Pharmaceutical Statistics

WILEY

MAIN PAPER

Drift Parameter Based Sample Size Determination in Multi-Stage Bayesian Randomized Clinical Trials

Yueyang Han[1] | Haolun Shi[1] | Jiguo Cao[1] | Ruitao Lin[2]

[1]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada | [2]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA

Correspondence: Haolun Shi (haolun_shi@sfu.ca)

Received: 25 November 2024 | **Revised:** 23 May 2025 | **Accepted:** 20 July 2025

Funding: This work was supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN/02963-2021). The research was also supported by the Collaborative Research Team sponsored by CANSSI.

Keywords: Bayesian design | multi-stage design | phase II clinical trial | posterior probability | randomized trial | sample size

ABSTRACT

Sample size determination in Bayesian randomized phase II trial design often relies on computationally intensive search methods, presenting challenges in terms of feasibility and efficiency. We propose a novel approach that greatly reduces the computing time of sample size calculations for Bayesian trial designs. Our approach innovatively connects group sequential design with Bayesian trial design and leverages the proportional relationship between sample size and the squared drift parameter. This results in a faster algorithm. By employing regression analysis, our method can accurately pinpoint the required sample size with significantly reduced computational burden. Through theoretical justification and extensive numerical evaluations, we validate our approach and illustrate its efficiency across a wide range of common trial scenarios, including binary endpoint with Beta-Binomial model, normal endpoint, binary/ordinal endpoint under Bayesian generalized linear model, and survival endpoints under Bayesian piecewise exponential models. To facilitate the use of our methods, we create an R package named "BayesSize" on GitHub.

1 | Introduction

Phase II clinical trials are pivotal in the drug development process, serving as the bridge between initial safety assessments in Phase I and the large-scale efficacy evaluations of Phase III. Traditional Phase II trials often employ a frequentist statistical approach, whereas more recently, the Bayesian approach offers a more flexible alternative for designing Phase II trials, particularly with more diverse and complex endpoints, prior information, and adaptive decision rules. Under the Bayesian framework, a prior probability distribution is specified for the unknown parameter to represent our ignorance or uncertainty about its value; see Berry [1, 2], Berger and Berry [3], and Efron [4]. The Bayesian approach enables continuous parameter updating and learning by incorporating new data into the posterior distribution, thereby enhancing monitoring and decision-making in adaptive clinical trials.

With the advent of the Bayesian paradigm, a number of adaptive designs have been proposed to model a variety of endpoints in Phase II clinical trials. Murray et al. [5, 6] develop a Bayesian utility-based sequential trial design with multinomial endpoints and propose methods for controlling type I and II error rates. Yu et al. [7] propose a Bayesian sequential design with adaptive randomization for two-sided hypothesis tests. Zhang et al. [8] present a Bayesian Phase II proof-of-concept design for clinical trials that incorporates longitudinal endpoints, offering a novel approach to evaluate treatment effects over time. Zhou et al. [9] propose a Bayesian optimal Phase II clinical trial design that employs an exponential-inverse gamma model to specifically

model the time-to-event endpoint, and Zhao et al. [10] extend the design for two-arm randomized trials that is adaptable to handle single, multiple primary, and coprimary endpoints across both superiority and noninferiority trials. Guo and Liu [11] introduce a flexible Bayesian optimal Phase II predictive probability design that simplifies the evaluation of both binary and complex endpoints in clinical trials, using a Dirichlet-multinomial model to calculate interim predictive probabilities of success. Unlike the frequentist design, where sample sizes can be calculated via a formula-based method, the sample size determination of these Bayesian designs remains a major challenge for practitioners of the Bayesian paradigm. Most of the existing Bayesian designs use numerical simulation to approximate the power and search for the optimal sample size by repeatedly conducting simulations under varying numbers of patients. Such a process is lengthy and time-consuming. For Bayesian sequential designs, there lacks an efficient alternative to the brute-force numerical search approach for controlling the power. While controlling the frequentist operating characteristics for the Bayesian designs has been highly advocated, for example, a recently finalized guideline on complex trial designs by the US Food and Drug Administration (FDA) that particularly highlights the importance of type I error and power control for complex Bayesian designs [12], methods for controlling the power of Bayesian sequential design are rarely elaborated in the literature.

We propose a new approach based on a proportional relationship of sample size with the squared drift parameter [13, 14]. Such a proportional relationship leads to an accurate initial guess of the sample size that locates the neighborhood of the true solution, after which a line search can be employed to accurately pinpoint the required sample size. Compared with the numerical search method, our proposed method leads to significant time savings. We provide theoretical justification for our method and conduct extensive numerical studies on its performance, including binary endpoint with Beta-Binomial model, normal endpoint, binary/ordinal endpoint under Bayesian generalized linear model, and survival endpoints under Bayesian piecewise exponential models. Our method serves as a complement to the theoretical posterior probability boundaries for controlling the type I error rate [15]. These theoretical posterior probability boundaries do not depend on the sample size, which avoids iteratively calculating the type I and II error rates, and thus greatly simplifies the process of design calibration. Taken together, we develop an effective framework for controlling the operating characteristics of Bayesian multi-stage design that uniformly works for various types of endpoints, as well as design schemes (single-arm/double-arm).

The rest of this article is organized as follows. Section 2 introduces some previous work on Bayesian sample size determination methods proposed by other researchers. In Section 3, we develop the methodology by introducing the concept of the drift parameter from the group sequential design and deriving a theoretical method that utilizes the proportional relationship between the sample size and the squared drift parameter. In Section 4, we conduct several case studies using the proposed method, including the applications to binary endpoint with beta-binomial model, normal endpoint, binary endpoint with logistic model, ordinal endpoint, and survival endpoint with piecewise exponential model. Finally, Section 5 concludes the article with some remarks.

2 | Literature Review

Sample size determination under the Bayesian paradigm has been a subject of extensive research. Santis [16] introduces a Bayesian approach to sample size determination, focusing on robustness by analyzing the range and bounds of posterior quantities under varying priors. M'lan et al. [17] consider and compare several interval-based methods for Bayesian sample size determination for binomial proportion. Whitehead et al. [18] introduce a straightforward Bayesian method for determining clinical trial sample sizes, which parallels frequentist approaches under certain conditions, and applies to binary and normal outcomes. Zaslavsky [19] compares Bayesian power calculations with frequentist approaches for binary and Poisson endpoints, and evaluates error rates through credible and confidence limits. Psioda and Ibrahim [20] present a simulation-based Bayesian approach to sample size determination in clinical trials, incorporating historical data to calibrate prior informativeness and ensure adequate power and type I error control. Zheng et al. [21] formulate Bayesian sample size calculations for experiments comparing two groups, incorporating pre-experimental information to support design and analysis. Golchi [22] proposes an efficient estimation method for design operating characteristics of Bayesian trials and applies the method to sample size determination for ordinal outcomes. Most of the proposals for Bayesian sample size determination are based on credible interval widths or decision theory, and do not consider frequentist operating characteristics, which may not meet the stringent requirements of regulatory bodies. A few existing bodies of work that evaluate frequentist operating characteristics [21, 22] have yet to fully explore the complexities of sequential testing scenarios, where hypothesis tests are conducted more than once. To address this gap, we propose a sample size determination method for Bayesian sequential trials that also considers controlling the frequentist operating characteristics, offering a more practical and comprehensive approach that aligns with regulatory standards and the needs of clinical trials.

3 | Methodology

3.1 | Preliminaries

Let θ_E and θ_S denote the effect parameters for the experimental arm and standard arm, respectively. We are interested in testing the one-sided hypotheses

$$H_0:\ \theta_E \leqslant \theta_S \quad \text{versus} \quad H_A:\ \theta_E > \theta_S$$

Under the frequentist paradigm, a type I error occurs when we reject H_0, given H_0 being true; a type II error occurs when we fail to reject H_0, given H_A being true. We define the type I error rate and the type II error rate as.

$$\text{Type I error rate} = \Pr\left(\text{Reject } H_0 \mid H_0 \text{ is true}\right)$$

$$\text{Type II error rate} = \Pr\left(\text{Do not reject } H_0 \mid H_A \text{ is true}\right)$$

As the trial accrues samples, based on the posterior distributions, a series of Bayesian sequential tests on the superiority of the experimental treatment can be constructed. Specifically, we repeatedly conduct Bayesian hypothesis tests after enrolling every m subjects to each arm; and a total of Km subjects will be enrolled over K stages. At the kth stage, we would stop the trial and declare the superiority of the experimental treatment if

$$\Pr\left(\theta_E > \theta_S \mid \text{Data}\right) \geqslant c_k, \qquad k = 1, \ldots, K$$

where $\Pr\left(\theta_E > \theta_S \mid \text{Data}\right)$ can be computed based on the posterior distribution and c_k is a cutoff of the posterior probability at stage k. Otherwise, we continue to enroll the next group of m subjects to each arm and conduct another analysis at the end of stage $k+1$, or if $k = K$, that is, we have reached the end of the trial, we declare the futility of the experimental treatment. In the same vein, a series of decision rules for futility stopping can be similarly constructed.

To control the type I error rate, using the methods proposed by Shi and Yin [15], the values of c_k can be calibrated to be

$$c_k = \Phi\left(a_k\right)$$

where a_k denotes the stopping boundaries for the frequentist group sequential design under the same type I error rate constraint and α-spending function, and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

Recall that m is the number of subjects randomized to each arm in each stage. To quantify the design power, we define a rejection probability function for the kth analysis as $R_k\left(\theta_E, \theta_S, m\right)$, which depicts the probability that treatment superiority is declared at the kth analysis, given the value of θ_E and θ_S. Because the conclusion on superiority is declared at either of the K analyses, we may further define the rejection probability function for the whole sequential trial as $R\left(\theta_E, \theta_S, m\right) = \sum_{k=1}^{K} R_k\left(\theta_E, \theta_S, m\right)$.

In the calibration of the design parameters, we first specify a null scenario where H_0 is true and an alternative scenario where H_A is true. Under the null scenario, for example, $\theta_E = \theta_S = \theta_{\text{null}}$, the probability of rejecting H_0, which equals $R\left(\theta_E = \theta_{\text{null}}, \theta_S = \theta_{\text{null}}, m\right)$, is simply the type I error rate. We can calibrate the value of c_k such that $R\left(\theta_E = \theta_{\text{null}}, \theta_S = \theta_{\text{null}}, m\right) \leqslant \alpha$, using either massive simulation, or theoretical approaches proposed in Shi and Yin [15].

After the values of c_k are determined, to control the power of the design, we specify the alternative scenario where H_A is true, for example, $\theta_E = \theta_{\text{null}} + \delta$ and $\theta_S = \theta_{\text{null}}$, where δ is a specified treatment difference. Under such an alternative scenario, the probability of rejecting H_0, which equals $R\left(\theta_E = \theta_{\text{null}} + \delta, \theta_S = \theta_{\text{null}}, m\right)$, is the power of the design. For notational brevity, we abbreviate $R\left(\theta_E = \theta_{\text{null}} + \delta, \theta_S = \theta_{\text{null}}, m\right)$ as $R\left(\theta_E, \theta_S, m\right)$ in the remainder of this paper. Typically, the larger the sample size, the more powerful the design; and thus the $R\left(\theta_E, \theta_S, m\right)$ is monotonically increasing on m.

The conventional method of calibrating m is based on either a line search or a bisectional search, which requires repeated computation of $R\left(\theta_E, \theta_S, m\right)$ based on varying values of m. The line search enumerates all the possible values in incremental steps until the sample size m leads to a power value greater than $1-\beta$. The bisectional search uses the monotonic relationship between m and $R\left(\theta_E, \theta_S, m\right)$, repeatedly dividing the search interval into halves and retaining the half where the solution lies.

When the exact computation of $R_k\left(\theta_E, \theta_S, m\right)$ is either impossible or too complex to compute, an approximation for $R_k\left(\theta_E, \theta_S, m\right)$ can be computed by simulating a large number of trials. Let L denote the total number of trial simulations. The approximated rejection probability function for the kth analysis is defined as

$$\hat{R}_k\left(\theta_E, \theta_S, m\right) = \frac{1}{L}\sum_{\ell=1}^{L}\left\{ I\left(P\left(H_A \mid D_k^{(\ell)}\right) > c_k\right)\prod_{j=1}^{k-1} I\left(P(H_A \mid D_j^{(\ell)}) \leqslant c_j\right)\right\}$$

where $D_k^{(\ell)}$ denotes the accumulated data up to stage k in the ℓ th simulation replication. Similarly, the approximated rejection probability function for the whole sequential design is defined as $\hat{R}\left(\theta_E, \theta_S, m\right) = \sum_{k=1}^{K} \hat{R}_k\left(\theta_E, \theta_S, m\right)$.

Depending upon the complexity of the Bayesian model, the computation of $\hat{R}\left(\theta_E, \theta_S, m\right)$ could be time-consuming. This is especially true for Bayesian models that use the Markov chain Monte Carlo method to compute the posterior probability $P\left(H_A \mid D_k^{(\ell)}\right)$, where a total of LK rounds of computations need to be performed. As an example, for a sequential design with $K = 4$ analyses, if it takes around 10 s to run 1 round of MCMC to compute a single value of $\hat{R}_k\left(\theta_E, \theta_S, m\right)$, it would take more than 50 h with $L = 5000$ simulation replications. Furthermore, a line search or a bisectional search of the sample parameter m that meets the target equation $\hat{R}\left(\theta_E, \theta_S, m\right) = 1-\beta$ usually requires repeated computation of $\hat{R}\left(\theta_E, \theta_S, m\right)$ under varying values of m, leading to a heavy computational burden.

3.2 | Sample Size Calibration: Drift Parameter-Based Approach

3.2.1 | Proposed Algorithm

We propose a fast algorithm for calibrating the value of m that satisfies the power constraint. The algorithm is suitable for Bayesian sequential design with various types of endpoints. Recall that θ_E and θ_S denote the effect parameters for the experimental arm and standard arm, respectively; the rejection probability function for the kth analysis $R_k\left(\theta_E, \theta_S, m\right)$ is the probability of rejecting H_0 given the effect parameters θ_E and θ_S.

A regression-based algorithm for calibrating sample size is as follows.

Step 1. Conduct simulations under the alternative hypothesis to compute values of $\hat{R}\left(\theta_E, \theta_S, m\right)$, for at least 3 candidate sample size parameter m. To speed up calculation, the number of replications in each simulation only needs to be 1 / 5 or 1 / 10 of the usual value of L required for approximation precision.

Step 2. Denote the drift parameter of a group sequential design as $\xi(\alpha, \beta, K)$, which can be viewed as a function of α, β, K. For sample size m, compute the squared drift parameter as

$\xi_m^2 = \xi\left(\alpha, 1-\widehat{R}(\theta_E, \theta_S, m), K\right)^2$. Fit a regression line that passes through the origin and all the points (ξ_m^2, m). Denote the resultant coefficient as γ.

Step 3. The proposal sample size is $\widetilde{m} = \gamma\xi(\alpha, \beta, K)^2$. A refined line search can be conducted around $\widetilde{m}$ until the solution m leads to $\widehat{R}(\theta_E, \theta_S, m)$ being greater than and closest to $1-\beta$.

A major advantage of utilizing the squared drift parameter to calibrate the sample size is that the value of the drift parameter is not contingent on the model, but only depends upon the type I and II error rates α and β and the number of analyses K. Therefore, the proposed method is generalizable to various types of endpoints and models.

To enable access to our method, we develop an R package entitled "BayesSize" on GitHub that implements our proposed methods, which is available at https://github.com/haoluns/Bayessize. After specifying the type I and II error rates α and β and the number of analyses K, the software only requires the user to input a few pair values of sample size and simulated power, and will output the proposal sample size $\widetilde{m}$.

In the following subsections, we introduce ideas and explanation of the drift parameter $\xi(\alpha, \beta, K)$, presents methods for its computation, and establish a linear relationship between the maximum sample size and the squared drift parameter for Bayesian sequential designs.

3.2.2 | Drift Parameter

Let $\theta = \theta_E - \theta_S$ denote the parameter of interest that corresponds to the treatment difference. Under the frequentist group sequential design, the usual Wald test statistic at stage k is

$$Z_k = \frac{\widehat{\theta}_k}{\left\{\mathrm{Var}\left(\widehat{\theta}_k\right)\right\}^{1/2}}$$

where $\widehat{\theta}_k$ is the maximum likelihood estimator of θ based on the accumulated data up to the kth analysis.

As the trial proceeds, the information at the kth analysis can be characterized by the Fisher information,

$$\mathcal{I}_k = \left\{\mathrm{Var}\left(\widehat{\theta}_k\right)\right\}^{-1}$$

We may define an internal time $t_k = \mathcal{I}_k / \mathcal{I}_K$ to characterize the progress of information accrual in the trial. The Wald test statistic at stage k can be rewritten as

$$Z(t_k) = \widehat{\theta}_k\sqrt{\mathcal{I}_k}$$

At the final analysis, the maximum information is $\mathcal{I}_K$, which increases from 0 to 1 as the trial proceeds. At time $t_K = 1$, the trial achieves the maximum information $\mathcal{I}_K$. Let N_k denote the cumulative sample size up to stage k. When the variance of the estimator $\mathrm{Var}\left(\widehat{\theta}_k\right)$ is inversely proportional to the sample size (e.g., for binary and normal endpoints), we have

$$t_k = N_k / N_K = \mathcal{I}_k / \mathcal{I}_K$$

Across all the K analyses, the test statistics $Z(t_1), \ldots, Z(t_K)$ jointly follow a multivariate normal distribution, commonly referred to as the canonical distribution [23]. The marginal distribution of $Z(t_k)$ is

$$Z(t_k) \sim N\left(\theta\sqrt{\mathcal{I}_k}, 1\right)$$

Lan and Demets [13] proposed converting the problem of solving for the boundaries of a frequentist group sequential test into a discrete boundary-crossing problem for a Brownian motion process. We define $W(t_k) = \sqrt{t_k}Z(t_k)$. Since $Z(t_k) \sim N\left(\theta\sqrt{\mathcal{I}_k}, 1\right)$, we have $\sqrt{t_k}Z(t_k) \sim N\left(\theta\sqrt{\mathcal{I}_k t_k}, t_k\right)$. By substituting the expression of $\mathcal{I}_K = \frac{\mathcal{I}_k}{t_k}$ into the distribution of $W(t_k) = \sqrt{t_k}Z(t_k)$, we show that

$$W(t_k) \sim N\left(\theta\sqrt{\mathcal{I}_K}t_k, t_k\right)$$

We define the drift parameter as $\xi = \theta\sqrt{\mathcal{I}_K}$, and thus the $W(t_k)$ follows a Brownian motion with drift ξ,

$$W(t_k) \sim N(\xi t_k, t_k)$$

Based on its definition, $\mathcal{I}_K = \left(\frac{\xi}{\theta}\right)^2$, that is, $\mathcal{I}_K \propto \xi^2$. Recall that we have shown that $\frac{N_k}{N_K} = \frac{\mathcal{I}_k}{\mathcal{I}_K}$, thus for cases where the maximum information $\mathcal{I}_K$ is proportional to the total sample size N_K, as $\xi = \theta\sqrt{\mathcal{I}_K}$, Kim and Demets [14] established the relationship

$$N_K \propto \xi^2$$

that is, the sample size as well as the total information is proportional to the square of the drift parameter.

3.2.3 | Computing Drift Parameter: Recursive Integration

For a group sequential design, given the values of α, β, K, and the α-spending function, the corresponding value of ξ can be calculated through recursive integration [14, 24]. Recall that $W(t_k) = \sqrt{t_k}Z(t_k)$ follows the Brownian motion, that is, $W(t_k) \sim N(\xi t_k, t_k)$, we introduce $Y(t_k) = W(t_k) / \sqrt{t_k}$. The drift parameter ξ must satisfy the following two conditions:

$$\Pr\left(Y(t_1) \leqslant \zeta_1, Y(t_2) \leqslant \zeta_2, \cdots, Y(t_{k-1}) \leqslant \zeta_{k-1}, Y(t_k) > \zeta_k \mid H_0\right) = \alpha(t_k) - \alpha(t_{k-1})$$
$$\text{for each in } \{1, \cdots, K\}$$

and

$$\Pr\left(Y(t_1) \leqslant \zeta_1, Y(t_2) \leqslant \zeta_2, \cdots, Y(t_K) \leqslant \zeta_K \mid H_A\right) = \beta$$

where $\alpha(t)$ is the specified α-spending function and thus $\alpha(t_k)$ is the probability that a type I error is made before or at t_k. The $\zeta_1, \ldots, \zeta_K$ are a series of boundary value for declaring treatment superiority. The first condition requires that the spending of the type I error rate at stage k is $\alpha(t_k) - \alpha(t_{k-1})$, and the second condition constrains the design power.

To compute the drift parameter, we use the following procedure via recursive integration. Essentially, this yields the drift parameter as a function of α, β, K, and the α-spending function. The detailed steps are as follows.

Step 1: Denote $\alpha_k = \alpha(t_k)$, where $k = 1, \ldots, K$. We first determine a series of boundary vectors ζ_k to ensure that the trial satisfies the type I error rate requirement. For $k = 1$, $\zeta_1 = \Phi^{-1}(1-\alpha_1)$, where $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution. For each $k = 2, \cdots, K$, the value of ζ_k are sequentially solved to be the value of x that satisfies

$$\int_{-\infty}^{\zeta_1} \cdots \int_{-\infty}^{\zeta_{k-1}} \int_{x}^{\infty} f(y_1, \cdots, y_k) dy_1 \cdots dy_k = \alpha_k - \alpha_{k-1}$$

where $(Y_1, \cdots, Y_k)$ jointly follows a multivariate normal distribution with a zero mean vector and a covariance matrix $\Gamma_{i,j} = \frac{t_i + t_j - |t_i - t_j|}{2\sqrt{t_i t_j}}$, $i, j \in (1, \ldots, k)$.

Step 2: Given boundary vector ζ_k, where $k = 1, \ldots, K$, the drift parameter is such that the $(Y_1, \cdots, Y_K)$, which jointly follows a multivariate normal distribution with mean vector $\left(\xi\sqrt{1/K}, \xi\sqrt{2/K}, \cdots, \xi\sqrt{K/K}\right)$ and a covariance matrix $\Gamma_{i,j} = \frac{t_i + t_j - |t_i - t_j|}{2\sqrt{t_i t_j}}$, $i, j \in (1, \ldots, K)$, satisfy the following type II error rate constraint.

$$\int_{-\infty}^{\zeta_1} \cdots \int_{-\infty}^{\zeta_K} f(y_1, \cdots, y_K) dy_1 \cdots dy_K = \beta$$

3.2.4 | Asymptotic Connection

So far, the linear relationship between the squared drift parameter and the sample size has been established under the frequentist test statistics. Due to the asymptotic connection between the posterior probability and the frequentist test statistic, such a proportional relationship between N_K and ξ^2 also holds true for Bayesian sequential designs. Dudley and Haughton [25] studied the asymptotical normality of the posterior probability of half-spaces, establishing a connection between the Bayesian method based on posterior probabilities and the frequentist approach. A half-space H is a set satisfying a linear inequality,

$$\mathcal{H} = \{\theta: \mathbf{a}^\top \theta > b\}$$

where $\boldsymbol{\theta}$ is the unknown vector of interest, $\mathbf{a}$ is a specified vector and b is a specified scalar. Let ∂H denote the boundary hyperplane of H,

$$\partial H = \{\theta: \mathbf{a}^\top \theta = b\}$$

Let D_k denote the accumulated data up to the kth analysis. The likelihood ratio statistic for the null hypothesis $H_0: \boldsymbol{\theta} \in \partial H$ is

$$\Delta_k = 2\log L\left(\hat{\theta} | D_k\right) - 2\log L\left(\tilde{\theta} | D_k\right)$$

where $\log L(\boldsymbol{\theta})$ is the log-likelihood function of $\boldsymbol{\theta}$, and $\hat{\theta}$ and $\tilde{\theta}$ are the unconstrained maximum likelihood estimator for $\boldsymbol{\theta}$ and the constrained maximum likelihood estimator for $\boldsymbol{\theta} \in \partial H$, respectively. Let S_k denote the signed root likelihood ratio statistic, that is, if $\hat{\theta} \notin \mathcal{H}$, $S_k = -\sqrt{\Delta_k}$; otherwise, $S_k = \sqrt{\Delta_k}$.

Let $\pi_k(\mathcal{H}) = \Pr\left(\boldsymbol{\theta} \in \mathcal{H} | D_k\right)$ denote the posterior probability of the half space given the data up to stage k. We provide a heuristic argument on the asymptotical relationship between sample size and squared drift parameter ξ^2 for $\pi_k(\mathcal{H})$. Under the regularity conditions of Dudley and Haughton [25], the joint statistics $\{\Phi^{-1}(\pi_1(\mathcal{H})), \ldots, \Phi^{-1}(\pi_K(\mathcal{H}))\}$ converge in distribution to $\{S_1, \ldots, S_K\}$, which asymptotically follows a multivariate normal distribution [15]. Moreover, as the test statistics $\{S_1, \ldots, S_K\}$ are asymptotically equivalent to $\{Z(t_1), \ldots, Z(t_K)\}$, the same linear relationship between the maximum sample size and the squared drift parameter ξ^2 holds true asymptotically for a sequential design based on $\{S_1, \ldots, S_K\}$. Because such a relationship is a distributional property of the joint canonical test statistics $\{S_1, \ldots, S_K\}$, it thus also holds true asymptotically for a Bayesian sequential design that uses $\{\Phi^{-1}(\pi_1(\mathcal{H})), \ldots, \Phi^{-1}(\pi_K(\mathcal{H}))\}$ for decision making.

4 | Case Studies

4.1 | Design Specification

4.1.1 | Binary Endpoint With Beta-Binomial Model

We first consider the control of error rates in a two-arm randomized group sequential trial with a binary endpoint, where the Bayesian posterior probability is used for decision-making at the trial completion. In a two-arm Bayesian sequential trial with a binary endpoint, we perform an analysis up to a total of K analyses every time after we fairly randomize $2m$ additional subjects to the standard and experimental arms. At the kth analysis, the cumulative number of subjects accrued in each arm is km. Let p_E and p_S denote the response rates of the experimental drug and the standard drug, respectively. We are interested in testing the one-sided hypotheses

$$H_0: \ p_E \leqslant p_S \quad \text{versus} \quad H_A: \ p_E > p_S$$

Under the Bayesian paradigm, we assume that p_E and p_S follow the beta prior distributions, that is, $p_E \sim \text{Beta}(a_E, b_E)$ and $p_S \sim \text{Beta}(a_S, b_S)$, where $(\alpha_E, \beta_E, \alpha_S, \beta_S)$ are hyperparameters. Let y_E and y_S denote the number of patients with successful outcomes treated with the experimental drug and the standard drug, respectively. Assuming the equal randomization, based on the conjugate property of the Beta distribution, the posterior distributions of p_E and p_S are $p_E \mid y_E \sim \text{Beta}(a_E + y_E, b_E + km - y_E)$ and $p_S \mid y_S \sim \text{Beta}(a_S + y_S, b_S + km - y_S)$, respectively.

We conduct the Bayesian sequential tests at the interim analyses as follows. Let c_k be the cutoff of the posterior probability at stage k, $k = 1, \ldots, K$. At the kth stage, we would stop the trial and declare the superiority of the experimental treatment if

$$\Pr\left(p_E > p_S | y_E, y_S\right) \geqslant c_k$$

where $\Pr\left(p_E > p_S | y_E, y_S\right)$ can be computed based on the posterior distribution. Otherwise, we continue to enroll the next

group of m subjects to each arm and conduct another analysis at the end of the next stage. We calibrate the values of c_k using the method introduced in Section 3.1.

Let PoP_k be the posterior probability of $p_E > p_S$ given D_k, the accumulated data up to stage k, and let $b(y; m, p)$ denote the binomial probability mass function. Recall that $R_k(p_E, p_S, m)$ denotes the rejection probability function for the kth analysis. For the Bayesian sequential design with a binary endpoint, $R_k(p_E, p_S, m)$ can be expressed as

$$R_k(p_E,p_S,m)=\sum_{y_{E,1}=0}^{m}\sum_{y_{S,1}=0}^{m}\sum_{y_{E,2}=0}^{m}\sum_{y_{S,2}=0}^{m}\cdots\sum_{y_{E,k}=0}^{m}\sum_{y_{S,k}=0}^{m}\left\{I(PoP\geqslant c_k)\prod_{j=1}^{k-1}I(PoP_j<c_j)\prod_{h=1}^{k}\prod_{i=1}^{k}b(y_{E,h};m,p_E)b(y_{S,i};m,p_S)\right\}$$

The rejection probability function $R_k(p_E, p_S, m)$ depicts the probability that treatment superiority is declared at the kth analysis, given the value of p_E and p_S. The rejection probability function for the whole sequential design can be calculated as $R(p_E, p_S, m) = \sum_{k=1}^{K} R_k(p_E, p_S, m)$. A numerical evaluation of the method is presented in Section 3.2.1.

4.1.2 | Normal Endpoint

We consider a two-arm Bayesian sequential trial with normal endpoints. We are interested in comparing the mean of the outcome of the experimental treatment versus that of the standard treatment. Let y_{Ei} and y_{Si} denote the response of the ith subject in the experimental and standard arms, respectively. We assume that each y_{Ei} or y_{Si} is independent and identically distributed as $y_{Ei} \sim N(\mu_E, \sigma^2)$, $y_{Si} \sim N(\mu_S, \sigma^2)$, with unknown means μ_E and μ_S, but a known variance $\sigma^2 = 1$.

We denote $\theta = \mu_E - \mu_S$ as the true difference in treatment effect. At the kth interim analysis, the estimated treatment difference is $\hat{\theta} = \bar{y}_E - \bar{y}_S$, where $\bar{y}_E = \sum_{i=1}^{N_k} y_{Ei}/N_k$ and $\bar{y}_S = \sum_{i=1}^{N_k} y_{Si}/N_k$ are the sample means. We are interested in testing the one-sided hypotheses,

$$H_0:\ \theta \leqslant 0 \quad \text{versus} \quad H_A:\ \theta > 0$$

Under a normal prior distribution $\theta \sim N(0, \sigma_0^2)$, the posterior distribution of θ is $\theta \mid D \sim N(\tilde{\mu}, \tilde{\sigma}^2)$, where

$$\tilde{\mu} = \frac{n\hat{\theta}\sigma_0^2}{n\sigma_0^2+2}, \quad \tilde{\sigma}^2 = \frac{2\sigma_0^2}{n\sigma_0^2+2}$$

At the kth analysis, the posterior probability of H_A can be derived as,

$$\Pr(\theta > 0 \mid D_k) = \Phi\left(\hat{\theta}(N_k/2)\sqrt{\frac{1}{N_k/2+1/\sigma_0^2}}\right)$$

If $\Pr(\theta > 0 \mid D_k) \geqslant c_k$, we would stop the trial and declare the superiority of the experimental treatment. We control the type I error rate using the method by Shi and Yin [15], and set $c_k = \Phi(a_k)$. To control the power of the design, we use the algorithm in Section 3.2 to determine the sample size. A numerical evaluation of the method is presented in Section 3.2.2.

4.1.3 | Binary Endpoint With Logistic Model

We consider a Bayesian sequential design for a logistic model, which is used for biomarker trials [26] to analyze the effect of an experimental treatment on a group of patients with a certain biomarker. Upon enrollment into the trial, a patient's biomarker profile is obtained. Let B denote the number of biomarkers. Let X_{ib} denote the biomarker expression of bth biomarker for the ith patient, which equals 1 if the expression is positive, and 0 otherwise. Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{iB})$ denote the vector of biomarker expressions of the ith patient. There are a total of J experimental treatments where the patients are equally randomized across all the treatment arms. Let $\mathbf{T}_i = (T_{i1}, \ldots, T_{iJ})$ denote the vector of treatment assignment, where $T_{ij} = 1$ if the ith patient receives the jth experimental treatment. If all the T_{ij} are zero, the patient receives the standard treatment.

Let Y_i denote the binary efficacy endpoint for the ith patient. A Bayesian generalized linear model is used to fit the efficacy endpoint Y_i on the treatment assignment vector and the biomarker vector, as well as their interaction, which is represented as

$$Y_i \sim \text{Bernoulli}(p_i),$$
$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^{J}\beta_j T_{ij} + \sum_{b=1}^{B}\gamma_b X_{ib} + \sum_{b=1}^{B}\sum_{j=1}^{J}\delta_{jb}X_{ib}T_{ij}$$

where β_0 is the intercept, β_j the effect of the jth treatment, γ_b the effect of the bth biomarker, and δ_{jb} the effect of the interaction between the bth biomarker and the jth treatment. We specify non-informative normal prior distributions for all the coefficients.

A series of hypotheses are tested throughout the trial. For the jth experimental treatment on patients with positive biomarker b, we are interested in testing the hypothesis,

$$H_A^{(j,b)}: \beta_j + \delta_{jb} > 0, \qquad j > 0$$

which represents the superiority of the jth experimental treatment over the control treatment in the bth biomarker group.

Moreover, for the jth experimental treatment on patients with no positive biomarker expression, we are interested in testing the hypothesis,

$$H_A^{(j,0)}: \beta_j > 0$$

which represents the superiority of the jth experimental treatment over the control treatment in the negative biomarker group.

At the kth analysis, if the posterior probability $\Pr(H_A^{(j,b)} \mid D_k)$ exceeds c_k, where $b = 0, \ldots, B$, the superiority of the jth experimental treatment is declared for the bth biomarker group. To

control the type I error rate, we adopt the method by Shi and Yin [15], setting $c_k = \Phi(a_k)$. To control the power of the design, we use the generalized algorithm in Section 3 to determine the sample size. A numerical evaluation of the method is presented in the Supporting Information.

4.1.4 | Ordinal Endpoint

We consider a two-arm Bayesian sequential clinical trial with ordinal endpoints. We are interested in comparing the effectiveness of the experimental treatment method with that of the standard treatment method. Let Y_i denote the ordinal outcome of the patient i. Under the non-proportional odds assumption, we assume that

$$\log\left(\frac{\Pr(Y_i = y \mid Y_i \geqslant y)}{1 - \Pr(Y_i = y \mid Y_i \geqslant y)}\right) = \beta_{0,0} + \beta_{1,0} T_i + \sum_{j=1}^{J} \beta_{0,j} x_j + \sum_{j=1}^{J} \beta_{1,j} x_j T_i$$

where $T_i = 1$ if the ith patient receives the experimental treatment, and $T_i = 0$ if the ith patient receives the standard treatment. Here, x_j's are the covariates that have effects on the probability, and β's are the coefficients associated with the covariates. If we would like to have the proportional odds assumption instead, we simply ignore the $\sum_{j=1}^{J} \beta_{1,j} x_j T_i$, terms in the above expression.

To quantify the effectiveness of a treatment, we may use the expected utility for the experimental and the standard treatment, denoted as μ_E and μ_S, respectively,

$$\mu_E = \sum_{y=1}^{Y} U(y) \times \Pr(Y_i = y \mid T_i = 1)$$
$$\mu_S = \sum_{y=1}^{Y} U(y) \times \Pr(Y_i = y \mid T_i = 0)$$

where $U(\cdot)$ is a specified utility function that quantifies how good the clinical outcome y is regarding the patient's overall health status, which can be derived by consulting the clinicians.

Given the posterior distribution, we can perform the hypothesis testing based on the expected utility of each treatment method calculated above. We are interested in the following hypothesis testing:

$$H_0: \mu_E \leqslant \mu_C \text{ versus } H_A: \mu_E > \mu_C$$

At the kth interim analysis, if $\Pr(\mu_E > \mu_C \mid \text{Data}) \geqslant c_k$, then we stop the trial and recommend the experimental treatment method over the standard treatment method. At the end of the trial, if $\Pr(\mu_E > \mu_C \mid \text{Data}) \geqslant c_K$, then we conclude that the experimental treatment is superior compared with the standard treatment. Otherwise, we do not reject the null hypothesis. A numerical evaluation of the method is presented in the Supporting Information.

4.1.5 | Survival Endpoint With Piecewise Exponential Model

We consider a Bayesian sequential design for the piecewise exponential model. We are interested in comparing the survival time, which is measured as a continuous variable, of patients in the experimental group with that of patients in the control group. We assume that the period of observation has been partitioned into J time intervals with cutpoints $0 = \tau_0 < \tau_1 < \cdots < \tau_J$. Let $h(t)$ denote the hazard function evaluated at time t. For each group of patients, we assume that the hazard is a constant on each time interval j, denoted by λ_{Ej} for the experimental group and λ_{Cj} for the control group. Mathematically,

$$h(t) = \lambda_{Ej} \text{ if } \tau_{j-1} < t < \tau_j \text{ for the experimental group}$$

$$h(t) = \lambda_{Cj} \text{ if } \tau_{j-1} < t < \tau_j \text{ for the control group}$$

Let $S(t)$ denote the survival function evaluated at time t, which is related to the hazard function as

$$S(t) = \exp\left(-\int_0^t h(s)ds\right)$$

Since the survival function is defined as the probability that the patient survives past time t, we can easily calculate the median survival time of each treatment group from the survival function of the treatment group. To do so, we can find the time t such that the probability that a patient in the treatment group survives past time t is 0.5. Let ρ_E and ρ_C be the median survival time of the experimental group and that of the control group, respectively. We are interested in the following hypothesis testing:

$$H_0: \rho_E \leqslant \rho_C \text{ versus } H_A: \rho_E > \rho_C$$

There are K planned analyses when the sample size reaches some specific number. At interim analysis k, if $\Pr(\rho_E > \rho_C \mid \text{Data}) \geqslant c_k$, then we reject the null hypothesis and conclude that the experimental treatment is superior compared with the standard treatment, and we stop the trial. At the end of the trial, if $\Pr(\rho_E > \rho_C \mid \text{Data}) \geqslant c_K$, then we reject the null hypothesis and conclude that the experimental treatment is superior compared with the standard treatment. Otherwise, we do not reject the null hypothesis. Here, c_k's are some cutoff values that should be calibrated so that the design has good operating characteristics. A numerical evaluation of the method is presented in the Supporting Information.

4.2 | Numerical Evaluation

4.2.1 | Binary Endpoint With Beta-Binomial Model

We conduct a numerical study on the performance of the proposed method for the binary endpoint with beta-binomial model described in Section 4.1.1. As an example, we consider a double-arm Bayesian sequential trial with a binary endpoint and $K = 4$ analyses. The target type I error rate is $\alpha = 0.1$ and the Pocock-type α-spending function is used. We adopt a noninformative prior distribution Beta (0.5,0.5) for p_E and p_S. The stopping boundaries for the posterior probabilities are set as $c_k = \Phi(a_k)$, where a_k is the boundary value for kth interim analysis in the frequentist group sequential design. For the power computation, we fix the scenario where $p_E = 0.4$ and $p_S = 0.2$, and simulate one million trials by generating random draws from the binomial distribution.

The proportion of times where the experimental treatment is declared as superior is defined as the approximated power or $R(p_E, p_S, m)$. To compute the drift term ξ, we adopt the recursive algorithm proposed by Lai [24]. Figure 1 shows the relationship between the group sample size m and the squared drift term $\xi(\alpha, 1-R(p_E, p_S, m), K)^2$. The relationship is evidently linear, confirming the validity of our methodological proposal.

To examine the effectiveness of our approach, we conduct comprehensive experiments on the proposed algorithm. We consider $K = 2, 4$, $\alpha = 0.05, 0.1, 0.2$, $\beta = 0.1, 0.2$ and both Pocock-type and O'Brien-Fleming type α-spending functions. It is worth noting that the linear relationship is only asymptotical and not exactly precise, and thus the initial sample size guess $\widetilde{m}$ calculated using the algorithm in Section 2 could deviate slightly from the sample size that exactly controls the design power, which is then further corrected and calibrated via a line search around the initial guess. Table 1 shows the comparison of $\widetilde{m}$ and the exact sample size m_{exact}, which is computed by a thorough numerical enumeration. We observe that the initial guess $\widetilde{m}$ is reasonably close to m_{exact}. This indicates that the initial step of the proposed method can locate the search region around the neighbourhood of the true solution, after which a line search can then be employed to accurately pinpoint the required sample size. Moreover, the final output of the proposed algorithm $\widehat{m}$ is equal to the exact solution m_{exact}.

Furthermore, we conduct a benchmark study that compares the computational time of the proposed theoretical approach with the bisectional search method. We experiment with K ranging from 2 to 5 and set $\alpha = 0.1$ and $\beta = 0.2$. Figure 2 shows a comparison of the computational time of the two approaches. Compared with the bisectional search method, our proposed method leads to significant time savings ranging from 80% to 90%.

4.2.2 | Normal Endpoint

We conduct a numerical study to assess the accuracy of the proposed method for the normal endpoint described in Section 4.1.2. We consider a double-arm Bayesian sequential trial with a normal endpoint and $K = 4$ analyses. The target type I error rate is $\alpha = 0.1$ and the Pocock-type α-spending function is used. We assume the prior distribution $\theta \sim N(0, \sigma_0^2)$ where $\sigma_0 = 1000$. We set the stopping boundaries for the posterior probabilities as $c_k = \Phi(a_k)$. For the power computation, we consider the alternative scenario where $\theta = 0.5$, and approximate the power as the proportion of times when the experimental treatment is declared as superior based on 10,000 trial simulations. We use the recursive algorithm proposed by Lai [24] to compute the drift term ξ. Figure 3 shows the relationship between the group sample size m and the squared drift term, which is evidently linear.

To assess the effectiveness of our approach, we apply the proposed algorithm to compute the sample size and compare the solution with the one based on a thorough numerical enumeration. We experiment with $K = 2, 4$, $\alpha = 0.05, 0.1, 0.2$, $\beta = 0.1, 0.2$,

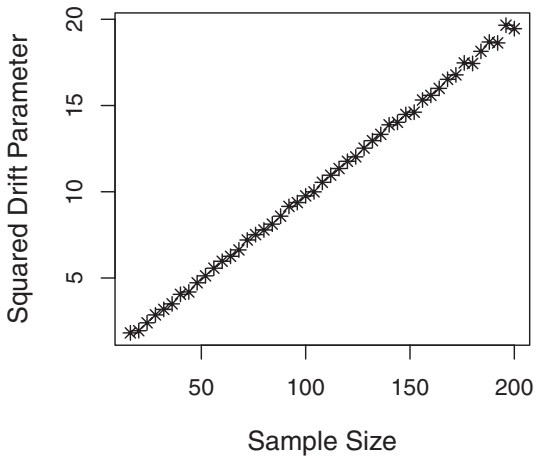

FIGURE 1 | The relationship between the squared drift parameter and the group sample size for a Bayesian sequential design with a binary endpoint with $K = 4$, a type I error rate of $\alpha = 0.1$, and a Pocock-type α-spending function.

TABLE 1 | Comparison of $\widetilde{m}$, the initial group sample size proposal; m_{exact}, the group sample size that exactly controls the design power; and $\widehat{m}$, the final output from the proposed algorithm, under various α -spending functions, K, α, and β for a Bayesian sequential design with a binary endpoint.

α-spending function	K	α	β	$\widetilde{m}$	m_{exact}	$\widehat{m}$
O'Brien-Fleming	2	0.05	0.1	87	90	90
			0.2	65	64	64
		0.1	0.1	68	70	70
			0.2	47	48	48
		0.2	0.1	47	50	50
			0.2	31	32	32
	4	0.05	0.1	90	92	92
			0.2	65	68	68
		0.1	0.1	70	72	72
			0.2	49	52	52
		0.2	0.1	49	52	52
			0.2	32	32	32
Pocock	2	0.05	0.1	97	98	98
			0.2	71	70	70
		0.1	0.1	77	76	76
			0.2	53	52	52
		0.2	0.1	53	52	52
			0.2	34	32	32
	4	0.05	0.1	103	104	104
			0.2	78	80	80
		0.1	0.1	83	84	84
			0.2	57	60	60
		0.2	0.1	58	60	60
			0.2	38	40	40

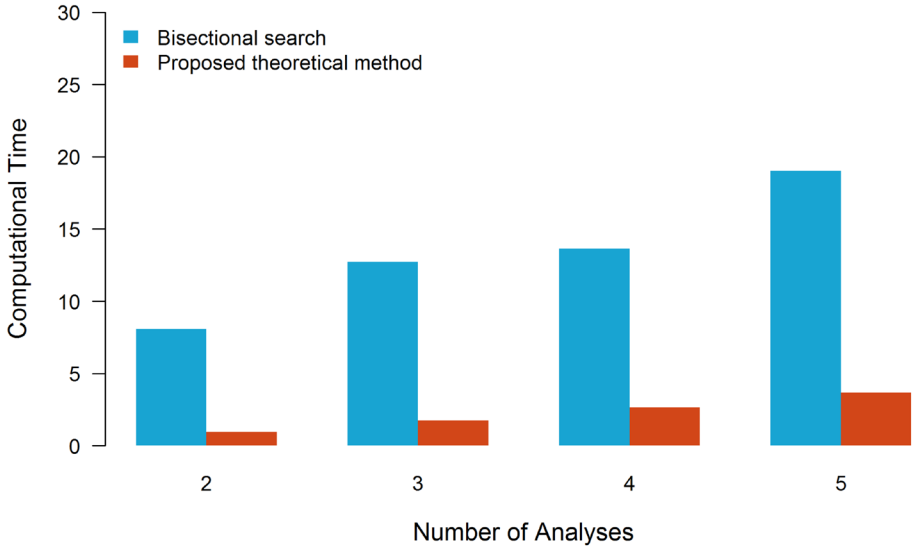

FIGURE 2 | Comparison of the computational time (in seconds) between the bisectional search method and the proposed theoretical method under various values of K, of a Bayesian sequential design with a binary endpoint and $\alpha = 0.1$, $\beta = 0.2$, and a Pocock-type α-spending function.

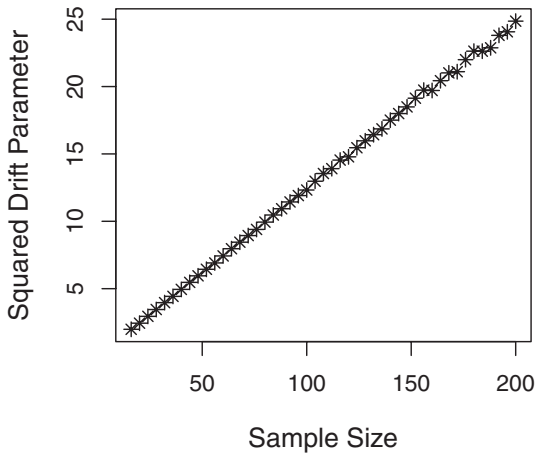

FIGURE 3 | The relationship between the squared drift parameter and the group sample size for a Bayesian sequential design with a normal endpoint with $K = 4$, a type I error rate of $\alpha = 0.1$, and a Pocock-type α-spending function.

and consider both the Pocock-type and O'Brien-Fleming type α-spending functions. Due to the approximation error of the value of power and the linear relationship being only asymptotical and not exactly precise, the initial group sample size guess $\widetilde{m}$ could be slightly different from the exact sample size that leads to accurate control of power. Table 2 shows the comparison of the initial group sample size guess $\widetilde{m}$ and the exact group sample size m_{exact}. It is evident that the initial guess $\widetilde{m}$ is close to m_{exact}, which suggests that the initial sample size proposal of the algorithm is already within the neighborhood of the true solution. The final output of the proposed algorithm $\widetilde{m}$ is exactly equal to the solution m_{exact}.

5 | Discussion

Determining the sample size needed to control the power of trial designs that involve multiple analyses based on the posterior probabilities is an unexplored aspect in Bayesian sequential designs. Our proposed method connects the properties of sample size in Bayesian designs with those under the framework of frequentist group sequential designs. The major advantage of our proposed method is that it utilizes the proportional relationship between the sample size and the squared drift parameter to pinpoint a relatively accurate initial guess of the sample size, which greatly saves computational time. In contrast, existing approaches mostly rely on numerical search, which generally has a very long computation time. The numerical evaluations in the case studies confirm the proportional relationship between the sample size and the squared drift parameter and show that the initial sample size proposal of the algorithm is close to the true sample size that exactly controls the design power. The numerical evaluations also show the reduction of the computation time using our method compared with existing methods. However, a limitation of the proposed method is that the selection of candidate sample sizes is somewhat arbitrary, although the method remains robust to this choice.

Our proposed method is motivated by the sample size calculation method used by Murray et al. [27], which also utilizes the linear relationship between the sample size and a value that only depends on some pre-specified values including the target type-I error rate and the target type-II error rate. Both methods first use simulations to compute the empirical power at certain sample sizes and then use the linear relationship, which passes through the origin, between the sample size and the value calculated from pre-specified values to estimate the sample size required to achieve the target power. However, there are differences in how to begin the search process. In Murray's paper, a frequentist formula for sample size calculation is used to find the initial sample size, and then simulations under the Bayesian setting are used to update the sample size. In our proposed method, instead of utilizing the frequentist formula, we conduct simulations for at least three candidate sample sizes under the Bayesian setting, and then use the simulated results to fit the regression line which describes the linear relationship between the sample size and the squared drift parameter. To choose the

TABLE 2 | Comparison of $\tilde{m}$, the initial group sample size proposal; m_{exact}, the group sample size that exactly controls the design power; and $\hat{m}$, the final output of the proposed algorithm, under various α-spending functions, K, α, and β for a Bayesian sequential design with a normal endpoint.

α-spending function	K	α	β	$\tilde{m}$	m_{exact}	$\hat{m}$
O'Brien-Fleming	2	0.05	0.1	71	70	70
			0.2	50	52	52
		0.1	0.1	54	54	54
			0.2	36	38	38
		0.2	0.1	39	40	40
			0.2	24	26	26
	4	0.05	0.1	69	72	72
			0.2	52	52	52
		0.1	0.1	56	56	56
			0.2	38	40	40
		0.2	0.1	39	40	40
			0.2	25	28	28
Pocock	2	0.05	0.1	75	78	78
			0.2	55	56	56
		0.1	0.1	60	60	60
			0.2	41	42	42
		0.2	0.1	41	40	40
			0.2	26	26	26
	4	0.05	0.1	81	84	84
			0.2	61	64	64
		0.1	0.1	65	64	64
			0.2	46	48	48
		0.2	0.1	46	48	48
			0.2	29	28	28

candidate sample sizes, one of the candidate sample sizes can be an initial guess of the sample size using frequentist methods or other statistical methods, and other candidate sample sizes can be far from each other to some extent, so that we can have a good estimate of the regression coefficient γ.

One area of future research is to investigate how the adoption of an informative prior or other types of adaptive priors would affect the performance of the proposed method. In cases where an informative prior is adopted, calibration of the decision boundaries c_k and power would be more complicated and may depend upon the intrinsic model assumption. An informative prior would carry with it a substantial "effective sample size" [28], which would naturally result in a smaller required sample size. It would be of interest to determine whether there exists any approximate numerical relationship between the effective sample size and the required sample size for various types of endpoints. Nevertheless, it has been theoretically shown that the power gain is limited for informative Bayesian priors when the type I error rate constraint is applied [20, 29, 30].

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant RGPIN/02963-2021.

The research is also supported by the Collaborative Research Team sponsored by CANSSI.

We thank Thomas Murray for his insightful comments and discussion about this paper.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. D. A. Berry, "Bayesian Clinical Trials," *Nature Reviews Drug Discovery* 5 (2006): 218–226.

2. D. A. Berry, "Adaptive Clinical Trials in Oncology," *Nature Reviews Clinical Oncology* 9 (2011): 199–207.

3. J. O. Berger and D. A. Berry, "Statistical Analysis and the Illusion of Objectivity," *American Scientist* 76 (1988): 159–165.

4. B. Efron, "Why Isn't Everyone a Bayesian?," *American Statistician* 40 (1986): 1–5.

5. T. A. Murray, P. F. Thall, Y. Yuan, S. McAvoy, and D. R. Gomez, "Robust Treatment Comparison Based on Utilities of Semi-Competing Risks in Non-Small-Cell Lung Cancer," *Journal of the American Statistical Association* 112 (2016): 11–23, https://doi.org/10.1080/01621459.2016.1176926.

6. T. A. Murray, P. F. Thall, and Y. Yuan, "Utility-Based Designs for Randomized Comparative Trials With Categorical Outcomes," *Statistics in Medicine* 35 (2016): 4285–4305.

7. Q. Yu, L. Zhu, and H. Zhu, "A Bayesian Sequential Design With Adaptive Randomization for Two-Sided Hypothesis Tests," *Pharmaceutical Statistics* 16 (2017): 451–465, https://doi.org/10.1002/pst.1830.

8. W. Zhang, G. Laird, J. Chen, and Y. Yuan, "A Bayesian Phase II Proof-Of-Concept Design for Clinical Trials With Longitudinal Endpoints," *Statistics in Medicine* 43, no. 1 (2023): 173–183.

9. H. Zhou, C. Chen, L. Sun, and Y. Yuan, "Bayesian Optimal Phase II Clinical Trial Design With Time-To-Event Endpoint," *Pharmaceutical Statistics* 19 (2020): 776–786.

10. Y. Zhao, B. Yang, J. J. Lee, L. Wang, and Y. Yuan, "Bayesian Optimal Phase II Design for Randomized Clinical Trials," *Statistics in Biopharmaceutical Research* 14, no. 4 (2022): 423–432.

11. B. Guo and S. Liu, "An Optimal Bayesian Predictive Probability Design for Phase II Clinical Trials With Simple and Complicated Endpoints," *Biometrical Journal* 62 (2020): 339–349.

12. FDA, "Guidance for Industry: Interacting With the FDA on Complex Innovative Trial Designs for Drugs and Biological Products Guidance for Industry" (2020), https://www.fda.gov/media/130897/download.

13. G. K. K. Lan and D. L. Demets, "Discrete Sequential Boundaries for Clinical Trials," *Biometrika* 70 (1983): 659–663.

14. K. Kim and D. L. Demets, "Sample Size Determination for Group Sequential Clinical Trials With Immediate Response," *Statistics in Medicine* 11 (1992): 1391–1399.

15. H. Shi and G. Yin, "Control of Type I Error Rates in Bayesian Sequential Designs," *Bayesian Analysis* 14 (2019): 399–425.

16. F. Santis, "Sample Size Determination for Robust Bayesian Analysis," *Journal of the American Statistical Association* 101 (2006): 278–291.

17. C. M'lan, L. Joseph, and D. Wolfson, "Bayesian Sample Size Determination for Binomial Proportions," *Bayesian Analysis* 3 (2008): 269–296, https://doi.org/10.1214/08-BA310.

18. J. Whitehead, E. Valdés-Márquez, P. Johnson, and G. Graham, "Bayesian Sample Size for Exploratory Clinical Trials Incorporating Historical Data," *Statistics in Medicine* 27 (2008): 2307–2327.

19. B. Zaslavsky, "Bayesian Sample Size Estimates for One Sample Test in Clinical Trials With Dichotomous and Countable Outcomes," *Statistics in Biopharmaceutical Research* 4 (2012): 76–85.

20. M. A. Psioda and J. G. Ibrahim, "Bayesian Clinical Trial Design Using Historical Data That Inform the Treatment Effect," *Biostatistics* 20 (2019): 400–415.

21. H. Zheng, T. Jaki, and J. Wason, "Bayesian Sample Size Determination Using Commensurate Priors to Leverage Preexperimental Data," *Biometrics* 79 (2020): 669–683, https://doi.org/10.1111/biom.13649.

22. S. Golchi, "Estimating Design Operating Characteristics in Bayesian Adaptive Clinical Trials," *Canadian Journal of Statistics* 50 (2022): 417–436.

23. C. Jennison and B. W. Turnbull, "Group Sequential Analysis Incorporating Covariate Information," *Journal of the American Statistical Association* 92 (1997): 1330–1341.

24. D. Lai, "Sample Size Determination for Group Sequential Test Under Fractional Brownian Motion," *Electronic Journal of Statistics* 7 (2013): 1957–1967.

25. R. M. Dudley and D. Haughton, "Asymptotic Normality With Small Relative Errors of Posterior Probabilities of Half-Spaces," *Annals of Statistics* 30 (2002): 1311–1344.

26. J. M. S. Wason, S. E. Abraham, R. D. Baird, et al., "A Bayesian Adaptive Design for Biomarker Trials With Linked Treatments," *British Journal of Cancer* 113 (2015): 699–705.

27. T. A. Murray, P. F. Thall, F. Schortgen, P. Asfar, S. Zohar, and S. Katsahian, "Robust Adaptive Incorporation of Historical Control Data in a Randomized Trial of External Cooling to Treat Septic Shock," *Bayesian Analysis* 16 (2021): 825–844.

28. S. Morita, P. Müller, and P. F. Thall, "Determining the Effective Sample Size of a Parametric Prior," *Biometrics* 64 (2008): 595–602.

29. A. Kopp-Schneider, S. Calderazzo, and M. Wiesenfarth, "Power Gains by Using External Information in Clinical Trials Are Typically Not Possible When Requiring Strict Type I Error Control," *Biometrical Journal* 62 (2020): 361–374.

30. H. Quan, B. Zhang, Y. Lan, X. Luo, and X. Chen, "Bayesian Hypothesis Testing With Frequentist Characteristics in Clinical Trials," *Contemporary Clinical Trials* 87 (2019): 105858.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** Supporting Information.