The New England Journal of Statistics in Data Science Volume 2, 54–71 (2024)
DOI: https://doi.org/10.51387/23-NEJSDS48

Bayesian Interim Analysis in Basket Trials

Cheng HUANG[1,3], Chenghao CHU[2], Yimeng LU[3], Bingming YI[2], and Ming-Hui CHEN[1,*]

Abstract

Basket trials have captured much attention in oncology research in recent years, as advances in health technology have opened up the possibility of classification of patients at the genomic level. Bayesian methods are particularly prevalent in basket trials as the hierarchical structure is adapted to basket trials to allow for information borrowing. In this article, we extend the Bayesian methods to basket trials with treatment and control arms for continuous endpoints, which are often the cases in clinical trials for rare diseases. To account for the imbalance in the covariates which are potentially strong predictors but not stratified in a randomized trial, our models make adjustments for these covariates, and allow different coefficients across baskets. In addition, comparisons are drawn between two-stage design and one-stage design for the four Bayesian methods. Extensive simulation studies are conducted to examine the empirical performance of all models under consideration. A real data analysis is carried out to further demonstrate the usefulness of the Bayesian methods.

keywords and phrases: Bayesian hierarchical model, Calibrated Bayesian hierarchical model, Covariates adjustment, Mixture of finite mixtures, One-minimum power.

1. INTRODUCTION

The recent advancement in genomics sequencing and molecular biology has enabled a detailed classification of patients based on genomics alterations or molecular profiles, thus inspiring the establishment of targeted therapies [1, 5] and precision medicine [2, 3, 14]. U.S. Food and Drug Administration (FDA) provides the *master protocol* [29, 11] guidance that facilitates the methodological innovation that coordinates efforts to investigate treatments in more than one patient population or disease types within one master protocol [42, 30, 13]. Among the master protocols, basket trials [28, 25, 13] have become a popular design since it allows the evaluation of an investigational treatment in multiple disease cohorts in parallel and hence expedite the efficiency of clinical research. U.S. Food and Drug Administration (FDA)[38] defines a basket trial protocol as follows:

> "A master protocol designed to test a single investigational drug or drug combination in different populations defined by different cancers, disease stages for a specific cancer, histologies, number of prior therapies, genetic or other biomarkers, or demographic characteristics is commonly referred to as a *basket trial*."

Basket trials are typically conducted in a phase II trial to provide preliminary proof-of-concept evidence for clinical validation. Some of the other unique purposes of basket studies and their application examples are discussed by Cunanan et al. [8] and Tao et al. [35]. Basket trials are predominantly designed for oncology studies. This design is utilized in [34], [16], [15], [22], [27] and [41], to name a few.

In a basket trial, patients from different baskets may be expected to have similar responses because they may share common features, such as disease stages or molecular alterations. This provides the basis for information borrowing across baskets, which is one of the key statistical advantages of basket trials. Note that different information borrowing approaches can be applied. As a well-known example, Vemurafenib [12], initially approved for melanoma with V600E BRAF mutations, was approved by the US Food and Drug Administration (FDA) for treating BRAF V600 mutation-positive Erdheim-Chester disease (ECD). The BRAF basket trial used the Simon Two-Stage design [32] that treated the responses of each basket independently as they were from separate studies and hence no information was borrowed across baskets. In another example, Vitrakvi was granted accelerated approval for treating locally advanced or metastatic solid tumors harboring a neurotrophic tyrosine kinase receptor (NTRK) gene fusion. The Vitrakvi basket trial used full information borrowing by pooling data together across all baskets, assuming that the response to the drug was homogeneous across baskets. These are the two extreme strategies for information borrowing. With the emergence of basket trials, FDA issued the general guidance on grouping strategies, stating FDA's position on the generalizability of the results of multiple basket studies [37] and providing guidance on the overall design of master protocols, as exemplified by the BRAF V600 basket trial [38]. An excellent review and discussion on the guidance of basket trials can be found from [31]. A wide range of statistical

*Corresponding author.
[1]Department of Statistics, University of Connecticut.
[2]Department of Biostatistics, Vertex Pharmaceuticals.
[3]Department of Biostatistics, Vir Biotechnology.

research has been devoted to exploring more flexible methods of information borrowing. These methods dynamically or partially borrow information across baskets based on the trial data.

One class of methods attempts to borrow across all baskets. A Bayesian sequential monitoring method is presented by Simon et al. [33]. It involves multiple interim looks, with the amount of data borrowing at each interim determined by the posterior probability of homogeneity. Thall et al. [36] and Berry et al. [4] proposed the basket trial design using the full Bayesian Hierarchical Model (BHM), with the treatment effects of the baskets modeled by a distribution $\mathcal{F}$. Due to a small number of baskets in actual basket trials, the variance in $\mathcal{F}$ is difficult to be reliably estimated, leading to potentially substantial Type I error inflation [9]. To circumvent this issue, Chu and Yuan [6] suggested a Calibrated Bayesian Hierarchical model (CBHM) that provides a more reliable estimate of the variance in $\mathcal{F}$ by pre-determine the shrinkage parameter. Their results suggest that Type I errors are better controlled than BHM, especially for the cases where the effect of the basket is heterogeneous.

There is another class of methods, in which similar baskets are first clustered and the information among similar baskets is fully borrowed. Unlike the BHM model which assumes the treatment effect for each basket is exchangeable and localizes around a common mean, the exchangeability-nonexchangeability (EXNEX) model of Neuenschwander et al. [21] assumes that some baskets may be predefined to be exchangeable (EX), while some may not be exchangeable (NEX). Thus, each basket is assigned to either an EX component, which allows within component parameter to be partially exchangeable; or an NEX component which is nonexchangeable from others. The assignment is based on pre-specified probabilities. Chu and Yuan [7] and Zhou and Ji [45] proposed a more sophisticated method, in which a latent class variable is employed to group baskets into clusters and hence avoid pre-specifying the aforementioned probabilities in the EXNEX model. According to their approach, the treatment effects within each cluster are assumed to be centralized such that information is borrowed locally using the BHM; a Dirichlet distribution prior is applied to the weighting probability that assigns the baskets into the clusters; and the number of latent clusters and the basket memberships are inferred by the data through a Dirichlet Process Mixture Model (DPMM) model [19, 20]. The Mixture models have a common difficulty in choosing the number of clusters. Within the Bayesian framework, one could consider the number of clusters as an unknown parameter and specify it with a prior distribution. This kind of model is referred to as the Mixture of Finite Mixtures (MFM). Miller and Harrison [18] proved many characteristics of the DPMM are also exhibited by MFM. By applying the Markov chain Monte Carlo (MCMC) sampling algorithm similar to DPMM but with feasible alteration, this sampling algorithm exhibits higher efficiency than the reversible jump technique. Geng and Hu [10] applied MFM onto basket trials for binary endpoint.

Basket trials can have a variety of designs. First of all, it may or may not have the concurrent control arms [26]. It can also be applied for various types of endpoints: a continuous outcome design is presented by Zheng and Wason [44] and a time to event design is presented by Xu et al. [43]. Some settings also include the multiple covariates [44, 24]. In prospective circumstances, the treatment effect of each basket may depend on several therapeutic indicating covariates, and the coefficients of one covariate can vary across baskets. Interim Analyses is also popular when designing clinical trials. In this article, we primarily focus on the evaluation of a few Bayesian approaches, including Bayesian Hierarchical Model (BHM), Calibrated Bayesian Hierarchical Model (CBHM), and Mixture of Finite Mixtures (MFM), with extended the assumptions: (i) continuous endpoints; (ii) with a concurrent controlled arm; (iii) interim analysis designs and (iv) adjusting covariates effects across different baskets.

The trials to which the Bayesian approaches are applied are mostly without a concurrent placebo control. In oncology trials, non-controlled trials are usually adopted due to ethical reasons. But in many other situations, for example, when (i) side-effects and tumor size shrinkage are limited so blindness can be maintained, and (ii) patients receive the investigational treatment and placebo in addition to a current curative treatment so there are no ethical concern, randomized trials are advantageous to provide more definitive answer to treatment effects and safety to the investigational product. To emphasize the importance and the principles of estimating treatment effects and their sensitivity analysis, the US Food and Drug Administration (FDA) released "E9(R1) Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials"[39]. This implies that assessing the magnitude of the treatment effect potentially will be of greater concern in many trial designs. As such, this article explores the performance of Bayesian methods in basket trials with a control arm and continuous endpoint. However, these methods can be readily extended to different designs.

The remaining part of this article is organized as follows. The trial setting and assumptions are described in Section 2. In Section 3, the considered Bayesian methods are introduced and discussed in detail. To assess and compare the performance of these methods, simulation studies and results are presented in Section 4. An application example is provided in Section 5. Conclusion and possible extension of the current work are addressed in Section 6. The additional results, sensitivity analysis and technical details including derivations are given in the Supplementary Material.

2. TRIAL SETTING

Consider a controlled basket trial with continuous endpoints and with K baskets. If interim analysis is planned and the study has multiple stages, for simplicity it is as-

sumed that equal number of subjects are enrolled in each stage. For $k = 1, \ldots, K$ baskets, let n_{ck} and n_{tk} denote the control and treatment arm sample sizes for basket k in each stage, respectively. Let Y denote the outcomes, with subscripts c and t denote the control arm and treatment arm. Write $i = 1, \ldots, n_{ck}$ and $i' = 1, \ldots, n_{tk}$ as individual subjects in the control and treatment groups. In addition, the outcomes are assumed to be normally distributed, and the subject level variation is the same across the whole population, denoted by σ^2.

The majority of the literature discussed information borrowing methods in basket trials without considering the possible diversity of patients across the baskets. It is reasonable to speculate that the outcomes of basket trials may depend on the participants' baseline characteristics, such as gender and age. Also, it is known that randomization is a common feature of clinical trials; however, complete randomization may not be achievable in reality for a variety of reasons. U.S. Food and Drug Administration [40] provides general guidance on which covariates should be included in the modeling of a clinical trial: including the adjustment of stratification factors for randomization and important baseline characteristics that are strong predictors for the outcome. In such cases, adjustments for covariates are necessary, as discussed in [44] and [43]. In this paper, we also model the outcomes by including covariates effects, with the following assumptions and notation.

Assume the outcome depends on r covariates. In the k-th basket, denote τ_k as the treatment effect. For the k-th basket, the covariates vectors are $\boldsymbol{x}_{cki}$ and $\boldsymbol{x}_{tki'}$, respectively, for individual i in the control arm and for individual i' in the treatment arm, where $\boldsymbol{x}_{cki} = (1, x_{cki1}, \ldots, x_{ckir})'$ for the control arm and $\boldsymbol{x}_{tki'} = (1, x_{tki'1}, \ldots, x_{tki'r})'$ for the treatment arm are $(r+1) \times 1$ vectors. The coefficient vectors $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \ldots, \beta_{kr})'$ are assumed to be different among baskets. Note the "β_{k0}" is interpreted as the control group effect for k-th basket (after adjusting covariates). Denote $\{(\boldsymbol{x}_{ck1}, y_{ck1}), \ldots, (\boldsymbol{x}_{ckn_{ck}}, y_{ckn_{ck}})\}$ and $\{(\boldsymbol{x}_{tk1}, y_{tk1}), \ldots, (\boldsymbol{x}_{tkn_{tk}}, y_{tkn_{tk}})\}$ for $k = 1, \ldots, K$ to be the data observations. The outcomes are assumed to follow

$$\begin{aligned} &Y_{cki}|\boldsymbol{\beta}_k, \sigma^2 \overset{iid}{\sim} N(\boldsymbol{x}'_{cki}\boldsymbol{\beta}_k, \sigma^2), \\ &\text{with} \quad k = 1, \ldots, K; \quad i = 1, \ldots, n_{ck}. \\ &\text{and independently,} \\ &Y_{tki'}|\boldsymbol{\beta}_k, \tau_k, \sigma^2 \overset{iid}{\sim} N(\boldsymbol{x}'_{tki'}\boldsymbol{\beta}_k + \tau_k, \sigma^2), \\ &\text{with} \quad k = 1, \ldots, K; \quad i' = 1, \ldots, n_{tk}. \end{aligned} \tag{2.1}$$

If there is no covariate that need to be adjusted, the above formulae will be simplified with $\boldsymbol{x}_{cki} = 1$, $\boldsymbol{x}_{tki'} = 1$ and the coefficients $\boldsymbol{\beta}_k = \beta_{k0}$.

The main objective is to detect whether the treatment is superior to the control in any of the K baskets. This corresponds to testing the following hypotheses:

$$H_0 : \tau_k = \delta \text{ versus } H_1 : \tau_k > \delta, \tag{2.2}$$

for $k = 1, \ldots, K$. The value of δ represents the magnitude of improvement over the control arm needed to declare a clinical benefit of the new treatment [44] and needs to be specified in advance. Implicitly, a positive τ_k indicates the drug or treatment is effective.

We are piloting two study designs: a one-stage design without any Interim Analysis (IA) and a two-stage design with one IA. Let "$\mathcal{D}$" denote the available data collected at the time point of the analysis. Let $\mathcal{D}_{IA}$ denote all available data at the interim analysis within the two-stage design; and $\mathcal{D}_{FA}$ denote all available data at the Final Analysis (FA). For trials with the two-stage design, we adopt similar decision rule proposed by Mehta and Pocock [17]: at the IA, a basket could stop early for futility or efficacy, or enter the "promising zone": meaning the recruitment continues and it is evaluated in the next stage. In the final stage, a basket could be claimed as futility or efficacy.

The inference is performed using the posterior probability based on the data available at the current stage, i.e.

- $Pr(\tau_k > \delta|\mathcal{D}_{IA})$ for Interim Analysis;
- $Pr(\tau_k > \delta|\mathcal{D}_{FA})$ for Final Analysis.

Two cut-off values $0 < q_1 < q_2 < 1$ are pre-set for futility stopping and efficacy stopping, respectively. Below describes the decision criteria for the stopping rules:

1. At IA,

- If $Pr(\tau_k > \delta|\mathcal{D}_{IA}) \leq q_1$, the basket is claimed ineffective and stops for futility;
- If $Pr(\tau_k > \delta|\mathcal{D}_{IA}) > q_2$, the basket is claimed effective and stops for efficacy;
- Otherwise (i.e. $q_1 < Pr(\tau_k > \delta|\mathcal{D}_{IA}) \leq q_2$) the basket enters the "promising zone": continues accrual and enters the next stage;

2. At FA,

- If $Pr(\tau_k > \delta|\mathcal{D}_{FA}) \leq q_1$ the basket is claimed ineffective;
- Otherwise the basket is claimed to be effective.

There criteria are illustrated graphically in Figure 1.

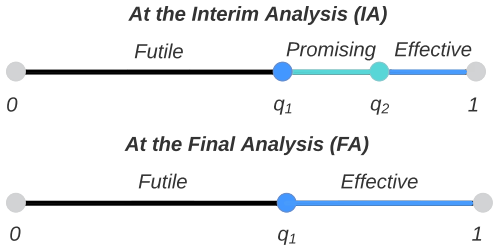

Figure 1: Stopping Criteria.

There are various methods to determine the cut-off values q_1 and q_2. Usually, they shall imply the percentage level required to certify that the treatment is compellingly improved over the control. In our simulation study, we link

q_1 and q_2 by a parameter $0 \leq \eta < 1$: $1 - q_2 = \eta(1 - q_1)$. η is essentially the ratio between the "effective zone" and the "effective zone" plus "promising zone", thus controlling alpha-splitting in the two stages. Specially, when $\eta = 0$, there is no chance claiming success at IA: only futility rule is applied. For η between 0 and 1, there will be a "promising zone" and both efficacy and futility rules are applied. Further, the cut-off values are determined by allowing the trial to comply with certain statistical constraints: under the scenario of the global null (GN) (i.e. the treatment effect is zero in every basket), controlling Family-Wise Error Rate (FWER) at the final stage to be 10%. The details of setting the rules are given in Section 4.3.

The sample sizes are assumed to be equally assigned in the two stages. The sample sizes of the two stages can be different in principle, and the allocation between stages is related to alpha-slitting. The calculation of overall sample size as well as allocation between stages is not the focus of this article, but may be a topic worth exploring.

3. METHODS

A major concern of the basket trial is that the sample size is often too small to achieve a desirable power. This relates to the accuracy of the variance estimates. Under our assumption that the response outcomes share the same population variation, a favorable set up is estimating the population variance by aggregating all the data across individual baskets. Assume conjugate priors for the normal likelihood function to facilitate the posterior derivation and save computation time. Thus in our modeling, data are assumed to be normal with conjugate priors and the population variance is given an Inverse Gamma prior. These assumptions are universal for all the methods applied in this article.

3.1 General Notation

The following notations are consistently used throughout this article:

1. $\boldsymbol{y}_{ck} = (y_{ck1}, \ldots, y_{ckn_{ck}})'$, $\boldsymbol{y}_{tk} = (y_{tk1}, \ldots, y_{tkn_{tk}})'$, $\boldsymbol{X}_{ck} = (\boldsymbol{x}_{ck1}, \ldots, \boldsymbol{x}_{ckn_{ck}})'$, $\boldsymbol{X}_{tk} = (\boldsymbol{x}_{tk1}, \ldots, \boldsymbol{x}_{tkn_{tk}})'$, $\boldsymbol{y}_k = \begin{pmatrix} \boldsymbol{y}_{ck} \\ \boldsymbol{y}_{tk} \end{pmatrix}$, $\boldsymbol{X}_k = \begin{pmatrix} \boldsymbol{X}_{ck} \\ \boldsymbol{X}_{tk} \end{pmatrix}$.
   They denote data observations from the k-th basket.
2. Vectors of parameters: $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)'$, $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \ldots, \beta_{kr})'$ for $k = 1, \ldots, K$.
   Denote $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K\}$.
3. $\mathbf{1}_{n_{tk}} = (1, \ldots, 1)'$ is an n_{tk} length vector of 1's.
4. Denote $\pi(\cdot)$ as a prior distribution, $\mathcal{L}(\cdot)$ as the likelihood function, and $p(\cdot)$ as the posterior distribution. Denote $\mathcal{D}$ as the collection of available data observations at the time of analysis.

3.2 Separate Model (SEP) with Full Conjugacy

We would like to compare the performance of the Bayesian methods with a separate analysis under the Bayesian framework. Under models adjusting for covariates, the linear model is

$$\begin{aligned} \boldsymbol{y}_{ck} &= \boldsymbol{X}_{ck}\boldsymbol{\beta}_k + \boldsymbol{\epsilon}_{ck}, \quad k = 1, \ldots, K, \\ \boldsymbol{y}_{tk} &= \boldsymbol{X}_{tk}\boldsymbol{\beta}_k + \tau_k \mathbf{1}_{n_{tk}} + \boldsymbol{\epsilon}_{tk}, \quad k = 1, \ldots, K, \end{aligned} \tag{3.1}$$

where the error terms $\boldsymbol{\epsilon}_{ck}$ and $\boldsymbol{\epsilon}_{tk}$ follow $MVN_{n_{ck}}(0, \boldsymbol{I}_{n_{ck}}\sigma^2)$ and $MVN_{n_{tk}}(0, \boldsymbol{I}_{n_{tk}}\sigma^2)$, respectively. In the k-th basket, the likelihood functions are given by

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\beta}_k, \sigma^2 | \boldsymbol{y}_{ck}, \boldsymbol{X}_{ck}) = \\ &\frac{1}{(\sqrt{2\pi\sigma^2})^{n_{ck}}} \exp\Big(- \frac{1}{2\sigma^2} \|\boldsymbol{y}_{ck} - \boldsymbol{X}_{ck}\boldsymbol{\beta}_k\|^2 \Big), \\ &\mathcal{L}(\boldsymbol{\beta}_k, \tau_k, \sigma^2 | \boldsymbol{y}_{tk}, \boldsymbol{X}_{tk}) = \\ &\frac{1}{(\sqrt{2\pi\sigma^2})^{n_{tk}}} \exp\Big(- \frac{1}{2\sigma^2} ||\boldsymbol{y}_{tk} - \tau_k \mathbf{1}_{n_{tk}} - \boldsymbol{X}_{tk}\boldsymbol{\beta}_k||^2 \Big). \end{aligned} \tag{3.2}$$

The complete likelihood function is

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\tau}, \sigma^2 | \mathcal{D}) = \prod_k \Big[\mathcal{L}(\boldsymbol{\beta}_k, \sigma^2 | \boldsymbol{y}_{ck}, \boldsymbol{X}_{ck}) \mathcal{L}(\boldsymbol{\beta}_k, \tau_k, \sigma^2 | \boldsymbol{y}_{tk}, \boldsymbol{X}_{tk}) \Big]. \tag{3.3}$$

Assuming conjugate priors for parameters $\boldsymbol{\beta}$, $\boldsymbol{\tau}$, σ^2,

$$\begin{aligned} \boldsymbol{\beta}_k | \sigma^2 &\overset{\text{iid}}{\sim} MVN_{r+1}(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda}_k^{-1}), \quad k = 1, \ldots, K, \\ \boldsymbol{\tau} | \sigma^2 &\sim MVN_K(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda}_{\boldsymbol{\tau}}^{-1}), \\ \sigma^2 &\sim IG(a_0, b_0). \end{aligned} \tag{3.4}$$

If no prior knowledge is available, non-informative priors are assumed, for example, by setting $\boldsymbol{\Lambda}_k = \text{diag}_{r+1}(\frac{1}{10000})$, $k = 1, \ldots, K$, and $\boldsymbol{\Lambda}_{\boldsymbol{\tau}} = \text{diag}_K(\frac{1}{10000})$.

The full posterior distribution is

$$p(\boldsymbol{\beta}, \boldsymbol{\tau}, \sigma^2 | \mathcal{D}) \propto \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\tau}, \sigma^2 | \mathcal{D}) \prod_k \pi(\boldsymbol{\beta}_k | \sigma^2) \pi(\boldsymbol{\tau} | \sigma^2) \pi(\sigma^2). \tag{3.5}$$

The posterior for $\boldsymbol{\tau}$ can be derived by integrating out the nuisance parameters:

$$p(\boldsymbol{\tau} | \mathcal{D}) \propto \iint \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\tau}, \sigma^2 | \mathcal{D}) \prod_k \pi(\boldsymbol{\beta}_k | \sigma^2) \pi(\boldsymbol{\tau} | \sigma^2) \pi(\sigma^2) d\boldsymbol{\beta} d\sigma^2. \tag{3.6}$$

Hence, the posterior distribution of $\boldsymbol{\tau}$ is a K-dimensional multivariate t distribution with degrees of freedom $\nu = 2a_0 + \sum_k n_{ck} + \sum_k n_{tk}$. Several quantities are defined below in order to express the location parameters and the scale

matrix of the posterior multivariate t distribution:

$$\begin{aligned}
&\boldsymbol{\Lambda}_{Dk} = \boldsymbol{\Lambda}_k + \boldsymbol{X}_k'\boldsymbol{X}_k,\\
&\boldsymbol{\mu}_k^* = \boldsymbol{X}_k'\boldsymbol{y}_k,\\
&c_k = \mathbf{1}_{n_{tk}}'\mathbf{1}_{n_{tk}} - \mathbf{1}_{n_{tk}}'\boldsymbol{X}_{tk}\boldsymbol{\Lambda}_{Dk}^{-1}\boldsymbol{X}_{tk}'\mathbf{1}_{n_{tk}},\\
&d_k = \boldsymbol{y}_{tk}'\mathbf{1}_{n_{tk}} - \boldsymbol{\mu}_k^{*'}\boldsymbol{\Lambda}_{Dk}^{-1}\boldsymbol{X}_{tk}'\mathbf{1}_{n_{tk}};\\
&\text{and denote diagonal matrix } \boldsymbol{C} \text{ and vector } \boldsymbol{d}\\
&\boldsymbol{C} = \mathrm{diag}(c_1,\ldots,c_K),\\
&\boldsymbol{d} = (d_1,\ldots,d_K)'.
\end{aligned} \tag{3.7}$$

The location parameter for the posterior of $\boldsymbol{\tau}$ is

$$\boldsymbol{\mu}_\tau = (\boldsymbol{C} + \boldsymbol{\Lambda}_{\boldsymbol{\tau}})^{-1}\boldsymbol{d}. \tag{3.8}$$

The scale matrix is

$$\boldsymbol{\Sigma} = \frac{2B}{\nu(\boldsymbol{C} + \boldsymbol{\Lambda}_{\boldsymbol{\tau}})}, \tag{3.9}$$

where $B = b_0 + \frac{1}{2}\sum_k \left(\boldsymbol{y}_k'\boldsymbol{y}_k - \boldsymbol{\mu}_k^{*'}\boldsymbol{\Lambda}_{Dk}^{-1}\boldsymbol{\mu}_k^*\right) - \frac{1}{2}\boldsymbol{d}'(\boldsymbol{C} + \boldsymbol{\Lambda}_{\boldsymbol{\tau}})^{-1}\boldsymbol{d}$. Given the above forms, τ_k's independently follow shifted and scaled t-distributions, and their posterior probabilities can be computed directly.

3.3 Bayesian Hierarchical Model (BHM)

Berry et al. [4] and Thall et al. [36] introduced a Bayesian adaptive design with frequentist interim analyses and hierarchical modeling across the patient subgroups. In addition to model (3.1), a additional structure is built to model τ_k's, i.e., $\tau_k \sim N(\mu_\tau, \sigma_\tau^2)$. Non-informative normal and inverse gamma priors are given to μ_τ and σ_τ^2, respectively. Since all τ_k's are sharing the same mean parameter μ_τ in their priors, one can expect that each estimate of them is pulled towards μ_τ. The shrinkage parameter σ_τ^2 indicates the amount of information borrowing intensity between the baskets: larger σ_τ^2 indicates less borrowing, and smaller σ_τ^2 indicates stronger borrowing. To control the shrinkage parameter, we can define $\sigma_\tau^2 = \phi\sigma^2$ and set a large value of ϕ to form a non-informative prior for τ_k's.

The model is described in (3.1). The likelihood function is described in (3.2) and (3.3). The priors are assumed to be

$$\begin{aligned}
&\boldsymbol{\beta}_k \overset{\text{ind}}{\sim} MVN_{r+1}(\mathbf{0}, \sigma^2\boldsymbol{\Lambda}_k^{-1}), \quad k = 1,\ldots,K,\\
&\tau_k \overset{\text{iid}}{\sim} N(\mu_\tau, \sigma_\tau^2), \quad k = 1,\ldots,K,\\
&\mu_\tau \sim N(0, \phi\sigma_\tau^2),\\
&\sigma_\tau^2 \sim IG(a_0', b_0'),\\
&\sigma^2 \sim IG(a_0, b_0).
\end{aligned} \tag{3.10}$$

To specify non-informative priors, set $\boldsymbol{\Lambda}_k = \mathrm{diag}_{r+1}(\frac{1}{10000})$ for $k = 1,\ldots,K$. The Gibbs collapsed sampling procedure is used to obtain the posterior samplers.

3.4 Calibrated Bayesian Hierarchical Model (CBHM)

The challenge of achieving advantageous results for BHM to basket trials is that the number of baskets, K, is often small so the parameter σ_τ^2 cannot be estimated robustly and the results are very sensitive for the prior setting. The detailed discussion can be found in [4]. To overcome this issue, Chu and Yuan [6] proposed a Calibrated Bayesian Hierarchical model (CBHM) for binary endpoints. Instead of giving a prior to σ_τ^2, they proposed to estimate σ_τ^2 as a monotonic increasing function of a Chi-square test statistic T, where T is a quantity that measures the similarity of the treatment effects across baskets (i.e. Chi-square test statistics):

$$\sigma_\tau^2 = \exp(a + b\log T), \tag{3.11}$$

where a and $b > 0$ are pre-calculated constants obtained by simulation. The information is borrowed more if the treatment effects are similar across baskets, and less otherwise. To determine the values of a and b, one needs to pre-specify the two cases of σ_τ^2's representing the data being heterogeneous and homogeneous (e.g., $\sigma_\tau^2 = 80$ versus $\sigma_\tau^2 = 1$). In the simulation, one also needs to first decide on some scenarios where information should be borrowed (e.g., the treatment effects are the same for all baskets) across baskets and some other scenarios where information should not be borrowed (e.g., only one basket is truly efficacious), and then record the median of their Chi-square statistics. The constants a and b could be calculated by solving (3.11) under these two scenarios.

We extend this idea to the linear regression models, by replacing the Chi-square test statistic with the p-value of a F test statistic. The F test statistic is to test whether any treatment effect difference are heterogeneity from other baskets when the covariates are included in the model. Since the F test statistics is associated with different degrees of freedom, which depends on the sample size of the basket, number of covariates adjusted and the number of baskets being assessed, we use their p-values instead of the test statistics. See Section S.4 of the Supplementary Material for the formulae of the testing procedure. The main challenge may be to set a suitable function that relates the value of σ_τ^2 to the p-value ($pval$). Note that the p-value is always between 0 and 1, and σ_τ^2 should monotonically decrease with the p-value. We simply apply the following relation:

$$\sigma_\tau^2 = a + b \times (1 - pval), \tag{3.12}$$

where $a, b \geq 0$ are the tuning parameters. To find a robust function between σ_τ^2 and p-values, some other functions are also tested, including the original exponential function linkage as well as the logit function of $pval$. The property of the exponential function may arise large variances in our setting and thus cause Type I error inflation. The logit function performs best on Type I error rate control, whereas the power

improvement may not be as satisfactory. Compared with these functions, the above linear function performs quite robustly, and has good Type I error rate control and power improvement. A simplified procedure is used by adjusting the a and b directly instead of running simulations. Using these procedures, comparable results are produced with those by Chu and Yuan [6] in terms of power and Type I error rate comparison with other methods.

The model is described in (3.1). The likelihood function is described in (3.2) and (3.3). The priors for the parameters are given as

$$\begin{aligned}
&\boldsymbol{\beta}_k \overset{\text{ind}}{\sim} MVN_{r+1}(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda}_k^{-1}), \quad k = 1, \ldots, K, \\
&\tau_k \sim N(\tau, \hat{\sigma}_\tau^2), \quad k = 1, \ldots, K, \\
&\tau \sim N(0, \sigma_0^2), \\
&\sigma^2 \sim IG(a, b).
\end{aligned} \tag{3.13}$$

Non-informative priors are specified by setting $\boldsymbol{\Lambda}_k = \text{diag}_{r+1}(\frac{1}{10000})$ for $k = 1, \ldots, K$. The Gibbs collapsed sampling procedure to obtain the posterior samples is given in Algorithm 1.

3.5 Mixture of Finite Mixtures (MFM)

Inferences from simple pooling or separate analyses are recognized to be inferior to hierarchical modeling methods, which allow adaptive borrowing of information across subgroups. However, the full Bayesian model bears the risk of too much shrinkage and excessive borrowing. In recent years, more and more attention is given to approaches based on local borrowing through mixture models. In such models, baskets can be grouped to clusters so that the information is borrowed within each cluster. Choosing or modeling the number of clusters M is critical when the mixture model is applied. It is even more important when applied to basket trials, as there are usually a limited number of baskets, and hence the overall model performance heavily relies on the model of M. We attempt to apply the Mixture of Finite Mixtures (MFM) approach [18], for its flexible and targeted control of modeling M.

The full MFM model can be specified as usual:

$$\begin{aligned}
&M \sim p(m): \text{ truncated Poisson p.m.f.} \\
&\text{on positive integer } m \text{ with parameter } \lambda, \\
&\pi_1, \ldots, \pi_M \sim Dirichlet_M(\gamma, \ldots, \gamma), \\
&P(z_k = m | M) = \pi_m, m = 1, \ldots, M, k = 1, \ldots, K,
\end{aligned} \tag{3.18}$$

where z_k, the latent variable, denotes cluster that the k-th basket belongs to, and π_m is the corresponding probability $P(z_k = m|M)$. The truncated Poisson prior is chosen because of its convenience in sampling from the posterior distribution [18]. Several parameters remain unspecified in the above model. The value of λ indicates the prior choice of the number of clusters. Setting λ equal to K seems to help alleviate the excessive borrowing. γ affects the probability of assigning the basket to a distinct new cluster, hence influencing the borrowing strength across baskets.

Algorithm 1 CBHM (models adjusting for covariates) MCMC Sampling Algorithm-Collapsed Gibbs Sampler.

1: **procedure** CBHM (WITH COVARIATES) SAMPLING
2: Initial $\boldsymbol{\beta}_k, k = 1, \ldots, K$, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K), \tau$ and σ^2.
3: **for** i =1 to N **do**
4: Sample $\boldsymbol{\beta}_k, k = 1, \ldots, K$ conditional on $\boldsymbol{\tau}$, σ^2 by

$$\begin{aligned}
&\boldsymbol{\beta}_k | \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\tau}, \sigma^2, \sim MVN_{r+1}(\boldsymbol{\mu}_k^*, \sigma^2 \boldsymbol{\Lambda}_{Dk}^{-1}), \\
&\text{where } \boldsymbol{\Lambda}_{Dk} = \boldsymbol{X}_k' \boldsymbol{X}_k + \boldsymbol{\Lambda}_k, \\
&\text{and } \boldsymbol{\mu}_k^* = \boldsymbol{\Lambda}_{Dk}^{-1} (\boldsymbol{X}_k' \boldsymbol{y}_k - \boldsymbol{X}_{tk}' \mathbf{1}_{n_{tk}} \tau_k);
\end{aligned} \tag{3.14}$$

5: Sample σ^2 conditional on $\boldsymbol{\beta}$, $\boldsymbol{\tau}$ by

$$\begin{aligned}
&\sigma^2 | \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\tau} \sim IG\Big(a^* = a_0 + \sum_k \frac{n_{ck} + n_{tk}}{2} + (r+1)K/2, \\
&b^*\Big), \text{ where } b^* = b_0 + \frac{1}{2} \sum_k \Big[||\boldsymbol{y}_{tk} - \boldsymbol{X}_{tk} \boldsymbol{\beta}_k - \mathbf{1}_{n_{tk}} \tau_k||^2 + \\
&||\boldsymbol{y}_{ck} - \boldsymbol{X}_{ck} \boldsymbol{\beta}_k||^2 + (\boldsymbol{\beta}_k - \boldsymbol{\mu}_k)' \Lambda_k (\boldsymbol{\beta}_k - \boldsymbol{\mu}_k) \Big];
\end{aligned} \tag{3.15}$$

6: Sample $\tau_k, k = 1, \ldots, K$ conditional on $\boldsymbol{\beta}$, σ^2 and τ by

$$\begin{aligned}
&\tau_k | \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\beta}_k, \sigma^2, \mu_\tau \sim \\
&N\left[\frac{\hat{\sigma}_\tau^2 (\boldsymbol{y}_{tk} - \boldsymbol{X}_{tk} \boldsymbol{\beta}_k)' \mathbf{1}_{n_{tk}} + \mu_\tau \sigma^2}{\hat{\sigma}_\tau^2 \mathbf{1}_{n_{tk}}' \mathbf{1}_{n_{tk}} + \sigma^2}, \frac{\hat{\sigma}_\tau^2 \sigma^2}{\hat{\sigma}_\tau^2 \mathbf{1}_{n_{tk}}' \mathbf{1}_{n_{tk}} + \sigma^2} \right];
\end{aligned} \tag{3.16}$$

7: Collapse sample μ_τ conditional on $\boldsymbol{\beta}$, σ^2 by integrating out τ_k's:

$$\begin{aligned}
&\mu_\tau | \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\beta}, \sigma^2 \sim \\
&N\left[\frac{\sum_k \frac{(\boldsymbol{y}_{tk} - \boldsymbol{X}_{tk} \boldsymbol{\beta}_k)' \mathbf{1}_{n_{tk}}}{\mathbf{1}_{n_{tk}}' \mathbf{1}_{n_{tk}} \hat{\sigma}_\tau^2 + \sigma^2}}{\sum_k \frac{\mathbf{1}_{n_{tk}}' \mathbf{1}_{n_{tk}}}{\mathbf{1}_{n_{tk}}' \mathbf{1}_{n_{tk}} \hat{\sigma}_\tau^2 + \sigma^2} + \frac{1}{\sigma_0^2}}, \frac{1}{\sum_k \frac{\mathbf{1}_{n_{tk}}' \mathbf{1}_{n_{tk}}}{\mathbf{1}_{n_{tk}}' \mathbf{1}_{n_{tk}} \hat{\sigma}_\tau^2 + \sigma^2} + \frac{1}{\sigma_0^2}}} \right].
\end{aligned} \tag{3.17}$$

8: **end for**
9: **end procedure**

The model is described in (3.1). The likelihood function is described in (3.2) and (3.3) by replacing τ_k by τ_{z_k}. The priors for the parameters are given as follows:

$$\begin{aligned}
&\boldsymbol{\beta}_k \overset{\text{iid}}{\sim} MVN_{r+1}(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda}_k^{-1}), \quad k = 1, \ldots, K, \\
&\tau_m \overset{\text{iid}}{\sim} N(0, \phi \sigma^2), \quad m = 1, \ldots, M, \\
&\sigma^2 \sim IG(a_0, b_0),
\end{aligned} \tag{3.19}$$

with $\boldsymbol{\Lambda}_k = \text{diag}_{r+1}(\frac{1}{\xi}, \frac{1}{10000}, \ldots, \frac{1}{10000})$. For ϕ and ξ, they together control the amount of information to be borrowed across the baskets and therefore need to be carefully speci-

Algorithm 2 MFM (models adjusting for covariates) MCMC Sampling Algorithm-Gibbs Sampler.

procedure MFM (WITH COVARIATES) SAMPLING
  Initial $\boldsymbol{z} = (z_1, \ldots, z_K), \boldsymbol{\beta}_k$, $k = 1, \ldots, K$, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_M)$ and σ^2.
  for i =1 to N **do**
    Sample $\boldsymbol{\beta}_k, k = 1, \ldots, K$ conditional on $\boldsymbol{\tau}$ and σ^2 by

$$\begin{aligned} &\boldsymbol{\beta}_k|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\tau}, \sigma^2 \sim MVN_{r+1}(\boldsymbol{\mu}_k^*, \sigma^2 \boldsymbol{\Lambda}_{Dk}^{-1}), \\ &\text{where } \boldsymbol{\Lambda}_{Dk} = \boldsymbol{X}_k' \boldsymbol{X}_k + \boldsymbol{\Lambda}_k, \\ &\text{and } \boldsymbol{\mu}_k^* = \boldsymbol{\Lambda}_{Dk}^{-1}(\boldsymbol{X}_k' \boldsymbol{y}_k - \boldsymbol{X}_{tk}' \mathbf{1}_{n_{tk}} \tau_{z_k}); \end{aligned} \tag{3.20}$$

    Sample σ^2 conditional on $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ by

$$\begin{aligned} &\sigma^2|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\tau} \sim IG(a^* = a_0 + \sum_k \frac{n_{ck} + n_{tk}}{2} + (r+1)K/2 + M/2, b^*) \\ &\text{where } b^* = b_0 + \frac{1}{2} \sum_{z_k=1}^{M} \tau_{z_k}^2/\phi + \frac{1}{2} \sum_k \Big(||\boldsymbol{y}_{ck} - \boldsymbol{X}_{ck}\boldsymbol{\beta}_k||^2 + ||\boldsymbol{y}_{tk} - \mathbf{1}_{n_{tk}} \tau_{z_k} - \boldsymbol{X}_{tk}\boldsymbol{\beta}_k||^2 + \boldsymbol{\beta}_k' \Lambda_k \boldsymbol{\beta}_k \Big); \end{aligned} \tag{3.21}$$

    Sample $\boldsymbol{\tau}$ conditional on $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{z}$ by

$$\tau_m|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}, \sigma^2 \sim N\left[\frac{\phi \sum_{k:z_k=m} (\boldsymbol{y}_{tk} - \boldsymbol{X}_{tk}\boldsymbol{\beta}_k)' \mathbf{1}_{n_{tk}}}{\phi \sum_{k:z_k=m} \mathbf{1}_{n_{tk}}' \mathbf{1}_{n_{tk}} + 1}, \frac{\phi\sigma^2}{\phi \sum_{k:z_k=m} \mathbf{1}_{n_{tk}}' \mathbf{1}_{n_{tk}} + 1} \right]; \tag{3.22}$$

    Update $\boldsymbol{z} = (z_1, \ldots, z_K)$ conditional on $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_M)$, $\boldsymbol{\beta}_k, k = 1, \ldots, K$ and σ^2 for each k in $\{1, \ldots, K\}$, based on close form of $P(z_k = g|\boldsymbol{z_{-k}}, \boldsymbol{\beta}, \boldsymbol{\tau}, \sigma^2, \mathcal{D})$:

$$\propto \begin{cases} [|g| + \gamma] \mathcal{L}(\boldsymbol{X}_{ck}, \boldsymbol{y}_{ck}|\boldsymbol{\beta}_k, \sigma^2) \mathcal{L}(\boldsymbol{X}_{tk}, \boldsymbol{y}_{tk}, \tau_g|\boldsymbol{\beta}_k, \sigma^2), & \text{in an existing cluster } g, \\ \frac{V_K(|\mathcal{G}_{-k}|+1)}{V_K(|\mathcal{G}_{-k}|)} \gamma m(\mathcal{D}|\boldsymbol{\beta}_k, \sigma^2), & \text{in a new cluster,} \end{cases} \tag{3.23}$$

    where $|\mathcal{G}_{-k}|$ denotes the number of clusters obtained by removing the k-th basket, $|g|$ denotes the cluster size of cluster labeled g, and $\mathcal{L}(\boldsymbol{y}_{ck}, \boldsymbol{X}_{ck}|\boldsymbol{\beta}_k, \sigma^2)\mathcal{L}(\boldsymbol{y}_{tk}, \boldsymbol{X}_{tk}, \tau_g|\boldsymbol{\beta}_k, \sigma^2)$ are the likelihood functions from the normal distributions of $\boldsymbol{y}_{ck}$ and $\boldsymbol{y}_{tk}$ presented in formula (3.2) conditioning on the updated values of σ^2 and $\boldsymbol{\beta}_k$.

    $m(\mathcal{D}|\boldsymbol{\beta}_k, \sigma^2)$ can be obtained by the closed expression:

$$m(\mathcal{D}|\boldsymbol{\beta}_k, \sigma^2) = \mathcal{L}(\boldsymbol{X}_{ck}, \boldsymbol{y}_{ck}|\boldsymbol{\beta}_k, \sigma^2) \frac{1}{(\sqrt{2\pi\sigma^2})^{n_{tk}}} \frac{1}{\sqrt{\phi \mathbf{1}_{n_{tk}}' \mathbf{1}_{n_{tk}} + 1}} \exp\left\{ \frac{\phi[(\boldsymbol{y}_{tk} - \boldsymbol{X}_{tk}\boldsymbol{\beta}_k)' \mathbf{1}_{n_{tk}}]^2}{2\sigma^2(\phi \mathbf{1}_{n_{tk}}' \mathbf{1}_{n_{tk}} + 1)} - \frac{1}{2\sigma^2} ||\boldsymbol{y}_{tk} - \boldsymbol{X}_{tk}\boldsymbol{\beta}_k||^2 \right\}. \tag{3.24}$$

  end for
end procedure

fied. A programmable algorithm to fit the MFM with covariates adjustment is given in Algorithm 2. In the Algorithm, as *a priori* the basket k is placed in

$$\begin{cases} \text{an existing cluster } g \in \mathcal{G}_{-k} \text{ with probability} \propto |g| + \gamma, \\ \text{a new cluster with probability} \propto \dfrac{V_K(t+1)}{V_K(t)} \gamma, \end{cases} \tag{3.25}$$

where $t = |\mathcal{G}_{-k}|$ is the number of clusters obtained by removing the k-th basket, and $V_K(t)$ needs to be precomputed as

$$V_K(t) = \sum_{m=1}^{\infty} \frac{m_{(t)}}{(\gamma m)^{(K)}} p_M(m). \tag{3.26}$$

Here $x^{(t)} = x(x+1)\ldots(x+t-1)$, and $x_{(t)} = x(x-1)\ldots(x-t+1)$. By convention, $x^{(0)} = 1$ and $x_{(0)} = 1$. Meanwhile, $p_M(m)$ is the mass density function of the truncated Poisson distribution on $\{1, 2, \ldots\}$ with parameter λ.

The derivation of $m(\mathcal{D}|\boldsymbol{\beta}_k, \sigma^2)$ in (3.24) is given in Section S.3 of the Supplementary Material.

4. SIMULATION STUDIES

4.1 Data Generation

Consider the case that the outcomes are influenced by two covariates, one continuous (X_1) and one binary (X_2). Following previous notation, the outcomes are modeled as

$$\begin{aligned} &Y_{cki} = \beta_{k0} + \beta_{k1} X_{cki1} + \beta_{k2} X_{cki2} + \epsilon_{cki}, \\ &\text{with} \quad i = 1, \ldots, n_{ck}, \\ &Y_{tki'} = \beta_{k0} + \tau_k + \beta_{k1} X_{tki'1} + \beta_{k2} X_{tki'2} + \epsilon_{tki'}, \\ &\text{with} \quad i' = 1, \ldots, n_{tk}, \end{aligned} \tag{4.1}$$

where $k = 1, \ldots, 4$. The number of subjects in control and treatment arm are assumed to be the same ($n_{ck} = n_{tk} \doteq n_k$). For baskets 1 to 4, they are assumed to be $n_1 = 30$, $n_2 = 30$, $n_3 = 20$, and $n_4 = 20$, and the same sample sizes are assumed for each stage (if the basket is available for analysis in that stage). Let β_{k0} denotes the intercept in the control arm for the k-th basket. τ_k denotes the difference in the intercept between treatment and control for the k-th basket. Each of the error terms $\epsilon_{cki}(\epsilon_{tki'}) \sim N(0, 1)$. The

coefficients are allowed to be different across baskets, and set to be $(\beta_{11}, \beta_{21}, \beta_{31}, \beta_{41}) = (0.2, 0.4, 0.6, 0.8)$. The continuous covariates X_{cki1} in the control arm are simulated from $logN(0, 0.9^2)$, and the covariates $X_{tki'1}$ in the treatment arm are simulated from $N(1.5, 2.8)$. The binary covariates X_{cki2} in the control arm are simulated from $Bin(0.4)$, and the covariates $X_{tki'2}$ in the treatment arm are simulated from $Bin(0.6)$. The values of the binary coefficients are assumed to differ by basket to be $(\beta_{12}, \beta_{22}, \beta_{32}, \beta_{42}) = (0.4, 0.3, 0.2, 0.1)$.

4.2 Scenarios and Evaluation

To evaluate the performance of these various methods, the data is generated assuming the control effects to be $(0, 0, 0, 0)$. Consider 6 scenarios of the true treatment effects, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_4)$, which are summarized in Table 1. Note that the first scenario corresponds to the global null (GN).

Table 1. True effects in the treatment arm.

Scenarios	τ_1	τ_2	τ_3	τ_4
1 GN	0	0	0	0
2	0.4	0	0	0
3	0.4	0.4	0	0
4	0.4	0.4	0.4	0
5	0.4	0.4	0.4	0.4
6	0	0.2	0.4	0.6

Under each scenario, 1000 trials are simulated. The following characteristics of all methods are evaluated under each scenario:

1. FWER: The percentage of trials in which any ineffective basket is wrongly selected as efficacious.
2. One-minimum power (P1): The percentage of trials in which at least one effective basket is correctly selected as efficacious.
3. Correct power (P2): The percentage of trials in which at least one effective basket is correctly selected as efficacious, and none of the ineffective baskets is selected.
4. Exact correct power (P3): The percentage of trials in which all effective baskets are correctly selected as efficacious, and none of the ineffective baskets is selected.
5. % Rejection: The percentage of trials in which the specific basket is selected as efficacious.
6. RMSE: Calculated as $E[(\hat{\tau}_k - \tau_k)^2]^{1/2}$ for each basket k, where $\hat{\tau}_k$ is the Posterior mean. The expectation is calculated as the average over the simulated trials.
7. Average Enrollment: The number of participants consumed in the specific basket, for each arm, averaged over all simulated trials.

Another set of simulation analyses is provided when the outcome data is affected by covariates but models are incorrectly designed as having no covariate. They are referred to as Mis-specified Models.

4.3 Value of Tuning Parameters

For all methods, if not specified, the values of ξ and ϕ are 10,000, and the hyper-parameters $a_0 = 0.5$ and $b_0 = 0.05$, forming non-informative priors for the corresponding parameters. BHM, CBHM and MFM require sampling procedures. The MCMC sample size is 1300, with 300 burn-in. Trace plots are monitored to ensure convergence.

For CBHM, the tuning parameters in equation (3.12) are set to be $a = 0$ and $b = 0.2$. These values are picked based on the data observations and they yield the best overall performance.

For MFM, the truncated Poisson parameter is set to be $\lambda = 4$ and the Dirichlet parameter is set to be $\gamma = 100$. This reflects the original belief that the basket are separately analyzed, which essentially helps reduce FWER in the simulation. To determine the values of ϕ and ξ, we run the simulation for various combinations and select the following values that provide the best balance between FWER inflation and power: $\phi = 0.1$ and $\xi = 0.16$. In particular, some settings actually provide even a larger power but those are not chosen because they also have a larger FWER. The results of the sensitivity analysis for other combinations of ϕ and ξ are reported in Tables S.4 to S.6 for models without covariates and Tables S.7 to S.9 for models adjusting for covariates in Section S.2 of the Supplementary Material.

All the methods are compared in One-Stage Design (i.e., no interim analysis) and two-stage designs (i.e., with one interim analysis). For two-stage design, we set $q_2 = 1 - \eta(1 - q_1)$ as one of the conditions to determine q_1 and q_2. When $\eta = \frac{1}{2}$, both early efficacy and futility rules are adopted; when $\eta = 0$, no early stopping for efficacy. For a fixed value of η, the value of q_1 (and hence q_2) is determined via simulation to control the FWER under the global null to be no more than 10%. The values of q_1 under different stage design are provided in Table 2 for Models adjusting for covariates; in Table S.2 for Models without covariates.

Table 2. Value of q_1 for Models Adjusting Covariates.

Method	One-Stage	Two-Stage, both stopping ($\eta = \frac{1}{2}$)	Two-Stage, futility stopping only ($\eta = 0$)
SEP	0.9772	0.9649	0.9434
BHM	0.9590	0.9400	0.9240
CBHM	0.9660	0.9480	0.9305
MFM	0.9180	0.8920	0.8790

4.4 Simulation Results

In this section, simulation results for the models adjusting for covariates when the covariates are correctly identified are presented. For simulation data without covariates, the simulation results are presented in Section S.1 of the Supplementary Material. The results for models with covariate adjustment and models without covariates are similar. This

is because both models are "correctly specified", i.e. the inclusion of covariates is consistent with how the data were generated. to assess the performance of these methods, the 6 scenarios summarized in Table 1 are applied. In Scenario 1, the treatment is ineffective in any of the baskets; In Scenario 2, the treatment is effective for basket 1 only; In Scenario 3, the treatment is effective for baskets 1 and 2; In Scenario 4, the treatment is effective for baskets 1, 2 and 3; In Scenario 5, the treatment is effective for all baskets. The effect size of 0.4 is equivalent for all baskets in Scenarios 2 to 5; In Scenario 6, the treatment is effective for baskets 2, 3 and 4 with effect sizes of 0.2, 0.4, and 0.6, respectively. Scenario 1 is used to determine the thresholds for all models to have the FWER of less or equal to 10% at the end of the final stage under the global null.

4.4.1 One-Stage Design

The simulation results for one-stage design are shown in Table 3. See Table S.3 for models without covariates in Supplementary Material.

With both BHM and CBHM, information is borrowed through a common mean parameter shared by various treatment effects in different baskets. As a result, the estimated treatment effects of all baskets tend to be pulled toward the average. For Scenario 2, CBHM and MFM performed better than SEP and BHM by an increase in P1 without lowering P2 and P3. CBHM and MFM has greater P1, which is the probability of identifying the only basket (basket 1) responsive to the treatment. Compared with SEP, BHM shows similar P1 but worse P2 and P3, as the result of excessive borrowing. In Scenarios 3 to 6, with the increase of the number of baskets responsive to the treatment, the gain in power increases with BHM, CBHM and MFM, from approximately 3% to 12%. In all scenario, including Scenario 6, where the treatment effect varies across all baskets, MFM performs uniformly better than all other methods, regarding P1, P2 and P3.

All three methods resulted in greater FWER than SEP as the result of information borrowing for scenarios where at least one but not all baskets are responsive to treatment. FWER with BHM is uniformly greater than all other methods due to extensive borrowing. Since information is borrowed depending on the similarities of the baskets, CBHM yields lower FWER than BHM without lowering in power. These results are consistent with the research by Chu and Yuan [6]. In all scenarios, MFM performed better than BHM and CBHM in keeping the FWER around 10%. Unlike BHM and CBHM, MFM allows localized borrowing by grouping baskets into new clusters thereby controlling FWER.

Among all three methods, BHM is the least favorable with the highest FWER and lowest P1, P2 and P3 across all scenarios. MFM is the most preferable with the lowest FWER and a substantial increase of P1 in all scenarios ranging from 5% to 14%. With MFM, P2 and P3 are also uniformly better than other methods for all scenarios.

For all borrowing methods, the estimation of treatment effects may be biased. However, biased estimates in basket trial may contribute to an overall power improvement. Thereby, bias is not regarded as an adequate assessment tool for performance; as an alternative perspective, the RMSE can be used as the examine criteria. RMSE accounts for both bias and variance. In all scenarios, MFM, BHM and CBHM perform better than SEP in reducing the RMSE. MFM is again the most preferable methods in reducing RMSE than BHM and CBHM.

4.4.2 Two-Stage Design with Efficacy and Futility Stopping

The simulation results for this design are reported in Table 4 for models adjusting for covariates. Performance of the methods for models without covariates are reported in Table S.4 in the Supplementary Material.

Unlike their performance in one-stage design, BHM and CBHM has higher P1 than SEP in all scenarios. This difference may be the result of adding the efficacy stopping at the first stage, thereby increase the power by selecting the basket responsive to the treatment at the interim analysis, preventing the information borrowing of the responsive baskets from other non-efficacious baskets at the second stage. With the increase of the number of responsive baskets, the gain in P1 increase with BHM and CBHM, from 1% to 11%. Like what is observed in one-stage design, MFM shows greater P1, and P2 compared with BHM, CBHM and SEP in all scenarios. Overall, MFM, BHM, CBHM perform better in two-stage design than one-stage design.

Like one-stage design, MFM, BHM, CBHM resulted in greater FWER than SEP in all scenarios. FWERs with BHM and CBHM in two-stage design are like those in one-stage design. However, FWER with MFM in two-stage design is lower than that with BHM, CBHM in two-stage design. Compared with MFM in one-stage design, FWER is lower in two-stage design. The early futility stopping helps dropping the baskets not responsive to the treatment before the second stage.

Similar to what is observed in one-stage study design, the MFM, BHM, CBHM produced lower RMSEs lower than SEP in all scenarios. The column "Average Enrollment" displays the average number of participants consumed. These numbers show advantage of MFM in retaining the baskets that are responsive to the treatment and dropping those that are not.

4.4.3 Two-Stage Design with Futility Stopping Only

For this design, the simulation results are reported in Table 5 for models adjusting for covariates and Table S.5 for models without covariates in the Supplementary Material.

Unlike the two-stage design with both futility and efficacy stopping at the interim, two-stage design with futility but without efficacy stopping lowers P1 with BHM and CBHM compared with SEP in most scenarios. The loss of power may be caused by the lower for futility and therefore fewer

Table 3. Performance of SEP, BHM, CBHM and MFM, one-stage design (models adjusting for covariates).

Scn	Method	FWER	P1	P2	P3	% Reject				100 x RMSE			
1						1	2	3	4	1	2	3	4
	SEP	10.0	0.0	0.0	0.0	2.1	2.8	2.7	2.7	26.1	26.4	34.0	32.8
	BHM	9.6	0.0	0.0	0.0	2.7	4.1	1.9	3.2	18.7	18.8	21.4	20.6
	CBHM	10.0	0.0	0.0	0.0	5.1	2.6	2.7	2.5	20.6	20.5	23.7	22.6
	MFM	9.9	0.0	0.0	0.0	3.4	3.9	2.2	2.2	13.7	13.8	14.3	13.7
2						[1]	2	3	4	[1]	2	3	4
	SEP	8.1	35.9	32.8	32.8	35.9	2.8	2.7	2.7	26.1	26.4	34.0	32.8
	BHM	13.4	32.3	24.0	24.0	32.3	6.2	4.9	5.2	22.9	19.9	22.7	22.4
	CBHM	10.3	40.0	33.7	33.7	40.0	4.3	4.1	4.2	22.4	21.3	25.4	24.1
	MFM	11.5	40.7	33.2	33.2	40.7	5.5	4.1	3.6	22.8	14.8	15.5	15.0
3						[1]	[2]	3	4	[1]	[2]	3	4
	SEP	5.4	56.8	53.6	10.2	35.9	31.7	2.7	2.7	26.1	26.4	34.0	32.8
	BHM	16.2	59.8	47.2	11.8	41.5	33.6	9.1	9.5	20.4	21.0	24.8	25.2
	CBHM	12.4	59.7	48.7	10.8	46.2	29.4	8.1	7.9	21.2	22.1	27.3	25.9
	MFM	11.7	68.9	58.4	20.9	49.0	45.5	6.6	6.4	20.6	21.3	17.0	16.4
4						[1]	[2]	[3]	4	[1]	[2]	[3]	4
	SEP	2.7	67.1	64.9	2.9	35.9	31.7	24.8	2.7	26.1	26.4	34.0	32.8
	BHM	13.2	74.1	61.9	7.9	47.4	37.8	38.8	13.2	19.2	19.6	22.5	28.0
	CBHM	11.9	74.0	62.5	6.3	50.4	37.1	34.6	11.9	20.7	21.2	24.2	28.1
	MFM	8.3	78.7	70.8	12.9	53.8	48.6	38.8	8.3	19.2	19.9	22.0	17.5
5						[1]	[2]	[3]	[4]	[1]	[2]	[3]	[4]
	SEP	0.0	73.7	73.7	0.5	35.9	31.7	24.8	23.6	26.1	26.4	34.0	32.8
	BHM	0.0	85.5	85.5	10.3	52.9	42.3	47.1	52.4	18.7	18.8	21.4	20.6
	CBHM	0.0	82.5	82.5	14.8	52.7	44.7	45.1	44.6	20.6	20.5	23.7	22.6
	MFM	0.0	87.6	87.6	13.2	60.0	56.3	45.3	43.6	17.9	18.6	20.3	20.0
6						1	[2]	[3]	[4]	1	[2]	[3]	[4]
	SEP	2.1	62.2	60.9	1.4	2.1	11.3	24.8	45.3	26.1	26.4	34.0	32.8
	BHM	11.9	68.2	57.1	4.7	11.9	19.9	35.4	55.2	23.7	19.7	23.7	27.0
	CBHM	11.8	63.8	53.2	3.2	11.8	16.5	31.2	50.5	24.8	21.2	24.7	27.1
	MFM	8.2	69.3	61.9	7.3	8.2	21.1	35.7	54.8	16.5	15.8	23.0	33.5

Table 4. Performance of SEP, BHM, CBHM and MFM, two-stage design and $\eta = \frac{1}{2}$ (models adjusting for covariates).

Scn	Method	FWER	P1	P2	P3	% Reject				100 x RMSE				Average Enrollment			
1						1	2	3	4	1	2	3	4	1	2	3	4
	SEP	10.0	0.0	0.0	0.0	1.9	3.2	2.6	2.6	25.3	25.8	33.3	31.9	30.7	30.7	20.4	20.4
	BHM	9.9	0.0	0.0	0.0	2.7	4.5	2.3	2.7	18.3	18.6	21.2	20.1	31.0	30.9	20.5	20.5
	CBHM	10.1	0.0	0.0	0.0	3.1	3.3	2.9	3.1	19.9	19.9	23.2	22.3	30.9	31.0	20.5	20.4
	MFM	9.8	0.0	0.0	0.0	3.4	3.8	2.3	2.3	13.5	13.5	14.0	13.4	31.3	31.2	20.7	20.6
2						[1]	2	3	4	[1]	2	3	4	[1]	2	3	4
	SEP	8.2	40.2	36.6	36.6	40.2	3.2	2.6	2.6	26.4	25.8	33.3	31.9	33.2	30.7	20.4	20.4
	BHM	13.6	41.6	32.3	32.3	41.6	6.8	4.9	4.9	23.6	19.5	22.3	21.7	34.4	31.2	20.8	20.8
	CBHM	11.1	41.4	34.3	34.3	41.4	4.7	4.9	4.1	23.6	20.6	24.6	23.2	32.6	31.2	20.7	20.6
	MFM	10.9	46.4	38.2	38.2	46.4	5.3	3.9	3.0	22.6	14.5	15.0	14.3	35.2	31.3	21.0	21.0
3						[1]	[2]	3	4	[1]	[2]	3	4	[1]	[2]	3	4
	SEP	5.2	62.1	58.7	12.4	40.2	35.0	2.6	2.6	26.4	26.6	33.3	31.9	33.2	33.1	20.4	20.4
	BHM	14.5	71.4	57.7	23.0	52.1	50.8	8.4	7.9	21.0	21.4	23.9	23.9	33.8	33.8	21.1	21.3
	CBHM	12.6	68.6	57.1	17.7	49.3	43.3	8.7	7.9	22.0	22.7	25.6	24.7	32.7	34.7	21.2	20.9
	MFM	9.6	74.4	65.6	27.2	54.8	50.7	6.0	4.3	20.4	21.2	16.1	15.4	35.5	35.9	21.3	21.2
4						[1]	[2]	[3]	4	[1]	[2]	[3]	4	[1]	[2]	[3]	4
	SEP	2.6	73.1	70.7	3.8	40.1	35.0	28.8	2.6	26.4	26.6	33.9	31.9	33.2	33.1	21.6	20.4
	BHM	11.8	80.9	69.3	20.1	60.5	57.4	43.8	11.8	19.6	19.9	23.4	26.2	33.9	33.3	22.6	21.9
	CBHM	13.8	79.9	66.4	11.9	55.0	48.8	44.1	13.8	21.1	21.7	25.0	27.1	32.3	33.8	22.6	21.0
	MFM	6.3	83.7	77.7	16.8	59.3	55.1	42.5	6.3	19.1	20.0	22.1	16.4	35.0	36.2	23.6	21.5
5						[1]	[2]	[3]	[4]	[1]	[2]	[3]	[4]	[1]	[2]	[3]	[4]
	SEP	0.0	78.9	78.9	1.3	40.1	35.0	28.8	26.4	26.4	26.6	33.9	32.6	33.2	33.1	21.6	21.7
	BHM	0.0	89.7	89.7	25.4	67.4	64.9	51.3	56.8	19.2	19.4	22.1	21.5	33.8	33.8	22.2	23.2
	CBHM	0.0	88.3	88.3	27.5	63.0	59.3	53.6	55.8	20.5	20.9	24.3	23.3	32.2	33.0	22.2	22.5
	MFM	0.0	91.0	91.0	16.2	64.7	60.4	48.2	48.7	17.9	18.7	20.7	20.3	35.0	35.7	24.0	24.2
6						1	[2]	[3]	[4]	1	[2]	[3]	[4]	1	[2]	[3]	[4]
	SEP	1.9	67.4	66.0	2.3	1.9	11.6	28.8	52.5	25.3	25.9	33.9	33.3	30.7	31.6	21.6	22.5
	BHM	10.7	74.9	64.7	8.9	10.7	25.4	39.2	65.7	22.4	19.8	24.4	26.7	32.4	32.6	22.1	23.1
	CBHM	9.3	73.5	64.9	6.2	9.3	19.4	39.3	61.7	23.0	21.0	25.5	27.3	30.9	32.5	22.2	22.6
	MFM	6.3	75.5	69.6	8.7	6.3	22.2	40.2	61.7	15.5	16.1	23.1	32.1	32.1	34.0	23.6	24.2

Table 5. Performance of SEP, BHM, CBHM and MFM, two-stage design and $\eta = 0$ (models adjusting for covariates).

Scn	Method	FWER	P1	P2	P3	% Reject				100 x RMSE				Average Enrollment			
1						1	2	3	4	1	2	3	4	1	2	3	4
	SEP	10.0	0.0	0.0	0.0	2.2	2.7	2.9	2.5	24.0	24.1	31.1	29.5	32.0	31.9	21.3	21.3
	BHM	10.0	0.0	0.0	0.0	3.1	3.7	2.7	2.7	17.7	18.0	20.9	19.2	32.0	31.9	21.0	21.2
	CBHM	10.0	0.0	0.0	0.0	3.3	3.7	2.8	2.5	18.5	19.2	22.2	20.9	32.8	32.2	21.2	21.2
	MFM	10.0	0.0	0.0	0.0	3.6	3.8	2.8	2.5	13.2	13.2	13.8	13.2	32.0	31.9	21.0	20.9
2						[1]	2	3	4	[1]	2	3	4	[1]	2	3	4
	SEP	8.0	46.6	42.5	42.5	46.6	2.9	2.9	2.4	22.6	24.1	31.1	29.5	45.2	31.9	21.3	21.3
	BHM	14.1	36.3	30.9	30.9	36.3	5.7	5.3	5.4	24.6	18.5	21.5	19.9	43.5	33.1	21.8	22.2
	CBHM	11.4	46.9	41.3	41.3	46.9	4.7	4.2	3.9	23.3	19.4	23.1	21.8	46.1	32.7	21.7	21.7
	MFM	11.9	47.8	39.8	39.8	47.8	5.2	4.6	3.7	23.0	14.0	14.9	14.0	45.5	32.6	21.5	21.5
3						[1]	[2]	3	4	[1]	[2]	3	4	[1]	[2]	3	4
	SEP	5.3	69.0	65.0	18.7	46.6	41.9	2.9	2.4	22.6	23.2	31.1	29.5	45.2	43.8	21.3	21.3
	BHM	14.6	62.8	53.2	11.2	39.3	37.7	7.9	8.2	22.0	22.4	22.1	20.9	46.2	43.5	23.0	23.5
	CBHM	12.6	68.4	58.5	12.3	48.7	35.3	7.9	6.6	21.7	23.0	23.7	22.1	48.2	44.8	23.1	22.9
	MFM	11.5	74.1	64.0	25.2	54.4	49.3	7.6	5.1	20.8	21.4	16.0	14.9	47.9	46.9	22.5	22.1
4						[1]	[2]	[3]	4	[1]	[2]	[3]	4	[1]	[2]	[3]	4
	SEP	2.4	80.2	78.1	6.7	46.6	41.9	34.1	2.4	22.6	23.2	28.8	29.5	45.2	43.8	27.8	21.3
	BHM	11.0	75.7	67.1	6.2	42.4	34.4	39.5	11.0	20.4	21.0	23.2	22.5	48.0	44.8	30.2	24.5
	CBHM	9.3	76.1	68.0	6.8	46.8	34.3	33.9	9.3	20.8	22.0	23.7	23.5	49.1	45.8	30.3	23.9
	MFM	7.1	85.1	78.3	17.8	57.7	55.1	46.5	7.1	19.4	20.1	22.2	15.8	49.0	48.5	30.2	22.7
5						[1]	[2]	[3]	[4]	[1]	[2]	[3]	[4]	[1]	[2]	[3]	[4]
	SEP	0.0	86.7	86.7	2.4	46.8	41.8	34.0	31.7	22.6	23.2	28.8	27.4	45.2	43.8	27.8	27.5
	BHM	0.0	87.4	87.4	5.6	45.1	32.8	41.8	52.7	19.5	20.1	21.2	19.9	49.1	46.2	31.8	33.2
	CBHM	0.0	81.3	81.3	7.5	43.3	38.0	38.7	38.2	20.7	21.3	23.1	22.3	49.5	47.2	31.7	31.7
	MFM	0.0	90.9	90.9	19.4	63.5	60.5	51.9	52.4	18.1	18.8	20.6	20.3	50.6	50.1	31.5	31.4
6						1	[2]	[3]	[4]	1	[2]	[3]	[4]	1	[2]	[3]	[4]
	SEP	2.4	77.5	75.5	4.0	2.4	14.7	34.6	61.6	24.0	22.8	28.8	28.5	32.0	35.8	27.8	32.7
	BHM	10.8	70.7	62.7	3.8	10.8	19.8	30.6	50.4	19.9	19.2	24.8	27.5	36.4	39.3	29.3	33.8
	CBHM	7.8	67.5	62.4	3.6	7.8	19.2	33.0	41.1	19.8	19.1	24.4	28.0	36.5	38.8	29.2	33.4
	MFM	6.6	77.2	70.9	9.3	6.6	24.8	42.9	61.9	14.9	16.1	23.4	30.4	34.3	39.4	29.3	33.5

baskets get dropped in the first stage compared with the two-stage design with both futility and efficacy stopping. Some baskets not responding to treatment entered the second stage and consequently increased FWER and lower the power, resulting from information borrowing. These results suggest the importance of an early efficacy stopping rule for hierarchical models used in two-stage design to keep FWER and boost the power and lower RMSE. In contrast, MFM exhibits a very robust performance under this study design with high power than BHM, CBHM and SEP in all scenarios.

This study design also increases the RMSE with BHM and CBHM. Moreover, the average enrollment is increased considerably within this study design compared with the two-stage design with both futility and efficacy stopping.

4.4.4 Summary of Simulation Results

The performance comparisons of these methods are graphically shown in Figure 2 for models adjusting for covariates and Figure S.1 for models without covariates in the Supplementary Material.

BHM and CBHM demonstrate greater power in both the one-stage study design and the two-stage study design with both futility and efficacy stopping, but not in the two-stage study with futility stopping only; Adding the early stop for efficacy helps the borrowing methods to identify baskets responding to treatment before the second stage. With hierarchical methods, early efficacy stopping helps increase in power while controlling FWER and saving resource. Compared with BHM and CBHM, MFM yields more robust power gains in all study designs while controlling FWER in a model level in almost all scenarios.

4.5 Simulation Results for Mis-specified Models

In the above simulations, we assume that the covariates are strong predictors of the outcomes and need to be adjusted in the model. The following tables reanalyze the simulated data in the presence of covariate effects, using a model with no adjustment for covariates. Since the effects of covariates are not included in the analytical model, the independent simulations used to determine the tuning parameters also assume no covariate effect, i.e. the tuning parameter values in these additional simulations are the same as those in Table 2.

This simulation is referred to "Mis-specified Model 1" and their results are summarized in Table 6. As expected, the performance of all methods is affected by the mis-specified analytical model. The proposed methods still provide significant power gains in most scenarios, but all methods have greatly increased FWER inflation. The FWER inflation for MFM is drastic. The effect on SEP seems to be minimal,

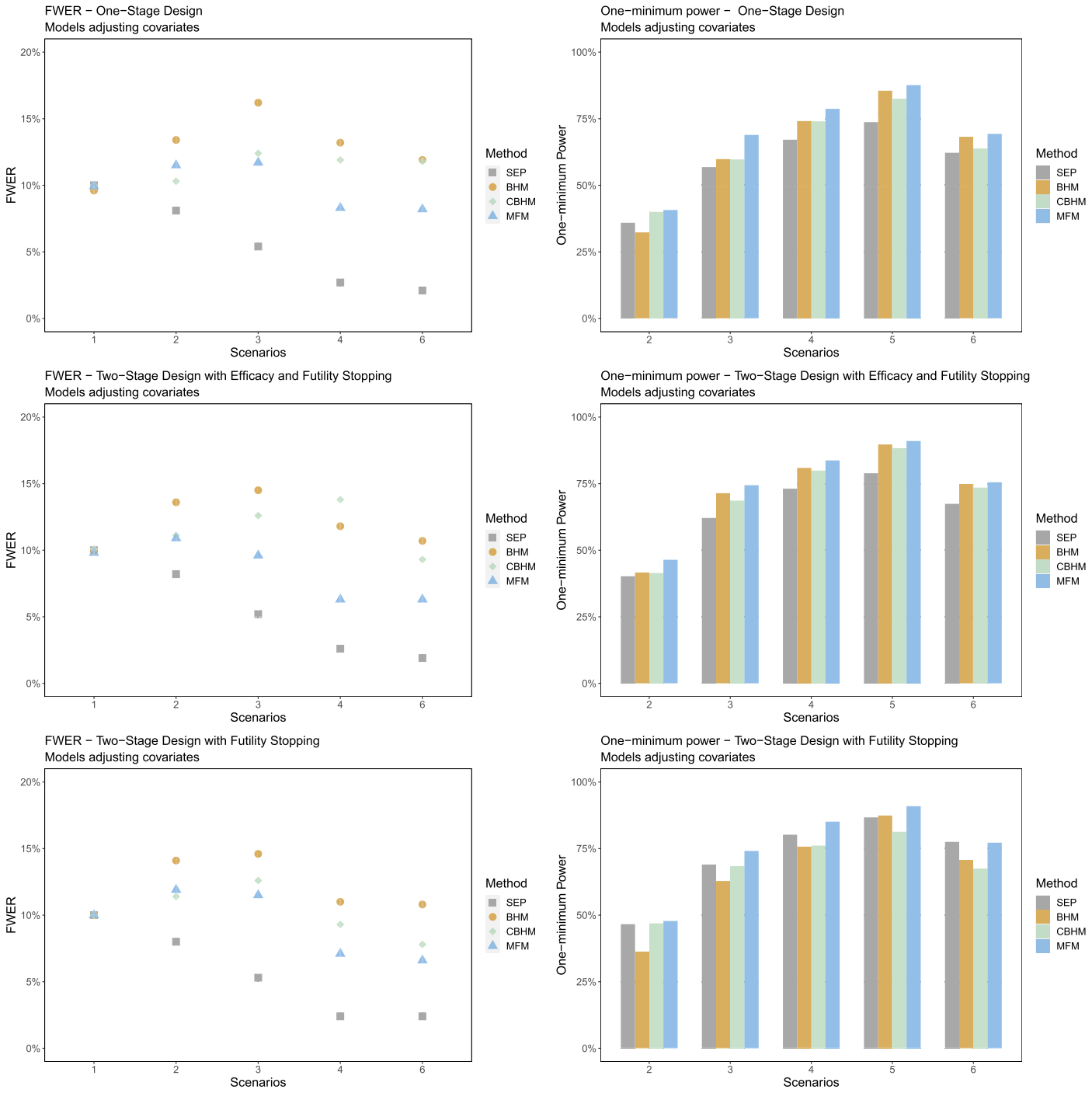

Figure 2: One-minimum Power and FWER for Bayesian Methods, models adjusting for covariates.

however, it may also depend on the magnitude of the covariates.

Another type of Mis-specified Model is those adjusting covariates of which the outcome is independent. These are referred to as "Mis-specified Model 2" and their results are summarized in Table 7. Our simulation results show that including negligible covariates does not impact the performance of our methods. Therefore, it is advantageous to include at least some potential predictors as the covariates.

4.6 Additional Simulation with Unbalanced Sample Sizes

As an extension our current simulation results, additional simulation is done with the doubled sample size of the treat-

Table 6. Performance of SEP, BHM, CBHM and MFM (Mis-specified Model 1).

		One-Stage Design				Two-Stage Design, $\eta = \frac{1}{2}$				Two-Stage Design, $\eta = 0$			
Scn	Method	FWER	P1	P2	P3	FWER	P1	P2	P3	FWER	P1	P2	P3
1	SEP	15.1	0.0	0.0	0.0	14.5	0.0	0.0	0.0	15.2	0.0	0.0	0.0
	BHM	16.3	0.0	0.0	0.0	16.7	0.0	0.0	0.0	16.5	0.0	0.0	0.0
	CBHM	13.6	0.0	0.0	0.0	13.6	0.0	0.0	0.0	13.4	0.0	0.0	0.0
	MFM	44.3	0.0	0.0	0.0	42.7	0.0	0.0	0.0	41.7	0.0	0.0	0.0
2	SEP	13.8	28.3	23.4	23.4	13.2	36.1	30.6	30.6	13.0	43.8	37.3	37.3
	BHM	22.2	33.5	20.2	20.2	23.0	42.3	24.4	24.4	22.3	37.3	25.4	25.4
	CBHM	19.4	23.5	13.3	13.3	19.6	34.0	21.4	21.4	14.0	34.5	26.7	26.7
	MFM	50.1	60.1	23.7	23.7	47.2	70.4	31.8	31.8	46.8	70.3	33.4	33.4
3	SEP	10.8	45.9	40.3	6.8	9.9	55.3	49.4	10.2	10.1	65.2	57.9	15.6
	BHM	24.0	59.5	38.8	15.1	23.0	67.9	46.2	20.1	23.6	61.0	41.8	8.5
	CBHM	21.6	44.7	27.4	5.5	22.3	56.4	36.9	10.6	15.5	54.0	41.7	6.1
	MFM	50.0	85.8	38.4	20.0	48.0	90.7	44.2	26.0	47.0	89.2	44.0	23.1
4	SEP	7.0	55.3	51.6	2.1	6.3	64.9	60.9	3.5	5.8	73.9	69.3	5.4
	BHM	19.6	74.2	55.8	10.6	18.3	80.3	62.4	16.7	17.3	75.0	58.9	8.4
	CBHM	20.8	65.3	46.4	4.9	24.1	74.2	51.0	8.3	11.3	66.9	56.6	4.4
	MFM	39.8	93.0	54.2	17.7	35.9	95.5	60.3	24.2	32.8	94.4	62.1	24.9
5	SEP	0.0	66.5	66.5	0.5	0.0	74.8	74.8	0.8	0.0	79.6	79.6	1.7
	BHM	0.0	83.7	83.7	11.8	0.0	88.7	88.7	24.4	0.0	82.6	82.6	5.5
	CBHM	0.0	73.9	73.9	13.0	0.0	83.4	83.4	26.1	0.0	72.7	72.7	4.1
	MFM	0.0	96.4	96.4	34.4	0.0	97.3	97.3	38.2	0.0	96.5	96.5	35.7
6	SEP	1.6	57.3	56.4	1.0	1.8	60.7	59.6	1.4	3.1	63.1	61.1	2.0
	BHM	16.8	61.8	46.9	5.6	14.4	68.8	55.2	9.3	15.8	62.6	51.0	4.3
	CBHM	8.6	56.8	48.7	3.5	9.5	65.5	56.3	5.8	9.2	57.8	50.4	2.1
	MFM	19.0	88.6	70.3	16.8	17.8	91.1	73.8	19.5	19.2	88.2	70.4	15.3

Table 7. Performance of SEP, BHM, CBHM and MFM (Mis-specified Model 2).

		One-Stage Design				Two-Stage Design, $\eta = \frac{1}{2}$				Two-Stage Design, $\eta = 0$			
Scn	Method	FWER	P1	P2	P3	FWER	P1	P2	P3	FWER	P1	P2	P3
1	SEP	10.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0
	BHM	9.6	0.0	0.0	0.0	9.3	0.0	0.0	0.0	10.0	0.0	0.0	0.0
	CBHM	10.0	0.0	0.0	0.0	10.1	0.0	0.0	0.0	10.0	0.0	0.0	0.0
	MFM	9.7	0.0	0.0	0.0	9.4	0.0	0.0	0.0	10.1	0.0	0.0	0.0
2	SEP	8.1	35.9	32.8	32.8	8.2	40.2	36.6	36.6	8.0	46.6	42.5	42.5
	BHM	13.4	32.3	24.0	24.0	12.7	38.6	29.8	29.8	14.1	36.3	30.9	30.9
	CBHM	10.3	40.0	33.7	33.7	11.1	41.4	34.3	34.3	11.4	46.9	41.3	41.3
	MFM	11.3	40.4	33.1	33.1	10.6	46.3	38.3	38.3	11.9	47.7	39.7	39.7
3	SEP	5.4	56.8	53.6	10.2	5.2	62.1	58.7	12.4	5.3	69.0	65.0	18.7
	BHM	16.2	59.8	47.2	11.8	13.9	68.8	56.1	22.4	14.6	62.8	53.2	11.2
	CBHM	12.4	59.7	48.7	10.8	12.6	68.6	57.1	17.7	12.6	68.4	58.5	12.3
	MFM	11.3	69.2	58.9	20.8	9.8	74.0	65.0	27.0	11.6	73.8	63.6	25.6
4	SEP	2.7	67.1	64.9	2.9	2.6	73.1	70.7	3.8	2.4	80.2	78.1	6.7
	BHM	13.2	74.1	61.9	7.9	11.5	80.5	69.5	17.5	11.0	75.7	67.1	6.2
	CBHM	11.9	74.0	62.5	6.3	13.8	79.9	66.4	11.9	9.3	76.1	68.0	6.8
	MFM	8.1	78.4	70.6	13.0	6.5	83.7	77.3	17.4	7.4	84.5	77.4	18.3
5	SEP	0.0	73.7	73.7	0.5	0.0	78.9	78.9	1.3	0.0	86.7	86.7	2.4
	BHM	0.0	85.5	85.5	10.3	0.0	90.3	90.3	22.4	0.0	87.4	87.4	5.6
	CBHM	0.0	82.5	82.5	14.8	0.0	88.3	88.3	27.5	0.0	81.3	81.3	7.5
	MFM	0.0	87.0	87.0	12.7	0.0	90.8	90.8	16.1	0.0	91.1	91.1	19.1
6	SEP	2.1	62.2	60.9	1.4	1.9	67.4	66.0	2.3	2.4	77.5	75.5	4.0
	BHM	11.9	68.2	57.1	4.7	10.2	75.2	65.4	9.6	10.8	70.7	62.7	3.8
	CBHM	11.8	63.8	53.2	3.2	9.3	73.5	64.9	6.2	7.8	67.5	62.4	3.6
	MFM	7.9	69.3	62.2	7.4	6.6	75.4	69.5	9.0	6.6	77.1	70.6	9.0

ment arm. In one stage, sample sizes for treatment arm are assumed to be $n_{t1} = 60$, $n_{t2} = 60$, $n_{t3} = 40$, and $n_{t4} = 40$; sample sizes for control arm are assumed to be $n_{c1} = 30$, $n_{c2} = 30$, $n_{c3} = 20$, and $n_{c4} = 20$. Their results are summarized in Table 8, Table 9 and Table 10. A larger cross model performance difference is observed. As expected, there was a significant power increase for all methods. For CBH and BHM, the inflation of type I errors is moderate, if not severe, as a cost of increased power. Similar to previously reported simulation results, implementing the early efficacy rule helps Bayesian methods in terms of power increase and Type I error control.

Table 8. Performance of SEP, BHM, CBHM and MFM, one-stage design (2:1 allocation).

Scn	Method	FWER	P1	P2	P3	% Reject				100 x RMSE			
1						1	2	3	4	1	2	3	4
	SEP	10.0	0.0	0.0	0.0	3.1	2.5	1.9	2.8	23.4	23.1	28.9	28.2
	BHM	10.0	0.0	0.0	0.0	4.8	3.7	2.1	3.8	15.6	15.3	16.7	16.3
	CBHM	9.9	0.0	0.0	0.0	6.1	2.5	2.0	3.0	18.1	17.4	19.4	18.9
	MFM	9.8	0.0	0.0	0.0	4.2	3.0	1.9	2.4	13.1	12.7	12.8	12.6
2						[1]	2	3	4	[1]	2	3	4
	SEP	7.2	43.0	39.8	39.8	43.0	2.5	1.9	2.8	23.4	23.1	28.9	28.2
	BHM	16.5	44.3	32.1	32.1	44.3	8.6	5.4	8.2	21.2	16.7	17.8	18.4
	CBHM	9.3	52.8	45.4	45.4	52.8	4.1	3.3	4.4	19.8	18.5	21.3	20.4
	MFM	11.7	50.9	42.2	42.2	50.9	5.7	3.4	3.6	20.7	14.2	14.4	14.2
3						[1]	[2]	3	4	[1]	[2]	3	4
	SEP	4.7	68.5	65.4	16.6	43.0	43.1	1.9	2.8	23.4	23.1	28.9	28.2
	BHM	22.6	77.8	56.7	24.3	58.0	57.1	12.9	15.0	17.9	18.0	20.6	22.2
	CBHM	13.5	76.9	64.1	26.0	61.9	50.6	8.4	9.7	18.6	18.7	23.7	22.5
	MFM	10.7	80.4	70.4	33.5	59.4	60.6	5.4	6.3	18.4	18.3	16.1	15.7
4						[1]	[2]	[3]	4	[1]	[2]	[3]	4
	SEP	2.8	78.8	76.6	5.4	43.0	43.1	30.5	2.8	23.4	23.1	28.9	28.2
	BHM	24.4	88.7	64.7	19.5	66.5	62.8	54.8	24.4	16.2	16.1	18.3	26.0
	CBHM	14.4	87.3	73.1	15.6	66.8	57.9	44.5	14.4	18.2	17.9	19.9	25.1
	MFM	8.4	90.5	82.4	22.8	64.7	64.6	51.6	8.4	17.1	16.9	19.3	17.1
5						[1]	[2]	[3]	[4]	[1]	[2]	[3]	[4]
	SEP	0.0	84.3	84.3	2.0	43.0	43.1	30.5	27.7	23.4	23.1	28.9	28.2
	BHM	0.0	95.6	95.6	30.6	71.0	65.8	66.2	76.4	15.6	15.3	16.7	16.3
	CBHM	0.0	93.8	93.8	29.7	74.1	67.4	58.6	58.9	18.1	17.4	19.4	18.9
	MFM	0.0	96.9	96.9	22.9	71.0	70.7	58.8	58.9	15.7	15.5	17.5	17.3
6						1	[2]	[3]	[4]	1	[2]	[3]	[4]
	SEP	3.1	75.6	73.4	2.7	3.1	15.2	30.5	58.0	23.4	23.1	28.9	28.2
	BHM	21.3	85.5	64.4	14.8	21.3	41.3	48.6	77.1	21.9	16.3	19.7	24.6
	CBHM	17.5	79.4	63.1	6.5	17.5	27.0	39.3	65.5	23.1	18.2	20.4	23.8
	MFM	10.4	85.1	75.1	12.0	10.4	30.3	48.4	70.8	16.2	14.7	20.2	29.5

Table 9. Performance of SEP, BHM, CBHM and MFM, two-stage design and $\eta = \frac{1}{2}$ (2:1 allocation).

Scn	Method	FWER	P1	P2	P3	% Reject				100 x RMSE				Average Enrollment			
1						1	2	3	4	1	2	3	4	1	2	3	4
	SEP	10.0	0.0	0.0	0.0	2.9	2.4	2.1	2.9	22.8	22.5	28.3	27.8	91.3	91.5	60.8	60.5
	BHM	9.9	0.0	0.0	0.0	4.3	3.1	2.0	3.2	15.4	15.1	16.6	16.0	91.9	91.9	60.8	61.3
	CBHM	10.0	0.0	0.0	0.0	5.4	2.7	2.4	2.6	17.4	17.1	19.0	18.3	92.6	91.4	60.8	61.0
	MFM	9.9	0.0	0.0	0.0	4.1	3.2	2.0	2.2	12.7	12.6	12.7	12.4	92.5	91.7	61.0	60.9
2						[1]	2	3	4	[1]	2	3	4	[1]	2	3	4
	SEP	7.4	49.2	45.5	45.5	49.2	2.4	2.1	2.9	23.7	22.5	28.3	27.8	96.8	91.5	60.8	60.5
	BHM	13.7	50.5	39.2	39.2	50.5	6.1	4.0	6.8	21.3	16.0	17.1	17.6	100.1	93.9	62.1	62.7
	CBHM	9.1	59.4	51.7	51.7	59.4	4.1	3.6	4.3	20.1	18.1	20.2	19.7	99.8	92.2	62.0	61.7
	MFM	12.2	60.9	50.9	50.9	60.9	5.9	3.8	3.8	20.1	13.7	13.9	13.7	101.8	93.8	62.0	61.8
3						[1]	[2]	3	4	[1]	[2]	3	4	[1]	[2]	3	4
	SEP	5.0	75.3	71.8	22.9	49.2	50.4	2.1	2.9	23.7	23.5	28.3	27.8	96.8	97.0	60.8	60.5
	BHM	16.4	82.0	66.5	36.5	65.3	66.3	7.7	11.2	17.9	18.1	19.0	20.6	100.2	101.2	63.8	64.8
	CBHM	11.6	82.4	71.3	31.6	67.0	56.0	7.2	7.2	18.8	19.5	21.6	21.2	98.4	103.4	63.6	62.7
	MFM	9.4	86.4	77.4	43.9	68.3	68.2	5.1	5.1	18.0	18.0	15.1	15.0	99.9	100.6	62.9	62.4
4						[1]	[2]	[3]	4	[1]	[2]	[3]	4	[1]	[2]	[3]	4
	SEP	2.9	85.4	82.8	8.4	49.2	50.4	36.2	2.9	23.7	23.5	29.0	27.8	96.8	97.0	63.7	60.5
	BHM	16.5	90.6	74.3	30.8	75.4	73.6	56.7	16.5	16.1	16.2	19.3	23.5	99.2	98.4	68.7	67.0
	CBHM	12.1	91.5	79.5	22.7	74.1	64.1	51.2	12.1	18.4	18.6	21.0	23.6	98.6	103.0	68.7	63.2
	MFM	7.2	94.9	87.8	31.7	73.8	72.9	58.0	7.2	16.6	16.7	19.5	16.0	100.3	99.9	68.2	63.0
5						[1]	[2]	[3]	[4]	[1]	[2]	[3]	[4]	[1]	[2]	[3]	[4]
	SEP	0.0	90.2	90.2	2.9	49.2	50.4	36.2	33.7	23.7	23.5	29.0	28.2	96.8	97.0	63.7	64.2
	BHM	0.0	97.5	97.5	44.2	82.9	81.2	66.3	75.1	15.8	15.4	17.6	17.2	97.6	96.2	67.4	68.0
	CBHM	0.0	97.1	97.1	35.8	80.8	72.6	62.5	65.3	18.4	17.9	20.4	19.7	97.3	99.7	67.5	67.1
	MFM	0.0	98.2	98.2	31.1	78.0	77.2	64.9	64.4	15.5	15.4	17.8	17.6	99.5	97.9	67.3	67.5
6						1	[2]	[3]	[4]	1	[2]	[3]	[4]	1	[2]	[3]	[4]
	SEP	2.9	80.8	78.6	3.4	2.9	16.8	36.2	63.8	22.8	22.6	29.0	28.4	91.3	93.8	63.7	63.9
	BHM	16.8	87.0	70.5	17.5	16.8	40.9	51.2	80.2	20.7	16.7	20.6	22.7	97.0	99.2	68.1	67.8
	CBHM	14.3	85.1	71.3	8.8	14.3	28.3	47.9	72.6	21.7	18.1	21.5	23.1	95.4	99.8	69.4	67.9
	MFM	10.3	89.9	79.9	15.7	10.3	35.2	55.2	80.4	15.4	15.0	20.3	27.7	94.8	100.6	68.4	66.8

Table 10. Performance of SEP, BHM, CBHM and MFM, two-stage design and $\eta = 0$ (2:1 allocation).

Scn	Method	FWER	P1	P2	P3	% Reject				100 x RMSE				Average Enrollment			
1						1	2	3	4	1	2	3	4	1	2	3	4
	SEP	10.0	0.0	0.0	0.0	3.3	2.9	1.8	2.6	21.3	21.1	26.5	25.7	94.5	94.3	62.6	62.6
	BHM	10.0	0.0	0.0	0.0	4.3	3.1	1.7	3.3	14.9	14.7	16.3	15.6	94.0	93.2	61.3	62.4
	CBHM	9.9	0.0	0.0	0.0	5.0	2.5	2.0	2.5	16.2	16.5	18.3	17.9	95.8	92.5	61.6	61.7
	MFM	10.0	0.0	0.0	0.0	4.2	3.2	2.1	2.2	12.4	12.2	12.4	12.2	94.5	93.4	61.6	61.5
2						[1]	2	3	4	[1]	2	3	4	[1]	2	3	4
	SEP	7.1	59.0	55.0	55.0	59.0	2.8	1.8	2.8	20.6	21.1	26.5	25.7	126.7	94.3	62.6	62.6
	BHM	15.1	46.8	38.5	38.5	46.8	6.4	4.6	6.8	22.8	15.3	16.6	16.5	121.5	97.0	63.5	64.8
	CBHM	9.5	59.5	52.7	52.7	59.5	3.5	3.5	4.2	20.5	17.2	19.4	18.8	127.3	94.4	63.2	63.0
	MFM	12.2	57.7	49.6	49.6	57.7	6.0	4.0	3.9	20.4	13.2	13.7	13.3	127.5	96.4	63.0	63.0
3						[1]	[2]	3	4	[1]	[2]	3	4	[1]	[2]	3	4
	SEP	4.3	83.6	80.0	32.4	59.1	58.6	1.8	2.7	20.6	20.6	26.5	25.7	126.7	126.9	62.6	62.6
	BHM	18.9	78.7	63.1	19.4	48.5	56.4	9.3	12.1	19.2	19.1	17.9	18.7	128.9	128.3	66.7	68.8
	CBHM	11.2	81.7	71.3	26.6	62.8	51.3	6.8	5.9	18.5	19.4	20.1	19.6	132.4	126.9	66.7	65.2
	MFM	10.4	80.3	71.4	30.9	58.0	57.8	5.4	5.9	18.0	18.0	14.7	14.5	132.2	132.4	64.9	64.2
4						[1]	[2]	[3]	4	[1]	[2]	[3]	4	[1]	[2]	[3]	4
	SEP	2.7	92.1	89.7	14.3	59.1	58.6	43.8	2.7	20.6	20.6	24.9	25.7	126.7	126.9	79.2	62.6
	BHM	20.2	89.0	70.1	13.9	55.3	50.1	56.4	20.2	16.9	17.1	19.2	20.7	133.5	130.3	85.3	73.3
	CBHM	8.7	87.9	79.4	15.2	62.2	52.9	45.9	8.7	17.5	18.2	19.3	20.9	135.4	130.9	83.2	67.6
	MFM	7.7	90.3	82.9	22.7	61.4	60.5	54.3	7.7	16.6	16.7	19.2	15.4	134.8	134.6	85.4	65.6
5						[1]	[2]	[3]	[4]	[1]	[2]	[3]	[4]	[1]	[2]	[3]	[4]
	SEP	0.0	95.7	95.7	6.7	59.1	58.8	43.7	43.2	20.6	20.6	24.9	24.5	126.7	126.9	79.2	79.0
	BHM	0.0	96.5	96.5	13.1	56.6	44.3	63.2	73.7	15.9	16.1	16.7	15.3	135.7	131.3	89.0	92.8
	CBHM	0.0	92.4	92.4	19.1	62.1	56.6	57.7	51.2	17.0	17.2	18.0	18.5	137.0	134.5	88.0	87.3
	MFM	0.0	95.0	95.0	24.2	65.8	64.7	61.6	60.4	15.2	15.2	17.5	17.3	137.3	137.1	88.0	87.7
6						1	[2]	[3]	[4]	1	[2]	[3]	[4]	1	[2]	[3]	[4]
	SEP	3.4	88.1	85.4	6.2	3.4	20.4	43.8	73.2	21.3	19.6	24.9	25.1	94.5	106.0	79.2	89.6
	BHM	17.2	85.7	69.6	11.1	17.2	38.6	43.1	68.6	18.2	16.4	21.3	22.7	105.7	117.6	82.6	93.0
	CBHM	11.0	82.3	72.4	4.9	11.0	27.5	46.5	50.0	18.3	16.1	20.0	24.1	103.9	111.5	82.0	89.8
	MFM	9.5	87.2	78.7	11.5	9.5	34.6	54.4	64.2	14.4	15.0	20.4	25.0	100.3	115.0	84.4	92.0

5. AN APPLICATION EXAMPLE

Note that as the assumptions of the basket trial are extended in many aspects, it is currently difficult to find a published study dataset of basket trials that satisfies these assumptions at the same time. We do find one example mentioned by Ouma et al. [23]: a study trial comparing pulse rates after participants performed three kinds of exercises between different dietary interventions. This study is similar to a basket trial with a controlled arm. Two diet types (low-fat and non low-fat) are considered, and low-fat is set as the control. The data source included two sub-studies that can be regarded as two stages. Each sub-study involved 30 participants randomly assigned to three exercises and two diets. Pulse rate measurement is the continuous endpoint of interest.[1]

We attempted to explore two model designs, the first assuming that the outcome estimates (pulse rate after exercise) are independent of any covariates; the second assessing the change from the baseline and model the outcome with adjusting the baseline values as the covariate. The results are summarized in Tables 11, 12 and 13 for one-stage design, two-stage design with futility and efficacy, and two-stage design with futility only, respectively. Also, note the one-stage design applied all data from the two sub-studies instead of only from the first sub-study.

Table 11. Bayesian methods applied to the case study One-Stage Design.

Non-Cov Model	q_1(%)	Estimates (95% HPD Interval) 1	2	3
SEP		2.7 (-3.2, 8.8)	4.4 (-1.4, 10.6)	26.4 (20.3, 32.2)
BHM		3.9 (-2.2, 11.5)	5.6 (0.1, 11.5)	24.1 (18.1, 29.9)
CBHM		3.3 (-2.7, 9.8)	5.1 (-1.1, 10.5)	24.9 (18.6, 30.7)
MFM		3.6 (-1.1, 8.4)	3.6 (-1.1, 8.4)	26.4 (20.0, 32.7)
		Posterior Probability (%)		
SEP	97.4	81.6	92.8	100.0
BHM	95.3	89.8	97.1	100.0
CBHM	95.0	85.6	95.2	100.0
MFM	90.0	94.7	94.7	100.0
Cov-Adj Model	q_1(%)	Estimates (95% HPD Interval) 1	2	3
SEP		0.3 (-4.5, 5.4)	-0.3 (-5.4, 5.2)	28.4 (22.8, 33.8)
BHM		0.6 (-4.1, 6.0)	0.0 (-5.0, 4.6)	27.3 (21.1, 32.1)
CBHM		1.0 (-4.3, 5.9)	0.5 (-4.6, 5.8)	26.9 (21.0, 32.3)
MFM		0.2 (-3.2, 3.9)	0.2 (-3.2, 3.9)	28.3 (23.0, 33.1)
		Posterior Probability (%)		
SEP	98.1	55.4	45.8	100.0
BHM	95.7	59.4	51.5	100.0
CBHM	96.7	64.9	56.3	100.0
MFM	92.9	54.3	54.3	100.0

The performance of the model may depend on the tuning parameters in our simulations, for illustration purposes we simply placed the non-informative prior in this example. The estimates of treatment effects and their 95% Highest

[1]The dataset can be found at https://stats.oarc.ucla.edu/r/seminars/repeated-measures-analysis-with-r.

Table 12. Bayesian methods applied to the case study Two-Stage Design with Futility and Efficacy.

Non-Cov Model	q_1(%)	Estimates (95% HPD Interval) 1	2	3
SEP		3.2 (-4.1, 11.3)	1.4 (-6.6, 8.8)	29.2 (21.4, 36.6)
BHM		4.7 (-3.3, 13.2)	3.3 (-4.0, 11.7)	25.9 (17.8, 33.6)
CBHM		4.3 (-3.4, 12.3)	3.0 (-4.5, 10.7)	26.4 (18.9, 34.2)
MFM		2.4 (-4.6, 8.6)	2.4 (-4.6, 8.6)	29.5 (20.4, 37.8)
		Posterior Probability (%)		
SEP	96.9	79.9	64.4	100.0
BHM	93.9	87.9	79.9	100.0
CBHM	94.0	85.9	78.4	100.0
MFM	89.4	75.6	75.6	100.0
Cov-Adj Model	q_1(%)	Estimates (95% HPD Interval) 1	2	3
SEP		0.8 (-5.6, 7.7)	-1.4 (-8.7, 5.4)	31.2 (23.7, 38.3)
BHM		1.5 (-5.3, 8.2)	-0.6 (-7.5, 5.4)	29.1 (21.0, 36.7)
CBHM		2.0 (-5.6, 8.5)	0.3 (-7.0, 7.7)	28.2 (20.1, 35.8)
MFM		-0.2 (-4.8, 4.0)	-0.2 (-4.8, 4.0)	31.0 (23.8, 38.1)
		Posterior Probability (%)		
SEP	97.4	59.1	34.4	100.0
BHM	94.6	66.9	42.7	100.0
CBHM	95.7	72.2	50.9	100.0
MFM	90.1	46.2	46.2	100.0

Table 13. Bayesian methods applied to the case study Two-Stage with Futility Only.

Non-Cov Model	q_1(%)	Estimates (95% HPD Interval) 1	2	3
SEP		3.2 (-4.1, 11.3)	1.4 (-6.6, 8.8)	26.4 (20.0, 32.8)
BHM		4.7 (-3.3, 13.2)	3.3 (-4.0, 11.7)	26.4 (20.0, 32.8)
CBHM		4.3 (-3.4, 12.3)	3.0 (-4.5, 10.7)	26.4 (20.0, 32.8)
MFM		2.4 (-4.6, 8.6)	2.4 (-4.6, 8.6)	26.4 (20.0, 32.8)
		Posterior Probability (%)		
SEP	96.9	79.9	64.4	100.0
BHM	93.9	87.9	79.9	100.0
CBHM	94.0	85.9	78.4	100.0
MFM	89.4	75.6	75.6	100.0
Cov-Adj Model	q_1(%)	Estimates (95% HPD Interval) 1	2	3
SEP		0.8 (-5.6, 7.7)	-1.4 (-8.7, 5.4)	28.4 (21.3, 35.8)
BHM		1.5 (-5.3, 8.2)	-0.6 (-7.5, 5.4)	28.4 (21.3, 35.8)
CBHM		2.0 (-5.6, 8.5)	0.3 (-7.0, 7.7)	28.4 (21.3, 35.8)
MFM		-0.2 (-4.8, 4.0)	-0.2 (-4.8, 4.0)	28.4 (21.3, 35.8)
		Posterior Probability (%)		
SEP	94.1	59.1	34.4	100.0
BHM	91.5	66.9	42.7	100.0
CBHM	92.9	72.2	50.9	100.0
MFM	86.1	46.2	46.2	100.0

Posterior Density (HPD) Interval are summarized in the tables. The cut-off value is firstly determined based on the simulated global null (GN) scenario. The GN scenario was simulated by bootstrapping the control arm data. The cut-off values of q_1 controlled the FWER to be 10% under GN and they are summarized in the result Tables. The posterior probability $Pr(\tau_k > 0|\mathcal{D})$ are also summarized to be compared with q_1.

For one-stage design, since all data are included in the analysis, borrowing resulted in the overestimate of the effects in small-effect-baskets such as baskets 1 and 2. In contrast, a two-stage design allows baskets that meet the futility and/or efficacy stopping rule to drop at interim analysis. Thus the two-stage designs helped reducing Type I Error Rate inflation and the estimates of baskets 1 and 2 are not much overestimated. BHM tend to have the largest estimates for baskets 1 and 2, which is also reflected in the high Type I error. In Table 13, the reason that all methods have the same estimates for basket 3 is that the other two baskets are dropped in stage one, and only basket 3 enters the second stage and it is evaluated the same way as SEP. In all designs, MFM tends to clustered baskets 1 and 2 together as they both have small effects. Comparing the methods discussed so far, a similar pattern is observed in our simulation results.

6. DISCUSSION

Basket trials study design are innovative for clinical development, which can bring effective treatment to patients faster. To achieve full strength of basket trials, sophisticated statistical analysis methods are needed. In the context of identifying true efficacious treatment for further evaluation, we pilot four Bayesian methods when apply to 3 study designs, and have demonstrated the feasibility of adding a control arm as well as adding covariates to the modeling.

Modifications to CBHM and MFM are proposed. For CBHM, we extend [6] to the continuous outcomes with a few changes to obtain similar performance. For MFM, tuning parameters are introduced into the prior settings to optimize the borrowing effect. Based on our simulation, each of CBHM and MFM performs better than others in some but not all scenarios. In general, MFM has more robust performance, and it has more power while lower inflated FWER than other methods in a wide range of scenarios.

For the one-stage design, the CBHM is an improvement over the BHM in terms of better control of the FWER while sacrificing a small amount of power. Both models surpass the power of the Bayesian Separate Model. For the two-stage design with only a futility rule, CBHM and BHM lose some of their power improvements because their full borrowing strategy creates a tendency to borrow from the futility basket in the second stage, dragging down the efficacy basket estimates. When the early efficacy stopping rule is incorporated in the two-stage design, CBHM and BHM gain improvement because the efficacy stopping rule prevents the truly active baskets from being mixed with the ineffective baskets. Meanwhile, MFM shows robustness in all study designs due to its allowance for local borrowing and deliberate prior parameters setting. The potential downside is the computation time when sample size getting large. For moderate sample size, as what this paper has considered, computation time does not seem to be an issue. It only take a few hours to complete 1000 simulations. Another downside is the lack of standard software for implementation. One needs to derive the complicated formulas to run model for different applications.

In our research, many assumptions were set at the beginning in order to test the modeling performance and study design. However, many of the assumptions can be either removed or made more flexible. First, all methods may assume that the baskets have different within basket variances

rather than a unified variance. To implement this, update the Gibbs sampling algorithm for BHM, CBHM and MFM, e.g. update σ_k^2 instead of σ^2 in (3.15) and (3.21). Second, the sample size can be different between the control and treatment arms, and between stages. In addition, the sample size determination is not our focus thus not mentioned in this article. Moreover, we arbitrarily set $\delta = 0$ for all stages; and set the parameter $\eta = 1/2$ that links the two cut-off values of futility and efficacy stopping rules. These settings may affect the alpha-split in two-stage trials. Many different considerations of these settings appear in the literature. All of the aforementioned may not be the main research focus in this work, but may be potential extensions for future work.

Another important extension of the work is the consideration of other types of endpoints, for example, count data and time-to-event data. Some of the models that can be considered are the Poisson regression model and the proportional hazards model. In a similar way to a continuous endpoint, a Bayesian hierarchical structure can be added to the treatment effect part of the model. When non-conjugacy modeling is considered, other sampling algorithms can be explored for both CBHM and MFM. Some references can be found in articles by Miller and Harrison [18, 19], and Neal [20].

SUPPLEMENTARY MATERIAL

The Supplementary Material includes the models without covariates supplementary results (Section S.1), sensitivity analysis of MFM (Section S.2), derivation of MFM sampling algorithm formula (Section S.3), and testing treatment difference among baskets for CBHM (Section S.4).

ACKNOWLEDGEMENTS

The authors would like to thank the Section Editors, the Associate Editor, and a reviewer for their useful comments and suggestions, which have led to a much improved version of our paper.

Accepted 5 August 2023

REFERENCES

[1] Amado, R. G., Wolf, M., Peeters, M., Van Cutsem, E., Siena, S., Freeman, D. J., Juan, T., Sikorski, R., Suggs, S., Radinsky, R. et al. (2008). Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *Journal of Clinical Oncology* **26**(10) 1626–1634.

[2] Ashley, E. A. (2015). The precision medicine initiative: a new national effort. *Jama* **313**(21) 2119–2120.

[3] Ashley, E. A. (2016). Towards precision medicine. *Nature Reviews Genetics* **17**(9) 507–522.

[4] Berry, S. M., Broglio, K. R., Groshen, S. and Berry, D. A. (2013). Bayesian hierarchical modeling of patient subpopulations: efficient designs of phase II oncology clinical trials. *Clinical Trials* **10**(5) 720–734.

[5] Chapman, P. B., Hauschild, A., Robert, C., Haanen, J. B., Ascierto, P., Larkin, J., Dummer, R., Garbe, C., Testori, A., Maio, M. et al. (2011). Improved survival with Vemurafenib in melanoma with BRAF V600E mutation. *New England Journal of Medicine* **364**(26) 2507–2516.

[6] Chu, Y. and Yuan, Y. (2018). A Bayesian basket trial design using a calibrated Bayesian hierarchical model. *Clinical Trials* **15**(2) 149–158.

[7] Chu, Y. and Yuan, Y. (2018). BLAST: Bayesian latent subgroup design for basket trials accounting for patient heterogeneity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **67**(3) 723–740. https://doi.org/10.1111/rssc.12255. MR3787974

[8] Cunanan, K. M., Gonen, M., Shen, R., Hyman, D. M., Riely, G. J., Begg, C. B. and Iasonos, A. (2017). Basket trials in oncology: a trade-off between complexity and efficiency. *Journal of Clinical Oncology* **35**(3) 271.

[9] Freidlin, B. and Korn, E. L. (2013). Borrowing Information across Subgroups in Phase II Trials: Is It Useful? Borrowing Information across Subgroups. *Clinical Cancer Research* **19**(6) 1326–1334.

[10] Geng, J. and Hu, G. (2020). Mixture of finite mixtures model for basket trial. *arXiv preprint arXiv:2011.04135*.

[11] Hirakawa, A., Asano, J., Sato, H. and Teramukai, S. (2018). Master protocol trials in oncology: review and new trial designs. *Contemporary Clinical Trials Communications* **12** 1–8.

[12] Hyman, D. M., Puzanov, I., Subbiah, V., Faris, J. E., Chau, I., Blay, J. -Y., Wolf, J., Raje, N. S., Diamond, E. L., Hollebecque, A. et al. (2015). Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations. *New England Journal of Medicine* **373**(8) 726–736.

[13] Kaizer, A. M., Koopmeiners, J. S., Kane, M. J., Roychoudhury, S., Hong, D. S. and Hobbs, B. P. (2019). Basket designs: statistical considerations for oncology trials. *JCO Precision Oncology* **3** 1–9.

[14] Kumar-Sinha, C. and Chinnaiyan, A. M. (2018). Precision oncology in the age of integrative genomics. *Nature Biotechnology* **36**(1) 46–60.

[15] Li, B. T., Shen, R., Buonocore, D., Olah, Z. T., Ni, A., Ginsberg, M. S., Ulaner, G. A., Offin, M., Feldman, D., Hembrough, T. et al. (2018). Ado-trastuzumab emtansine for patients with HER2-mutant lung cancers: results from a phase II basket trial. *Journal of Clinical Oncology* **36**(24) 2532.

[16] Lopez-Chavez, A., Thomas, A., Rajan, A., Raffeld, M., Morrow, B., Kelly, R., Carter, C. A., Guha, U., Killian, K., Lau, C. C. et al. (2015). Molecular profiling and targeted therapy for advanced thoracic malignancies: a biomarker-derived, multiarm, multihistology phase II basket trial. *Journal of Clinical Oncology* **33**(9) 1000.

[17] Mehta, C. R. and Pocock, S. J. (2011). Adaptive increase in sample size when interim results are promising: a practical guide with examples. *Statistics in Medicine* **30**(28) 3267–3284. https://doi.org/10.1002/sim.4102. MR2861612

[18] Miller, J. W. and Harrison, M. T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association* **113**(521) 340–356. https://doi.org/10.1080/01621459.2016.1255636. MR3803469

[19] Neal, R. M. (1992). Bayesian mixture modeling. In *Maximum Entropy and Bayesian Methods* 197–211 Springer.

[20] Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**(2) 249–265. https://doi.org/10.2307/1390653. MR1823804

[21] Neuenschwander, B., Wandel, S., Roychoudhury, S. and Bailey, S. (2016). Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharmaceutical Statistics* **15**(2) 123–134.

[22] Oaknin, A., Friedman, C. F., Roman, L. D., D'Souza, A., Brana, I., Bidard, F. -C., Goldman, J., Alvarez, E. A., Boni, V., ElNaggar, A. C. et al. (2020). Neratinib in patients

with HER2-mutant, metastatic cervical cancer: Findings from the phase 2 SUMMIT basket trial. *Gynecologic oncology* **159**(1) 150–156.

[23] Ouma, L. O., Grayling, M. J., Wason, J. M. and Zheng, H. (2022). Bayesian modelling strategies for borrowing of information in randomised basket trials. *Journal of the Royal Statistical Society Series C: Applied Statistics* **71**(5) 2014–2037. https://doi.org/10.1111/rssc.12602. MR4511139

[24] Pan, J., Bunn, V., Hupf, B. and Lin, J. (2022). Bayesian Additive Regression Trees (BART) with covariate adjusted borrowing in subgroup analyses. *Journal of Biopharmaceutical Statistics* **32**(4) 613–626.

[25] Park, J. J., Siden, E., Zoratti, M. J., Dron, L., Harari, O., Singer, J., Lester, R. T., Thorlund, K. and Mills, E. J. (2019). Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols. *Trials* **20**(1) 1–10.

[26] Park, J. J., Hsu, G., Siden, E. G., Thorlund, K. and Mills, E. J. (2020). An overview of precision oncology basket and umbrella trials for clinicians. *CA: A Cancer Journal for Clinicians* **70**(2) 125–137.

[27] Patel, S. P., Othus, M., Chae, Y. K., Giles, F. J., Hansel, D. E., Singh, P. P., Fontaine, A., Shah, M. H., Kasi, A., Baghdadi, T. A. et al. (2020). A Phase II Basket Trial of Dual Anti–CTLA-4 and Anti–PD-1 Blockade in Rare Tumors (DART SWOG 1609) in Patients with Nonpancreatic Neuroendocrine TumorsIpilimumab and Nivolumab in Rare Tumors S1609: Neuroendocrine. *Clinical Cancer Research* **26**(10) 2290–2296.

[28] Redig, A. J. and Jänne, P. A. (2015). Basket trials and the evolution of clinical trial design in an era of genomic medicine. *J Clin Oncol* **33**(9) 975–977.

[29] Redman, M. W. and Allegra, C. J. (2015). The master protocol concept. In *Seminars in Oncology* **42** 724–730. Elsevier.

[30] Renfro, L. and Sargent, D. (2017). Statistical controversies in clinical research: basket trials, umbrella trials, and other master protocols: a review and examples. *Annals of Oncology* **28**(1) 34–43.

[31] Schuck, R. N., Pacanowski, M., Woodcock, J. and Zineh, I. (2018). Regulatory policy for the development of targeted therapies for low-frequency molecular subtypes of disease. *Nature Reviews Drug Discovery* **17**(2) 79–80.

[32] Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* **10**(1) 1–10.

[33] Simon, R., Geyer, S., Subramanian, J. and Roychowdhury, S. (2016). The Bayesian basket design for genomic variant-driven phase II trials. In *Seminars in Oncology* **43** 13–18. Elsevier.

[34] Sleijfer, S., Bogaerts, J. and Siu, L. L. (2013). Designing transformative clinical trials in the cancer genome era. *Journal of Clinical Oncology* **31**(15) 1834–1841.

[35] Tao, J. J., Schram, A. M. and Hyman, D. M. (2018). Basket studies: redefining clinical trials in the era of genome-driven oncology. *Annual Review of Medicine* **69** 319.

[36] Thall, P. F., Wathen, J. K., Bekele, B. N., Champlin, R. E., Baker, L. H. and Benjamin, R. S. (2003). Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. *Statistics in Medicine* **22**(5) 763–780.

[37] U. S. Food and Drug Administration (2018). U.S. Food And Drug Administration: Developing Targeted Therapies in Low-Frequency Molecular Subsets of a Disease. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/developing-targeted-therapies-low-frequency-molecular-subsets-disease.

[38] U. S. Food and Drug Administration (2018). U.S. Food And Drug Administration: Master Protocols: Efficient Clinical Trial Design Strategies To Expedite Development of Oncology Drugs and Biologics. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/master-protocols-efficient-clinical-trial-design-strategies-expedite-development-oncology-drugs-and.

[39] U. S. Food and Drug Administration (2021). E9(R1) Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e9r1-statistical-principles-clinical-trials-addendum-estimands-and-sensitivity-analysis-clinical.

[40] U. S. Food and Drug Administration (2023). Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products. https://www.fda.gov/media/148910/download.

[41] Wen, P. Y., Stein, A., van den Bent, M., De Greve, J., Wick, A., de Vos, F. Y., von Bubnoff, N., van Linde, M. E., Lai, A., Prager, G. W. et al. (2022). Dabrafenib plus trametinib in patients with BRAFV600E-mutant low-grade and high-grade glioma (ROAR): a multicentre, open-label, single-arm, phase 2, basket trial. *The Lancet Oncology* **23**(1) 53–64.

[42] Woodcock, J. and LaVange, L. M. (2017). Master protocols to study multiple therapies, multiple diseases, or both. *New England Journal of Medicine* **377**(1) 62–70.

[43] Xu, Y., Müller, P., Tsimberidou, A. M. and Berry, D. (2019). A nonparametric Bayesian basket trial design. *Biometrical Journal* **61**(5) 1160–1174. https://doi.org/10.1002/bimj.201700162. MR4013340

[44] Zheng, H. and Wason, J. M. (2022). Borrowing of information across patient subgroups in a basket trial based on distributional discrepancy. *Biostatistics* **23**(1) 120–135. https://doi.org/10.1093/biostatistics/kxaa019. MR4366039

[45] Zhou, T. and Ji, Y. (2021). RoBoT: a robust Bayesian hypothesis testing method for basket trials. *Biostatistics* **22**(4) 897–912. https://doi.org/10.1093/biostatistics/kxaa005. MR4325734

Cheng Huang. Department of Statistics, University of Connecticut, USA.
E-mail address: cheng.huang@uconn.edu

Chenghao Chu. Biostatistics Department, Vertex Pharmaceuticals, USA.
E-mail address: chenghao_chu@vrtx.com

Yimeng Lu. Biostatistics Department, Vir Biotechnology, USA.
E-mail address: ylu@vir.bio

Bingming Yi. Biostatistics Department, Vertex Pharmaceuticals, USA.
E-mail address: bingming_yi@vrtx.com

Ming-Hui Chen. Department of Statistics, University of Connecticut, USA.
E-mail address: ming-hui.chen@uconn.edu