Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

BAGS: A Bayesian Adaptive Group Sequential Trial Design With Subgroup-Specific Survival Comparisons

Ruitao Lin, Peter F. Thall & Ying Yuan

To cite this article: Ruitao Lin, Peter F. Thall & Ying Yuan (2021) BAGS: A Bayesian Adaptive Group Sequential Trial Design With Subgroup-Specific Survival Comparisons, Journal of the American Statistical Association, 116:533, 322-334, DOI: 10.1080/01621459.2020.1837142

To link to this article: https://doi.org/10.1080/01621459.2020.1837142

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 116, NO. 533, 322–334: Theory and Methods Special Issue on Precision Medicine and Individualized Policy Discovery
https://doi.org/10.1080/01621459.2020.1837142

BAGS: A Bayesian Adaptive Group Sequential Trial Design With Subgroup-Specific Survival Comparisons

Ruitao Lin, Peter F. Thall, and Ying Yuan

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX

ABSTRACT

A Bayesian group sequential design is proposed that performs survival comparisons within patient subgroups in randomized trials where treatment–subgroup interactions may be present. A latent subgroup membership variable is assumed to allow the design to adaptively combine homogeneous subgroups, or split heterogeneous subgroups, to improve the procedure's within-subgroup power. If a baseline covariate related to survival is available, the design may incorporate this information to improve subgroup identification while basing the comparative test on the average hazard ratio. General guidelines are provided for calibrating prior hyperparameters and design parameters to control the overall Type I error rate and optimize performance. Simulations show that the design is robust under a wide variety of different scenarios. When two or more subgroups are truly homogeneous but differ from the other subgroups, the proposed method is substantially more powerful than tests that either ignore subgroups or conduct a separate test within each subgroup. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received June 2019
Accepted August 2020

KEYWORDS
Bayesian analysis; Group sequential; Piecewise exponential model; Randomized comparative trial; Response heterogeneity; Survival comparison

1. Introduction

This article describes a Bayesian adaptive design for randomized clinical trials with heterogeneous patients that performs subgroup-specific group sequential (GS) tests to compare survival between an experimental treatment, E, and a control, C. While the proposed methodology is quite general, it may be motivated by a recent randomized, open-label, phase III trial in patients with non-small-cell lung cancer (OAK) (Rittmeyer et al. 2017). In the OAK trial, patients were randomized to receive either docetaxel, which has been the standard of care for second-line or third-line treatment, or atezolizumab, an antibody that targets the humanized antiprogrammed death-ligand 1 (PD-L1) pathway. Docetaxel has been demonstrated to have efficacy against lung cancer, but it has substantial toxic effects. As an immunotherapy, atezolizumab is generally safer and has been shown to be promising in phase II studies. The objective of the OAK trial was to perform a confirmatory comparison of survival between atezolizumab and docetaxel. Patients were stratified into four subgroups based on PD-L1 expression: subgroup 1 (TC0 or IC0) was defined as PD-L1 expression on $\leq$1% of tumor cells (TC) or tumor-infiltrating immune cells (IC); subgroup 2 (TC1 or IC1) was defined as PD-L1 expression on $\geq$1% but $\leq$5% of these cells; subgroup 3 (TC2 or IC2) was defined as PD-L1 expression on $\geq$ 5% but $\leq$50% (or 10%) of TC (or IC); and the remaining patients were stratified into subgroup 4 (TC3 and IC3). Heterogeneity between subgroups was observed in Rittmeyer et al. (2017): according to the Kaplan–Meier curves of overall survival (Figure 1), for patients treated with atezolizumab, median survival time was 12.6 months in the TC0 or IC0 patients, compared to 20.5 months in the TC3 or IC3 subgroup. In contrast, for patients treated with docetaxel, both of these subgroups had median overall survival 8.9 months. Thus, the estimated atezolizumab-versus-docetaxel differences in median survival were 3.7 months in subgroup 1, and 11.6 months in subgroup 4. Additional post hoc analyses were reported by Rittmeyer et al. (2017) in which the data from different subgroups were combined iteratively. While the above numerical results may seem compelling because they suggest a much larger atezolizumab-versus-docetaxel effect in the TC3 or IC3 subgroup, they are a typical example of post hoc estimation or testing of comparative effects within subgroups. This common practice has been the subject of much debate for many years. It may be criticized as data dredging, because the probability of a false positive conclusion increases with the number of analyses and a large estimated subgroup-specific effect detected in this way may be due to chance. Bechhofer, Santner, and Goldsman (1995) and Hsu (1996) provided general developments of selection and multiple testing, and practical recommendations for subset analysis are given by Pocock et al. (2002) and Wang et al. (2007), among many others.

Still, in the era of precision medicine, heterogeneity may be defined by a wide variety of biomarkers, and the goal of identifying subgroups where targeted agents are most effective is a central issue. For example, in oncology, due to diverse tumor genomics, cancer cells may interact differentially with their surrounding microenvironment, see Seoane and De Mattos-Arruda (2014). This may cause patients in different biomarker

CONTACT Peter F. Thall rex@mdanderson.org Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TA 77030.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

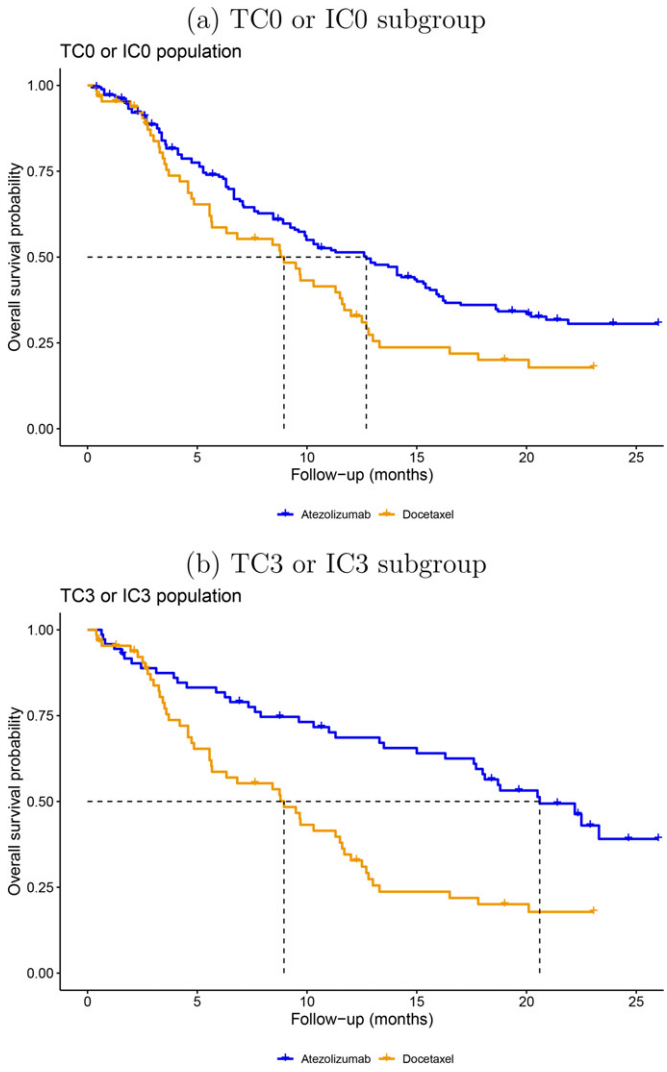

Figure 1. Reconstructed Kaplan–Meier survival curves for overall survival under the atezolizumab and docetaxel arms in the trial reported by Rittmeyer et al. (2017). Two subgroups are present: the TC0 or IC0 subgroup, defined as PD-L1 expression on less than 1% of tumor cells or tumor-infiltrating immune cells; and the TC3 or IC3 subgroup, defined as PD-L1 expression on 50% or more of tumor cells or on 10% of tumor-infiltrating immune cells. The patients treated with docetaxel had median survival 8.9 months in both subgroups. In contrast, for patients treated with atezolizumab, median survival time was 12.6 months in the TC0 or IC0 subgroup, compared to 20.5 months in the TC3 or IC3 subgroup.

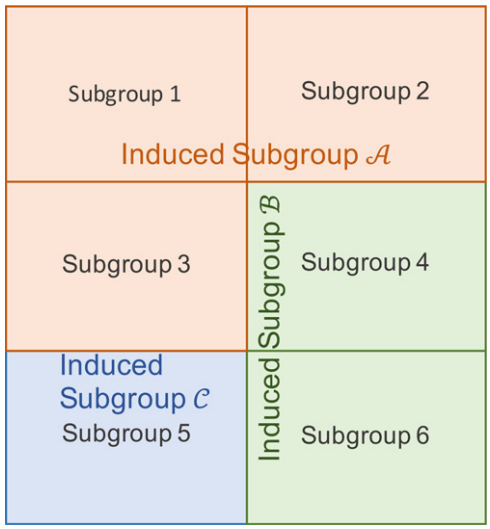

Figure 2. Illustration on induced subgroup classification based on six subgroups. In this example, there are a total of three induced subgroups. In particular, subgroups 1–3 are homogeneous and thus they are in the induced subgroup $\mathcal{A}$, subgroups 4 and 6 are in induced subgroup $\mathcal{B}$, and induced subgroup $\mathcal{C}$ only contains subgroup 5. Borrowing the notation described in Section 2.2, we have three induced subgroups $\mathcal{S} = \{1, 4, 5\}$ with the latent subgroup variables $z_1 = z_2 = z_3 = 1$, $z_4 = z_6 = 4$, and $z_5 = 5$.

subgroups to respond differently to a targeted treatment. Such biological heterogeneity brings many new challenges to clinical trial design and analysis (Garralda et al. 2019), including planning a trial where the subgroups are given a priori but heterogeneity of treatment effects between subgroups is unknown (Thall et al. 2003; Chapple and Thall 2018; Murray et al. 2018); addressing the multiple testing issue if heterogeneity is known (Rosenblum et al. 2016); and identifying a subgroup that may respond more fully to a new agent (Simon and Simon 2013).

We consider the problem of designing a randomized trial to compare survival time of E versus C in settings where, a priori, the patient population has been partitioned into subgroups, often defined biologically, but it is not known whether E will have effects of different magnitudes in the subgroups. Figure 2 illustrates a case where six subgroups are prespecified, but the data show that, in terms of the E effect on survival, they can be combined into the three induced subgroups, $\mathcal{A} = \{1, 2, 3\}$, $\mathcal{B} = \{4, 6\}$, and $\mathcal{C} = \{5\}$. Combining subgroups 1, 2, and 3 into $\mathcal{A}$ increases the power of the subgroup-specific test while, in contrast, subgroups 4 and 5 have different treatment effects, so they are not combined. Our proposed GS procedure will address two main statistical goals: (1) adaptively determining a subpartition of empirically induced subgroups for which E (or C) has similar effects on survival within each induced subgroup, but different effects between induced subgroups, while controlling the misclassification rate, and (2) carrying out GS comparisons that control within-subgroup and overall false positive rates and provide high subgroup-specific power. Our procedure will address these goals prospectively, thus avoiding the problems that arise from doing post hoc subgroup analyses, as in the OAK trial.

Formally, for prespecified subgroups $g = 1, \ldots, G$, we denote the survival distributions for subgroup g in arms E and C by $S_{g,E}(t)$ and $S_{g,C}(t)$. For each g, we consider the hypothesis

$$H_{g,0}: S_{g,E}(t) = S_{g,C}(t), \quad \text{for all } t > 0, \tag{1}$$

with alternative hypothesis $H_{g,1}: S_{g,E}(t) \neq S_{g,C}(t)$ for some t. This problem is addressed most commonly by either doing a "one size fits all" test of one global null hypothesis for all G subgroups based on the combined data, which potentially may lead to an inflated Type I error rate, or using a multiple testing procedure to analyze the data from the different subgroups independently, which also may result in a loss of power (Robertson and Wason 2019). An enormous number of different approaches to the issue of heterogeneity in survival analysis have been proposed. For instance, Schumacher, Olschewski, and Schmoor (1987) discussed the effect of ignoring population heterogeneity when comparing survival time distributions by considering two heterogeneous populations. Aalen (1988) used mixture distributions to model the impact of patient heterogeneity. Cécilia-Joseph et al. (2015) used an unobserved random

frailty in the hazard to reflect individual heterogeneity. Smith, Williamson, and Marson (2005) reviewed some popular methods that account for heterogeneity. Existing work has focused mainly on survival analysis based on heterogeneity explained by prognostic variables (or biomarkers) under a parametric proportional hazard (PH) assumption (Cox 1972), or under flexible non-PH survival models (De Iorio et al. 2009; Sparapani et al. 2016; Xu et al. 2019). Fitting such survival models with subgroup–treatment interactions may partially address the potential heterogeneity in testing (1). However, this approach sometimes requires strong association assumptions and does not perform data-adaptive subgroup identification (i.e., information borrowing). Conventionally, if subgroups are specified prospectively, treatment–subgroup interactions are determined manually by fitting several possible models and choosing a final model based on some goodness-of-fit criterion. In addition to its subjectivity, this approach suffers from the limitations that the number of possible models may be large and, typically, only a limited number of models are examined. In contrast, we will provide a flexible survival model and a GS trial design that includes adaptive subgroup combination, also allowing subsequent resplitting, with this done automatically via the latent subgroup membership variable structure, including testing within the induced subgroups.

To provide a basis for adaptive subgroup combination that allows subgroups that have been combined to be separated later, we take a Bayesian hierarchical latent variable approach that facilitates subgroup-specific survival comparisons. At each Markov chain Monte Carlo (MCMC) step of the posterior computation, we update a vector of latent variables that indicate each subgroup's "true" (empirically induced) subgroup. In this way, the data from different subgroups are combined or split adaptively, based on whether the observed survival data show that they are in the same subgroup or different subgroups. To obtain robustness and allow broad applicability, we assume a piecewise constant hazard function for survival time in each treatment arm, with a three-level Bayesian hierarchical Markov-gamma process prior. Under this model, we propose an adaptive group-sequential trial design that accommodates the changing latent subgroup classification during the trial. We optimize the interim decision rules by explicitly controlling the family-wise Type I error rate and maximizing the subgroup-specific power, to obtain desirable frequentist operating characteristics.

The problem that we address here is closely related to that addressed by basket trials, which investigate the treatment effect of one drug on a single mutation in a variety of cancer subtypes, as described by Simon et al. (2016). Recently, basket trials have received a great deal of attention, in part because they borrow strength across disease subtypes, which may improve the trial's efficiency in terms of sample size and trial duration. Simon et al. (2016) proposed a Bayesian basket design for phase II trials for binary efficacy endpoints. Trippa and Alexander (2016) applied adaptive randomization to basket trials. Cunanan et al. (2017) developed a two-stage design by assessing heterogeneity interimly. Chu and Yuan (2018) proposed a Bayesian basket design that use a calibrated Bayesian hierarchical model for adaptive information borrowing. Additional basket trial designs are given by Hobbs and Landin (2018) and Psioda et al. (2019). In contrast with basket trials, which usually use a simple binary endpoint in a single-arm phase II trial, we consider the confirmatory randomized comparative trials with survival time as the endpoint. In addition, we calibrate our design using simulations to explicitly control the family-wise Type I error rate under the global null hypothesis $\cap_{g=1}^{G} H_{g,0}$ in a weak sense, but with the possibility to be extended to control in a more stringent sense. Most existing basket trial designs only control subgroup-specific Type I errors. Our approach also is relevant for the class of enrichment designs (Simon and Simon 2013; Chen et al. 2016; Rosenblum et al. 2016; Lai, Lavori, and Tsang 2019; Mehta, Liu, and Theuer 2019), because our design includes multiple interim analyses to decide whether to terminate unpromising subgroups early. Unlike existing enrichment strategies, which mostly deal with binary or continuous outcomes and dichotomize the population into two subgroups, treatment-sensitive and nontreatment sensitive patients, our design treats overall survival time as the primary endpoint and determines multiple subgroups adaptively and repeatedly.

The remainder of the article is organized as follows. In Section 2, we present the data structure and propose a flexible Bayesian hierarchical latent variable model for survival time. In Section 3, we propose a GS design that adaptively tests for a survival difference between treatment arms within each empirically induced subgroup, and we discuss how to calibrate prior hyperparameters and design parameters to obtain a design with desirable properties. In Section 4, we apply the proposed design to the motivating OAK trial, and conduct simulation studies to evaluate the design's performance. We close with a brief discussion in Section 5.

2. Probability Model

2.1. Data Structure

Let $g_{i,j} \in \{1, \ldots, G\}$ denote the subgroup of patient $i = 1, \ldots, n_j$ in arm $j = E, C$, where n_j is the number of patients in arm j. We focus on two-arm randomized trials where it is desired to test for possible interactive treatment–subgroup effects on survival time. If the subgroups refer to different cancer subtypes having a common genetic mutation, then the trial may be called a "basket trial." If, instead, the subgroups are defined in terms of one or more prespecified biomarkers, then the trial may be called "biomarker-stratified." In our illustrative example, patients were stratified into $G = 4$ subgroups based on PD-L1 expression, as described in Section 1.

We denote survival time of the ith patient in arm j by $T_{i,j}$, and assume that each patient is followed until death or administrative right-censoring. For current trial time t where a decision is made, denote $r_{i,j} = I(T_{i,j} < t)$, so $r_{i,j} = 1$ indicates that death was observed before time t and $r_{i,j} = 0$ denotes the event that $T_{i,j}$ was administratively censored at t. Thus, the observed time to death or censoring is $Y_{i,j}(t) = \min(T_{i,j}, t)$. We also allow the possibility that a covariate, $X_{i,j}$, that is associated with $T_{i,j}$, may be available at enrollment. Here, $X_{i,j}$ refers to additional patient characteristics other than those utilized for defining subgroups. For example, for oncology trials in patients with solid tumors, $X_{i,j}$ may be a biomarker related to $T_{i,j}$. In immunotherapy trials, $X_{i,j}$ may be a binary or real-valued immune response variable. Specifically, in the OAK trial

example, $X_{i,j}$ can be defined as the expression of the T-effector gene signature such as interferon gamma (IFNG). In addition to regression of $T_{i,j}$ on $X_{i,j}$, we allow the possibilities that the distributions of $T_{i,j}$ and $X_{i,j}$ each may differ between two or more subgroups, and may include treatment–subgroup interactions. The efficiency and accuracy of our proposed adaptive subgroup identification process may be improved by borrowing strength from $X_{i,j}$. To focus on the main ideas of our proposed design, we will consider a one-dimensional continuous $X_{i,j}$ and assume that $X_{i,j}$ can be observed quickly. Our proposed method can be generalized easily, however, to accommodate cases where $X_{i,j}$ is multivariate or late-onset.

2.2. Model Specification

We will assume a Bayesian hierarchical model for the survival time distribution, including a latent subgroup membership variable that we will exploit to adaptively collapse homogeneous subgroups or identify heterogeneous subgroups. We consider the data from the C and E groups to be independent, and we do not borrow information between treatment arms. This independent modeling procedure is generally flexible, and has good performance under nonproportional hazard structures (Berry et al. 2004).

For most oncology trials, a common assumption is that the standard of care under C induces response distributions that are homogeneous across subgroups. Following this, we assume that the distributions of $T_{i,C}$ and $X_{i,C}$, $i = 1, \ldots, n_C$, are identical for all $g = 1, \ldots, G$. In contrast, we assume potential heterogeneity of the distributions of $T_{i,E}$ and $X_{i,E}$ between subgroups. However, we make this assumption to simplify the presentation, and our proposed approach is not restricted to this homogeneity assumption under C and can readily be extended to the general case where the survival distributions for the control patients also may be heterogeneous. Discussion about such a generalization will be given at the end of this section, and in the supplementary materials.

For illustrative purposes, we first present our modeling assumptions for $(T_{i,E}, X_{i,E} \mid g)$, and then describe our simpler model for $(T_{i,C}, X_{i,C})$. We temporarily suppress the treatment arm index $j = E$ to reduce notation. To account for patient heterogeneity in arm E and define the spike-and-slab priors that will be discussed later, we introduce a latent subgroup membership variable $z_g \in \{1, \ldots, G\}$ for each patient subgroup g, and denote the indicator $\xi_g = I(z_g = g)$. Thus, if $z_g = z_{g'} = g'$ for $g \neq g'$, then $\xi_g = 0$ and $\xi_{g'} = 1$, subgroups g and g' are homogeneous, and these two subgroups are combined into one induced subgroup g'. As a result, the distributions of (T_i, X_i) and all parameters associated with subgroups g and g' are identical. To deal with the label switching issue arising from data-adaptive clustering, when subgroups g and g' are homogeneous, we artificially take $z_g = z_{g'} = \min(g, g')$ as an identifiability constraint. If $z_1 = \cdots = z_G = 1$, then this corresponds to the completely homogeneous case. At the other extreme, if $z_g = g$ for all $g \in \{1, \ldots, G\}$, then the G subgroups are fully heterogeneous. Taking $z_g = z_{g'}$ allows the data from these two subgroups to be combined as one induced subgroup in the likelihood function, reducing the number of treatment–subgroup interaction parameters. If the two subgroups are truly homogeneous, such a combination in turn increases the power of the subgroup-specific E-versus-C survival comparison in the combined subgroup. However, there always is uncertainty in identification of these subgroups, especially when the sample signal-to-noise ratio is low, and incorrect subgroup combinations may lead to an increase in false positive rates as well as a decrease in true positive rates. In Section 3, we will discuss a general design calibration procedure, that accounts for this identification uncertainty, to ensure a low (misclassification) error rate and a desired true positive rate.

Let $\mathcal{S} = \{g : z_g = g\}$ denote the set of induced subgroups, with $|\mathcal{S}|$ denoting the cardinality of $\mathcal{S}$, so $|\mathcal{S}| \leq G$. As an illustration for $G = 4$, if $\mathcal{S} = \{1, 2\}$ then subgroups 1 and 2 are heterogeneous and are in different induced subgroups, while each of subgroups 3 and 4 belongs to either induced subgroup 1 or induced subgroup 2, and $|\mathcal{S}| = 2$. As another example, for the partitions in Figure 2, there are three induced subgroups $\mathcal{S} = \{1, 4, 5\}$ with latent subgroup variables $z_1 = z_2 = z_3 = 1$, $z_4 = z_6 = 4$, and $z_5 = 5$.

Mimicking Chapple and Thall (2018), we assume the following distribution for (ξ_g, z_g), which will be utilized at each MCMC posterior sampling step for adaptive subgroup combination:

$$\begin{aligned} \xi_g &\sim \text{Bernoulli}(p_g), \\ z_g \mid \xi_g &\sim \xi_g \delta_g(z_g) + (1 - \xi_g)\text{Cat}(\mathcal{S}), \end{aligned} \tag{2}$$

where $p_g = \Pr(\xi_g = 1) = \Pr(z_g = g)$, $\delta_g(\cdot)$ is the Dirac distribution with point mass on g, and $\text{Cat}(\mathcal{S})$ is a uniform categorical distribution with $\Pr(z_g = g') = 1/|\mathcal{S}|$ for $g' \in \mathcal{S}$. As a result, when $\xi_g = 1$, subgroup g is in the its own induced subgroup g. If $\xi_g = 0$, then $z_g = g' \neq g$ for some $g' \in \mathcal{S} \backslash \{g\}$, and in this case subgroups g and g' are homogeneous and both are in the induced subgroup g'. According to the prior (2), different subgroups are likely to be collapsed into the same induced subgroup if the observed data indicate strong evidence for collapsing. This would lead to the model dimension being changed adaptively. When the number of subgroups is relatively large, it generally is not feasible to enumerate all possible models. To tackle the problem of repeatedly changing model dimension, we adopt a Bayesian reversible jump MCMC approach to adaptively identify the latent subgroup indicators based on the observed data (Green 1995). Details of the MCMC sampling steps are provided in the supplementary materials.

We next discuss how to jointly model the marker and survival outcomes. For simplicity, we consider a continuous marker X_i following a normal distribution,

$$X_i \mid g_i = g \sim \text{N}(\mu_g, \sigma_x^2), \tag{3}$$

where μ_g is the population mean for patients in subgroup g and σ_x^2 is a common variance. While we assume the same variance σ_x^2 for all subgroups, if desired, our proposed model and method can be generalized easily to accommodate different variances across subgroups. Because the sample size of a randomized confirmatory study typically is large, any vague prior distribution for σ_x^2 will work well. For example, one may assume an inverse gamma prior $\sigma_x^2 \sim \text{IG}(a_0, b_0)$, where (a_0, b_0) are fixed hyperparameters. To account for cases where some subgroups are

homogeneous while others are heterogeneous with respect to X_i, given the latent subgroup membership prior (2) on (z_g, ξ_g), we assume a spike-and-slab prior on μ_g,

$$\mu_g \mid z_g, \xi_g \sim \xi_g \mathrm{N}(\mu_0, \sigma^2_{\mu_0}) + (1-\xi_g)\delta_{\mu_{z_g}}(\mu_g), \qquad (4)$$

where $(\mu_0, \sigma^2_{\mu_0})$ are prespecified hyperparameters. When there are some homogeneous subgroups, that is, some $\xi_g = 0$, this prior facilitates adaptively borrowing information across subgroups that belong to the same induced subgroup.

Let $h_{g,i}(t)$ denote the hazard function of patient i in subgroup g. We assume a Cox PH model to quantify association between T_i and the marker X_i, with hazard function

$$h_{g,i}(t) = \tilde{h}_g(t)\exp(\beta x_i),$$

where $\tilde{h}_g(t)$ is the baseline hazard for subgroup g, and β is the log-hazard-ratio regression parameter. Any vague prior distribution can be placed on β, for example, $\beta \sim \mathrm{N}(\beta_0, \sigma^2_{\beta_0})$ with hyperparameters $(\beta_0, \sigma^2_{\beta_0})$, for a suitably large $\sigma^2_{\beta_0}$. Here, we only use the PH assumption to characterize the marker effects on T_i. In contrast, no PH assumption is imposed on the E-versus-C treatment effect because we fit the survival models for E and C independently. In general, covariate effects can be modeled using more flexible Bayesian nonparametric survival models, such as those proposed by De Iorio et al. (2009), Sparapani et al. (2016), and Xu et al. (2019). However, such an extension, which is technically more interesting, and would require further investigation and is beyond the scope of this article.

To construct flexible survival distributions, we assume a piecewise constant hazard (Ibrahim, Chen, and Sinha 2001) by partitioning the time scale $(0, \infty)$ into L intervals, $0 = s_0 < s_1 < \cdots < s_L = \infty$, and assuming a constant hazard $\lambda_{g,l}$ on interval $[s_{l-1}, s_l)$ for each subgroup $g = 1, \ldots, G$, and interval $l = 1, \ldots, L$. In subgroup g, denoting $\boldsymbol{\lambda}_g = (\lambda_{g,1}, \ldots, \lambda_{g,L})$, the baseline piecewise hazard function is $\tilde{h}_g(t \mid \boldsymbol{\lambda}_g) = \sum_{l=1}^{L} \lambda_{g,l} I(s_{l-1} \le t < s_l)$, and the baseline survival function is $S_{0,g}(t \mid \boldsymbol{\lambda}_g) = \exp\left\{-\sum_{l=1}^{L} \lambda_{g,l}\Delta(t, s_{l-1}, s_l)\right\}$, $t > 0$, where $\Delta(t, s_{l-1}, s_l) = \max\{0, \min(t, s_l) - s_{l-1}\}$. It follows that the survival function for patient i in subgroup g is $S_g(t \mid x_i, \boldsymbol{\lambda}_g) = \exp\left\{-\sum_{l=1}^{L} \lambda_{g,l}\Delta(t, s_{l-1}, s_l)\exp(\beta x_i)\right\}$.

To obtain a survival model that is robust and tractable, we assume a spike-and-slab prior on the each subinterval's hazard,

$$\lambda_{g,l} \mid z_g, \xi_g \sim \xi_g \pi(\lambda_{g,l}) + (1-\xi_g)\delta_{\lambda_{z_g,l}}(\lambda_{g,l}), \qquad (5)$$

where $\pi(\lambda_{g,l})$ denotes the following three-level hierarchical Markov gamma process (HMGP) (Nieto-Barajas and Walker 2002):

$$\begin{aligned} \lambda_{g,l} \mid \gamma_{g,l-1}, \eta_{g,l-1} &\sim \mathrm{Gamma}(a_{g,l} + \gamma_{g,l-1}, b_{g,l} + \eta_{g,l-1}), \\ \gamma_{g,l} \mid \lambda_{g,l}, \eta_{g,l} &\sim \mathrm{Poisson}(\eta_{g,l}\lambda_{g,l}), \\ \eta_{g,l} \mid w_g &\sim \mathrm{Gamma}(1, w_g), \quad l = 1, \ldots, L \\ w_g &\sim \mathrm{Gamma}(c_g, d_g), \quad g = 1, \ldots, G. \end{aligned} \qquad (6)$$

Here, $a_{g,l}, b_{g,l}, c_g$, and d_g are prespecified hyperparameters. The assumed priors (5) and (6) for the piecewise hazard facilitate borrowing information across homogeneous subgroups within the same induced subgroup, and also between adjacent subintervals in the partition of the survival time domain. Conditional on $\lambda_{g,l-1}$ and w_g, the prior mean of $\lambda_{g,l}$ under (6) is $\frac{w_g}{w_g+1/b_{g,l}}\frac{a_{g,l}}{b_{g,l}} + \left(1 - \frac{w_g}{w_g+1/b_{g,l}}\right)\lambda_{g,l-1}$. Thus, the parameter w_g controls the smoothness of the estimated piecewise hazard. As $w_g \to 0$, the conditional prior mean of $\lambda_{g,l}$ converges to $\lambda_{g,l-1}$ and the prior variance converges to zero. As w_g increases, information borrowing between each pair of adjacent intervals decreases. As default values, we recommend $a_{g,1} = \cdots = a_{g,L} = 1/L$, $b_{g1} = \cdots = b_{gL} = a_{g,l}/\lambda_0$, $c_g = 0.5$, and $d_g = c_g/\lambda_0$ with λ_0 being the prior mean for all $\lambda_{g,l}$'s. Based on this hyperparameter specification, the conditional prior mean of $\lambda_{g,l} \mid \lambda_{g,l-1}$ is $\frac{1}{L+1}\lambda_0 + \frac{L}{L+1}\lambda_{g,l-1}$. This piecewise baseline hazard and hierarchical prior formulation leads quite generally to reliable convergence of the MCMC computations, a robust survival model, and good performance of the proposed design.

Let $\mathcal{D}_n = \{(y_i, x_i, g_i, r_i), i = 1, \ldots, n\}$ denote the observed data for the first n patients enrolled in the trial. Under the proposed model, the likelihood function takes the form

$$\begin{aligned} &L(\mu_1, \ldots, \mu_G, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_G, \beta, \sigma_x \mid \mathcal{D}_n) \\ &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left\{-\frac{(x_i - \mu_{g_i})^2}{2\sigma_x^2}\right\} \prod_{l=1}^{L} \left\{\lambda_{g,l}\exp(\beta x_i)\right\}^{\delta_{i,l}} \\ &\quad \exp\left\{-\lambda_{g,l}\exp(\beta x_i)\Delta(y_i, s_{l-1}, s_l)\right\}, \end{aligned}$$

where $\delta_{i,l} = 1$ if $r_i = 1$ and $y_i \in [s_{l-1}, s_l)$, and $\delta_{i,l} = 0$ otherwise. In the proposed method, both the marker and survival outcomes contribute to the identification of subgroups. We implement a Gibbs sampler and reversible jump MCMC to compute posteriors. The detailed full conditional distributions, and the posterior sampling algorithms, are provided in the supplementary materials.

In what follows, we will turn the task of subgroup identification into a Bayesian model selection problem. To do this, we first denote $\boldsymbol{z} = (z_1, \ldots, z_G)$ with posterior distribution $f(\boldsymbol{z} \mid \mathcal{D}_n)$. Essentially, different combinations of $\boldsymbol{z}$ create different induced subgroups for the effects of E, and thus different models. By enumerating all possible combinations of $\boldsymbol{z}$ based on the posterior samples, we can assign each unique combination a model index, $M_1, M_2, \ldots, M_K$, with K denoting the total number of unique models sampled from the posterior. Denote the value of $\boldsymbol{z}$ under M_k by $\boldsymbol{z}^{(k)}$. We use maximum a posteriori (MAP) estimation to select the most plausible model M_{k^*}, that is, the most likely possible induced subgroup combination,

$$k^* = \underset{k=1,\ldots,K}{\mathrm{argmax}}\, f(\boldsymbol{z}^{(k)} \mid \mathcal{D}_n). \qquad (7)$$

Letting $|M_k|$ denote the number of induced subgroups under model M_k, it follows that $|M_{k^*}| = \sum_{g=1}^{G} I(z_g^{(k^*)} = g) = \sum_{g=1}^{G} \xi_g^{(k^*)}$. For each subgroup g with $\xi_g^{(k^*)} = 1$, we also can enumerate the members of the induced subgroup, given by $\{g' : z_{g'}^{(k^*)} = g\}$.

We now reintroduce the treatment index $j = E, C$, and describe the model for C. We assume complete homogeneity across subgroups in C, so all subgroups in C should share the same distribution, and there is no need to introduce latent

subgroup variables. In particular, we assume $\mu_{1,C} = \cdots = \mu_{G,C} = \mu_C$ and $\boldsymbol{\lambda}_{1,C} = \cdots = \boldsymbol{\lambda}_{G,C} = \boldsymbol{\lambda}_C$, and replace the spike-and-slab priors (4) and (5) utilized for the subgroup-specific distributions under E with the simpler piecewise hazard distribution priors $\mu_C \sim \mathrm{N}(\mu_0, \sigma^2_{\mu_0})$, $\lambda_{C,l} \sim \pi(\lambda_{C,l})$, $l = 1, \ldots, L$, where $\pi(\lambda_{C,l})$ denotes the HMGP (6). As aforementioned, when response homogeneity is not guaranteed across subgroups in C, then the same estimation procedure for E can be applied to C. Although data-adaptive subgroups can be identified in C, a *caveat* is that, potentially, it would lead to a loss of power when the subgroups in C are truly homogeneous, which is often the case for standard control treatments. This is because the subgroup identification may suffer from moderate uncertainty, especially when the sample size is limited. Investigation of the proposed method without the stringent homogeneity assumption for C is provided in the supplementary materials.

3. Trial Design

3.1. Group Sequential Survival Comparison

Our Bayesian adaptive GS survival comparative test, which we call BAGS, relies on the model introduced in Section 2, with the goal to sequentially compare the survival difference between E and C for each subgroup $g = 1, \ldots, G$. In general, GS designs (Pocock 1977; Jennison and Turnbull 1999) provide a flexible, practical way to repeatedly examine observed data as it accumulates for comparing treatments. In our setting, using BAGS provides a way to reliably combine similar subgroups, identify promising subgroups, and drop subgroups having small E-versus-C effects much earlier than the fixed-sample design. GS designs also can be generalized to accommodate multiple-endpoints (Kosorok, Yuanjun, and DeMets 2004; Ye et al. 2013), and trials with more than two treatment arms (Maurer and Bretz 2013; Urach and Posch 2016).

The BAGS design records each patient's subgroup g_i and covariate X_i at enrollment, randomizes them between E and C, and follows them for survival time. At each interim decision time, the BAGS procedure has two steps: the first step identifies the induced subgroups $\mathcal{S} = \{g : z_g = g\}$ based on all currently available data, and the second step then tests the null hypothesis for each induced subgroup $g \in \mathcal{S}$. Let $\mathcal{D}_n = \{(y_{i,j}, x_{i,j}, g_{i,j}, r_{i,j}), i = 1, \ldots, n_j, j = C, E\}$ denote the observed data for the first n patients. Let N be the maximum sample size, and denote interim sample sizes by n_E, n_C with $n = n_E + n_C$. To classify the G subgroups in E into induced subgroups $\mathcal{S}$ based on $\mathcal{D}_n$, we compute posterior samples of $\boldsymbol{z}$ and evaluate k^* using the MAP formula (7). According to the prior distributions (3) and (5), if $z_g^{(k^*)} = z_{g'}^{(k^*)}$, then subgroups g and g' are homogeneous and in the same induced subgroup, and thus they share the same survival distribution, $S_g(t) = S_{g'}(t)$. Under the homogeneity assumption for C, this means that our procedure is essentially testing $|M_{k^*}|$ hypotheses, each of which corresponds to an induced subgroup under E, where the value $|M_{k^*}|$ is random at each decision-making time. On the other hand, when the subgroups under C are heterogeneous, the essential hypotheses to be tested should account jointly for the induced subgroups under C and E.

In the second step, we use the average hazard ratio (AHR) (Kalbfleisch and Prentice 1981; Schemper, Wakounig, and Heinze 2009) to sequentially test the survival difference between E and C for each induced subgroup $g \in \mathcal{S}$. The AHR is a more valid measure of treatment effect than the standard hazard ratio under nonproportional hazards, and the corresponding test provides greater power than the standard logrank test (Rauch et al. 2018). For arm $j = C, E$, let $h_{g,j}(t)$ denote the hazard function for subgroup g, and let $f(\boldsymbol{\lambda}_{g,j}, \mu_{g,j}, \beta_j \mid M_{k^*}, \mathcal{D}_n)$ denote the conditional posterior distribution of the parameters $(\boldsymbol{\lambda}_{g,j}, \mu_{g,j}, \beta_j)$ under model M_{k^*} given $\mathcal{D}_n$. Given $(\boldsymbol{\lambda}_{g,j}, \mu_{g,j}, \beta_j)$, the average arm E to "total" hazard ratio (Kalbfleisch and Prentice 1981) is

$$\begin{aligned}
\theta_{g,E} &= -\int_0^\infty \frac{h_{g,E}(t)}{h_{g,E}(t) + h_C(t)} \mathrm{d}S_{g,E}^{1/2}(t) S_{g,C}^{1/2}(t) \\
&= -\int_0^\infty S_{g,C}^{1/2}(t) \mathrm{d}S_{g,E}^{1/2}(t) \\
&= \sum_{l=1}^{L} \frac{\lambda_{gEl} e^{\beta_E \mu_{g,E}}}{\lambda_{gEl} e^{\beta_E \mu_{g,E}} + \lambda_{gCl} e^{\beta_C \mu_{gC}}} \\
&\quad \left[\exp\left\{ -\frac{\lambda_{g,E,l} e^{\beta_E \mu_{g,E}} + \lambda_{g,C,l} e^{\beta_C \mu_{g,C}}}{2} s_{l-1} \right\} \right. \\
&\quad \left. - \exp\left\{ -\frac{\lambda_{g,E,l} e^{\beta_E \mu_{g,E}} + \lambda_{g,C,l} e^{\beta_C \mu_{g,C}}}{2} s_l \right\} \right],
\end{aligned}$$

for each subgroup $g = 1, \ldots, G$. We define the average arm C to "total" hazard ratio, $\theta_{g,C}$, similarly. Based on these definitions of AHR for E and C, if $S_{g,E}(t) = S_{g,C}(t)$, then $\theta_{g,E} = \theta_{g,C} = 0.5$. Ideally, one may expect $\theta_{g,E} + \theta_{g,C} = 1$, but this equality does not always hold for the piecewise exponential structure. To accommodate this, we define standardized versions of the AHRs,

$$\tilde{\theta}_{g,E} = \frac{\theta_{g,E} + (1 - \theta_{g,C})}{2}, \quad \tilde{\theta}_{g,C} = \frac{\theta_{g,C} + (1 - \theta_{g,E})}{2},$$

which guarantees that $\tilde{\theta}_{g,E} + \tilde{\theta}_{g,C} = 1$. If $\tilde{\theta}_{g,E} = 0.5$, then there is no survival difference between the two arms. If $\tilde{\theta}_{g,E} < 0.5$, then E is superior to C in terms of survival; and if $\tilde{\theta}_{g,E} > 0.5$, then E is inferior. We also note that $\tilde{\theta}_{g,E} < 0.5$ is also equivalent to AHR < 1 with AHR $= \theta_{g,E}/\theta_{g,C}$ denoting the average hazard ratio.

Denote the conditional posterior distribution of $\tilde{\theta}_{g,j}$ by $f(\tilde{\theta}_{g,j} \mid M_{k^*}, \mathcal{D}_n)$, which can be computed based on $f(\boldsymbol{\lambda}_{g,j}, \mu_{g,j}, \beta_j \mid M_{k^*}, \mathcal{D}_n)$. When $|M_{k^*}| > 1$, the issue of multiple testing arises, and a multiplicity adjustment is needed to control the family-wise Type I error rate. To do this in our BAGS design, at each interim analysis we adopt a Holm-like sequential testing procedure to gain more power. The general idea of our sequential testing procedure is as follows: at each test for induced subgroup $g \in \mathcal{S}$, we compute the specified "Bayesian test statistic" and the number of active hypotheses, denoted by m, and require the probability cutoff for the test statistic to be a decreasing function of m. As a result, the larger the value of m, that is, as the multiplicity increases, the more difficult it is to reject the hypothesis.

To do this, for each GS test we use a cutoff $c(n, m)$ that is dependent on both the current sample size n and number of active hypotheses, m. At each decision-making time, we first recalculate m, and then test the hypotheses for each induced subgroup $g \in \mathcal{S}$ with $\xi_g = 1$, which includes its homogeneous subgroups, that is, $\mathcal{S}_g = \{g'; z_{g'}^{(k^*)} = g\}$. The three possible subgroup-specific decisions are as follows:

1. Superiority of E: For each $g \in \mathcal{S}$, if $\Pr(\tilde{\theta}_{g,E} < 0.5 \mid \xi_g = 1, M_{k^*}, \mathcal{D}_n) > c(n, m)$, then reject the composite null (CN) hypothesis $\cup_{g' \in \mathcal{S}_g} H_{g',0}$ and conclude that arm E is superior to arm C in the induced subgroup $\mathcal{S}_g$.

2. Inferiority of E: For each $g \in \mathcal{S}$, if $\Pr(\tilde{\theta}_{g,E} > .5 \mid \xi_g = 1, M_{k^*}, \mathcal{D}_n) > c(n, m)$, then reject the CN hypothesis $\cup_{g' \in \mathcal{S}_g} H_{g',0}$ and conclude that E is inferior to C in the induced subgroup $\mathcal{S}_g$.

3. Inconclusive: Otherwise, there is insufficient evidence in the current data to reject the union of null hypotheses.

Additional design actions: If the CN hypothesis $\cup_{g' \in \mathcal{S}_g} H_{g',0}$ is rejected for some induced subgroup $\mathcal{S}_g$, then the trial stops recruiting patients from this induced subgroup, and also drops the union of hypotheses $\cup_{g' \in \mathcal{S}_g} H_{g',0}$ for the remainder of the trial. The trial continues recruiting patients in the remaining induced subgroups until the maximum sample size N has been enrolled. In this regard, BAGS is an enrichment design. If at some point there are no remaining induced subgroups, then the trial is terminated. Figure 3 provides a flowchart to illustrate the process of how a trial is conducted using the BAGS design.

The probability cutoff $c(n, m)$ plays a central role in the BAGS design, and it must be calibrated to ensure good operating characteristics in the multiple-testing framework. To facilitate calibration of $c(n, m)$, we assume the flexible two-parameter functional form

$$c(n, m) = 1 - \frac{\kappa}{m}(n/N)^{\epsilon},$$

where $\kappa > 0$ and $\epsilon > 0$ are tuning parameters. The cutoff function $c(n, m)$ has two prominent features. First, similar to the α-spending function for a standard GS design (Lan and DeMets 1983; Jennison and Turnbull 1999) and adaptive cut-off function utilized in Bayesian GS testing (Wathen and Thall 2008; Lin, Coleman, and Yuan 2020), $c(n, m)$ is monotonically decreasing in the interim sample size n. At the beginning of the trial, we impose a more stringent stopping rule to control the risk of false discoveries due to sparseness of the early data. As more patients are accrued and longer follow-ups are observed, more survival time information is accumulated and there is less uncertainty, and thus the changing values of $c(n, m)$ can graduate promising subgroups more reliably. A second feature of $c(n, m)$ is that it depends on the number of active hypotheses, m, which also must be updated for each GS decision.

For example, suppose there are $G = 4$ subgroups initially, and at an interim analysis with sample size n, the induced subgroups are $\{1, 2\}$, $\{3\}$, and $\{4\}$. Consider testing the CN hypothesis for induced subgroup $\{1, 2\}$ first. Because there are three induced subgroups, the number of active CN hypotheses is $m = 3$, and the cutoff is $c(n, 3)$. If the CN hypothesis for induced subgroup $\{1, 2\}$ is rejected, then the number of CN hypotheses is reduced to $m = 2$. As a result, provided that the induced subgroups $\{3\}$, and $\{4\}$ are not subsequently combined, the cutoff $c(n, 2)$ is utilized next to test the remaining two active CN hypotheses. Similarly, if subgroup $\{4\}$ is rejected next, then $c(n, 1)$ is utilized for testing the last null hypothesis for the final induced subgroup $\{2\}$. This sequential procedure is similar to the Holm multiple-testing procedure for frequentist GS designs

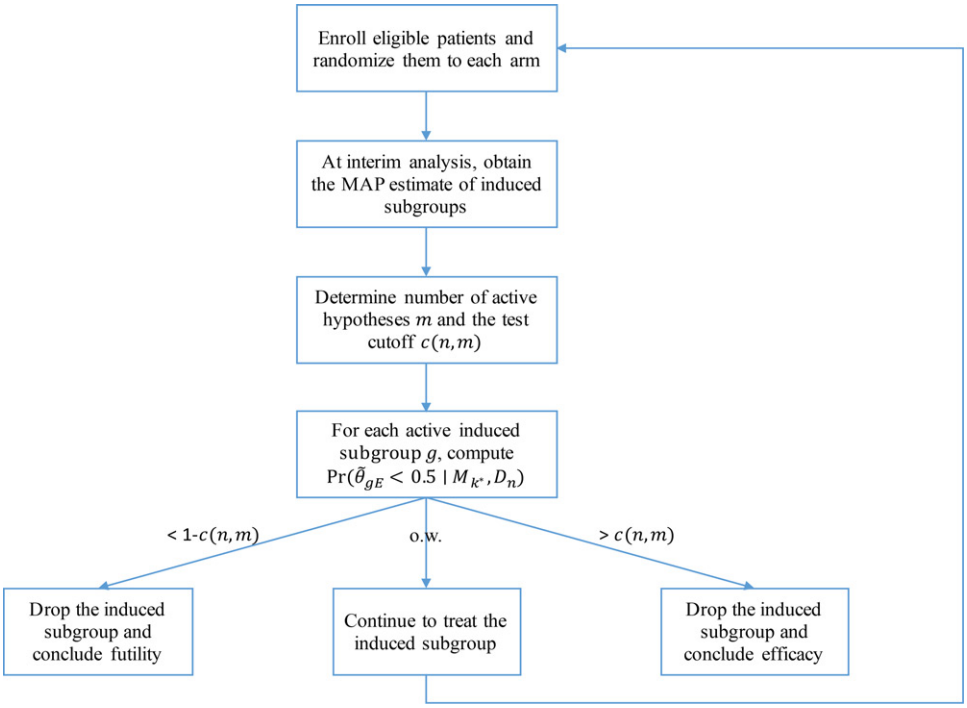

Figure 3. Flowchart of the proposed BAGS design, where n is the interim sample size, m is the number of active hypotheses, $c(n, m)$ is the probability cutoff, M_{k^*} is the most plausible model (i.e., subgroup classification), and homogeneous subgroups constitute an induced subgroup.

(Maurer and Bretz 2013; Ye et al. 2013), and is generally more powerful than the parallel testing procedure using the cutoff $c'(n,m) = 1 - \kappa(n/N)^{\epsilon}$ that ignores m.

There is an important difference between the decision-making procedure of the BAGS design and existing GS procedures. For an existing GS design with G hypotheses, the number of active hypotheses at an interim point in the trial equals G minus the number of previously rejected hypotheses. In contrast, since the BAGS procedure requires the induced subgroups to be identified before each test, the essential number of hypotheses is bounded by $|M_{k^*}|$. As a result, for the BAGS design, m depends on both $|M_{k^*}|$ and the number of previously rejected hypotheses. Consequently, the BAGS design adds another layer of randomness to the number of active hypotheses, which is one of its key features.

3.2. Design Calibration

As mentioned earlier, the BAGS procedure not only accounts for variation of parameter estimates, but also the uncertainty in data-adaptive subgrouping. The performance of BAGS is highly dependent on the classification accuracy induced by the subgroup combination priors in (2), (4), and (5). To obtain good operating characteristics, we propose a general simulation-based calibration procedure to establish the design parameters of BAGS. The use of simulations to optimize design parameters is well documented in the FDA's recent draft guidance (U.S. Food and Drug Administration 2019).

For each $H_{g,0}$, $g = 1, \ldots, G$, the subgroup-specific Type I error rate (SSER) is $\alpha_g = \Pr(\text{reject } H_{g,0} \mid H_{g,0})$. Because we simultaneously test $H_{g,0}$ for different subgroups in one trial, there also is a family-wise Type I error rate (FWER), defined as

$$\tilde{\alpha} = \Pr(\text{reject at least one } H_{g,0} \mid \cap_{g=1}^{G} H_{g,0}).$$

The FWER is very important for a subgroup-specific comparative trial, since it is important to control false discoveries when testing multiple hypotheses. On the other hand, if $H_{g,1}$ holds for some subgroup g, it also is desirable to have large subgroup-specific power (SSP), defined as $1 - \beta_g = \Pr(\text{reject } H_{g,0} \mid H_{g,1})$. The SSP quantifies a design's ability to correctly identify a promising subgroup. For BAGS, in addition to these test error probabilities, there also is a misclassification rate (MCR), since BAGS includes adaptive determination of induced subgroups as part of its GS procedure. The MCR is defined as $\alpha_c = \Pr(\text{at least one subgroup } g \text{ is misclassified})$. Because BAGS combines classification and testing in each GS step, we also define the generalized family-wise power (GFWP) as

$$1 - \tilde{\beta} = \Pr(\text{obtain correct induced subgroups and make correct test decisions}).$$

Having a high GFWP is a very important property for the BAGS design, since correct induced subgroup selection is a key element of the decision-making process. Obtaining a good GFWP depends on well-calibrated design parameters and a sufficiently large sample size. An important point is that the event in the definition of the GFWP is a subset of the event (make correct test decisions), so, for a given N, it is more difficult to achieve a large GFWP than a large Pr(make correct test decisions).

Next, we discuss how to choose the design parameters κ, ϵ, and N. The elicitation guidelines for other parameters, including the fixed prior hyperparameters in the Bayesian models and the partition scheme of the time scale, are provided in the supplementary materials. In general, a larger N gives larger SSP and GFWP, as well as a smaller MER. Determination of the number of interim analyses depends on the specific trial requirements and available resources. We suggest performing the first interim analysis at $n = N/2$. This choice can prevent a high false discovery rate caused by sparse data, while still preserving the efficiency of the trial.

The design parameters κ, ϵ, and N can be determined based on a grid search over all possible combinations, so that the performance of the BAGS design can be nearly optimized. In the motivating trial, we did this for the three values $\epsilon = 1, 2, 3$, six values $\kappa = 0.010, 0.015, 0.020, 0.025, 0.030, 0.035$, and five values $N = 600, 650, 700, 750, 800$. In total, this gave a grid of $3 \times 6 \times 5 = 90$ combinations to study. Our strategy was to choose (κ, ϵ, N) to maximize the GFWP of the BAGS design over the grid while controlling the FWER and SSP at prespecified levels. This can be done using the following steps:

Step 1: Ask the clinicians to specify desirable FWER and SSP values. Our procedure requires statisticians to work with clinicians to establish three sets of hypotheses: (1) Homogeneous null: Under $H_0 = \cap_{g=1}^{G} H_{g,0}$, all subgroups are homogeneous and there is no treatment effect. (2) Heterogeneous null: Under $H_0' = \cap_{g=1}^{G} H_{g,0}'$, the subgroups are heterogeneous and there is no treatment effect. (3) Alternative: Under H_1, there are two induced subgroups, with one subgroup containing the responders, defined as patients who have a lower death rate with E, while the other subgroup contains nonresponders. In addition, other trial parameters, such as patient accrual rate and follow-up time, also should be determined.

Step 2: Choose one sample size N, carry out simulation studies, and search all possible combinations of (κ, ϵ) that control the FWER under both H_0 and H_0'. According to our study, we found that, given a set of (κ, ϵ), the FWER is generally unchanged regardless of the sample size N. Therefore, one just needs to consider one value of N in this step.

Step 3: Among the admissible set of (κ, ϵ) values identified in Step 2, carry out simulation studies under H_1 for different values of N, and select the combination (N, κ, ϵ) that yields the desired SSP while maximizing the GFWP as the optimal design parameters.

Based on simulations, we simultaneously control the FWER under the null and maximize the GFWP under the alternative, which in turn can control the MCR to a satisfactory level. Although we consider using both the homogeneous null and the heterogeneous null for FWER control, due to the subgroup identification uncertainty, the aforementioned calibration procedure can only control the FWER in a weak sense, that is, $\Pr(\text{reject at least one } H_{g,0} \mid \cap_{g=1}^{G} H_{g,0})$ does not exceed the nominal level (Hochberg 1988). Because of misclassification, the aforementioned calibration procedure cannot guarantee a well-controlled false positive rate under situations where some subgroups benefit but others do not. However, when there is no

subgroup identification error, according to the closure principle and the proposed Holm-like sequential testing procedure, the oracle BAGS design (by assuming that the subgroup classification is always true) can control the FWER in a strong sense. More details about the FWER control of BAGS are explored in the simulation study. While the FWER control can be gradually enhanced with an increasing sample size and thus a decreasing MCR, it is also possible to obtain a relatively strong FWER control by controlling the SSER when only one $H_{g,0}$ is true and the other subgroups have homogeneous promising treatment effects. However, the accompanying high price to pay is a loss of power. See the supplementary materials for a discussion of the more stringent calibration procedure.

4. Numerical Studies

4.1. Comparative Study

In the simulation study, we aim to control the FWER ≤ 0.05 and the SSP ≥ 0.90. We utilized the published data from the OAK trial (Rittmeyer et al. 2017) to construct the simulation scenarios, and describe details of the data generating process in the supplementary materials. We assumed identical prevalence rates for the four subgroups, 0.25 each, although we examine the performance of the BAGS design for different true prevalences in additional sensitivity analyses. We specified a total of ten scenarios, given in Table 1, to thoroughly examine the performance of BAGS. We performed two interim analyses, when $n = N/2$ and $n = 3N/4$ patients had been accrued, and a final analysis at the end of follow-up. The calibration procedure described in Section 3.2 resulted in $N = 700$, $\kappa = 0.02$, and $\epsilon = 3$.

In our simulation study, we compared the BAGS design to several common approaches utilized in multiple subgroup settings. The first, referred to as the LR_{homo} design, assumes that all subgroups are homogeneous and implements a logrank test based on a GS design. The second approach, called the $\text{LR}_{\text{hetero}}$ design, does a separate logrank test in each subgroup and does not borrow information between subgroups. The third approach, referred to as $\text{LR}_{\text{enrich}}$, implements a logrank test based adaptive enrichment strategy in a three-stage design such that only the patients who are likely to benefit from E are enrolled in subsequent trial stages (Lai, Lavori, and Tsang 2019). As benchmarks, we also included two oracle designs, LR_{O} and BAGS_{O}, which always use the correct subgroup classification in their inferences. Furthermore, we considered a more general BAGS design, denoted by BAGS*, which conducts data-adaptive subgroup combination under C without assuming homogeneity across subgroups. For the logrank test based GS designs, we utilized O'Brien–Fleming boundaries (O'Brien and Fleming 1979) and applied the Holm procedure to adjust for multiplicity (Ye et al. 2013). The BAGS* and BAGS_{O} designs use the same set of design parameters as the original BAGS design. Since the original $\text{LR}_{\text{enrich}}$ design was not proposed for time-to-event outcomes, we outline the implementation procedure of $\text{LR}_{\text{enrich}}$ in the supplementary materials.

4.2. Simulation Results

Tables 2 and 3 report the operating characteristics of the designs based on 5000 simulated trials under each of the ten scenarios given in Table 1. Scenario 1 is the homogeneous null case, where all subgroups have the same distributions and Atezolizumab gives no survival benefit over Docetaxel. All seven designs control the probability of making one or more Type I errors near 5%. The BAGS design has a particularly low misclassification rate, as low as 0.3%. This nearly ignorable misclassification rate, together with the FWER, leads to a high GFWP for BAGS. On the other hand, the $\text{LR}_{\text{hetero}}$ design assumes complete heterogeneity among the subgroups and uses a multiplicity adjustment to control the FWER. As a result, the SSER of the $\text{LR}_{\text{hetero}}$ design is lower than that of the other designs. In scenario 2, where all subgroups are different, it is desirable to not borrow information across subgroups. The BAGS design preserves the FWER at the prespecified level while maintaining a low MCR. In the first two null scenarios, the performance of the more general BAGS* design, which conducts data-adaptive subgroup identifications under C, is very similar to that of the original BAGS design. This is related to the fact that BAGS* also possesses accurate subgroup identification under C, which can be inferred from the MCR of BAGS under scenario 1 by symmetry. In contrast, $\text{LR}_{\text{enrich}}$ selects the best promising subgroup at each interim analysis, which always leads to two induced subgroups. Thus, the MCR of $\text{LR}_{\text{enrich}}$ is almost 100% and the GFWP is close to 0. Scenario 3 is the alternative hypothesis, where there is one induced subgroup, {3, 4}, with responders and another

Table 1. Configurations of the ten simulation scenarios.

Scenario	Subgroups	$-\log(\lambda_{g,E})$	$\mu_{g,E}$	AHR_g
1	{1, 2, 3, 4}	(2.5, 2.5, 2.5, 2.5)	(0.5, 0.5, 0.5, 0.5)	(1, 1, 1, 1)
2	{1}, {2}, {3}, {4}	(2.5, 2.3, 2.1, 1.9)	(0.5, 1.3, 2.1, 2.9)	(1, 1, 1, 1)
3	{1, 2}, {3, 4}	(2.5, 2.5, 2.7, 2.7)	(0.5, 0.5, 1.22, 1.22)	(1, 1, 0.7, 0.7)
4	{1, 2, 3}, {4}	(2.5, 2.5, 2.5, 2.7)	(0.5, 0.5, 0.5, 1.22)	(1, 1, 1, 0.7)
5	{1}, {2, 3, 4}	(2.5, 2.7, 2.7, 2.7)	(0.5, 1.22, 1.22, 1.22)	(1, 0.7, 0.7, 0.7)
6	{1}, {2}, {3, 4}	(2.5, 2.1, 2.7, 2.7)	(0.5, 2.1, 1.22, 1.22)	(1, 1, 0.7, 0.7)
7	{1}, {2}, {3}, {4}	(2.5, 2.3, 2.1, 2.7)	(0.5, 1.3, 2.1, 1.22)	(1, 1, 1, 0.7)
8	{1}, {2}, {3}, {4}	(2.5, 3.2, 2.7, 2.5)	(0.5, 0.5, 1.22, 2.0)	(1, 0.5, 0.7, 0.7)
9	{1, 2, 3, 4}	(2.7, 2.7, 2.7, 2.7)	(1.22, 1.22, 1.22, 1.22)	(0.7, 0.7, 0.7, 0.7)
10	{1}, {2, 3, 4}	(2.2, 2.7, 2.7, 2.7)	(0, 1.22, 1.22, 1.22)	(1.5, 0.7, 0.7, 0.7)

NOTE: We simulated data for the Docetaxel patients with baseline covariates $X_{i,C} \overset{\text{iid}}{\sim} \text{N}(\mu_C, 1)$, survival times $T_{i,C} \mid X_{i,C} \overset{\text{iid}}{\sim} \text{Exp}(\lambda_C \exp(\beta_C X_{i,C}))$ with $\beta_C = -0.25$, $\mu_C = 0.5$ and $\lambda_C = \exp(-2.5)$, such that the median survival time under C mimics the published results (i.e., 9.6 months). Similarly, the data for the Atezolizumab patients were generated from $(X_{i,E} \mid g) \overset{\text{iid}}{\sim} \text{N}(\mu_{g,E}, 1)$, and $(T_{i,E} \mid g, X_{i,E}) \overset{\text{iid}}{\sim} \text{Exp}(\lambda_{g,E} \exp(\beta_E X_{i,E}))$ with $\beta_E = -0.25$. $\log(\lambda_{g,E})$ is the log baseline hazard rate for each subgroup $g = 1, 2, 3, 4$, $\mu_{g,E}$ is the population mean for the marker outcomes, AHR_g is the average hazard ratio between arms E and C.

Table 2. Operating characteristics of the proposed BAGS design and the comparators under the scenarios 1–5 given in Table 1.

Scenario	Method	% Reject $H_{g,0}$				FWP (FWER)	MCR	GFWP	Sample size
		1	2	3	4				
1	BAGS	5.3	5.4	5.3	5.4	5.6	0.4	94.2	696
	BAGS*	3.9	3.9	3.9	3.9	4.0	0.9	95.2	697
	$BAGS_O$	5.2	5.2	5.2	5.2	5.2	0.0	94.8	696
	LR_{homo}	6.2	6.2	6.2	6.2	6.2	0.0	93.8	695
	LR_{hetero}	1.5	1.3	1.7	1.6	5.8	100.0	0.0	700
	LR_{enrich}	3.7	3.6	3.4	3.5	5.4	98.5	0.0	696
	LR_O	6.2	6.2	6.2	6.2	6.2	0.0	93.8	695
2	BAGS	1.5	1.4	1.6	1.7	5.9	4.2	90.2	700
	BAGS*	1.4	0.9	1.0	1.1	3.0	4.4	92.0	700
	$BAGS_O$	1.9	1.9	1.7	1.9	6.6	0.0	93.4	700
	LR_{homo}	6.2	6.2	6.2	6.2	6.2	100.0	0.0	695
	LR_{hetero}	1.5	1.3	1.7	1.6	5.8	0.0	94.2	700
	LR_{enrich}	3.8	3.6	3.5	3.5	5.4	100.0	0.0	696
	LR_O	1.5	1.3	1.7	1.6	5.8	0.0	94.2	700
3	BAGS	8.2	8.2	92.3	92.5	93.6 (9.0)	9.2	80.6	693
	BAGS*	7.0	6.8	91.6	90.6	92.2 (7.3)	9.3	81.2	696
	$BAGS_O$	5.7	5.7	94.4	94.4	94.6 (5.7)	0.0	88.9	698
	LR_{homo}	66.5	66.5	66.5	66.5	66.5 (66.5)	100.0	0.0	647
	LR_{hetero}	1.7	1.7	64.2	63.3	84.1 (3.5)	100.0	0.0	700
	LR_{enrich}	52.8	52.8	86.6	86.6	91.2 (52.8)	66.0	29.8	658
	LR_O	4.4	4.4	92.6	92.6	92.8 (4.4)	0.0	90.3	699
4	BAGS	5.6	5.2	5.2	77.9	78.3 (7.0)	9.7	68.9	697
	BAGS*	5.3	5.0	4.9	76.1	76.5 (5.8)	9.3	67.5	698
	$BAGS_O$	4.6	4.6	4.6	79.2	79.6 (4.6)	0.0	75.4	699
	LR_{homo}	23.7	23.7	23.7	23.7	23.7 (23.7)	100.0	0.0	685
	LR_{hetero}	1.5	1.3	1.7	62.1	64.4 (4.0)	100.0	0.0	700
	LR_{enrich}	16.5	16.6	16.5	69.5	69.7 (18.6)	26.0	51.1	689
	LR_O	3.9	3.9	3.9	71.8	72.4 (4.3)	0.0	68.5	699
5	BAGS	21.1	96.0	95.1	95.1	97.2 (21.1)	23.2	69.1	675
	BAGS*	19.7	95.1	94.7	94.6	96.3 (19.7)	21.4	70.8	682
	$BAGS_O$	5.9	97.9	97.9	97.9	98.0 (5.9)	0.0	92.6	697
	LR_{homo}	93.7	93.7	93.7	93.7	93.7 (93.7)	100.0	0.0	583
	LR_{hetero}	3.1	68.3	67.2	66.8	91.0 (3.1)	100.0	0.0	700
	LR_{enrich}	87.4	91.7	91.9	91.8	94.8 (87.4)	96.0	0.0	591
	LR_O	4.7	97.6	97.6	97.6	97.7 (4.7)	0.0	95.2	699

NOTE: All values except sample sizes are given as percentages. $H_{g,0}$ is the subgroup-specific null hypothesis for $g = 1, \ldots, 4$, FWP(FWER) is the family-wise power or Type I error rate. When there is mixture of responder and nonresponder subgroups, the numbers in parentheses denote Pr(reject at least one $H_{g,0}$) for the nonresponder subgroups. MCR is the subgroup misclassification rate, and GFWP is the generalized family-wise power. BAGS is the proposed Bayesian adaptive subgroup-specific group sequential design; BAGS* is the more general version of BAGS that conducts data-adaptive subgrouping under C; LR_{homo} is the GS design based on the logrank test and the homogeneity assumption; LR_{hetero} is the GS design based on the logrank test and the heterogeneity assumption; LR_{enrich} is the adaptive enrichment design based on the logrank test; The oracle methods (i.e., $BAGS_O$ and LR_O) assume that the subgroup classifications are correct under all scenarios.

induced subgroup, {1,2}, with nonresponders. In this scenario, it is desirable to combine the data within each of the two induced subgroups to make inferences more efficient. In this scenario, (1) fully borrowing information across all subgroups, which is what the LR_{homo} does, leads to a high subgroup-specific error rate, SSER, which in this case is % Reject $H_{g,0}$ for each g, while (2) no borrowing of information, which is what the LR_{hetero} does, results in low power for detecting promising subgroups. In contrast, because the BAGS design adaptively determines subgroup heterogeneity or homogeneity and reliably creates induced combined subgroups, it can adaptively exploit borrowed information in its tests. In fact, BAGS is as powerful as the oracle designs in identifying the promising subgroup {3,4}. However, since there is a nonzero chance of misclassification (around 9.8%), the BAGS design has a slightly higher SSER and lower GFWP than the oracle designs. In contrast, the LR_{enrich} fails to identify the true underlying subgroups accurately, resulting in high SSERs for $g = 1, 2$ and low SSPs for $g = 3, 4$. In scenario 4, only patients in subgroup 4 have a higher response rate with Atezolizumab. We use this scenario to assess whether the decision for subgroup 4 based on BAGS would be contaminated by the majority of patients, which includes the nonresponder subgroups. Surprisingly, the BAGS design still is able to detect subgroup 4 reliably and achieve subgroup-specific power similar to that of oracle-BAGS, and better than that of oracle-logrank. In addition, the SSERs for subgroups 1–3 also are close to the 5% nominal level.

In scenario 5, the majority of subgroups are responders, which can be regarded as an opposite case of scenario 4. BAGS still yields a GFWP of 69.1% in scenario 5, compared to GFWP = 0 for LR_{homo}, LR_{hetero}, and LR_{enrich}. Scenarios 6 and 7 mimic scenarios 3 and 4, respectively. The only difference is that the nonresponder subgroups all are heterogeneous in scenarios 6 and 7. Scenario 8 mimics scenario 5 but with four heterogeneous subgroups under E. Scenarios 5–8 are included to examine whether the designs considered can control the FWER in a strong sense. We report Pr(reject at least one $H_{g,0}$) for the nonresponder subgroups, that is, $g = 1$, $g = 1, 2$, $g = 1, 2, 3$, and $g = 1$ in scenarios 5, 6, 7, and 8, respectively. The simulation results show that the LR_{hetero}, oracle LR_O, and $BAGS_O$ designs can maintain the 5% nominal error rate, while

Table 3. Operating characteristics of the proposed BAGS design and the comparators under the scenarios 6–10 given in Table 1.

Scenario	Method	% Reject $H_{g,0}$				FWP (FWER)	MCR	GFWP	Sample size
		1	2	3	4				
6	BAGS	3.8	2.8	91.6	90.8	92.2 (6.3)	2.6	83.1	700
	BAGS*	3.5	2.3	88.7	88.1	89.3 (5.5)	2.8	81.4	700
	$BAGS_O$	3.3	3.1	92.5	92.5	92.7 (5.8)	0.0	86.6	700
	LR_{homo}	64.8	64.8	64.8	64.8	64.8 (64.8)	100.0	0.0	651
	LR_{hetero}	2.0	1.7	64.3	63.3	82.8 (3.4)	100.0	0.0	700
	LR_{enrich}	52.8	52.8	86.6	86.6	92.0 (52.9)	100.0	0.0	655
	LR_O	2.5	2.1	89.6	89.6	89.6 (4.5)	0.0	85.5	700
7	BAGS	3.2	2.0	2.3	70.9	71.7 (7.1)	5.9	63.5	700
	BAGS*	2.6	1.5	1.1	64.6	65.3 (4.8)	5.4	59.4	700
	$BAGS_O$	2.0	2.3	2.2	72.2	73.4 (6.1)	0.0	67.2	700
	LR_{homo}	22.3	22.3	22.3	22.3	22.3 (22.3)	100.0	0.0	687
	LR_{hetero}	1.4	1.4	1.6	60.9	62.5 (3.9)	0.0	58.4	700
	LR_{enrich}	16.5	16.6	16.5	69.5	69.6 (20.1)	100.0	0.0	688
	LR_O	1.3	1.3	1.7	60.9	62.5 (4.0)	0.0	58.5	700
8	BAGS	27.3	98.8	81.7	81.9	99.7 (27.3)	28.0	41.9	686
	BAGS*	21.6	97.8	77.6	78.1	98.5 (21.6)	24.0	43.7	693
	$BAGS_O$	4.5	99.9	80.3	79.0	100.0 (4.5)	0.0	62.9	699
	LR_{homo}	99.3	99.3	99.3	99.3	99.3 (99.3)	100.0	0.0	524
	LR_{hetero}	3.9	99.8	71.3	69.4	99.9 (3.9)	0.0	52.0	700
	LR_{enrich}	97.9	99.8	98.4	98.4	99.8 (97.9)	100.0	0.0	525
	LR_O	3.9	99.8	71.3	69.4	99.9 (3.9)	0.0	52.0	700
9	BAGS	99.6	99.2	99.2	99.1	99.6	0.4	98.6	567
	BAGS*	99.3	98.9	99.0	99.0	99.3	0.6	98.3	603
	$BAGS_O$	99.6	99.6	99.6	99.6	99.6	0.0	99.6	569
	LR_{homo}	99.4	99.4	99.4	99.4	99.4	0.0	99.4	504
	LR_{hetero}	71.9	72.9	71.8	71.6	94.5	100.0	0.0	699
	LR_{enrich}	98.8	98.9	98.9	98.8	99.3	1.5	98.5	506
	LR_O	99.4	99.4	99.4	99.4	99.4	0.0	99.4	504
10	BAGS	91.3	97.9	96.2	96.2	99.8	8.7	83.7	678
	BAGS*	88.1	97.2	96.1	96.1	99.6	6.0	80.1	684
	$BAGS_O$	92.6	98.7	98.7	98.7	99.8	0.0	91.6	672
	LR_{homo}	79.3	79.3	79.3	79.3	79.3	100.0	0.0	637
	LR_{hetero}	87.8	75.3	74.4	73.9	99.1	100.0	0.0	698
	LR_{enrich}	56.7	70.9	70.9	70.2	79.8	80.0	0.0	663
	LR_O	95.5	98.4	98.4	98.4	99.9	0.0	93.9	649

NOTE: All values except sample sizes are given as percentages. $H_{g,0}$ is the subgroup-specific null hypothesis for $g = 1, \ldots, 4$, FWP(FWER) is the family-wise power or Type I error rate. When there is mixture of responder and nonresponder subgroups, the numbers in parentheses denote Pr(reject at least one $H_{g,0}$) for the nonresponder subgroups. MCR is the subgroup misclassification rate, and GFWP is the generalized family-wise power. BAGS is the proposed Bayesian adaptive subgroup-specific group sequential design; BAGS* is the more general version of BAGS that conducts data-adaptive subgrouping under C; LR_{homo} is the GS design based on the logrank test and the homogeneity assumption; LR_{hetero} is the GS design based on the logrank test and the heterogeneity assumption; LR_{enrich} is the adaptive enrichment design based on the logrank test; The oracle methods (i.e., $BAGS_O$ and LR_O) assume that the subgroup classifications are correct under all scenarios.

LR_{homo} and LR_{enrich} have severely inflated Type I error rates. Although BAGS and BAGS* can adaptively identify combined subgroups, they still cannot achieve strongly control of FWER. Especially in scenarios 5 and 8, the high MCR of BAGS (or BAGS*) leads to a higher SSER for subgroup 1. Scenario 8 is a difficult case for the BAGS design, because there are four heterogeneous subgroups, and the low sample size per subgroup tends to lead to a higher MCR, which in turn causes a higher SSER for subgroup 1. Scenario 9 is a completely homogeneous case where patients in all subgroups respond to Atezolizumab. Compared to the LR_{hetero} design, BAGS has much higher power and lower achieved sample size. In this case, the performances of BAGS or BAGS* are almost identical to those of the oracle designs. In scenario 10, subgroup 1 and subgroups 2–4 have opposite treatment effects, with Atezolizumab inferior to Docetaxel in subgroup 1, and Atezolizumab superior in subgroups 2–4. Despite this, the BAGS design still maintains a high GFWP.

Overall, this comparative simulation study indicates that, when the MCR is low, the operating characteristics of the BAGS design are nearly identical to those of its oracle counterpart. There are some scenarios that cause a higher MCR for BAGS, and thus a higher SSER. This is mainly due to the fact that sample sizes for misclassified subgroups typically are small. Nevertheless, the BAGS design shows greatly superior performance compared to the LR_{homo}, LR_{hetero}, and the LR_{enrich} designs, which are by far the most commonly utilized designs in real applications. Although similar to BAGS, the more general version, BAGS*, has slightly lower SSPs across all scenarios. This is because BAGS* elicits a noninformative prior for subgrouping under C and does not fully borrow information across subgroups. Therefore, the posterior distributions under C based on BAGS* are slightly flatter compared to those based on the original BAGS.

We provide more simulation studies in the supplementary materials. Specifically, (1) we additionally tested the sensitivity of the BAGS design to different sample sizes, subgroups prevalences, interval partitions, and also different data generating mechanisms, including model misspecification. (2) We

investigated a more stringent version of BAGS, which attempts to control FWER in a stronger way. The simulation results show that the more stringent BAGS design can maintain the false positive rates at the 5% level, but at the cost of a loss of power. (3) We also examined the BAGS* design when the subgroups under C have heterogeneous treatment effects and different survival curves. Overall, the additional studies further confirm the flexibility, generalizability, and robustness of the BAGS design.

5. Conclusions

We have proposed a Bayesian adaptive subgroup-specific GS design that addresses the challenges of making subgroup-specific survival comparisons when the true subgroup-specific effects are unknown and some predefined subgroups may have similar treatment effects. The proposed BAGS design accommodates a survival outcome and a baseline marker variable that may be related to survival time, with both variables having distributions possibly heterogeneous between subgroups. We introduce a latent subgroup variable to facilitate adaptive subgroup combination or splitting. As a result, when homogeneous subgroups are combined adaptively to form an induced subgroup, the resulting test for the induced subgroup is more efficient. At each interim analysis, the proposed design tests the subgroup-specific survival differences between two treatments using the posterior distribution of the average hazard ratio. To deal with control of Type I error rate in multiple testing, we also have developed a sequential Holm-like procedure which yields a higher power of detecting induced subgroups with responders. The simulation study given in Section 4 shows that, compared to standard methods for dealing with multiple subgroups, under a range of scenarios, the proposed BAGS design has much higher subgroup-specific power and generalized family-wise power, while the family-wise Type I error rate is maintained. In some cases, such as scenarios 5 or 8, there is a relatively large SSER for subgroup 1. One also can control the SSER by adding an extra step in the calibration procedure that adjusts the test cutoff $c(n, m)$ to obtain a smaller SSER, but it should be kept in mind that such a recalibration has the risk of decreasing the SSP for other subgroups.

While we have considered only the case of complete homogeneity for the control arm, generalization to the heterogeneous control case is straightforward, although this requires estimating the latent subgroup indicators $\{z_g\}$ for the control patients. By combining the estimated latent subgroups from both groups, the proposed subgroup-specific approach still can be applied. A potential limitation of the proposed design is that we assume that the marker outcome is observed before death. In immunotherapy trials, such an assumption may not hold if the marker outcome is an immune-response, which may be late-onset (Lin, Coleman, and Yuan 2020). In this case, late-onset marker outcomes may be censored by death. The BAGS design may be generalized to accommodate such late-onset variables by treating the unobserved marker values as missing data. For example, the Bayesian data augmentation approach of Liu, Yin, and Yuan (2013) may be used to sample both missing marker data and model parameters from their posterior full conditional distributions. In addition, the proposed method requires an adequate sample size within each subgroup for reliable estimation. In a case where a subgroup has very low prevalence, if this subgroup has a high death rate and the other subgroups have substantively lower death rates, it might be the case that parameter estimates in the low prevalence subgroup are unreliable and thus the FWER might be inflated.

Finally, the BAGS design is developed based on a parametric piecewise constant hazard assumption. In the MCMC computations to obtain the posterior at each group-sequential stage of the trial, we repeatedly iterate between (1) determining the subgroups using the latent variables and (2) model fitting conditional on the grouping. Once this posterior has been determined by this iterative MCMC process, the design bases its decisions and actions on the posterior. We do not include a treatment–subgroup interaction parameter in our model and we perform splitting and recollapsing within each arm, because we want our model to be robust against nonproportional hazards. As a future research project, under a robust Bayesian nonparametric approach (Xu et al. 2019), we may consider inclusion of treatment–subgroup interaction parameters, which would provide a more flexible model.

Supplementary Materials

Supplementary materials contain detailed MCMC sampling steps, the prior elicitation procedure, simulation configurations, and additional simulation results.

Acknowledgments

The authors are grateful to the editor, an associate editor, and two referees for their detailed and constructive comments.

Funding

This research was partially support by NIH/NCI grant P30 CA 016672.

References

Aalen, O. O. (1988), "Heterogeneity in Survival Analysis," *Statistics in Medicine*, 7, 1121–1137. [323]

Bechhofer, R. E., Santner, T. J., and Goldsman, D. M. (1995), *Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons*, New York: Wiley. [322]

Berry, S. M., Berry, D. A., Natarajan, K., Lin, C. S., Hennekens, C. H., and Belder, R. (2004), "Bayesian Survival Analysis With Nonproportional Hazards: Meta-Analysis of Combination Pravastatin–Aspirin," *Journal of the American Statistical Association*, 99, 36–44. [325]

Cécilia-Joseph, E., Auvert, B., Broët, P., and Moreau, T. (2015), "Influence of Trial Duration on the Bias of the Estimated Treatment Effect in Clinical Trials When Individual Heterogeneity Is Ignored," *Biometrical Journal*, 57, 371–383. [323]

Chapple, A. G., and Thall, P. F. (2018), "Subgroup-Specific Dose Finding in Phase I Clinical Trials Based on Time to Toxicity Allowing Adaptive Subgroup Combination," *Pharmaceutical Statistics*, 17, 734–749. [323,325]

Chen, C., Li, X., Yuan, S., Antonijevic, Z., Kalamegham, R., and Beckman, R. A. (2016), "Statistical Design and Considerations of a Phase 3 Basket Trial for Simultaneous Investigation of Multiple Tumor Types in One Study," *Statistics in Biopharmaceutical Research*, 8, 248–257. [324]

Chu, Y., and Yuan, Y. (2018), "A Bayesian Basket Trial Design Using a Calibrated Bayesian Hierarchical Model," *Clinical Trials*, 15, 149–158. [324]

Cox, D. R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society*, Series B, 34, 187–202. [324]

Cunanan, K. M., Iasonos, A., Shen, R., Begg, C. B., and Gönen, M. (2017), "An Efficient Basket Trial Design," *Statistics in Medicine*, 36, 1568–1579. [324]

De Iorio, M., Johnson, W. O., Müller, P., and Rosner, G. L. (2009), "Bayesian Nonparametric Nonproportional Hazards Survival Modeling," *Biometrics*, 65, 762–771. [324,326]

Garralda, E., Dienstmann, R., Piris-Gimenez, A., Brana, I., Rodon, J., and Tabernero, J. (2019), "New Clinical Trial Designs in the Era of Precision Medicine," *Molecular Oncology*, 13, 549–557. [323]

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732. [325]

Hobbs, B. P., and Landin, R. (2018), "Bayesian Basket Trial Design With Exchangeability Monitoring," *Statistics in Medicine*, 37, 3557–3572. [324]

Hochberg, Y. (1988), "A Sharper Bonferroni Procedure for Multiple Tests of Significance," *Biometrika*, 75, 800–802. [329]

Hsu, J. C. (1996), *Multiple Comparisons: Theory and Methods*, London: Chapman and Hall-CRC Press. [322]

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001), *Bayesian Survival Analysis*, New York: Springer. [326]

Jennison, C., and Turnbull, B. W. (1999), *Group Sequential Methods With Applications to Clinical Trials*, Boca Raton, FL: Chapman & Hall/CRC Press. [327,328]

Kalbfleisch, J. D., and Prentice, R. L. (1981), "Estimation of the Average Hazard Ratio," *Biometrika*, 68, 105–112. [327]

Kosorok, M. R., Yuanjun, S., and DeMets, D. L. (2004), "Design and Analysis of Group Sequential Clinical Trials With Multiple Primary Endpoints," *Biometrics*, 60, 134–145. [327]

Lai, T. L., Lavori, P. W., and Tsang, K. W. (2019), "Adaptive Enrichment Designs for Confirmatory Trials," *Statistics in Medicine*, 38, 613–624. [324,330]

Lan, K. K. G., and DeMets, D. L. (1983), "Discrete Sequential Boundaries for Clinical Trials," *Biometrika*, 70, 659–663. [328]

Lin, R., Coleman, R. L., and Yuan. Y. (2020), "TOP: Time-to-Event Bayesian Optimal Phase II Trial Design for Cancer Immunotherapy," *Journal of the National Cancer Institute*, 112, 38–45. [328,333]

Liu, S., Yin, G., and Yuan, Y. (2013), "Bayesian Data Augmentation Dose Finding With Continual Reassessment Method and Delayed Toxicity," *The Annals of Applied Statistics*, 7, 2138–2156. [333]

Maurer, W., and Bretz, F. (2013), "Multiple Testing in Group Sequential Trials Using Graphical Approaches," *Statistics in Biopharmaceutical Research*, 5, 311–32. [327,329]

Mehta, C. R., Liu, L., and Theuer, C. (2019), "An Adaptive Population Enrichment Phase III Trial of TRC105 and Pazopanib Versus Pazopanib Alone in Patients With Advanced Angiosarcoma (TAPPAS Trial)," *Annals of Oncology*, 30, 103–108. [324]

Murray, T. A., Yuan, Y., Thall, P. F., Elizondo, J. H., and Hofstetter, W. L. (2018), "A Utility-Based Design for Randomized Comparative Trials With Ordinal Outcomes and Prognostic Subgroups," *Biometrics*, 74, 1095–1103. [323]

Nieto-Barajas, L. E., and Walker, S. G. (2002), "Markov Beta and Gamma Processes for Modelling Hazard Rates," *Scandinavian Journal of Statistics*, 29, 413–424. [326]

O'Brien, P. C., and Fleming, T. R. (1979), "A Multiple Testing Procedure for Clinical Trials," *Biometrics*, 35, 549–556. [330]

Pocock, S. J. (1977), "Group Sequential Methods in the Design and Analysis of Clinical Trials," *Biometrika*, 64, 191–199. [327]

Pocock, S. J., Assmann, S. F., Enos, L. E., and Kasten L. E. (2002), "Subgroup Analysis, Covariate Adjustment and Baseline Comparisons in Clinical Trial Reporting: Current Practice and Problems," *Statistics in Medicine*, 21, 2917–2930. [322]

Psioda, M. A., Xu, J., Jiang, Q., Ke, C., Yang, Z., and Ibrahim, J. G. (2019), "Bayesian Adaptive Basket Trial Design Using Model Averaging," *Biostatistics* (in press), DOI: 10.1093/biostatistics/kxz014. [324]

Rauch, G., Brannath, W., Brückner, M., and Kieser, M. (2018), "The Average Hazard Ratio—A Good Effect Measure for Time-to-Event Endpoints When the Proportional Hazard Assumption Is Violated?," *Methods of Information in Medicine*, 57, 89–10. [327]

Robertson, D. S., and Wason, J. M. S. (2019), "Family-Wise Error Control in Multi-Armed Response-Adaptive Trials," *Biometrics*, 75, 885–894. [323]

Rittmeyer, A., Barlesi, F., Waterkamp, D., Park, K., Ciardiello, F., von Pawel, J., Gadgeel, S. M., Hida, T., Kowalski, D. M., Dols, M. C., and Cortinovis, D. L. (2017), "Atezolizumab Versus Docetaxel in Patients With Previously Treated Non-Small-Cell Lung Cancer (OAK): A Phase 3, Open-Label, Multicentre Randomised Controlled Trial," *The Lancet*, 389, 255–265. [322,323,330]

Rosenblum, M., Qian, T., Du, Y., Qiu, H., and Fisher, A. (2016), "Multiple Testing Procedures for Adaptive Enrichment Designs: Combining Group Sequential and Reallocation Approaches," *Biostatistics*, 17, 650–662. [323,324]

Schemper, M., Wakounig, S., and Heinze, G. (2009), "The Estimation of Average Hazard Ratios by Weighted Cox Regression," *Statistics in Medicine*, 28, 2473–2489. [327]

Schumacher, M., Olschewski, M., and Schmoor, C. (1987), "The Impact of Heterogeneity on the Comparison of Survival Times," *Statistics in Medicine*, 6, 773–784. [323]

Seoane, J., and De Mattos-Arruda, L. (2014), "The Challenge of Intratumour Heterogeneity in Precision Medicine," *Journal of Internal Medicine*, 276, 41–51. [322]

Simon, N., and Simon, R. (2013), "Adaptive Enrichment Designs for Clinical Trials," *Biostatistics*, 14, 613–625. [323,324]

Simon, R., Geyer, S., Subramanian, J., and Roychowdhury, S. (2016), "The Bayesian Basket Design for Genomic Variant-Driven Phase II Trials," *Seminars in Oncology*, 43, 13–18. [324]

Smith, C. T., Williamson, P. R., and Marson, A. G. (2005), "Investigating Heterogeneity in an Individual Patient Data Meta-Analysis of Time to Event Outcomes," *Statistics in Medicine*, 24, 1307–1319. [324]

Sparapani, R. A., Logan, B. R., McCulloch, R. E., and Laud, P. W. (2016), "Nonparametric Survival Analysis Using Bayesian Additive Regression Trees (BART)," *Statistics in Medicine*, 35, 2741–2753. [324,326]

Thall, P. F., Wathen, J. K., Bekele, B. N., Champlin, R. E., Baker, L. O., and Benjamin, R. S. (2003), "Hierarchical Bayesian Approaches to Phase II Trials in Diseases With Multiple Subtypes," *Statistics in Medicine*, 22 763–780. [323]

Trippa, L., and Alexander, B. M. (2016), "Bayesian Baskets: A Novel Design for Biomarker-Based Clinical Trials," *Journal of Clinical Oncology*, 35, 681–687. [324]

Urach, S., and Posch, M. (2016), "Multi-Arm Group Sequential Designs With a Simultaneous Stopping Rule," *Statistics in Medicine*, 35, 5536–555. [327]

U.S. Food and Drug Administration (FDA) (2019), "Draft Guidance for Industry: Interacting With the FDA on Complex Innovative Trial Designs for Drugs and Biological Products," available at *https://www.fda.gov/vaccines-blood-biologics/guidance-compliance-regulatory-information-biologics/biologics-guidances*. [329]

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., and Drazen, J. M. (2007), "Statistics in Medicine—Reporting of Subgroup Analyses in Clinical Trials," *New England Journal of Medicine*, 357, 2189–2194. [322]

Wathen, J. K., and Thall, P. F. (2008), "Bayesian Adaptive Model Selection for Optimizing Group Sequential Clinical Trials," *Statistics in Medicine*, 27, 5586–5604. [328]

Xu, Y., Thall, P. F., Hua, W., and Andersson, B. S. (2019), "Bayesian Non-Parametric Survival Regression for Optimizing Precision Dosing of Intravenous Busulfan in Allogeneic Stem Cell Transplantation," *Journal of the Royal Statistical Society*, Series C, 68, 809–828. [324,326,333]

Ye, Y., Li, A., Liu, L., and Yao, B. (2013), "A Group Sequential Holm Procedure With Multiple Primary Endpoints," *Statistics in Medicine*, 32, 1112–1124. [327,329,330]