a goldilocks approach to sample size selection. J Biopharm Stat. 2014;24(3):685–705.

112. Wang Y, Fu H, Kulkarni P, Kaiser C. Evaluating and utilizing probability of study success in clinical development. Clin Trials. 2013;10(3):407–13.

113. Chuang-Stein C. Sample size and the probability of a successful trial. Pharm Stat J Appl Stat Pharm Ind. 2006;5(4):305–9.

114. Dmitrienko A, Wang MD. Bayesian predictive approach to interim monitoring in clinical trials. Stat Med. 2006;25(13):2178–95.

115. Della Bella P, Baratto F, Vergara P, Bertocchi P, Santamaria M, Notarstefano P, et al. Does timing of ventricular tachycardia ablation affect prognosis in patients with an implantable cardioverter defibrillator?

Results from the multicenter randomized PARTITA trial. Circulation. 2022;145(25):1829–38.

116. Lan KG, Lachin JM, Bautista O. Over-ruling a group sequential boundary-a stopping rule versus a guideline. Stat Med. 2003;22(21):3347–55.

117. Zimetbaum P. Antiarrhythmic drug therapy for atrial fibrillation. Circulation. 2012;125(2):381–9.

118. Barber S, Jennison C. Symmetric tests and confidence intervals for survival probabilities and quantiles of censored survival data. Biometrics. 1999;55(2):430–6.

119. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(282):457–81.

120. Greenwood M. A Report on the Natural Duration of Cancer. Reports on Public Health and Medical Subjects. Ministry Health. 1926;33:iv26

121. Li X, Herrmann C, Rauch G. Optimality criteria for futility stopping boundaries for group sequential designs with a continuous endpoint. BMC Med Res Methodol. 2020;20:1–8.

122. Rossignol P, Hernandez AF, Solomon SD, Zannad F. Heart failure drug treatment. Lancet. 2019;393(10175):1034–44.

123. Spertus JA, Jones PG, Sandhu AT, Arnold SV. Interpreting the Kansas City Cardiomyopathy Questionnaire in clinical trials and clinical care: JACC state-of-the-art review. J Am Coll Cardiol. 2020;76(20):2379–90.

124. O'Brien PC. Procedures for comparing samples with multiple endpoints. Biometrics. 1984;40:1079–87.

125. Dmitrienko A, Bretz F, Westfall PH, Troendle J, Wiens BL, Tamhane AC, et al. Multiple testing methodology. In: Multiple testing problems in pharmaceutical statistics. New York: Chapman and Hall/CRC; 2009. pp. 53–116.

126. Dmitrienko A, D'Agostino RB Sr. Multiplicity considerations in clinical trials. N Engl J Med. 2018;378(22):2115–22.

127. Dmitrienko A, D'Agostino RB Sr, Huque MF. Key multiplicity issues in clinical drug development. Stat Med. 2013;32(7):1079–111.

128. U.S. Food and Drug Administration. Multiple endpoints in clinical trials guidance for industry. 2022. www.fda.gov/regulatory-information/search-fda-guidance-documents/multiple-endpoints-clinical-trials-guidance-industry. Accessed 23 Nov 2023.

129. Bretz F, Hothorn T, Westfall P. Multiple comparisons using R. Boca Raton: CRC Press; 2016.

130. Vickerstaff V, Omar RZ, Ambler G. Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomised controlled trials with multiple primary outcomes. BMC Med Res Methodol. 2019;19(1):1–13.

131. Hochberg Y. Multiple comparison procedures. New York: Wiley Series in Probability and Statistics; 1987.

132. Senn S, Bretz F. Power and sample size when multiple endpoints are considered. Pharm Stat J Appl Stat Pharm Ind. 2007;6(3):161–70.

133. Proschan MA, Waclawiw MA. Practical guidelines for multiplicity adjustment in clinical trials. Control Clin Trials. 2000;21(6):527–39.

134. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988;75(4):800–2.

135. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6:65–70.

136. Hommel G, Bretz F, Maurer W. Multiple hypotheses testing based on ordered p values-a historical survey with applications to medical research. J Biopharm Stat. 2011;21(4):595–609.

137. Kim K, Bretz F, Cheung YKK, Hampson LV. Handbook of statistical methods for randomized controlled trials. Boca Raton: CRC Press; 2021.

138. Lewis C, Thayer DT. Bayesian decision theory for multiple comparisons. Lect Notes-Monogr Ser. 2009;326–32.

139. Gelman A, Hill J, Yajima M. Why we (usually) don't have to worry about multiple comparisons. J Res Educ Eff. 2012;5(2):189–211.

140. Berry DA, Hochberg Y. Bayesian perspectives on multiple comparisons. J Stat Plann Infer. 1999;82(1–2):215–27.

141. Gopalan R, Berry DA. Bayesian multiple comparisons using Dirichlet process priors. J Am Stat Assoc. 1998;93(443):1130–9.

142. Zucker D, Schmid C, McIntosh M, D'agostino R, Selker H, Lau J. Combining single patient (N-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment. J Clin Epidemiol. 1997;50(4):401–10.

143. Takeda K, Liu S, Rong A. Constrained hierarchical Bayesian model for latent subgroups in basket trials with two classifiers. Stat Med. 2022;41(2):298–309.

144. Stunnenberg BC, Raaphorst J, Groenewoud HM, Statland JM, Griggs RC, Woertman W, et al. Effect of mexiletine on muscle stiffness in patients with nondystrophic myotonia evaluated using aggregated N-of-1 trials. Jama. 2018;320(22):2344–53.

145. Liu Y, Kane M, Esserman D, Blaha O, Zelterman D, Wei W. Bayesian local exchangeability design for phase II basket trials. Stat Med. 2022;41(22):4367–84.

146. Berry SM, Broglio KR, Groshen S, Berry DA. Bayesian hierarchical modeling of patient subpopulations: efficient designs of phase II oncology clinical trials. Clin Trials. 2013;10(5):720–34.

147. McGlothlin AE, Viele K. Bayesian hierarchical models. Jama. 2018;320(22):2365–6.

148. Lee SY. The Use of a Log-Normal Prior for the Student t-Distribution. Axioms. 2022;11(9):462.

149. Neuenschwander B, Wandel S, Roychoudhury S, Bailey S. Robust exchangeability designs for early phase clinical trials with multiple strata. Pharm Stat. 2016;15(2):123–34.

150. Efron B. The future of indirect evidence. Stat Sci Rev J Inst Math Stat. 2010;25(2):145.

151. Jones HE, Ohlssen DI, Neuenschwander B, Racine A, Branson M. Bayesian models for subgroup analysis in clinical trials. Clin Trials. 2011;8(2):129–43.

152. James W, Stein C. Estimation with quadratic loss. In: Breakthroughs in statistics: Foundations and basic theory. New York: Springer; 1992. pp. 443–60.

153. Agency EM. Guideline on Clinical Trials in Small Population. 2006. https://www.ema.europa.eu/en/clinical-trials-small-populations-scientific-guideline. Accessed 23 Nov 2023.

154. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika. 1934;26(4):404–13.

155. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. Biometrika. 1986;73(3):751–4.

156. Guo M, Heitjan DF. Multiplicity-calibrated Bayesian hypothesis tests. Biostatistics. 2010;11(3):473–83.

157. Thompson L, Chu J, Xu J, Li X, Nair R, Tiwari R. Dynamic borrowing from a single prior data source using the conditional power prior. J Biopharm Stat. 2021;31(4):403–24.

158. Gökbuget N, Dombret H, Ribera JM, Fielding AK, Advani A, Bassan R, et al. International reference analysis of outcomes in adults with B-precursor Ph-negative relapsed/refractory acute lymphoblastic leukemia. Haematologica. 2016;101(12):1524.

159. Gökbuget N, Kelsh M, Chia V, Advani A, Bassan R, Dombret H, et al. Blinatumomab vs historical standard therapy of adult relapsed/refractory acute lymphoblastic leukemia. Blood Cancer J. 2016;6(9):e473–e473.

160. Goring S, Taylor A, Müller K, Li TJJ, Korol EE, Levy AR, et al. Characteristics of non-randomised studies using comparisons with external controls submitted for regulatory approval in the USA and Europe: a systematic review. BMJ Open. 2019;9(2):e024895.

161. Campbell G. Bayesian statistics in medical devices: innovation sparked by the FDA. J Biopharm Stat. 2011;21(5):871–87.

162. Psioda MA, Ibrahim JG. Bayesian clinical trial design using historical data that inform the treatment effect. Biostatistics. 2019;20(3):400–15.

163. Duan Y, Ye K, Smith EP. Evaluating water quality using power priors to incorporate historical information. Environmetrics Off J Int Environmetrics Soc. 2006;17(1):95–106.

164. Pawel S, Aust F, Held L, Wagenmakers EJ. Normalized power priors always discount historical data. Stat. 2023;12(1):e591.

165. Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B. Robust meta-analytic-predictive priors in clinical trials with historical control information. Biometrics. 2014;70(4):1023–32.

166. Neuenschwander B, Branson M, Spiegelhalter DJ. A note on the power prior. Stat Med. 2009;28(28):3562–6.

167. Ye K, Han Z, Duan Y, Bai T. Normalized power prior Bayesian analysis. J Stat Plan Infer. 2022;216:29–50.

168. Galwey N. Supplementation of a clinical trial by historical control data: is the prospect of dynamic borrowing an illusion? Stat Med. 2017;36(6):899–916.

169. Nikolakopoulos S, van der Tweel I, Roes KC. Dynamic borrowing through empirical power priors that control type I error. Biometrics. 2018;74(3):874–80.

170. Edwards D, Best N, Crawford J, Zi L, Shelton C, Fowler A. Using Bayesian Dynamic Borrowing to Maximize the Use of Existing Data: A Case-Study. Ther Innov Regul Sci. 2023;58:1–10.
171. Baeten D, Baraliakos X, Braun J, Sieper J, Emery P, Van der Heijde D, et al. Anti-interleukin-17A monoclonal antibody secukinumab in treatment of ankylosing spondylitis: a randomised, double-blind, placebo-controlled trial. Lancet. 2013;382(9906):1705–13.
172. Saville BR, Berry SM. Efficiencies of platform clinical trials: a vision of the future. Clin Trials. 2016;13(3):358–66.
173. Adaptive platform trials. definition, design, conduct and reporting considerations. Nat Rev Drug Discov. 2019;18(10):797–807.
174. Bretz F, Koenig F, Brannath W, Glimm E, Posch M. Adaptive designs for confirmatory clinical trials. Stat Med. 2009;28(8):1181–217.
175. Brannath W, Koenig F, Bauer P. Multiplicity and flexibility in clinical trials. Pharm Stat J Appl Stat Pharm Ind. 2007;6(3):205–16.
176. Hackshaw A. Small studies: strengths and limitations. Eur Respiratory J. 2008;32(5):1141–3.
177. Stan Development Team. RStan: the R interface to Stan. 2024. R package version 2.32.5. https://mc-stan.org/. Accessed 23 Nov 2023.
178. Plummer M, et al. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: Proceedings of the 3rd international workshop on distributed statistical computing. vol. 124. Vienna; 2003. pp. 1–10.
179. R Core Team. R: A Language and Environment for Statistical Computing. Vienna; 2023. https://www.R-project.org/.
180. SAS Institute. SAS/IML software: usage and reference, version 6. New York: Sas Inst; 1990.

Publisher's Note