Received: 2 January 2023 | Revised: 17 July 2023 | Accepted: 5 October 2023

DOI: 10.1002/sim.9936

RESEARCH ARTICLE

Statistics in Medicine WILEY

BEATS: Bayesian hybrid design with flexible sample size adaptation for time-to-event endpoints

Dehua Bi[1] | **Meizi Liu**[2] | **Jianchang Lin**[2] | **Rachael Liu**[2]

[1]Department of Public Health Sciences, University of Chicago, Chicago, Illinois, USA

[2]Statistical & Quantitative Sciences, Takeda Pharmaceuticals, Cambridge, Massachusetts, USA

Correspondence
Meizi Liu and Rachael Liu, Statistical & Quantitative Sciences, Takeda Pharmaceuticals, Cambridge, MA, 02139, USA.
Email: meizi.liu@takeda.com and yue.liu@takeda.com

As the roles of historical trials and real-world evidence in drug development have substantially increased, several approaches have been proposed to leverage external data and improve the design of clinical trials. While most of these approaches focus on methodology development for borrowing information during the analysis stage, there is a risk of inadequate or absent enrollment of concurrent control due to misspecification of heterogeneity from external data, which can result in unreliable estimates of treatment effect. In this study, we introduce a Bayesian hybrid design with flexible sample size adaptation (BEATS) that allows for adaptive borrowing of external data based on the level of heterogeneity to augment the control arm during both the design and interim analysis stages. Moreover, BEATS extends the Bayesian semiparametric meta-analytic predictive prior (BaSe-MAP) to incorporate time-to-event endpoints, enabling optimal borrowing performance. Initially, BEATS calibrates the expected sample size and initial randomization ratio based on heterogeneity among the external data. During the interim analysis, flexible sample size adaptation is performed to address conflicts between the concurrent and historical control, while also conducting futility analysis. At the final analysis, estimation is provided by incorporating the calibrated amount of external data. Therefore, our proposed design allows for an approximation of an ideal randomized controlled trial with an equal randomization ratio while controlling the size of the concurrent control to benefit patients and accelerate drug development. BEATS also offers optimal power and robust estimation through flexible sample size adaptation when conflicts arise between the concurrent control and external data.

KEYWORDS

Bayesian borrowing, historical control, hybrid design, real-world data, sample size rebalance, semi-parametric meta-analytic-predictive prior

1 | INTRODUCTION

In evidence-based medicine, randomized controlled trials (RCT) have long been considered as the gold standard for evaluating the efficacy and safety of new investigational drugs.[1] Typically, equal randomization ratios are employed in RCTs, as unequal allocation often necessitates a larger sample size under the same power.[2] However, in many therapeutic areas with unmet medical needs, such as cancer and rare disease, ethical considerations may favor exposing patients less to

standard care or placebo in order to minimize disease progression and reduce patient burden.[3] Patients are also more willing to participate in clinical trials if there is a higher chance of being assigned to the experimental arm. Unequal randomization not only enhances patients' compliance with the trial[4] but also improves clinical data. Moreover, if the experimental group is anticipated to have a significant dropout rate, a larger number of patients receiving the experimental drug can increase power in the intent-to-treat analysis. As evidenced by Gupta et al,[5] there has been a noticeable increase in the utilization of unequal allocation in oncology clinical trials over the past decade. When unequal randomization does not result in a larger total sample size, it can lead to enhanced economic efficiency and reduced development timelines.[6]

To maximize the efficiency of unequal allocation without increasing the overall sample size, a promising solution is to adopt a hybrid design that incorporates external data sources, including real-world data (RWD) and historical trials, to augment the control arm. In various therapeutic areas such as oncology and rare disease, there is often a wealth of external data available on standard of care including historical trials, disease registry data, and other types of RWD specific to targeted patient populations.[7,8] In 2018, the FDA introduced a framework and provided guidance on using RWD in hybrid design.[9] These designs combine randomization and pragmatic elements, supplemented by RWD or novel data collection approaches, to capture long-term outcomes in real-world settings. By incorporating randomization, hybrid designs preserve the advantages of traditional RCTs while expediting the development of new treatments and optimizing operational efficiency. However, the successful implementation of hybrid designs relies on careful selection and calibration of external data, as pooling RWD or historical trials without accounting for heterogeneity among different data sources may introduce bias and yield misleading treatment benefit conclusions.

In recent years, several statistical methods under the Bayesian framework have been developed to integrate historical information in the trial analysis. Among those, power priors,[10] modified power priors,[11,12] and the dependent modified power prior[13] are based on a weighted likelihood approach where the weight parameter controls the degree of borrowing. Hobbs et al[14] proposed a commensurate prior, in which the priors for the model parameters of the current data are centered at the historical data. Another popular group of methods for incorporating external data is the hierarchical modeling approach. The meta-analytic-predictive (MAP) prior[15,16] is derived using a random-effects meta-analysis model, which accounts for the between-trial heterogeneity among the external and current data. The MAP prior assumes that the model parameters of the trials are exchangeable and follow a common prior distribution.[17] As a result, the MAP prior is unable to dynamically discount the amount of borrowing. Schmidli et al[18] proposed a robust MAP (R-MAP) prior by incorporating a weakly informative component to the MAP prior, which allows for discounting external data in the case of substantial prior-data conflict. The application of R-MAP prior requires pre-specification of the weight parameters on each component. Choosing and justifying appropriate weight parameters without observing any data from the current study is challenging. To address the potential limitations of the MAP and R-MAP approaches, Hupf et al[19] proposed the BaSe-MAP prior, which allows for trial-specific variance around the common mean by adopting the non-parametric Dirichlet process mixture model. The method provides a more flexible and robust approach to adaptively leverage historical data from multiple study sources without the need for parameter pre-specification. Furthermore, Liu et al[20] proposes a propensity-score-based MAP prior to select a subset of patients from external data who are similar to those in the current study in terms of their propensity scores and to stratify the selected patients together with those in the current study into more homogeneous strata.

Most of the aforementioned methods for borrowing information from external sources are applicable only at the data analysis stage. However, if the heterogeneity is not accurately estimated from the beginning, it can lead to insufficient planning of the concurrent control arm, resulting in underpowered analyses or excessive borrowing for biased estimation. To optimize the borrowing in hybrid study designs, it is essential to emulate a RCT with a 1:1 randomization ratio. Wen et al[21] proposed a Bayesian group sequential design for randomized biosimilar trials that adaptively borrow historical data using the elastic MAP (EMAP) prior to augment the control arm and reduce the sample size. This method incorporates borrowing information at both interim and final analyses, adjusting the borrowing extent based on the observed heterogeneity. This design is developed only for binary outcomes, while time-to-event outcomes, such as progression-free survival or overall survival, are commonly used as primary endpoints in real-world settings. Li et al[22] have presented a design utilizing a commensurate prior and propensity score weighting to incorporate a hybrid control arm from the historical IMblaze370 phase 3 study with a current trial for metastatic colorectal cancer. This approach requires patients' baseline covariates for propensity score weighting, which may not be available during the design stage. Consequently, it is preferable to calibrate the borrowing extent during the design stages to ensure appropriate planning of the concurrent control arm and sufficient statistical power. Furthermore, given the uncertainty surrounding the heterogeneity

between the concurrent control and external data, adaptability at interim analysis is crucial for reducing the risk associated with potential misspecification in the initial design. For instance, as clinical practices improve over time, the effect of standard care may be underestimated when relying solely on historical trials. The ability to adjust the borrowing extent and sample size requirements of the concurrent control arm based on observed heterogeneity would be highly advantageous.

In this work, we propose a Bayesian hybrid design with flexible sample size adaptation for RCT with time-to-event endpoints (BEATS) to adaptively leverage external data to augment the control arm. The calibration of borrowing extent will be performed at various stages, including the initial planning stage, interim analysis, and final analysis. We extend the Bayesian semiparametric meta-analytic predictive (BaSe-MAP) prior to incorporate time-to-event endpoints (eg, overall survival). Specifically, at the planning stage, we calibrate the required sample size of concurrent control with available historical data. The effective sample size is utilized to quantify the information borrowing from the historical data. The BEATS design allows for early stopping due to futility at interim, and if the interim decision is to continue the trial, the sample size can be re-balanced by adjusting the randomization ratio after incorporating the heterogeneity between concurrent control and historical information using BaSe-MAP. The sample size rebalance through randomization ratio adjustment will maximize the benefit of randomization, while rescuing the trial if the borrowing extent is mis-specified in the beginning.

The remainder of the article is organized as follows. In Section 2, we introduce the proposed extension of BaSe-MAP prior for time-to-event endpoints and describe the details of the BEATS design. We perform simulation studies in Section 3 to compare the operating characteristics of our proposed method with existing methods across different scenarios. In Section 4, we illustrate the application of the BEATS design via an example of confirmatory trial in breast cancer. Lastly, we conclude in Section 5 with a discussion on the proposed design as well as its merits and limitations.

2 | METHODS

2.1 | BaSe-MAP prior for time-to-event endpoints

We consider the randomized trial setting, with an experimental arm, denoted as T, against a control arm, denoted as C. Additionally, assume that $k = 1, \ldots, K$ historical trials are available for borrowing to augment the concurrent control arm. To overcome the limitations of the existing Bayesian borrowing methods, the BaSe-MAP prior, models the random effects in the MAP prior nonparametrically as a Dirichlet process mixture of Gaussian distributions. By relaxing the parametric assumption on the random effects, the BaSe-MAP prior adaptively learn the relationship between the historical data and current control data while still being able to discount the historical data in case of prior-data conflict. The full BaSe-MAP prior for binary endpoint proposed by Hupf et al[19] is given by:

$$
\begin{aligned}
y_j|p_j &\sim \text{Bin}\left(n_j, p_j\right), j = C, 1, \ldots, K, \\
\text{logit}\left(p_j\right) &= \mu + \delta_j, \\
\delta_j|\sigma_j^2 &\sim N\left(0, \sigma_j^2\right), \\
\sigma_j^2|G &\sim G, \\
G &\sim DP(G_0, \alpha),
\end{aligned}
\tag{1}
$$

where p_j is response rate of trial j, and p_C refers to the response rate of the concurrent control arm. BaSe-MAP fixes the means of the Gaussian distributions to a common value of μ, and δ_j is trial specific random effect. The hyper-prior for the variance of the random effects, σ_j^2 is assumed to follow a Dirichlet process, where α is the concentration parameter, and G_0 is the base distribution.

To adapt the BaSe-MAP method for time-to-event endpoints, we propose an extension by assuming an exponential distribution with a hazard rate λ_j to model the time-to-event outcome for each trial. More flexible survival models (eg, piecewise exponential distributions) can also be adopted in case the constant hazard assumption does not hold.[23]

Equivalently, for each trial, there are n_j patients at risk with a corresponding total exposure time E_j, of which r_j experience an event.[24] Then the count data follows a Poisson distribution

$$r_j|\lambda_j \sim \text{Poisson}\left(\lambda_j E_j\right), \quad j = C, 1, \ldots, K,$$

where λ_j is the hazard of trial j. And we denote the hazard of the concurrent control arm as λ_C. Therefore, the extension of the BaSe-MAP prior to time-to-event endpoints can be described as below:

$$\begin{aligned} r_j|\lambda_j &\sim \text{Poisson}\left(\lambda_j E_j\right), \quad j = C, 1, \ldots, K, \\ \theta_j &= \log\left(\lambda_j\right) = \mu + \delta_j, \\ \mu &\sim N\left(0, \sigma_\mu^2\right), \\ \delta_j|\sigma_j^2 &\sim N\left(0, \sigma_j^2\right), \\ \sigma_j^2|G &\sim G, \\ G &\sim DP(G_0, \alpha), \end{aligned} \tag{2}$$

where θ_j is the log-hazard parameters of trial j, μ is the common mean log-hazard and δ_j is trial specific random effect. The common mean log-hazard has a noninformative normal prior with a mean of 0 and a variance of σ_μ^2 set at 10 000. The hyper-prior for the variance of the random effects, σ_j^2 is assumed to follow a Dirichlet process, where α is the concentration parameter, and G_0 is the base distribution. Following Hupf et al,[19] α is assumed to be fixed at 1 for simplicity and G_0 is assumed to be a half-normal distribution with variance 1. Other possible base distributions for G_0 include half-t, lognormal, and inverse-gamma distributions.[16,17,25]

2.2 | Prior effective number of events

When leveraging external information to augment the planned patients in the current study through a prior, quantifying the amount of information borrowed is essential. It is often expressed as the prior effective sample size, or, in the time-to-event setting, prior effective number of events (ENE).[23] To compute ENE, various methods are proposed in literature, such as the variance-ratio method,[26] the Morita-Thall-Müller method,[27] and the expected local-information ratio (ELIR) method.[26] Among them, the ELIR method is defined as the expected ratio of prior information $i(p(\theta))$ to Fisher information $i_F(\theta)$:

$$\text{ENE} = E_\theta\left\{\frac{i(p(\theta))}{i_F(\theta)}\right\}, \tag{3}$$

where $i(p(\theta)) = -\frac{d^2}{d\theta^2}\log(p(\theta))$ and $i_F(\theta) = -E_{Y_1|\theta}\left\{\frac{d^2}{d\theta^2}\log(p(Y_1|\theta))\right\}$. It can be shown that, unlike the other methods, the ELIR method fulfills a basic prediction consistency criterion, which requires the expected posterior ENE (under the prior distribution) for a sample size of n to be the sum of the prior ENE and n. Hence, we adopt this method for the calculation of prior and posterior ENE in the proposed design.

Since the BaSe-MAP prior introduced in Section 2.1 is only available as an MCMC sample, the definition of ENE in Equations (3) cannot be directly calculated. Therefore, following a similar approach to Hupf et al,[19] approximations of the BaSe-MAP priors for the log-hazard parameters is performed using a mixture of normal distributions. Various software tools have been developed for fitting mixture distributions such as the R package RBesT.[28] We then use closed-form mixture representation of the priors to evaluate the ENE.

In general, the prior ENE is useful to determine the amount of borrowing at design stage of a study thus to calibrate the initial randomization ratio of experimental and control arms. The posterior ENE is of interest as well during interim and final analysis to update the information borrowed with observed data from the concurrent control. It can

be calculated by replacing the prior $p(\theta)$ in Equations (2) and (4) with the posterior $p(\theta|Y)$, where Y is the combined historical data and observed concurrent control data. Again, as no closed form of BaSe-MAP can be obtained, mixture distributions is fitted to evaluate the posterior ENE. Note that although prior and posterior ENE are used to quantify the information borrowed, we would still need the corresponding *equivalence sample size* (ESS) for study design. We assume simple exponential distributions for the time-to-event endpoints, and details of the conversion is included in Appendix A.

2.3 | The proposed BEATS design

The objective of the BEATS design is to adaptively augment the concurrent control arm of a RCT through the BaSe-MAP approach so that the RCT attains a certain target allocation ratio. For simplicity, we assume the target allocation ratio is 1:1, note that the BEATS design is applicable to achieve any fixed randomization ratio. The trial starts with an initial randomization ratio of $1 : R_0$ for the treatment arm and the control arm, where $0 \leq R_0 \leq 1$ such that the prior ESS introduced by BaSe-MAP prior plus the initial planned sample size for the control arm equals to the planned sample size for the treatment arm (ie, achieving 1:1 balanced allocation). Then at each interim, we perform Bayesian go/no-go decision making by comparing the investigational treatment arm against the augmented control, where historical data will be adaptively borrowed using the BaSe-MAP method. Moreover, if the stopping for futility is not satisfied at interim, we perform sample size adaptation to adjust the randomization ratio based on the updated posterior ESS of BaSe-MAP. We describe these procedures in details as below.

2.3.1 | Design stage

Recall that λ_T, λ_C are the unknown hazard rates for the treatment arm and control arm, respectively, and θ_T, θ_C are the log-hazards for the treatment and control arms. Let $\Delta = \theta_T - \theta_C = \log\left(\frac{\lambda_T}{\lambda_C}\right)$ denote the log-hazard ratio between the two arms, and assume the RCT is designed to test whether the investigational treatment has a lower hazard than the standard of care, that is,

$$H_0 : \Delta = 0 \qquad \text{vs} \qquad H_1 : \Delta < 0.$$

Suppose the trial is planned for N patients per arm, based on the standard sample size calculation with 1:1 randomization ratio to ensure a reasonable power (eg, 90%) and type I error rate (eg, 0.025). In other words, there should be N patients enrolled in the treatment arm and N-posterior ESS patients enrolled in the control arm at end of the trial, if the trial is not stopped early for futility. Moreover, assume that there are $I - 1$ interim analysis planned for the study and the *Ith* analysis is the final analysis, and n_{T_i} and $n_{C_i}, i = 1, \ldots, I$ are the number of patients enrolled in the treatment and control arms at the *ith* interim analysis. For ease of illustration, we assume the planned interims are equally spaced, but our method can also be applied to non-equally spaced interims.

In addition, assume after careful clinical and statistical evaluation, K available historical studies have been identified to augment the concurrent control arm. The BaSe-MAP prior for time-to-event endpoints can then be applied to the K historical data, and the prior ENE and its ESS can be computed as discussed in Section 2.2. The trial begin with enrolling $n_{T_1} = \frac{N}{I}$ in the treatment arm and $n_{C_1} = \frac{N\text{-prior ESS}}{I}$ in the control arm (ie, the initial randomization ratio is set as $1 : R_0$, where $R_0 = \frac{(N\text{-prior ESS})}{N}$).

2.3.2 | Interim analysis

At the *ith* interim analysis, we first apply the BaSe-MAP method given K historical data, denoted as $D_{H_1}, \ldots, D_{H_K}$, where $D_{H_k} = \{r_k, E_k\}$ and the cumulative interim data from the concurrent control, $D_{C_i} = \left\{r_{C_i}, E_{C_i}\right\}$, to derive the posterior of θ_C at the *ith* interim, denoted as $p\left(\theta_C|D_{H_k}, D_{C_i}, k = 1, \ldots, K\right)$. Since the BaSe-MAP prior adaptively integrates external historical data based on the degree of similarity with the concurrent control and maintain robustness in case of prior-data conflict, it is repeatedly applied at each interim to adjust the amount of borrowing from external sources.

For the treatment arm, with interim data $D_{T_i} = \{r_{T_i}, E_{T_i}\}$, the posterior distribution of θ_T at *ith* interim, $p(\theta_T|D_{T_i})$ can be computed with the following probability model,

$$\begin{aligned} r_T|\lambda_T &\sim \text{Poisson}(\lambda_T E_T), \\ \theta_T = \log(\lambda_T) &\sim N(\mu_T, \sigma_T^2), \end{aligned} \tag{4}$$

where r_T is the number of events, E_T is the total exposure time, and θ_T is the log-hazard of the treatment arm. As limited historical information is available for the treatment arm, μ_T is usually assigned a vague prior distribution and σ_T^2 is assumed to be fixed at some large number.

At interim, the proposed design allows for early stopping due to futility by evaluating the following criterion:

$$\text{Stop the trial early if } \Pr(\Delta \le 0|D_{H_k} D_{C_i}, D_{T_i}, k = 1, \ldots K) \le C_i;$$

Note that C_i is the probability threshold for futility stopping for *ith* interim, which can be calibrated through simulations for desirable operating characteristics. And the posterior probability of the treatment effect is given by,

$$\begin{aligned} &\Pr(\Delta \le 0|D_{H_k}, D_{C_i}, D_{T_i}, k = 1, \ldots K) \\ &= \iint I(\theta_T - \theta_C \le 0) p(\theta_T|D_{T_i}) p(\theta_C|D_{H_k}, D_{C_i}, k = 1, \ldots, K) d\theta_T d\theta_C. \end{aligned} \tag{5}$$

If the above Bayesian go/no-go criterion is met, terminate the trial, and conclude that the treatment is futile. Otherwise, continue the trial and adjust the randomization ratio between the treatment and the control arms based on the rules introduced next.

2.3.3 | Sample size adaptation

The objective of sample size adaptation is to maintain the same expected sample size between the treatment and the control arms. That is, the future number of patients to be enrolled in the concurrent control arm should equal to N minus the posterior *ESS* of the control data at interim, and again N is the total number of patients planned per arm for a standard 1:1 RCT. In this case, if the heterogeneity between concurrent control and external data is underestimated during the design stage, this is the opportunity to rescue a trial to allow more accurate treatment benefit estimation, improve the power or avoid overborrowing.

Specifically, we calculate the posterior *ESS* of the BaSe-MAP given the historical data and the cumulative interim data from the control as discussed in Section 2.2 and denote it as posterior ESS_i. In the case that observed control data seems to be consistent with the historical data, (ie, posterior $\text{ESS}_i \cong$ prior ESS $+ n_{C_i}$), then sample size adaptation is not necessary, and the initial randomization ratio is kept until the next interim analysis. However, when there is clear data conflict between the observed control data and the historical data, (ie, posterior $ESS_i <$ prior ESS $+ n_{C_i}$), then the number of patients to be enrolled in the control arm should increase accordingly to achieve the desired level of power at the final analysis. To continue the trial, enroll another $n_{T_{i+1}} - n_{T_i} = \frac{N}{I}$ patients to the treatment arm and $n_{C_{i+1}} - n_{C_i} = \frac{N(N\text{-posterior ESS}_i)}{I(N - n_{T_i})}$ patients to the control arm, such that the randomization ratio is adjusted to $1 : \frac{N\text{-posterior ESS}_i}{N - n_{T_i}}$. It is possible to see that posterior $\text{ESS}_i >$ prior ESS $+ n_{C_i}$, and if that is the case, we suggest not to reduce the number of patients to be enrolled in the control arm in the next stage and keep the initial randomization ratio to avoid power loss and regulatory concern.

At the end of a trial, let D_C, D_T denote the complete trial data collected from the treatment and control arms, respectively, we reject the null H_0 and conclude the treatment is effective using the following criterion:

$$\Pr(\Delta \le 0|D_{H_k}, D_C, D_T, k = 1, \ldots K) > C_f, \tag{6}$$

where C_f is the probability threshold for rejecting the null at the final analysis, and similar to thresholds C_i discussed earlier, it can be calibrated through simulation to ensure desirable operating characteristics (eg, type I error rate).

In summary, the proposed BEATS design consists of the following steps (see Figure 1):

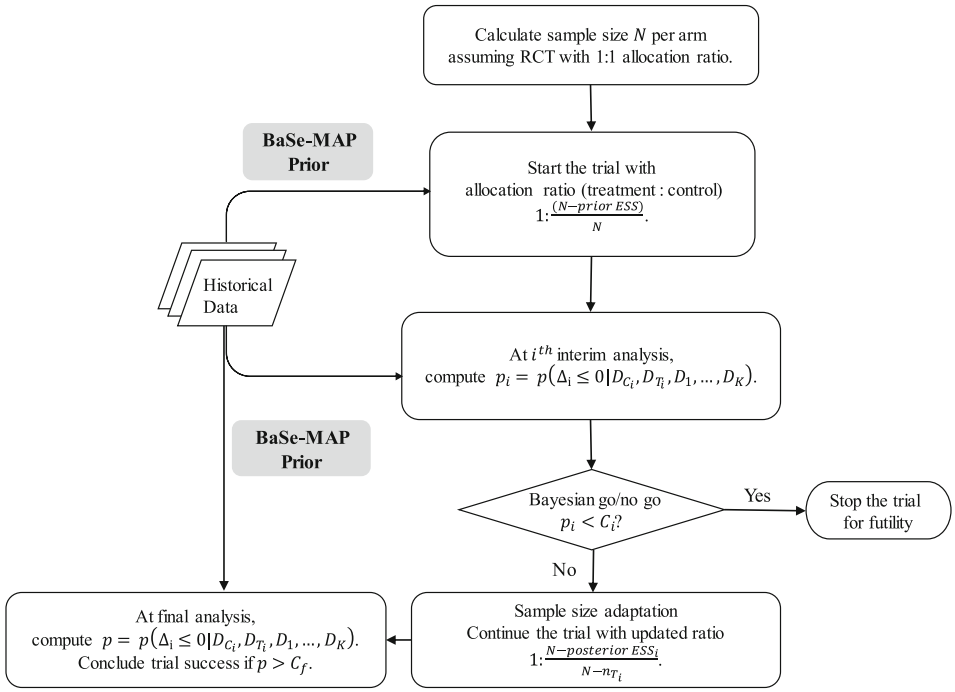

FIGURE 1 The BEATS design flowchart.

1. Calculate the sample size N patients per arm based on the standard RCT sample size calculation. Identify appropriate external/historical data sources and compute the prior ESS by applying the BaSe-MAP approach.
2. Start the trial with $1:\frac{(N\text{-prior ESS})}{N}$ randomization ratio.
3. At the *ith* interim analysis, apply BaSe-MAP to the interim control data and compute the posterior probability $p_i = \Pr\left(\Delta \le 0 | D_{H_k}, D_{C_i}, D_{T_i}, k = 1, \ldots K\right)$.
   i) If $p_i < C_i$, terminate the trial early for futility.
   ii) Otherwise, continue the trial with updated randomization ratio $1:\frac{N\text{-posterior } ESS_i}{N-n_{T_i}}$.
4. Stop the trial when there are N patients in the treatment arm and posterior $ESS_{I-1} + \left(n_{CI} - n_{C_{I-1}}\right)$ reaches N patients in the control arm, and all patients finish follow up. Compute the posterior probability $p = \Pr\left(\Delta \le 0 | D_{H_1}, \ldots, D_{H_K}, D_C, D_T\right)$. If $p > C_f$, we conclude the treatment is effective, otherwise, we conclude the investigational treatment is not superior than the control. Note that no additional sample size adaptation is performed at the final analysis, careful consideration should be given to the timing of the interim analyses to ensure that the posterior ESS_I at the final analysis closely aligns with the planned sample size of N.

3 | SIMULATION STUDY

We conduct simulation studies to investigate the operating characteristics of the proposed BEATS design in comparison with other existing methods.

3.1 | Simulation study setting

For simulation studies, we consider a two-arm randomized study comparing an investigational treatment with a control drug. The primary endpoint is time-to-event (eg, overall survival), and target hazard ratio (treatment/control) is 0.68. Assuming an exponential distribution with $\lambda_T = 0.27$ for the treatment and $\lambda_C = 0.4$ for the control, the standard sample size calculation method yields an event size of 273, or $2N = 500$ number of patients, in an RCT with 1:1 randomization and 90% power at one-sided significant level of 0.025. Moreover, we assume the trial has an accrual period of 3 years, and enrolled patients have 1 year of follow-up period. In practice, due to logistic considerations, more than one interim analysis is rarely considered. Thus, in simulation, we assume one interim analysis, and it is conducted when half of the

TABLE 1 Historical data under different simulation scenarios.

	Consistent scenarios (Sc 1-3)			Data conflict scenarios (Sc 4-6)		
Historical data k	**Number of events r**	**Total at-risk time E**	**Hazard λ**	**Number of events r**	**Total at-risk time E**	**Hazard λ**
1	137	350	0.391	161	350	0.460
2	142	350	0.406	151	350	0.431
3	145	350	0.414	168	350	0.480
4	138	350	0.394	147	350	0.420

planned event size is observed. We investigate the performance of our proposed BEATS design and compare it with the following existing designs:

1. **RCT**: a standard RCT design with 1:1 allocation ratio (ie, 250 patients in each arm) and no interim analysis.
2. **RCT-i**: a standard RCT design with 1:1 allocation ratio (ie, 250 patients in each arm) and 1 interim analysis.
3. **Hist-only**: a single-arm study with 250 patients enrolled in the treatment arm, no concurrent control, with historical studies as external control using the R-MAP approach.
4. **Hist-3:1**: an RCT study with 250 patients enrolled in the treatment arm, and 83 patients enrolled in the control arm. The control arm is augmented with information borrowed from historical data using the R-MAP approach.

Note that the threshold value $C_1 = 0.50$ is adopted for the BEATS and RCT-i design, and $C_f = 0.975$ is used for final analysis of all five designs. And the mixture weights of R-MAP prior are calibrated such that the prior ESS of R-MAP equals to or close to the number of patients in treatment arm (prior ESS = 250) in the Hist-only method and equals to 2/3 of the treatment patients (prior ESS = 167) in the Hist-3:1 method. We consider six simulation scenarios and simulate treatment and control data under different hypotheses.

- Null case (Scenarios 1 and 4): both the treatment and control data are simulated with $\lambda_T = \lambda_C = 0.40$.
- Alternative case (Scenarios 2 and 5): the treatment data is simulated with $\lambda_T = 0.27$ and the control data is simulated with $\lambda_C = 0.40$.
- Under power case (Scenarios 3 and 6): the treatment data is simulated with $\lambda_T = 0.30$ and the control data is simulated with $\lambda_C = 0.40$.

To evaluate the performance of various methods in the setting of prior-data conflict, we consider two groups of historical studies. Assume there are $K = 4$ historical trials available (see Table 1). In scenarios 1-3, we assume there is little discrepancies between the historical and control data, that is, the hazard rates of the historical trials are centered around 0.40. In practice, due to selection bias or improvement in medical care, patients in historical controls could have overall worse outcomes than a current control group.[29] Therefore, in scenarios 4-6, we assume the hazard rates of all the historical data are higher than that of concurrent control data to evaluate the robustness of the BaSe-MAP prior in the setting of prior-data conflict.

3.2 | Simulation results

Under each of the six scenarios discussed earlier, we simulate 1000 trials and compared the methods in terms of the frequentist operating characteristics including type I error, power, average sample size and bias. The simulation results are given in Table 2 and Figure 2.

Type I error or power, average sample size and percentage of early stopped trials are shown in Table 2. In scenarios 1-3, the historical data are consistent with the current control data, and less variability than the control. As a result, Hist-only, Hist-3:1 and the proposed BEATS design have lower type I error rates in comparison with RCT and RCT-i designs by leveraging historical data in scenario 1. In addition, our proposed design and RCT-i both yield early futility stopping (%ES) slightly lower than 50%, with our proposed design offering substantial sample size reduction (272 vs 380 patients). In scenarios 2 and 3, the BEATS and Hist-3:1 seem to keep comparable power with the RCT design under the alternative

TABLE 2 Operating characteristics of compared methods under six scenarios.

Scenarios	Method	Type I error/ power	Avg SS	%ES	Prior ESS	Scenarios	Type I error/ power	Avg SS	%ES	Prior ESS
1	RCT	0.025	500	-	-	4	0.025	500	-	-
	RCT-i	0.019	380	48.3%	-		0.019	380	48.3%	-
	Hist-only	0.014	250	-	251		0.035	250	-	250
	Hist-3:1	0.019	333	-	166		0.072	333	-	168
	BEATS	0.017	272	45.1%	156		0.037	310	33.2%	138
2	RCT	0.911	500	-	-	5	0.911	500	-	-
	RCT-i	0.883	495	2.1%	-		0.883	495	2.1%	-
	Hist-only	0.755	250	-	251		0.803	250	-	250
	Hist-3:1	0.891	333	-	166		0.913	333	-	168
	BEATS	0.889	352	2.6%	156		0.911	375	1.9%	138
3	RCT	0.681	500	-	-	6	0.681	500	-	-
	RCT-i	0.647	484	6.6%	-		0.647	484	6.6%	-
	Hist-only	0.502	250	-	251		0.665	250	-	250
	Hist-3:1	0.686	333	-	166		0.725	333	-	168
	BEATS	0.672	345	5.6%	156		0.707	368	5.3%	138

Note: Avg SS denotes average sample size across 1000 simulated trials, and %ES denotes the percentage of trials stopped early due to futility.

and under power scenarios. The Hist-only design with R-MAP prior tends to overborrow and has the lowest power in both scenarios.

Scenarios 4-6 allows for greater heterogeneity between the historical trials and the control data, and the results show clear advantages of the BEATS design with the robust BaSe-MAP prior. The Hist-3:1 method has higher power in scenarios 5 and 6 at the cost of inflated type I error rate (7.2%) in scenario 4. In contrast, the BEATS design shows comparable power with the RCT without significantly inflate the type I error rate (3.7%), which indicates that the BaSe-MAP approach is robust to prior data conflict. This trend can also be demonstrated by the higher calibrated power of the BEATS design than the other methods, which is discussed in the next section. In all six scenarios, the BEATS design demonstrates significantly sample size reduction in comparison with the standard RCT without sacrificing the frequentist operating characteristics.

Table 2 also reports the prior ESS measured from the R-MAP prior in Hist-only and Hist-3:1 and BaSe-MAP prior in the BEAT design across different scenarios. Note that the prior ESS in Hist-only and Hist-3:1 are intentionally fixed by tuning the mixture weight parameter in R-MAP prior to achieve the desired control sample size at the end of study. On the other hand, the BaSe-MAP prior does not require a pre-specified weight parameter, and it can borrow more when the historical trials are homogeneous and borrow less in case of prior data conflict. This can be shown in the decreased prior ESS in scenarios 4-6 compared to scenarios 1-3 (138 vs 156).

Figure 2 shows the bias of the estimated treatment effect across all methods. In scenarios 1-3, the BEATS and RCT-i methods seem to have very similar bias, whereas the BEATS design has significantly smaller sample size. Note that the bias of BEATS and RCT-i are greater in magnitude than the RCT, Hist-only, Hist-3:1 method in scenario 1 (null case) due to high percentage of early futility stopping. In the data-conflict scenarios 4-6, our proposed method has greatly reduced the bias among the three designs with historical borrowing, which again shows that the BaSe-MAP prior along with the sample size rebalance is able to appropriately leverage the historical information in case of prior data conflict. The bias and MSE results are also summarized in Appendix B.

3.3 | Calibrated power

We further evaluate the performance of different methods using calibrated power to account for the trade-off between type I error and power. To calculate calibrated power, we find the final cutoff C_f for each method such that the type I error

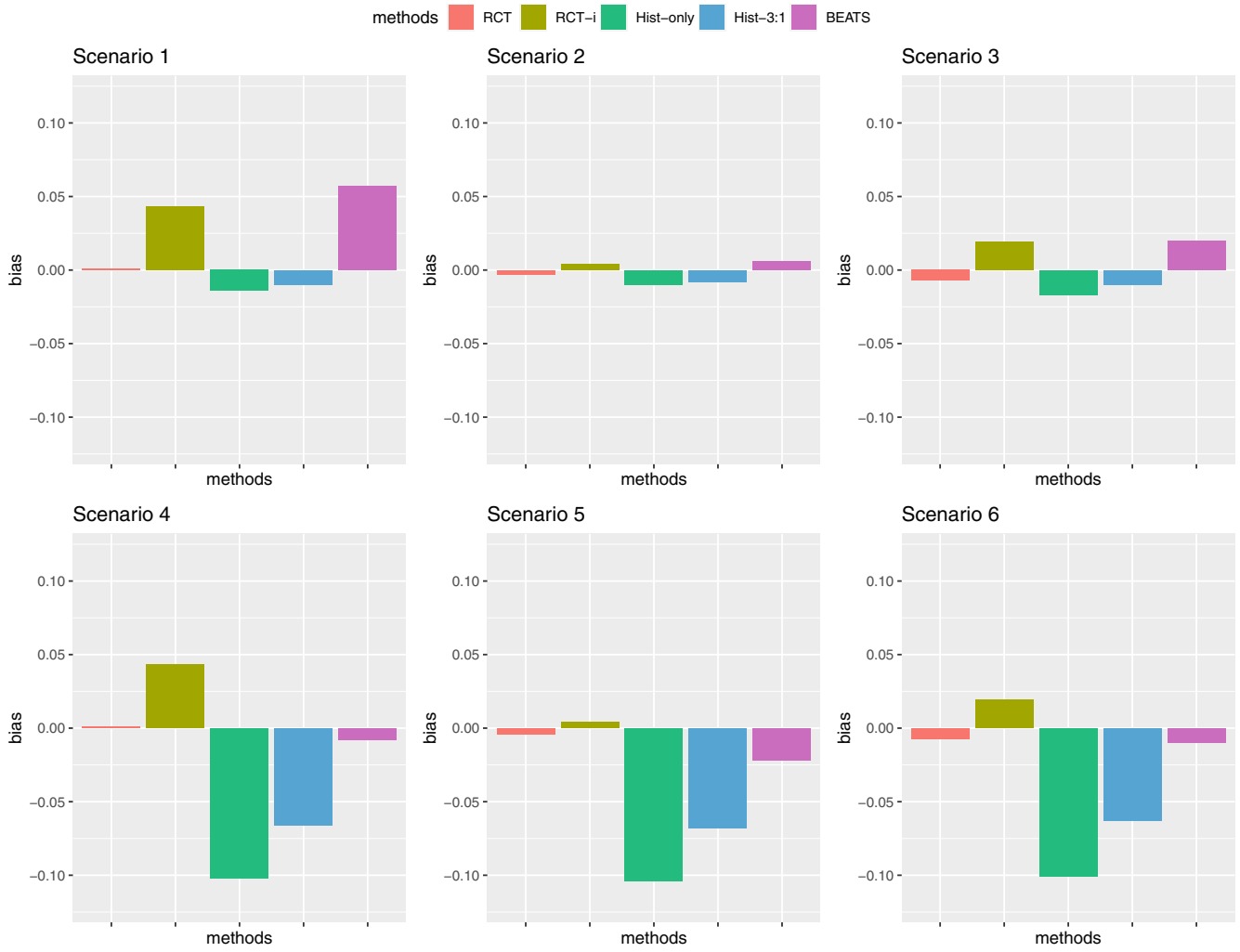

FIGURE 2 Simulation comparison via bias of the estimated treatment effect under six scenarios.

rate is fixed at approximately 2.5%, then the power for each method is evaluated using the same C_f under the alternative. The result for the calibrated power and average sample size is shown in Table 3.

Among all scenarios (2, 3, 5, and 6), the proposed BEATS design has higher calibrated power compared to Hist-only and Hist-3:1 designs, illustrating that our design is more robust to heterogeneity than other methods. Furthermore, the BEATS design has the highest calibrated power in scenarios 2 and 3, and comparable power with the RCT and RCT-i designs in scenarios 5 and 6, which suggests that our proposed design could dramatically reduce the sample size while maintaining desirable operating characteristics.

4 | CASE STUDY IN HER2-LOW ADVANCED BREAST CANCER

In this section, we illustrate the implementation of the BEATS design in a hypothetical randomized phase III trial involving patients with low expression of human epidermal growth factor receptor (HER2-low) metastatic breast cancer. Treating the HER2-low breast cancer has been a known challenge in the oncology community, and data shows that at least 55% of patients with breast cancer fall into the HER2-low category.[30]

The aim of this hypothetical two-armed randomized study is to compare a new treatment (T) against physician's choice as the control arm (C). The primary endpoint of the phase III study is overall survival (OS), defined as the time from randomization to death. For the control arm, the median OS is assumed to be 16.8 months (hazard rate 0.495) based on a recently published study.[31] Assuming a 28% reduction in OS (HR = 0.72) for the new treatment (hazard rate 0.355),

TABLE 3 Simulation comparison via calibrated power and average sample size.

	Method	Calibrated power	Average sample size
Scenario 2	RCT	0.911	500
	RCT-i	0.903	496
	Hist-only	0.836	250
	Hist-3:1	0.928	333
	BEATS	0.957	352
Scenario 3	RCT	0.681	500
	RCT-i	0.665	486
	Hist-only	0.697	250
	Hist-3:1	0.818	333
	BEATS	0.841	348
Scenario 5	RCT	0.911	500
	RCT-i	0.903	496
	Hist-only	0.790	250
	Hist-3:1	0.838	333
	BEATS	0.885	375
Scenario 6	RCT	0.681	500
	RCT-i	0.665	486
	Hist-only	0.607	250
	Hist-3:1	0.632	333
	BEATS	0.670	368

TABLE 4 Observed data of four hypothetical historical studies.

Historical studies	Median OS (months)	Number of events	Total at-risk times
1	16.5	126	250
2	16.1	155	300
3	15.8	184	350
4	15.4	216	400

a 1:1 randomization, a one-sided 2.5% level of significance and 90% power, the study would require $2N = 686$ patients with a 2-year accrual period and 1-year follow-up period.

Assume that the sponsor decided to leverage historical data to augment the control arm of the study to speed up the development of the investigational treatment and protect patients from disease progression. Also assume that after an extensive research, four historical trials for the active control are identified by the clinical team. In general, time-to-event endpoint summary of a trial is readily available from Kaplan-Meier plot of the published literature, various methods have been proposed to extract the summary data from the Kaplan-Meier plots.[32,33] Table 4 shows the observed number of events and exposure time of the four hypothetical trials.

Suppose one interim analysis is planned ($I = 2$) and after evaluation by the clinical team and simulation calibration, threshold values $C_1 = 0.6$ and $C_f = 0.975$ are specified for the study. In addition, the prior ENE and ESS of the BaSe-MAP prior derived from four historical trials are 123 and 201, respectively, given assumed control hazard of 0.495. The trial starts with an initial randomization ratio, $1 : \frac{(N\text{-prior ESS})}{N} \approx 5 : 2$, and $n_{T_1} = 172$ patients and $n_{C_1} = 71$ patients are allocated to the treatment and control arm, respectively.

At interim, suppose 28 events in the treatment arm and 11 events in the control arm are observed. According to the BEATS design, we apply the BaSe-MAP approach to the interim control data and check for early futility stopping by

computing the posterior probability that log hazard ratio is less than 0, p_1, given in Equation (5). Since $p_1 = 0.821$, which is greater than the futility early stopping threshold 0.6, the trial continues to the second stage with sample size adaptation.

Based on the proposed sample size adaptation process, the posterior ENE_1 is first computed for the updated BaSe-MAP prior with interim data as discussed in Section 2.2, and the posterior ESS_1 is generated with the interim observed hazard for the control arm. In this case, the posterior ENE_1 is 123 and ESS_1 is 211. Note that the posterior ESS_1-$n_{C_1} = 140$ is less than the prior ESS of 201, which implies that the prior ESS is over-estimated, and additional patients are needed in the control arm. Therefore, in the second stage, the randomization ratio is adjusted to $1 : \frac{N\text{-posterior } ESS_1}{N - n_{T_1}} \approx 4 : 3$. And $n_{T_2} = 172$ patients are allocated to the treatment arm, and $n_{C_2} = 132$ patients, increased from 71 patients in the first stage, are allocated to the control arm. Suppose at the end of the trial, 187 events out of 343 patients and 113 events out of 203 patients are observed in the treatment and control arm, respectively. The BaSe-MAP method is applied with the completed control arm and the posterior probability is computed. In this case, we obtained $Pr\left(\Delta < 0 | D_{H_1}, \ldots, D_{H_4}, D_C, D_T\right) = 0.984$, which is greater than $C_f = 0.975$. Hence, a significant treatment effect of the investigational treatment can be concluded. The final estimated hazard of the control arm is 0.505 (95% credible interval: [0.453, 0.556]), and the hazard of the treatment arm is 0.345 (95% credible interval: [0.297, 0.400]).

To compare with the proposed design BEATS, we have also implemented a Hist-3:1 design with the R-MAP prior. We have selected $w = 0.88$ to obtain a prior ESS of 232, thus we enroll 111 patients in the control arm and 343 patients in the treatment arm (3:1 randomization). The result is $Pr\left(\Delta < 0 | D_{H_1}, \ldots, D_{H_4}, D_C, D_T\right) = 0.998$. The estimated hazard of the control arm in this case is 0.539 (95% credible interval: [0.480, 0.633]) and the treatment hazard is 0.341 (95% credible interval: [0.292, 0.416]). Compared to our proposed design, we see that the treatment benefit is clearly overestimated in Hist-3:1 design.

Additionally, the case study demonstrates that the proposed BEATS design substantially reduces the required sample size of the concurrent control when compared to an ideal RCT design with 1:1 randomization. In this case study, the BEATS design only requires 546 patients, which is 140 fewer than the standard 1:1 RCT design. This benefits patients by enabling expedited access to reliable innovative treatment options. Moreover, the BEATS design clearly reduces the impact of conflicting historical data, surpassing the performance of a design utilizing R-MAP prior and a fixed randomization ratio.

5 | DISCUSSION

With the rapidly increasing availability of historical data and RWD, many novel methods have been proposed to leverage these external data for more efficient and ethical clinical trial design. While most designs focus on borrowing historical data at analysis stages, we have proposed a Bayesian hybrid design, the BEATS design, on time-to-event endpoints that adaptively borrow information from external data based on the BaSe-MAP prior method during both design and interim analyses stages. After the initial calibration of concurrent control arm sample size based on the heterogeneity from external data, the BEATS design comprises interim analyses, which make the "go/no-go" decisions based on a Bayesian early futility stopping criterion. If the decision is "go", we re-evaluate the sample size needed with the adaptation method for the control arm. Hence, in the final analysis, our trial mimics an ideal RCT design with 1:1 randomization, with the control arm augmented by information from the external data.

As observed in the simulation study, BEATS design outperforms other hybrid designs using the R-MAP prior by demonstrating controlled type I error rate, minimum bias, optimal power and greatest calibrated power especially in the presence of prior data conflict. Further, compared to standard RCT designs, the proposed method offers significant sample size reduction with comparable operating characteristics.

The proposed design requires user input cutoff values C_i and C_f. Simulation studies can be performed to calibrate these values by evaluating the operating characteristics under a range of threshold values. The choice of C_i should strike a balance between ensuring sufficient evidence of treatment efficacy while reducing the risk of prematurely stopping the trial. A higher value of C_i (eg, >0.6) may create a higher bar for passing interim analyses, ensuring more stringent evidence of efficacy. The choice of C_f at the final analysis should consider the tradeoff between power and false positive rate. A lower value of C_f increases the power but may also increase the risk of a false positive decision. Conversely, a higher value of C_f reduces the power but also decreases the risk of false positives. Ultimately, the determination of the cutoff values should be made case by case based on a comprehensive evaluation of the study objectives, regulatory requirements, and clinical considerations.

At the trial design phase of hybrid controlled trials, it is essential to carefully select the appropriate external data sources to reduce the heterogeneity among the historical data themselves as well as between the historical data and the control data of the current trial. Nevertheless, due to the continual advancements in the standard of care, it is likely that the current control shows improved benefit compared to the historical data. Consequently, a limitation of hybrid controlled trials is the potential inflation of the type I error rate when inconsistencies arise between the historical trials and the current control arm. However, our proposed design offers a potential remedy through the use of BaSe-MAP prior and adaptive sample size adjustments at interim analyses, effectively mitigating the inappropriate borrowing of information. Moreover, additional considerations such as conducting sensitivity analyses are crucial for assessing the robustness and generalizability of the results derived from the incorporation of external data.

One possible extension to our proposed design is to include patient-level covariate information. Methods such as the PS-MAP prior[20] can be used to better select the patients' subpopulation in historical data that are similar to patients in the current control. Other extensions include the consideration of surrogate endpoints to allow accelerated approval from regulatory agencies.[34,35] Our proposed design could also be extended to pediatric trial development to control the sample size while leveraging the adult data appropriately.

ORCID

Jianchang Lin https://orcid.org/0000-0002-9123-0690
Rachael Liu https://orcid.org/0000-0001-6033-4510

REFERENCES

1. Hariton E, Locascio JJ. Randomised controlled trials—the gold standard for effectiveness research: study design: randomised controlled trials. *BJOG*. 2018;125(13):1716.
2. Hey SP, Kimmelman J. The questionable use of unequal allocation in confirmatory trials. *Neurology*. 2014;82(1):77-79.
3. Avins AL. Can unequal be more fair? Ethics, subject allocation, and randomised clinical trials. *J Med Ethics*. 1998;24(6):401-408.
4. Dumville JC, Hahn S, Miles JN, Torgerson DJ. The use of unequal randomisation ratios in clinical trials: a review. *Contemp Clin Trials*. 2006;27(1):1-12.
5. Gupta S, Jain S, Yeh J, Guddati AK. Unequal allotment of patients in phase III oncology clinical trials. *Am J Cancer Res*. 2021;11(7):3735-3741.
6. Torgerson D, Campbell M. Unequal randomisation can improve the economic efficiency of clinical trials. *J Health Serv Res Policy*. 1997;2(2):81-85.
7. United States Food and Drug Administration. International Conference on Harmonization (ICH) E10: choice of control group and related issues in clinical trials. https://database.ich.org/sites/default/files/E10_Guideline.pdf. Accessed July 2000.
8. United States Food and Drug Administration. Real-world data: assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products. https://www.fda.gov/media/152503/download. Accessed September 2021.
9. United States Food and Drug Administration. Framework for FDA's real-world evidence program; 2018.
10. Ibrahim JG, Chen MH. Power prior distributions for regression models. *Stat Sci*. 2000;15(1):46-60.
11. Duan Y, Ye K, Smith EP. Evaluating water quality using power priors to incorporate historical information. *Environmetrics*. 2005;17(1):95-106.
12. Neuenschwander B, Branson M, Spiegelhalter DJ. A note on the power prior. *Stat Med*. 2009;28(28):3562-3566.
13. Banbeta A, Rosmalen JV, Dejardin D, Lesaffre E. Modified power prior with multiple historical trials for binary endpoints. *Stat Med*. 2019;38(7):1147-1169.
14. Hobbs BP, Carlin BP, Mandrekar SJ, Sargent DJ. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*. 2011;67(3):1047-1056.
15. Neuenschwander B, Capkun-Niggli G, Branson M, Spiegelhalter DJ. Summarizing historical information on controls in clinical trials. *Clin Trials*. 2010;7(1):5-18.
16. Spiegelhalter DJ, Abrams KR, Myles JP. Prior distributions. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. New York: Wiley; 2004:139-180.
17. Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Anal*. 2006;1(3):515-533.
18. 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):1023-1032.
19. Hupf B, Bunn V, Lin J, Dong C. Bayesian semiparametric meta-analytic-predictive prior for historical control borrowing in clinical trials. *Stat Med*. 2021;40:3385-3399.
20. Liu M, Bunn V, Hupf B, Lin J, Lin J. Propensity-score-based meta-analytic predictive prior for incorporating real-world and historical data. *Stat Med*. 2021;40(22):4794-4808.
21. Zhang W, Pan Z, Yuan Y. A Bayesian group sequential design for randomized biosimilar clinical trials with adaptive information borrowing from historical data. *J Biopharm Stat*. 2022;32(3):359-372.

22. Li C, Ferro A, Mhatre SK, et al. Hybrid-control arm construction using historical trial data for an early-phase, randomized controlled trial in metastatic colorectal cancer. *Commun Med*. 2022;2:90.
23. Roychoudhury S, Neuenschwander B. Bayesian leveraging of historical control data for a clinical. *Stat Med*. 2020;39:984-995.
24. Zhang H, Shen Y, Chiang AY, Li J. An empirical Bayes robust meta-analytical-predictive prior to adaptively leverage external data. arXiv preprint arXiv:2109.10237, 2021.
25. Polson NG, Scott JG. On the half-cauchy prior for a global scale parameter. *Bayesian Anal*. 2012;7(4):887-902.
26. Neuenschwander B, Weber S, Schmidli H, O'Hagan A. Predictively consistent prior effective sample sizes. *Biometrics*. 2020;76(2):578-587.
27. Morita S, Thall PF, Müller P. Determining the effective sample size of a parametric prior. *Biometrics*. 2008;64(2):595-602.
28. Weber S. RBesT: R Bayesian evidence synthesis tools; 2020. https://CRAN.R-project.org/package=RBesT
29. Ghadessi M, Tang R, Zhou J, et al. A roadmap to using historical controls in clinical trials—by drug information association adaptive design scientific working group (DIA-ADSWG). *Orphanet J Rare Dis*. 2020;15:69.
30. Tarantino P, Hamilton E, Tolaney SM, et al. HER2-low breast cancer: pathological and clinical landscape. *J Clin Oncol*. 2020;38(17):1951-1962.
31. Modi S, Jacot W, Yamashita T, et al. Trastuzumab Deruxtecan in previously treated HER2-low advanced breast cancer. *N Engl J Med*. 2022;387:9-20.
32. Guyot P, Ades A, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol*. 2012;12:9.
33. Liu N, Zhou Y, Lee JJ. IPDfromKM: reconstruct individual patient data from published Kaplan-Meier survival curves. *BMC Med Res Methodol*. 2021;21:111.
34. Li Q, Lin J, Liu M, Wu L, Liu Y. Using surrogate endpoints in adaptive designs with delayed treatment effect. *Stat Biopharm Res*. 2022;14(4):661-670.
35. Fleming TR. Surrogate endpoints and FDA's accelerated approval process. *Health Aff (Millwood)*. 2005;24(1):67-78.

How to cite this article: Bi D, Liu M, Lin J, Liu R. BEATS: Bayesian hybrid design with flexible sample size adaptation for time-to-event endpoints. *Statistics in Medicine*. 2023;42(30):5708-5722. doi: 10.1002/sim.9936

APPENDIX A. CONVERSION FROM ENE TO EQUIVALENCE SAMPLE SIZE

Suppose patients are enrolled to the current study uniformly over an accrual period A units of time, and there is a follow-up period for another F units of time. We assume an exponential distribution with a hazard rate, $\tilde{\lambda}$, for the time-to-event outcome, then the relationship between number of events, denoted d, and the number of patients, denoted n, is given by,

$$\begin{aligned} d = nPr(\text{event};\tilde{\lambda}) &= n\left(1-\frac{1}{A}\int_0^A S(a+F)da\right) \\ &= n\left(1-\frac{1}{\tilde{\lambda}A}\left[e^{-\tilde{\lambda}F}-e^{-\tilde{\lambda}(A+F)}\right]\right) \end{aligned} \tag{A1}$$

.

Therefore, we can calculate the equivalent sample size (ESS) given ENEs (prior or posterior) as,

$$\text{ESS} = \frac{\text{ENE}}{1-\frac{1}{\tilde{\lambda}A}\left[e^{-\tilde{\lambda}F}-e^{-\tilde{\lambda}(A+F)}\right]}. \tag{A2}$$

When computing the posterior ESS_i, we use Equation (A2), with $\tilde{\lambda}$ being the posterior $\hat{\lambda}_C = \log\left(\hat{\theta}_C\right)$ estimate at interim to reduce the potential bias if the control hazard rate we used at planning stage is different from the control hazard rate of the current trial.

APPENDIX B. SIMULATION COMPARISON VIA BIAS AND MSE OF ESTIMATED TREATMENT EFFECT

Scenario	Method	Bias*100	MSE*100	Scenarios	Bias*100	MSE*100
1	RCT	0.1	2.6	4	0.1	2.6
	RCT-i	4.3	4.4		4.3	4.4
	Hist-only	−1.4	6.7		−10.2	6.5
	Hist-3:1	−1.0	2.8		−6.6	4.1
	BEATS	5.7	3.0		−0.8	4.0
2	RCT	−0.3	3.0	5	−0.3	3.0
	RCT-i	0.4	3.8		0.4	3.8
	Hist-only	−1.0	8.8		−10.4	5.9
	Hist-3:1	−0.8	3.7		−6.8	3.6
	BEATS	0.6	2.7		−2.2	2.1
3	RCT	−0.7	2.9	6	−0.7	2.9
	RCT-i	1.9	3.8		1.9	3.8
	Hist-only	−1.7	6.6		−10.1	6.4
	Hist-3:1	−1.0	3.7		−6.3	3.9
	BEATS	2.0	4.0		−1.0	3.4