Computational Statistics and Data Analysis 127 (2018) 50–68

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Web-based statistical tools for the analysis and design of clinical trials that incorporate historical controls

Nan Chen [a], Bradley P. Carlin [b], Brian P. Hobbs [c,*]

[a] Department of Biostatistics, University of Texas M.D. Anderson Cancer Center, Houston, TX, United States
[b] Division of Biostatistics, University of Minnesota, Minneapolis, MN, United States
[c] Quantitative Health Sciences and The Taussig Cancer Institute, Cleveland Clinic, Cleveland, OH, United States

ARTICLE INFO

Article history:
Received 12 July 2017
Received in revised form 3 May 2018
Accepted 4 May 2018
Available online 15 May 2018

Keywords:
Adaptive design
Bayesian analysis
Clinical trials
Hierarchical modeling
Historical controls
MCMC
Simulation

ABSTRACT

A collection of web-based statistical tools (http://research.mdacc.tmc.edu/SmeeactWeb/) are described that enable investigators to incorporate historical control data into analysis of randomized clinical trials using Bayesian hierarchical modeling as well as implement adaptive designs that balance posterior effective sample sizes among the study arms and thus maximize power. With balanced allocation guided by "dynamic" Bayesian hierarchical modeling, the design offers the potential to assign more patients to experimental therapies and thereby enhance efficiency while limiting bias and controlling average type I error. The tools effectuate analysis and design for static (non-hierarchical Bayesian analysis) and two types of dynamic (hierarchical Bayesian inference using empirical Bayes and spike-and-slab hyperprior) methods for Gaussian data models, as well as a dynamic method for time-to-failure endpoints based on a piecewise constant hazard model. The site also offers interfaces to facilitate calibration of the model hyperparameters. These allow users to test different parameters in the presence of the historical data on the basis of their resultant frequentist properties, including bias and mean squared error. All calculations are performed on a central computational server. The user may upload data, choose trial settings, run computations in real-time, and review the results using only a web browser. The back-end web module, computation module, and MCMC sampling module are developed in the C#, R, and C++ languages, respectively, and a communication module is also available to ensure the continued connection between the client computer and the back-end server during the Bayesian computations. The statistical tools are described and demonstrated with examples.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

The quality of health care evolves through the continual endeavor to enhance the safety and effectiveness of current therapeutic strategies through clinical study. Yet, translating biomedical discoveries into clinical practice is inherently challenging. Beneficial therapeutic strategies are established through a gradual process devised to define the safety and efficacy profiles of new strategies in phases, over the course of a sequence clinical trials. Transitions between phases involve latency periods wherein the next study is designed and reviewed prior to initiation, introducing inefficiency. In oncology, such latency periods span a duration of nearly two years on average (Committee on Cancer Clinical Trials and the NCI

* Correspondence to: Cancer Biostatistics Section Head in The Taussig Cancer Institute; Associate Staff, Department of Quantitative Health Sciences in The Lerner Research Institute, Cleveland Clinic, 9500 Euclid Ave. CA-60, Cleveland, OH 44195, United States.
E-mail address: hobbsb@ccf.org (B.P. Hobbs).

https://doi.org/10.1016/j.csda.2018.05.002
0167-9473/© 2018 Elsevier B.V. All rights reserved.

Cooperative Group Program Board on Health Care Services, 2010). Moreover, each single successful study typically requires several years to achieve the targeted enrollment, and many studies fail due to low recruitment (Williams et al., 2015). This system produces redundancies, whereby similar treatment strategies are replicated, either as experimental or comparator standard-of-care therapies, across development phases and multiple studies. While systemic redundancy is necessary as sequential learning is needed to effectively devise prospective studies, given the nearly prohibitive cost of conducting clinical trials in humans, The Institute of Medicine recently advocated for the need to restructure the entire clinical trials system to avoid such redundancies as well as address other deficiencies that limit the effectiveness and efficiency of trials (Committee on Cancer Clinical Trials and the NCI Cooperative Group Program Board on Health Care Services, 2010). The initiative was recently re-affirmed with the 21st Century Cures Act, recently passed by the U.S. Congress and signed into law by President Obama in late 2016.

When planning a future trial supplemental information obtained from prior study is needed at the "design stage" to formulate plausible data-generating models that can be used to identify clinically meaningful effect sizes, as well as evaluate trial operating characteristics. Conventionally, data obtained from similar patient cohorts acquired from prior study is utilized formally to facilitate "comparative" evaluations of surrogate endpoints in single-arm phase II trials or conduct retrospective systematic literature reviews. Apprehension pertaining to formally incorporating data from historical studies into the comparative evaluations effectuated by randomized study is well-founded. Intrinsic to randomized design is the desire to infer causal relationships using random allocation strategies that attempt to balance the prognostic determinants (both known and unknown) which obscure the attribution of trends observed in the data to the studied interventions. The classical statistical tests that are conventionally used to compare study arms in randomized trials rely on exchangeable data sampling models. Pooling data from disparate studies using classical statistical tests yields statistical estimators that are sensitive to bias stemming from "trial effects." For example, trial effects stemming from differences in enrollment characteristics, patient surveillance, or clinical-care practice diminish the extent to which one can infer causal pathways, and thereby undermine the purpose of randomized study.

Statistical methods for integrating information from commensurate trials that relax the assumption of inter-cohort data exchangeability and leverage inter-trial redundancy have been developed. Pocock (1976) was first to propose using Bayesian models to incorporate supplemental information into the analysis of a primary data source through static, data-independent shrinkage estimators that require the extent of between-source variability to be prespecified. Numerous models have been discussed since, which involve prespecification of the amount of borrowing under different paradigms related to the power prior (Ibrahim and Chen, 2000; Hobbs et al., 2011; De Santis, 2006; Rietbergen et al., 2011) or inflating the standard error to downweight supplemental cohorts (Goodman and Sladky, 2005; French et al., 2012; Whitehead et al., 2008).

Hierarchical linear models and models which include adaptive down-weighting of data from supplemental cohorts have been extensively explored as well. For these models, the extent of shrinkage towards the supplemental sources is not predetermined but is estimated from the data. More strength is borrowed in the absence of evidence for inter-trial effects, which controls the extent of bias induced from using the supplemental information. One approach is the power prior of Ibrahim and Chen (2000) which can be constructed to discount supplemental sources relative to the primary data. Bayesian (Smith et al., 1995) and frequentist (Doi et al., 2011) methods which utilize hierarchical modeling have been developed to estimate between-source variability with univariate observables or repeated measures. Other authors have considered hyperprior specifications for Bayesian hierarchical models (Daniels, 1999; Natarajan and Kass, 2000; Spiegelhalter, 2001; Gelman, 2006; Browne and Draper, 2006; Kass and Natarajan, 2006). Recently, the use of Bayesian hierarchical modeling to leverage supplemental controls in analyses (Neuenschwander et al., 2010; Pennello and Thompson, 2008; Chen et al., 2011; Neelon and O'Malley, 2010) and trial designs (Hobbs et al., 2013) has been explored. Dynamic approaches to incorporating supplemental information using hierarchical modeling with sparsity inducing spike-and-slab hyperpriors and empirical Bayesian inference have also been described (Hobbs et al., 2011, 2012; Murray et al., 2014).

The impetus for leveraging historical controls, however, is often the desire to use fewer concurrent patients on previously studied control arms. While joint inference may lead to increased precision for estimation of the control (or null) effect, supplementing the control data alone creates imbalances in effective information and thereby impacts frequentist size and power only moderately. Consequently, the most effective utilization of historical control data actually occurs within the context of prospective adaptive trial design with allocation strategies devised to balance the extent of effective information among study arms.

In this article, we describe a collection of web-based statistical tools (http://research.mdacc.tmc.edu/SmeeactWeb/) hosted by M.D. Anderson Cancer Center that enable investigators to incorporate historical control data into conduct of randomized clinical trials using Bayesian methods. The interfaces use Bayesian hierarchical models to produce shrinkage estimators which can be used as the basis for integrating supplementary control data into an analysis of the primary trial source data (Hobbs et al., 2012, 2013). In addition to standard posterior summary statistics, the interfaces output the estimated effective historical sample size as well as the resultant randomization probability which should be used to assign the next trial patient to experimental versus control therapies when targeting balanced effective sample size at the trial's final analysis. This adaptive randomization method is based on the general concept of "multi-source adaptive randomization" first described in Hobbs et al. (2013).

The remainder of this article is organized as follows. Section 2 describes Bayesian models implemented by the interfaces, their derivations of effective historical sample size (EHSS), and the adaptive randomization method. Section 3 describes the computational infrastructure of the web-hosted software, while Section 4 describes implementation of the specific interfaces. Section 5 discusses a simulation study demonstrating the frequentist operating characteristics of multi-source adaptive randomization designs. Finally, Section 6 offers a brief closing discussion. A user manual is appended with the supplementary content.

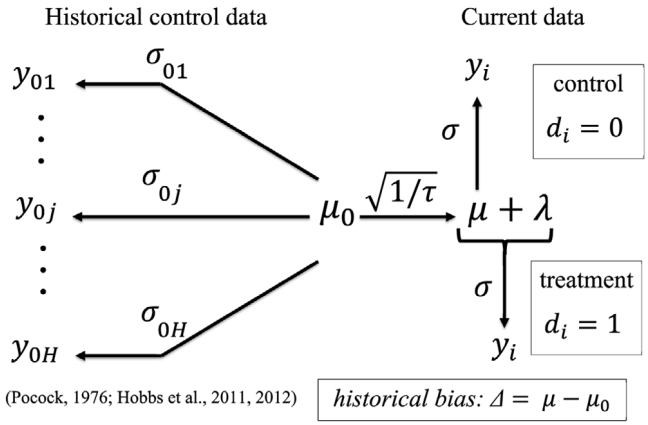


Fig. 1. Diagram depicting the utilized Bayesian hierarchical modeling strategy.

2. Methodology

Our web-hosted interfaces consider analysis and conduct of adaptive design for a primary (or "current") study designed to compare a novel, experimental intervention to an established control therapy for which observations have been acquired from a total of H supplemental or "historical" cohorts.

Thus, the trial's primary objective is to estimate the relative effectiveness of the experimental treatment. The motivation for a Bayesian analytical approach is to yield shrinkage estimators in the presence of the data to effectively integrate the historical control data into the analysis such that the extent of shrinkage is determined by the empirical evidence of historical bias. This offers the potential to enhance estimation of the novel treatment effect as well as balance the design's allocation through adaptive randomization, while avoiding considerable bias and type I error inflation. Given the sequential nature of health-care progression, wherein clinical practice guidelines for "standard of care therapies" are routinely updated on the basis of results from confirmatory studies which provide "baseline" measures of performance for the subsequent generation of innovations, this is perhaps the most useful context needed for the randomized study of biomedical therapies. In the remainder of this section, we describe the methodology for Bayesian analysis and adaptive design in the Gaussian case. Methodology for time-to-failure analysis as well as adaptive design have been described in detail in Hobbs et al. (2013).

2.1. Analysis methodology

Let $\boldsymbol{y}$ denote a vector of i.i.d. responses of length n from patients enrolled in the primary trial, such that $y_i \overset{iid}{\sim} N(\mu + d_i\lambda, \sigma^2)$ where d_i is an indicator of novel treatment. Characterizing the extent to which the mean response of the experimental cohort differs from the overall mean, μ, estimation of parameter λ determines the extent of treatment benefit for this model. Suppose that we have patient-level response data for patients assigned to the current control arm from H historical trials. Let $\boldsymbol{y}_{0,h}$ denote response vectors of length $n_{0,h}$ for the historical data, $\boldsymbol{y}_{0,h} \sim N(\mu_0, \sigma^2_{0,h})$, where $h = 1, \ldots, H$. Suppose that current trial's objective is to compare the novel treatment to the previously studied control, and thus the posterior distribution of λ is of primary interest for treatment evaluation.

2.1.1. Bayesian model

Fig. 1 illustrates the Bayesian hierarchical modeling strategy that is used to facilitate joint posterior inference, which has been described in Hobbs et al. (2012). The model assumes randomized study in the "sparse-information" setting most commonly encountered in design contexts wherein patient-level prognostic characteristics are unavailable (as when institutional barriers prohibit the sharing of all patient-level data), whence estimation of bias arises from trial effects ascertained through the response distributions. The diagram denotes observable random variables with ys. Historical control observables, y_0, arise as Gaussian variates assumed to have identical mean μ_0 and source specific scale parameters $\sigma_{0,h}$. Current control ($d = 0$) and experimental ($d = 1$) observables, y, arise with scale σ and mean μ and $\mu + \lambda$, respectively. Parameter $\Delta = \mu - \mu_0$ characterizes the extent of bias resulting from incorporating the historical controls. At second level of the hierarchical model, the prior distribution for μ is centered at historical mean μ_0 with variance $1/\tau$. As described in detail below, estimation of τ (either by maximizing the marginal density or via extending the model through hyperprior specification for τ) controls the effective sample size of posterior inference and thus should be calibrated in accordance with one's tolerance for bias. This is described in detail in the subsequent sections.

Having observed the historical control data and assumed a flat initial prior for μ_0, the prior distribution for the concurrent control mean emerges as its predictive distribution conditional on $\boldsymbol{y}_0 = (\boldsymbol{y}_{0,1}, \ldots, \boldsymbol{y}_{0,H})$,

$$p(\mu|\boldsymbol{y}_0, \sigma_0^2, \tau) = N\left(\sum_{h=1}^{H} \frac{\bar{y}_{0,h}/u_{0,h}^{-1}}{\sum_{h=1}^{H} u_{0,h}}, \ \frac{1}{\sum_{h=1}^{H} u_{0,h}} + \frac{1}{\tau}\right),$$

where cohort-level weight $u_{0,h} = n_{0,h}/\sigma_{0,h}^2$, and is thus independent of τ unlike the conventional single-source hierarchical model. Having observed concurrent data, $\boldsymbol{y}$, from the primary trial, the model yields the following posterior distributional form for μ

$$p(\mu|\tau, \sigma_0^2, \sigma^2, \boldsymbol{y}_0, \boldsymbol{y}) = N\left\{C\hat{\mu}_0 + (1-C)\hat{\mu}, \ (v_0 + \frac{1}{\tau})C\right\}, \tag{1}$$

with weight $C = v/(v + v_0 + \frac{1}{\tau})$ associated with $\hat{\mu}_0$, such that $v_0 = (\sum_{h=1}^{H} 1/u_{0,h})^{-1}$, $\hat{\mu}_0 = v_0\left(\sum_{h=1}^{H} \bar{y}_{0,h}/u_{0,h}\right)$, $\hat{\mu} = (n\bar{y} - \bar{y}_d \sum_{i=1}^{n} d_i)/(n - \sum_{i=1}^{n} d_i)$, and $v = \sigma^2/\left(n - \sum_{i=1}^{n} d_i\right)$. The resultant marginal density for τ can be written as

$$m(\boldsymbol{y}, \boldsymbol{y}_0|\tau) \propto N\left\{\hat{\mu} - \hat{\mu}_0 \ \middle| \ 0, \frac{\sigma^2}{n - \sum_{i=1}^{n} d_i} + v_0 + \frac{1}{\tau}\right\}.$$

Thus, parameter τ characterizes overdispersion associated with historical bias.

2.1.2. Effective historical sample size

Historical data is usually incorporated into clinical trial analysis with the intention of enhancing the efficiency of the resultant posterior estimators using all of the available information. The extent of gain in efficiency is conventionally characterized by a reduction in posterior variance. However, representing information on the variance domain can be highly non-intuitive. Alternatively, for many models one can derive (or approximate) the posterior's *effective sample size* (ESS), providing a highly interpretable metric for quantifying the extent of the *a posteriori* information that results from updating the information in the prior with the information contained in the likelihood on the sample size domain. The hierarchical model described above yields an explicit and tractable formulation of posterior ESS.

Let θ denote a general model parameter, and define $\mathcal{P} = \left[E_{\theta|D,D_0}\left\{\theta - E_{\theta|D,D_0}(\theta)\right\}^2\right]^{-1}$ as the posterior precision that is obtained when estimating parameter θ using both primary (D) and supplemental (D_0) source data. ESS is in essence the sample size that is required to attain $\mathcal{P}$ under the appropriate "reference" posterior excluding the supplemental data D_0. The reference posterior in this context refers to the posterior distribution obtained from updating the maximum entropy prior (MEP) distribution for θ with the current data likelihood D. One can derive an expression for the relationship between precision and sample size analytically for reference posteriors that arise from conjugate MEPs. A framework for characterizing the effective sample size of conjugate prior distributions was formalized by Morita et al. (2008) and later extended to multi-level modeling (Morita et al., 2012).

The relative gain in posterior precision that results from incorporating historical data can be mapped into a value that quantifies the additional ESS contributed to the joint posterior inference, which we refer to as the *effective historical sample size* (EHSS). Returning to our notation presented above, let $\mathcal{P}^*(\boldsymbol{y})$ denote the posterior precision of $\mu|\boldsymbol{y}$ under the reference model and $\mathcal{P}(\boldsymbol{y_0}, \boldsymbol{y})$, denote the posterior precision of $\mu|\boldsymbol{y_0}, \boldsymbol{y}$ under the joint model in (1), and let $n_c = n - \sum_i d_i$ represent the number of concurrent controls. The reference model yields the following simple relationship between posterior precision and sample size, $\mathcal{P}^*(\boldsymbol{y}) = g(n_c) = n_c/\sigma^2$. The ESS of the joint inference arises by inverting this relationship with respect to $\mathcal{P}(\boldsymbol{y_0}, \boldsymbol{y})$,

$$g^{-1}\{\mathcal{P}(\boldsymbol{y_0}, \boldsymbol{y})\} = \sigma^2\mathcal{P}(\boldsymbol{y_0}, \boldsymbol{y}) = n_c + \frac{\sigma^2}{v_0 + \frac{1}{\tau}}. \tag{2}$$

Following Hobbs et al. (2013), the effective historical sample size can be obtained by subtracting the observed current sample size n_C from ESS to obtain the effective historical sample size, $EHSS = ESS - n_C$. EHSS effectively maps the gain in posterior precision induced by incorporating the historical data to a sample size representing the number of *additional* effective current controls for the joint posterior inference.

It should be noted that, as a ratio of variance components, this estimator of EHSS may exceed the historical control sample size. As an example, consider the case of a single historical study that enrolled n_0 controls. EHSS for the joint model follows as $EHSS = \sigma^2/(\sigma_0^2/n_0 + 1/\tau)$. Thus, EHSS is maximal as $\tau \to \infty$, which implies that the current and historical cohorts yield "exchangeable" data. In the limit, $EHSS = n_0(\sigma^2/\sigma_0^2)$. Thus, the effective historical sample size is bounded above by the product of historical sample size and ratio of sampling-level variances. EHSS may exceed n_0 when observable responses in the historical cohort exhibit less deviation from their mean than observable responses assigned control in the current cohort. In this case, the extent of influence on posterior precision of each historical control sample is effectively larger than each current control.

The general approach provides a method for measuring the extent to which competing dynamic models "borrow strength" from the historical data. Note that the EHSS computation in Hobbs et al. (2013) involved an intractable reference

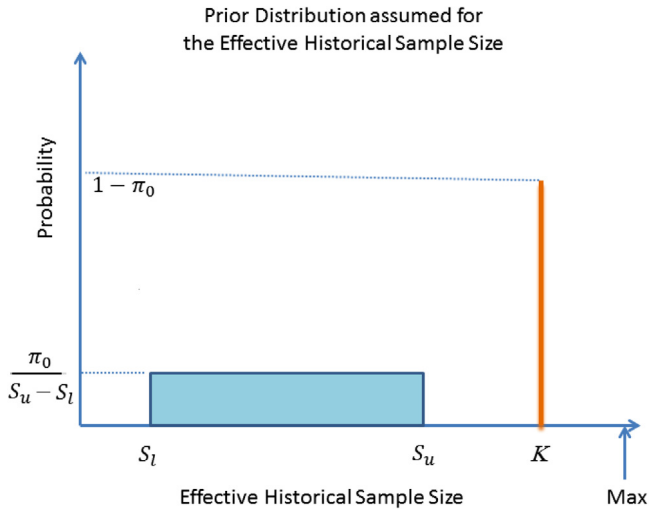


Fig. 2. Diagram depicting the implemented spike-and-slab prior distribution on domain of EHSS.

posterior for analysis of time-to-failure data. Therefore, the authors approximated EHSS with the assumption that precision is proportional to sample size under the reference posterior. This assumption is often satisfied for models that yield symmetric and unimodal posterior distributions, but also justified asymptotically through the Bayesian Central Limit Theorem (Carlin and Louis, 2009, Chapter 3).

2.1.3. Estimation of trial effects

Our web interfaces implement three types of Bayesian inference which differ with respect to estimation of τ. Non-hierarchical Bayesian analysis proceeds with pre-specification of parameter τ. Fixing τ to a value on the precision domain facilitates data-independent partial pooling of concurrent and historical information, wherein the extent to which the model borrows strength from the historical controls is pre-determined. Consequently, the total ESS is also determined *a priori* by the specification of τ and thus not a function of bias. Due to its lack of control for "trial effects" or cohort bias, this approach is referred to hereafter as *static*. This approach, which was initially proposed by Pocock (1976), yields an estimator with linear bias and parabolic risk under squared error loss, as will be demonstrated in Section 5.

Additionally, the interface implements two *dynamic* approaches described in Hobbs et al. (2012). These methods facilitate data-dependent partial pooling such that the extent to which data from the historical study is statistically exchangeable with the current control data is not pre-determined, but rather estimated. These are based on an *empirical Bayesian* (EB) method implementing conditional posterior inference after fixing τ, as well as a fully Bayesian method using a generalized spike-and-slab prior distribution assumed for the EHSS. The EB approach uses constrained optimization over a bounded domain such that τ is fixed to the value that maximizes the marginal likelihood as well as satisfies the constraint that it does not exceed a user specified maximum value for the EHSS, which we denote $EHSS^*$. Let $\hat{\Delta} = \hat{\mu} - \hat{\mu}_0$. Specifically, the interface conducts posterior inference conditional on the largest value of

$$\hat{\tau} = \left\{ \max\left(\hat{\Delta}^2 - \frac{\sigma^2}{n - n_d} - v_0,\ 0 \right)^{-1} \right\}, \tag{3}$$

such that $EHSS(\hat{\tau}) = \sigma^2/(v_0 + 1/\hat{\tau}) \leq EHSS^*$. Note that the marginal density is maximized at $\hat{\tau} = \infty$ when the estimated bias $\hat{\Delta} = 0$. EB inference based on the bounded marginal maximum likelihood estimator is devised to avoid over-smoothing given limited evidence for bias in the data. In addition to data analysis, the interfaces facilitate tools for computation of frequentist properties (discussed in Sections 4 and 5) which enable the user to interrogate hyperparameter specification. Our interface also permits fully Bayesian dynamic inference by adding a hyperprior for τ (equivalently EHSS). The Bayesian model formulation described in Hobbs et al. (2012) assumed a spike-and-slab (see e.g. Mitchell and Beauchamp, 1988) hyperprior distribution on the domain of τ. In the interest of enhancing the interpretability of model specification for practical implementation, however, the interface was devised instead to specify the spike-and-slab distribution with respect to EHSS. Fig. 2 provides a conceptual depiction of the prior's specification on the basis of four hyperparameters. Specifically, the distribution is locally uniform between $0 \leq \mathcal{S}_l < \mathcal{S}_u$ with an additional point mass discontinuity, $\mathcal{K} > \mathcal{S}_u$, such that $Pr(EHSS > \mathcal{S}_u) = Pr(EHSS = \mathcal{K}) = 1 - \pi_0$. Using this notation, π_0 denotes the prior probability assumed for observing EHSS within the slab component, $\mathcal{S}_l \leq EHSS \leq \mathcal{S}_u$. This prior is quite general, encompassing a two-point mixture (when $\mathcal{S}_u = \mathcal{S}_l$) as well as uniform density (given $\mathcal{S}_u = \mathcal{K}$ and $\pi_0 = 1$) as special cases.

In consideration of the manner in which Bayesian sequential updating impacts the marginal density of τ, justification for a sparsity-inducing prior formulation was provided in Hobbs et al. (2012). The rationale holds for specification with respect to the domain of EHSS.

Estimators of EHSS resulting from both dynamic approaches adapt in relation to the observable data as functions of empirical bias. As demonstrated in Section 5, this leads to shrinkage estimators that are robust in the presence of trial effects yielding desirable frequentist properties with flexible expected bias and mean squared error behavior. The interface for time-to-failure analysis utilizes only a spike-and-slab prior formulation as the hierarchical piecewise constant hazard model fails to yield an analytically tractable expression for the resultant marginal maximum likelihood estimation. The piecewise constant hazard model formulation is described in detail in Hobbs et al. (2013).

2.2. Adaptive design methodology

Adaptive designs refer to those that facilitate mid-trial modifications based on interim information (see e.g. Berry, 2006). Several types of designs incorporating adaptive features have been proposed. These methods tend to vary in complexity, extent of theoretical support, use of predictive probabilities, inclusion of decision theoretic arguments, sequential decisions, and more. Perhaps the most widely recognized type of adaptive designs for comparative, intermediate-phased trials use interim data to adjust the allocation or assignment of patients to therapies through adaptive randomization (AR) (see e.g. Friedman et al., 1998). Two types of AR procedures are often considered. *Baseline AR* designs use allocation rules that endeavor to balance the study arms, on average, with respect to patient-level prognostic factors that are available at baseline (Pocock and Simon, 1975; Anderson et al., 2011). Attempting to minimize confounding arising from clinical factors known to be strongly associated with the study endpoints in advance of the study enhances the quality of treatment comparisons. By way of contrast, several types of response-adaptive (or outcome-adaptive) randomization (RAR) methods (Cheung et al., 2006; Hu and Rosenberger, 2006; Huang et al., 2009; Lee et al., 2010; Sverdlov et al., 2011; Yin et al., 2012) were devised to imbalance designs on the basis of accumulating trial data in favor of putatively more effective or safer study arms. Owing to concerns of intrinsic selection bias stemming from population drift as well as overall attenuated operating characteristics for the resultant treatment comparisons (Korn and Freidlin, 2011; Thall et al., 2015), RAR techniques have elicited controversy in recent years.

The impetus for incorporating historical controls is often *efficiency*, with the desire to use fewer concurrent patients on previously studied control arms. In consideration of leveraging historical control data that satisfy Pocock's six "acceptability" criteria (Pocock, 1976, p.177), joint modeling offers increased precision. Our interfaces facilitate a third type of AR, which we refer to as *multi-source adaptive randomization*, first described by Hobbs et al. (2013), which is based on an allocation strategy devised to balance the extent of effective information among study arms. The resulting adaptive design attempts to maximize power on the basis of interim posterior estimates of EHSS. With balanced allocation guided by "dynamic" Bayesian hierarchical modeling, the design offers the potential to assign more patients to experimental therapies and thereby enhance efficiency while limiting bias and controlling average type I error.

The multi-source adaptive design attempts to balance the extent of effective information at the trial's completion, such that after the final enrollment, $n - n_c = EHSS + n_c$. Extending the notation presented in Section 2.1 to incorporate trial time, t, let $n(t)$ denote the total sample size for the current study observed at time t, $n_c(t) = n(t) - \sum_{i=1}^{n(t)} d_i$ represent the number of patients enrolled on control arm at time t, and $R(t)$ denote the number of remaining future enrollments at time t, out of a maximum of n. Let $\omega(t)$ denote the randomization probability in favor the experimental arm. Given an interim estimate of the effective historical sample size, $EHSS(t)$, a design targets balanced allocation through the following identity: $\{n(t) - n_c(t)\} + \omega(t)R(t) = EHSS(t) + n_c(t) + \{1 - \omega(t)\}R(t)$, yielding posterior assignment probabilities that follow from

$$\omega(t) = \frac{1}{2}\left(\frac{EHSS(t) + n_c(t) - \{n(t) - n_c(t)\}}{R(t)} + 1\right). \tag{4}$$

As described in Section 4, the tools provided in MD Anderson interfaces can be used to obtain estimates of both EHSS and $\omega(t)$ given inputs of interim data and $R(t)$. The induced frequentist operating characteristics are presented in a simulation study in Section 5. We note that when the novel therapy is inferior to control, (4) may lead to more patients being assigned to the less effective treatment, and thus is perhaps best utilized with sequential trial monitoring using the posterior probability outputted by the interface for novel treatment superiority.

3. Website design and computational infrastructure

To avoid irregularities arising from inter-platform heterogeneities, we decided to disseminate our statistical tools in the form of web-hosted interfaces. We designed the software website to be comprised of eight different function-specific web pages. The main index page is a menu prompting the user to select one of seven computational interfaces. The index page also includes URL links to related publications that provide more detailed information regarding the statistical and Bayesian computational techniques. Each subpage contains a URL linked to the index web page for ease of navigation. All calculations are performed on a central computational server to which the user may upload data, choose trial settings, run computations in real-time, and review the results using only a web browser and an internet connection. The back-end web module, computation module, and MCMC sampling module were developed in the C#, R, and C++ languages, respectively. A communication module was also developed to ensure continued connection between the client computer and the back-end server during the Bayesian computations. This section describes the computational infrastructure of our web interface.

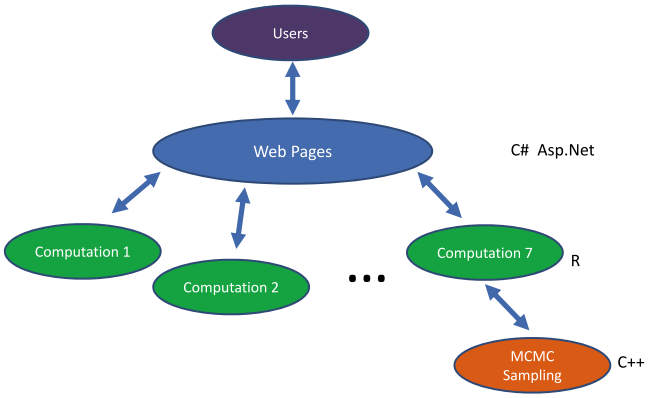


Fig. 3. Schematic Layer Structure of the Website.

3.1. Computational infrastructure

All computations are performed on a back-end server, coded using the `Asp.Net`, `R`, and `C++` languages. Most of the computational functions are written in `R`, with exception of the MCMC sampling algorithm for the piecewise constant hierarchical model which was optimized using `C++` to improve speed. An `Asp.Net` framework was used to generate each interface, as well as implement input/output data processing and communicating with the `R` computation modules.

The schematic structure of the website is presented in Fig. 3. After providing data (either typed or file upload) and specifying hyperparameters, the user clicks the "Submit" button on the computation web interface. Thereafter, the back-end web server receives the request and processes the input parameters. The first step of the back-end processing is to validate the input parameters. If the input parameter validation fails, an error message will be generated and the corresponding error web page will be displayed to the user. If the input data passes validation, it will be transferred to the appropriate computation module for analysis. Upon completion of the computation, either the results or an error message related to the potential computation failure will be transferred back from the computation module to the website back-end. Finally, based on the calculation results, a web page will be generated and passed back to the user.

3.2. Communication between the client user and the computational server

The computation time for submitted jobs varies from 20 seconds to 3 minutes depending on different calculation methods and input parameters. This far exceeds the normal web page processing time of 0.01–3 seconds, significantly increasing the complexity of the website design. Because it may take several minutes for a user to receive the calculation results after a job is submitted to the server, the network connection between the user computer and the back-end server may be disconnected by the network nodes during this latency period. For instance, a reverse proxy server between the user computer and the back-end server may automatically disconnect the network connection during the calculation due to long response time, whence the user will never receive the calculated results. To resolve this issue, we developed a method using multiple communications between the user browser and the back-end server which is performed during implementation of the Bayesian computation. The mechanism is illustrated in Fig. 4. After a user clicks the "Submit" button on a computation interface, a `Javascript` on the client side is launched that keeps communicating with the back-end server to monitor the computation job status at a specific frequency (say, 3–4 times/minute) until it receives the final results. During the entire computation process, the connection is maintained using the multiple communication process managed by the `Javascript` code on the client end and the `Asp.Net` code on the server. The multicommunication process is launched automatically and maintained during the entire period of computation.

4. Tool usage and examples

Our software website includes one static and two dynamic (empirical Bayes and spike-and-slab hyperprior) borrowing interfaces for implementing Bayesian analysis of Gaussian distributed response data, as well as one dynamic borrowing method for analysis of right-censored time-to-event (TTE) data based on the spike-and-slab hierarchical model of Hobbs et al. (2013). For each data analysis interface, the software estimates posterior distributions for the model parameters and estimates EHSS based on the model's derivation. Additionally, given that the user specifies the number of patients that remain to be randomized ($R(t)$ in notation presented in Section 2.2), the interfaces computes the resultant adaptive randomization probability that is recommended for assigning the next cohort of patients to the experimental therapy when targeting information balance, as demonstrated in Section 2.2. For each Gaussian data analysis interface, the user is required to input

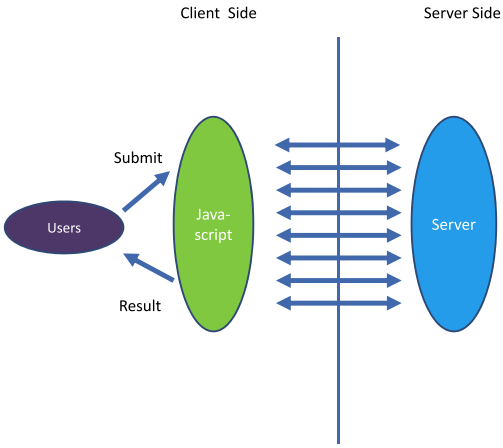
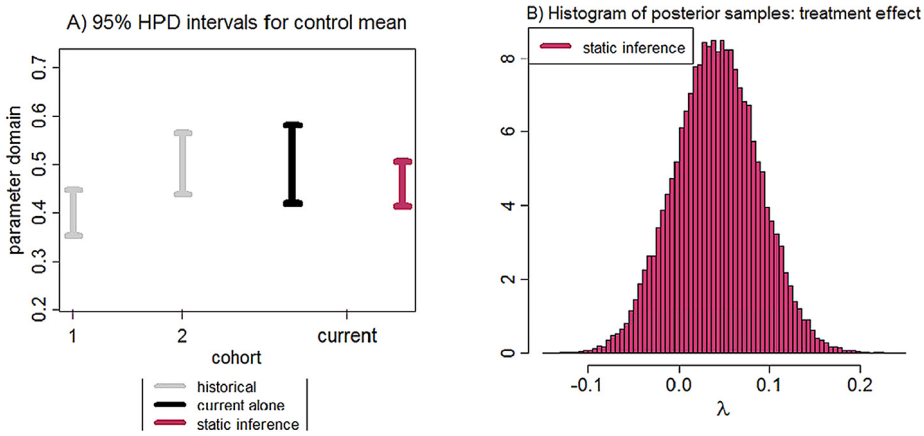


Fig. 4. Multiple communications between a user computer and the back-end server during calculation.



Posterior Inference

Sample Size	
# Current treated	25
# Current controls	25
# Historical controls	80
Effective Historical Sample Size	45.0 (Max possible = 102.6)
Randomization Probability	
# Remaining patients	100
Randomization probability to novel treatment	0.725
Posterior Summary of Lambda	
Mean	0.041
Standard deviation	0.047
95% HPD interval lower	-0.05
95% HPD interval upper	0.13
Posteror probability lambda > 0 | Data	0.81

Fig. 5. Example output table from the static Bayesian analysis interface.

the sample size, sample mean, and sample standard deviation for the concurrent control and novel treatment arms. The output for a typical data analysis page is illustrated in Fig. 5. The output includes the 95% highest posterior density (HPD) intervals of the posterior distribution of the control mean, the posterior distribution of the treatment effect, and a detailed table including the sample size, EHSS, adaptive randomization probability, and the summary of the treatment effect λ.

In addition to Bayesian data analysis of the user-provided historical and concurrent data, interfaces for computing frequentist properties were developed for each of the Gaussian methods, for which computation is feasible in this framework due to their tractable formulations as demonstrated in Hobbs et al. (2012). For each frequentist properties interface, one

Table 1
Example data scenarios.

Scenario	Hist. 1	Hist. 2	Hist. 3	Hist.4	No. Hist	Control	Novel	No. Pat	Std.
1	0.49	0.5	0.48	0.52	40	0.5	0.5	25	0.2
2	0.34	0.35	0.33	0.37	40	0.5	0.5	25	0.2
3	0.57	0.37	0.33	0.47	40	0.5	0.5	25	0.2

may explore hyperparameter values and monitor the corresponding extent of bias and mean squared error (MSE) (i.e., risk under squared error loss (SEL)) for the resultant posterior mean estimator of the novel treatment effect as a function of the true extent of bias, Δ, attributable to trial effects. Frequentist properties such as 95% highest posterior density (HPD) interval coverage and width are also available for the EB and static borrowing methods. Additionally, the software plots mean EHSS as a function of true bias Δ for the dynamic EB method, elucidating potential gains in posterior precision and adaptive allocation arising from the Bayesian model specification. Computations require that the user provide historical control data as well as the sample size and planned allocation between control and experimental arms for a future study.

The remainder of this section demonstrates implementation of the interfaces through comparisons based on three hypothetical data examples described in Table 1 and illustrations of frequentist properties that result from various model specifications. In the first scenario, the mean values of the historical data and the control data are near each other, offering little empirical evidence of bias. The second example presents a scenario wherein trial effects may be evident, as the mean of the current control cohort appears to be larger than each of the historical cohorts. In the third scenario, the mean values of the historical cohorts vary about the mean value observed from the current control, suggesting high dispersion or lack of statistical exchangeability among the cohorts.

4.1. Static borrowing approach

As described in Section 2.1.3, the extent to which the model borrows strength from the historical information is predetermined when using the static borrowing method based on pre-specification of EHSS among the other input parameters on the webpage. Thus, mean EHSS is omitted from the frequentist properties output as it is assumed to be constant here.

Our examples present four historical cohorts each contributing 40 patients. This section considers analyses based on a fixed EHSS value of 100, reflecting an analysis with preference for a high degree of shrinkage or an *a priori* belief in limited bias from inter-trial heterogeneities. The 95% HPD intervals that result from these analyses are plotted in the top row of Fig. 6 along with the interval estimators that result from cohort-specific analyses of the historical and concurrent control data. Comparing the black- and maroon-colored segments, the preference for a high degree of partial pooling leads to much more precise estimates of the current control mean in all scenarios. For Scenario 1, the results appear to be favorable, as the historical and current cohort contribute commensurate estimators with little evidence for bias. Under Scenario 2, however, the results are much less favorable as the Bayesian interval estimator nearly fails to intersect with the corresponding interval obtained from ignoring the historical data. Scenario 3 represents an intermediate result. While popular due to ease of interpretation and straightforward model specification, static Bayesian analyses are inflexible in the sense that they do not consider the empirical similarity of the historical and concurrent controls, often leading to gains in posterior precision that are partially or fully offset by unwelcome increases in posterior bias. Section A of the Appendix presents results obtained from the static borrowing frequentist properties interface for two different EHSS values (10.0 and 100.0). The results further elucidate the trade-offs and statistical properties offered by the static Bayesian method.

4.2. Dynamic borrowing methods

The section demonstrates the implementation of the dynamic approaches based on empirical Bayesian inference and fully Bayesian hierarchical modeling with a spike-and-slab hyperprior. These methods admit data-dependent estimators of experimental treatment effects, for which the amount of strength borrowed from the historical controls is a function of the empirical evidence for bias.

4.2.1. Empirical Bayesian

As described in Section 2, the EB analysis interface implements posterior inference conditionally on an estimate (3) of the borrowing hyperparameter τ, obtained via constrained optimization of the marginal density. This is achieved through a user-specified upper bound for the EHSS parameter. Apart from this, the interface inputs are structurally similar to those for static analysis. Example analyses for the EB method with a maximum EHSS=100 are presented in the second row of Fig. 6. When comparing to the interval estimators resulting from static Bayesian inference with EHSS=100, the EB interval estimators are sensitive to the empirical evidence for bias and absence of exchangeability. For example, EB offers a gain in posterior precision nearly identical to that of the static method in the favorable Scenario 1, while avoiding highly biased estimation in the less favorable Scenarios 2 and 3. The trends are even better elucidated through the frequentist properties presented in Section B of the Appendix with Fig. A.2, which depicts MSE and bias for estimation of the novel treatment effect as well as mean EHSS as a function of true historical bias.

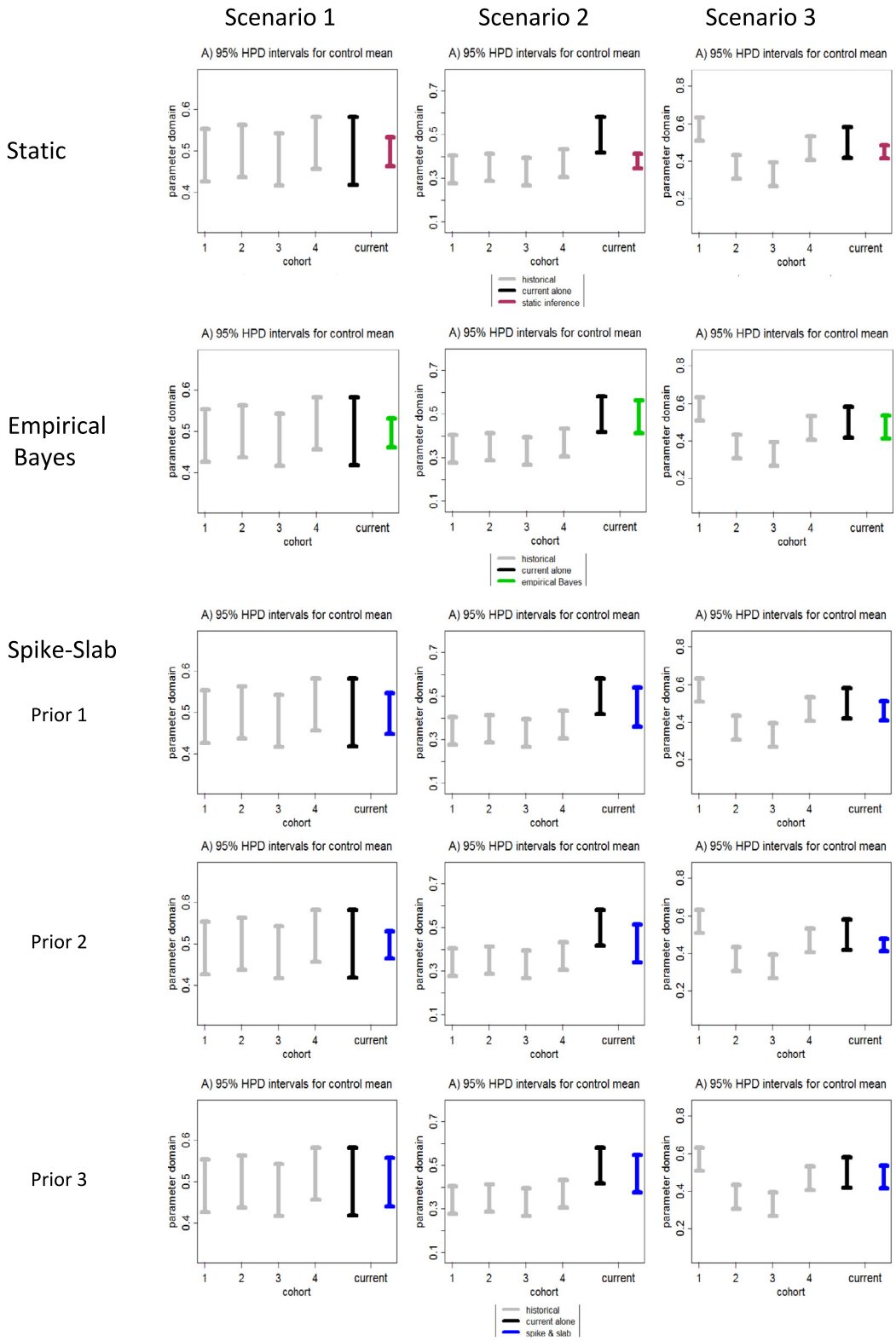


Fig. 6. Results of our example analyses for the methods of static, empirical Bayes, and the spike-and-slab. All model specifications impose a maximum value of EHSS=100.

Table 2
Slab–Spike hyperprior parameters.

Hyperprior	Slab S_L	Slab S_U	Spike	Slab probability
1	1	5	40	0.99
2	1	10	120	0.995
3	1	2	20	0.98

4.2.2. Spike-and-slab hierarchical model

Even more flexibility may be obtained using a fully Bayesian hierarchical model, incorporating full posterior inference for τ (equivalently, EHSS). As explained in Section 2, effective hierarchical modeling may necessitate a moderately informative hyperprior in the presence of a small number of cohorts. The spike-and-slab dynamic interfaces facilitate analysis based on a general Bayesian hierarchical model for EHSS, which includes two-point mixture and uniform densities as special cases. This method requires the user to specify the four spike-and-slab hyperparameters displayed in Fig. 2. Specifically, these include the upper and the lower bounds of the slab, the value of the spike, and the probability of the slab, all on the EHSS scale.

To illustrate, we consider three spike-and-slab hyperprior formulations, described in Table 2, in re-analyses of our three example data scenarios. The resulting interval estimators are presented in the last three rows of Fig. 6, which demonstrate the method's flexibility for imposing various bias–variance trade-offs through the spike-and-slab hyperprior specification. For example, Hyperprior 2, with its high EHSS spike $\mathcal{K} = 120$ and intermediate slab upper bound $\mathcal{S}_u = 10$, offers more gains in posterior precision under the borrowing-favorable Scenario 1 than the other hyperpriors, the EB and even the static approaches. However, this hyperprior also risks more bias under the less borrowing-friendly Scenarios 2 and 3, though slightly less than the static approach. Conversely, Hyperprior 3, which assumes a very low value of $\mathcal{K} = 20$, is more conservative, inducing less potential shrinkage under Scenario 1 but also rewarding us with less bias in the other scenarios. Frequentist properties depicted in Fig. A.3 of the Appendix further elucidate the bias–variance trade-offs that result from the spike-and-slab hyperprior formulations.

4.2.3. Right-censored time-to-event analysis

The interface also implements a dynamic Bayesian method for analysis of right-censored time-to-event data assuming a single historical cohort. As illustrated in Fig. 7, this model integrates partially informative historical control data into the analysis of a two-sample comparison using the piecewise exponential Bayesian hierarchical model described in Hobbs et al. (2013). For this model, sufficient summary statistics are not available. Implementation thus requires the user to upload files containing the historical and current data onto the server. The computation server validates the uploaded data, and performs the calculation using the joint commensurate prior model. Specific details about the data format and use of this interface are provided in the supplementary content. Estimation is conditional on the AIC-optimal time-axis partition, which is estimated in advance of MCMC-Bayesian inference. To avoid overfitting the data, the interface prompts the user to prespecify a minimum number of observed events to be required within each unique time interval of the AIC-optimal partition. Additionally, in order to conduct the analysis and calculate the resultant adaptive randomization probability, the user must indicate the number of patients that remain to be randomized in the study, as well as two hyperparameters for the spike-and-slab model specification of Hobbs et al. (2013). The interface outputs a table summarizing the posterior distributions of the treatment effect (hazard ratio), the posterior median failure time for the control and novel treatments, posterior mean EHSS value, and the optimal novel treatment adaptive randomization probability for the next subject. Additional output includes Kaplan–Meier curves for the historical data and concurrent data, as well as 95% pointwise HPD intervals for the survival function in the treatment and control groups.

We briefly demonstrate the interface using the hypothetical dataset depicted in Fig. 7, which is comprised of 200 historical controls, 200 concurrent controls, and 200 novel treatment patients. The only two hyperparameters that require specification are the upper slab parameter S_u and the slab prior probability p_0, which have been fixed at 2.0 and 0.5 in our analysis, respectively. In this interface, the lower slab parameter and the spike value are fixed at 0.1 and 5000, respectively.

Fig. 8 provides posterior summaries for the novel treatment effect, the median survival in the control and novel treatment groups, and the optimal adaptive randomization probability to the novel treatment for future patients. Given the lack of strong of evidence for bias in these data, the hierarchical model admits a relatively large EHSS of 60. Assuming 100 patients remaining to be randomized, the multisource adaptive design targeting effective information balance recommends an AR probability in favor the novel therapy of 0.798, again indicating a need for more novel treatment subjects since many effective patients can be borrowed from the historical controls.

5. Operating characteristics of the multi-source adaptive design

This section is intended to elucidate the properties of multi-source adaptive designs that result from using our interfaces for trial conduct at interim analyses to target effective information balance. Specifically, the simulation involves three historical control cohorts each consisting of 120 patients and one concurrent randomized trial enrolling a total of 120 patients. Observables in all cohorts are assumed to be normally distributed with variance of 1.0. For the two cohorts (control and experimental) in the concurrent randomized trial, we assume a mean response in the concurrent control cohort to

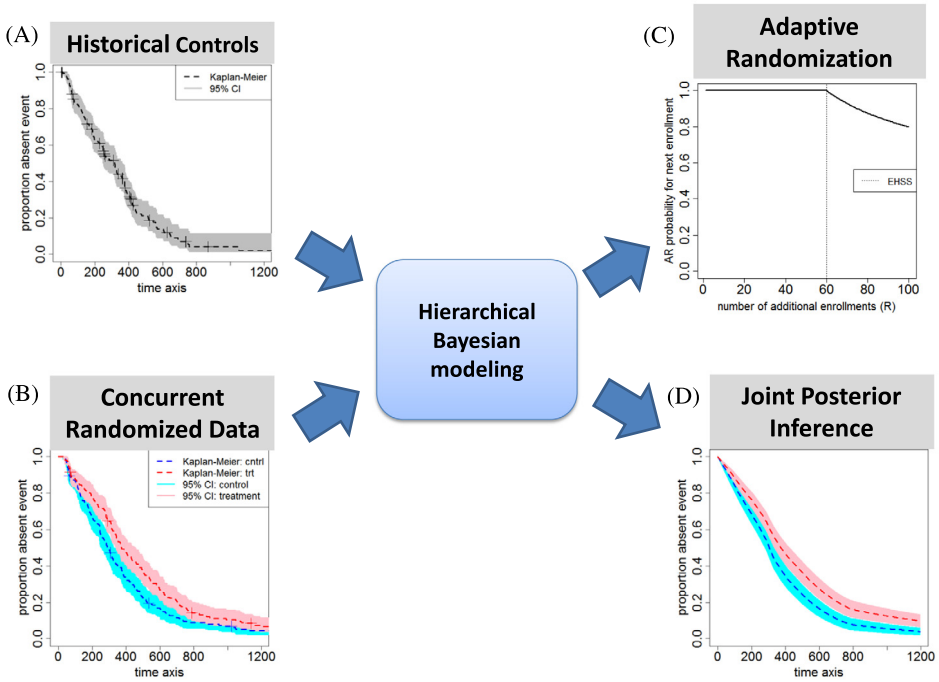


Fig. 7. Kaplan–Meier curves derived from the historical control data (upper) and current data (lower), with 95% log-transformed pointwise confidence intervals. Right-censored observations are marked by +. Posterior mean survival curves with 95% pointwise highest posterior density intervals derived from analysis of the historical and current data using the Bayesian hierarchical model. Adaptive randomization probability (π) for the next patient to enroll as a function of the number of patients remaining to enroll into the trial.

Posterior summary for novel treatment effect ξ, median survival for control and treatment; effective historical sample size (EHSS), and the probability used to randomized to treatment the next patient to enroll (AR probability)					
Parameter		**Mean**	**SD**	**Mean HR (95% HPD)**	**Pr(HR<1 | Data)**
Log-hazard ratio ξ		-0.372	0.098	0.72(0.57-0.87)	0.999
		Mean (95% HPD)			
Median	**Control**	305(279-333)			
	Treatment	373(330-420)			
# Current treated	200				
# Current controls	200				
# Historical controls	200				
EHSS	60.0				
AR probability	0.798				
Calculation Time (sec.)	5				
Note. SD = posterior standard deviation; HR = hazard ratio; HPD = highest posterior density interval; Pr(HR<1 | Data) = posterior probability < 1 for the hazard ratio indicating strength of evidence for an improvement associated with treatment arm					

Fig. 8. Example results for the right-censored time-to-event analysis.

be 0 for all scenarios. Then simulation scenarios were devised over a matrix of values characterizing the true extent of historical bias (mean of the current control minus the mean of the historical controls) denoted by Δ, and true treatment benefit of the experimental cohort in the concurrent trial denoted by λ. A positive test of novel treatment superiority used the posterior probability threshold that provides type I error of 0.1 and power $=$ 0.8 for true $\lambda = 0.39$ in the absence of the bias (true $\Delta = 0$) for the reference design ignoring historical controls with the "no borrowing" model. Motivating the need analytical frameworks that acknowledge potential heterogeneity when integrating data from multiple disparate source, Fig. 9 demonstrates the frequentist power and type I error that results from non-adaptive trial designs that use equal

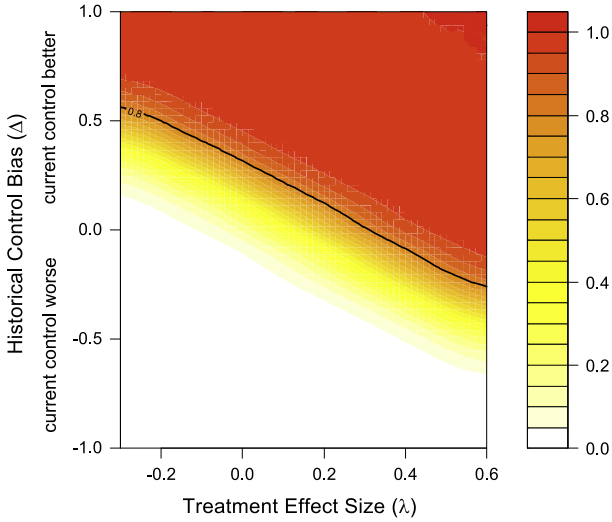


Fig. 9. Power surface obtained from assuming exchangeable historical and concurrent control cohorts.

randomization and pool concurrent and historical controls under the exchangeable data assumption. It should be noted that this example effectively underlies the predominant mode of intermediate-phased clinical testing, wherein single-arm trials are devised with respect to hypothesis tests that compare to a fixed null value derived from point estimation of data obtained in a prior study. Our example is actually less extreme in that it recognizes uncertainty in the current and historical control parameter estimators. In the absence of bias ($\Delta = 0$), assuming fully exchangeable controls facilitates improvement with respect to the reference design with power $= 0.8$ achieved for true $\lambda = 0.31$. Ignoring potential heterogeneity among controls, however, yields very poor frequentist properties for treatment comparison as was demonstrated for bias and MSE in Section 4. When the concurrent controls outperform historical, $\Delta > 0$, the paradigm is very likely to result in a positive test in the absence of true benefit. Specifically, our example settings yields type I error > 0.42 for true $\Delta > 0.14$ when true $\lambda = 0$. Critically, this scenario is likely commonplace in clinical oncology settings wherein projections from phase II trials have routinely overestimated the potential comparative effectiveness of novel therapeutic strategies based on single-arm trials perhaps by choosing a sufficiently low null hypothesis value or leveraging prior studies from patient populations with less favorable prognostic characteristics or older formulations of the control therapy. On the other hand, when $\Delta < 0$, naive exchangeable data modeling attenuates power for identifying true treatment benefits, e.g. with less than 0.55 power for true $\lambda = 0.39$ when $\Delta < -0.20$.

An ideal analytical framework controls type I error at an acceptable level when true $\lambda = 0$ with monotonically increasing power for $\lambda > 0$. Fig. 10 illustrates the frequentist power and type I error (top row) that result from multi-source adaptive designs that adjust allocation to target balanced enrollment after a single interim analysis based on interim estimation of EHSS. Additionally, mean EHSS and mean proportion allocated to novel therapy are depicted in the middle and bottom rows, respectively, which are shown to adapt as a function of true bias, Δ. Specifically, the adaptive designs were implemented in two-stages. An initial "burn-in" stage in which equal randomization was applied, and a subsequent adaptive randomization stage using the possibly unequal randomization allocation probability resulting from a single interim analysis to estimate EHSS based on the empirical Bayesian dynamic method. To balance total information between the control and experimental cohorts, the adaptive design assigns more patients to the experimental cohort after the burn-in stage of the trial. All plots are depicted for three burn-in periods (40, 60, and 80).

As shown in the middle and bottom graphs, for each burn-in period, less bias (Δ closer to zero) yields larger EHSS and thereby proportionally more dynamic allocation to novel therapies. This improves power when $\lambda > 0$ for a comparable type I error rate when $\lambda = 0$. To aid this comparison, we have added contours lines reflecting 0.8 power for the reference design (or no borrowing; green), multi-source adaptive design with EB (in black), as well as a contour line reflecting equal randomization during the trial with final analysis based on the same empirical Bayesian method (in blue). At $\Delta = 0$ we see that the multi-source adaptive design achieves power 0.8 at lower effect sizes (λ) than both the reference no-borrowing design and equal randomization with joint final analysis. In the presence of a sufficient extent of true bias ($|\Delta| > 0$), the Bayesian model is sensitive to bias. Thus, EHSS is effectively zero for the adaptive design, yielding little borrowing from the historical controls (and thus little dynamic allocation) which consequently protects the validity of the treatment comparison and validity for concluding attribution to novel therapeutic benefit. Note that at the boundaries of the vertical axes, the power surface that results from the multi-source adaptive design reflects minimal disruption of type I error and power when compared to the considerable skewness evident in Fig. 9. This is evident in Fig. 10 by the eventual overlap of green, black, and blue curves.

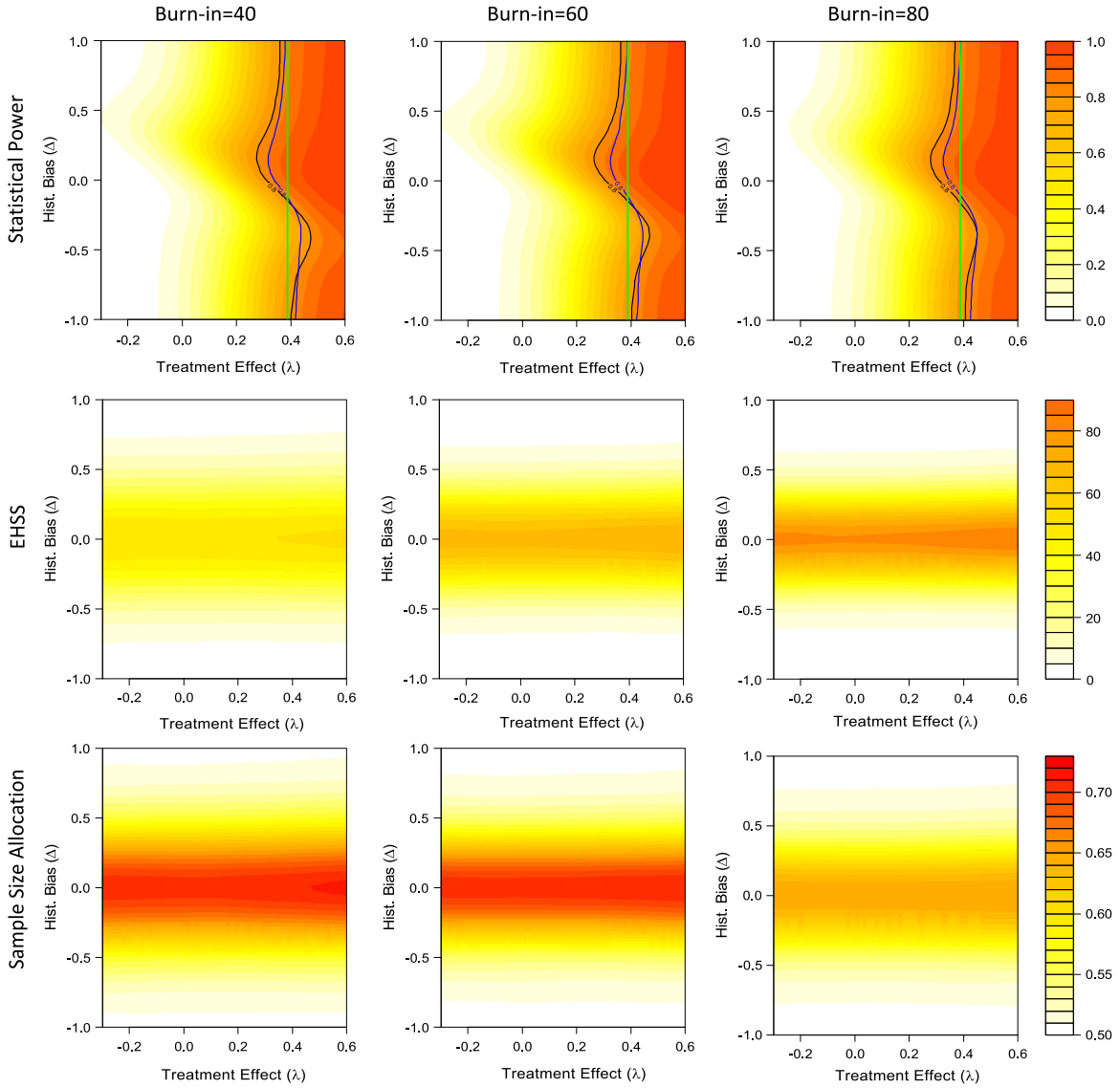


Fig. 10. Operating characteristics obtained from multi-source adaptive design with a single interim analysis based on the empirical Bayesian method. In the first row, rejection rates (power values) with different number of patients in the burn-in stage are illustrated. Green, black, and blue contour lines correspond to the 80% power of no borrowing, dynamic borrowing with AR, and dynamic borrowing with ER designs, respectively. In the third row, allocation rates to the experimental cohort are illustrated. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The third row of Fig. 10 depicts performance for the optimal allocation to experimental therapy using the adaptive design. Allocation to novel therapy is high when the values of Δ are close to 0, when the dynamic procedure yields higher estimated EHSS. Note that the burn-in period determines the quality of EHSS estimation. More burn-in provides less deviation and higher EHSS, on average, improving performance for the adaptive in relation to the non-adaptive ER design with final joint analysis. In fact, for a burn-in period of 80 patients, the AR power values are greater than or equal to the ER power values across the entire range of Δ. With fewer adaptively randomized patients, however, larger burn-in limits the extent to which the adaptive allocation procedure can balance effective information across the study arms. Fig. 11 illustrates the trade-off of burn-in versus allocation when $\Delta = 0$. For intermediate burn-in periods, the multi-source adaptive design facilitates more efficiency with the adaptive allocation of >70% of new patients to the novel therapy in the absence of true bias. For burn-in periods greater than 60, however, the number of patients that remain after the burn-in stage reduces to an extent that cannot compensate for the increase in EHSS (and thus adaptive randomization probability) causing an overall reduction of the experimental therapy allocation. The trends for all considered values of Δ are conveyed in the third row of Fig. 10.

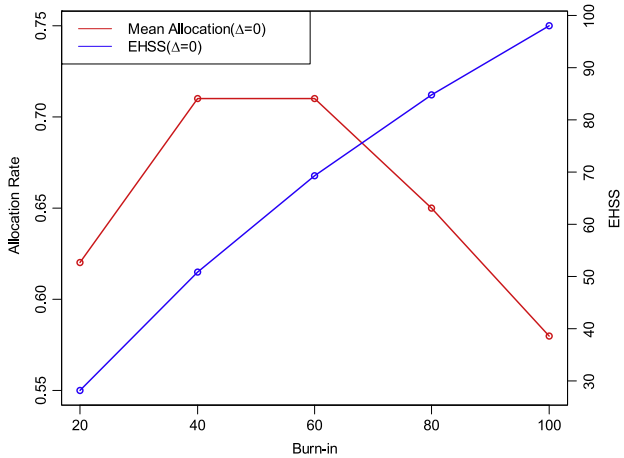


Fig. 11. Effect of burn-in period on the overall sample size allocation to the experimental arm.

6. Discussion

This article presented a collection of web-based tools for the design and analysis of clinical trials that incorporate historical information about the control therapy while adjusting for bias stemming from inter-trial heterogeneity. Hosted by the University of Texas M.D. Anderson Cancer Center, these tools are freely available at http://research.mdacc.tmc.edu/SmeeactWeb/. After a review of the Bayesian models underlying the methods, we described the website and its design, illustrated its use for both Gaussian and time-to-event likelihoods, and described their frequentist properties for both point estimation of treatment effects as well as power and size across a range of true historical control biases and treatment effects.

A few limitations should be noted. In particular, concerns of bias are intrinsic to any method which attempts to infer causal associations while integrating data from multiple heterogeneous populations. Decisions that pertain to which historical control sources should be considered for joint Bayesian analysis should be based on the study objectives, as well as centered on controlling bias arising from discernible sources, such as inconsistent eligibility criteria, application of the interventions, or manner of outcome measurement. Additionally, the type of study contributing historical controls, retrospective/observational versus prospective/randomized, impacts the quality and reliability of the resultant estimators, and thus design methodology should be taken into account when determining the quality of historical control sources. With the emergence of Bayesian nonparametric density estimation and multisource exchangeability models (Kaizer et al., 2018b), more flexible hierarchical modeling strategies are available and may improve shrinkage estimation in the presence of potential non-exchangeable historical source data. Furthermore, innovations in platform trial design (Hobbs et al., 2018) enable multiple therapeutics to enter and exit the trial seamlessly, improving efficiency when compared to conducting multiple independent two-arm trials. Kaizer et al. (2018a) have recently demonstrated that platform designs may benefit from multi-source adaptive randomization strategies to adjust for bias stemming from population drift.

Acknowledgments

We thank Clift Norris for his help on the web-site design, Sylvain Laroche for the server management. The work was supported in part by National Cancer Institute grants R01-CA157458 and P30-CA016672.

Appendix A

This technical appendix illustrates the frequentist properties interfaces and compares methods using the example data and model specifications discussed in Section 4.

A.1. Static borrowing frequentist properties example

This subsection presents results obtained from the static borrowing frequentist properties interface for two different EHSS values (10.0 and 100.0). The results, provided in Fig. A.1, further elucidate the trade-offs and statistical properties offered by the static Bayesian method. Specifically, the plots depict the frequentist properties for the experimental treatment effect estimator of bias, MSE, and 95% HPD coverage as a function of historical control bias using the historical control data of Scenario 3. The true extent of historical bias is determined in the sampling model via a fixed true value of Δ. No inter-cohort bias is present only for sampling models assuming a true value of $\Delta = 0$. MSE and HPD coverage are functions of the magnitude, but not direction, of historical bias, and thus are necessarily symmetric about 0. As benchmarks for evaluating

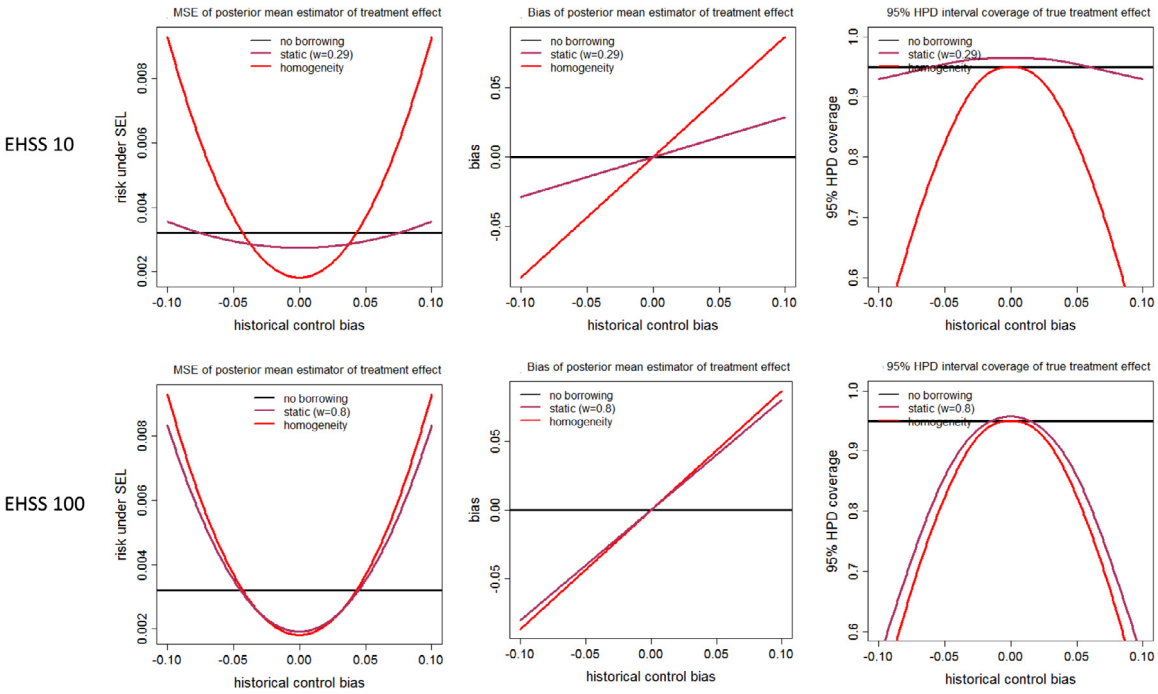


Fig. A.1. The calculated results of static borrowing frequentist properties for scenario 3 with EHSS=10.0 (top row) and EHSS=100.0 (bottom row).

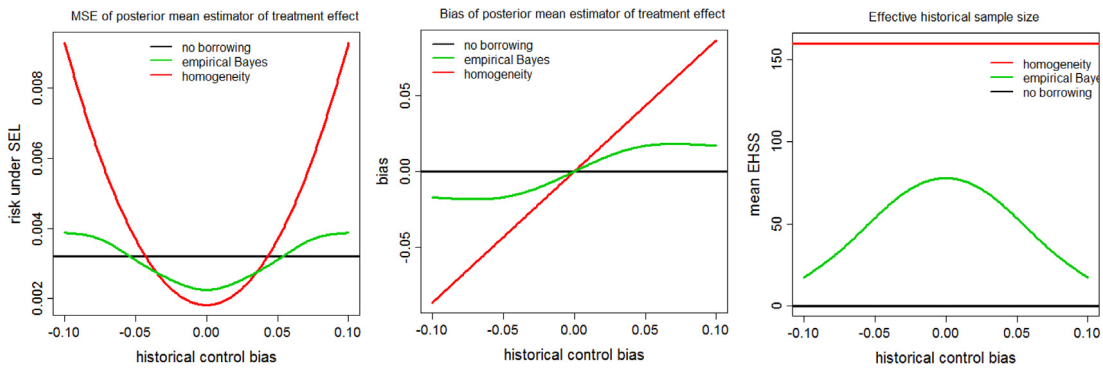


Fig. A.2. Frequentist properties of the empirical Bayes dynamic borrowing analysis for scenario 3 (Maximum EHSS=100).

performance, each plot also depicts the results that would be obtained from non-hierarchical Bayesian methods assuming exchangeable data or "homogeneity" (red) and independence or "no borrowing" (black). Notably, under independence the historical data is ignored and thus the frequentist properties are invariant as functions of the historical bias Δ. Conversely, the assumption that data acquired from the historical and concurrent cohorts is statistically exchangeable yields an estimator of the experimental treatment effect with linear trends in bias and parabolic trends in MSE and HPD coverage. The red curves demonstrate that incorporating the historical information without acknowledging the potential for trial effects offers a more favorable bias–variance trade-off (as characterized by MSE) within a more limited interval about 0 while risking considerable bias as $|\Delta|$ increases from 0.

The data-independent static Bayesian approach yields estimators of the experimental treatment effect that are structurally similar to the homogeneity analysis, with linear trends in bias and parabolic trends in MSE. The examples demonstrate that for the static method, the data-independent pre-specified value of EHSS determines the slope of the linear bias, as well as the depth of the MSE's vertex at $\Delta = 0$ and steepness with which it diverges from the vertex. Sufficiently large values of EHSS (such as 100 in our example setting) are insensitive to evidence for bias in the data which yields poor frequentist properties that mirror those of homogeneity. To adjust for this, considerable reductions in borrowing magnitude

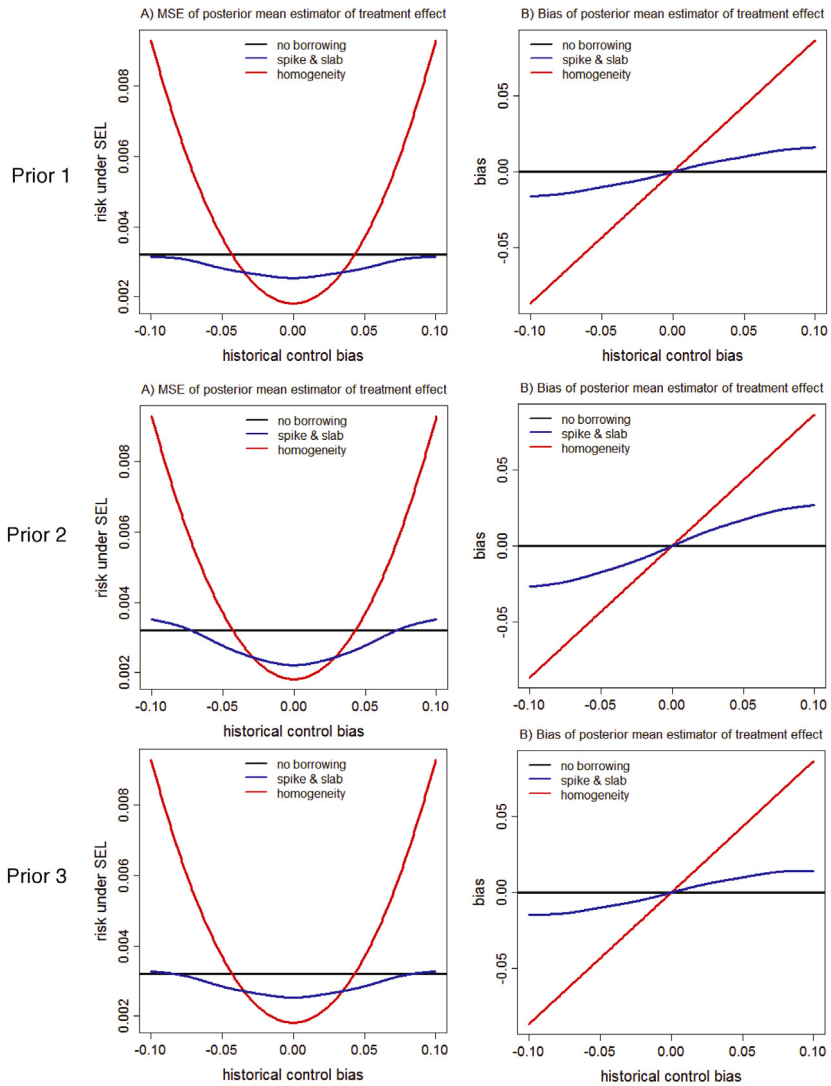


Fig. A.3. Frequentist properties of spike-and-slab hyperprior dynamic borrowing for scenario 3 with different prior parameters.

are necessary with the static model. For example, when compared to 100, an EHSS value of 10 attenuates the extent of bias within the domain of historical bias, yielding MSE that extends the domain by which the static method improves upon independence. In the absence of bias, however, diminishing borrowing in this way limits the extent to which the multi-source adaptive designs may effectively re-balance the allocation of future patients and thereby increase statistical power for the treatment comparisons.

A.2. Empirical bayesian frequentist properties example

This subsection presents frequentist properties for the empirical Bayesian method described in Section 4.2.1 with maximum EHSS $= 100$. Fig. A.2 depicts MSE and bias for estimation of the novel treatment effect as well as mean EHSS as a function of true historical bias. The figures demonstrate the method's sensitivity which is evident from non-monotonic trends of MSE and bias. In the absence of bias ($\Delta = 0$) the EB estimators offer nearly the extent of reductions in MSE attained by the naive exchangeability assumption. In the presence of bias, however, the treatment effect estimators are considerably less biased when estimated with EB. Moreover, the bias actually peaks and then attenuates as the data exhibit stronger empirical evidence for bias. This is evident from the plot of mean EHSS, which is maximal at 0 and functionally decreasing with increasing bias. Thus, unlike the static borrowing method, EB offers more potential to re-balance an adaptive design's accrual while protecting the integrity of the treatment comparison.

A.3. Spike-and-slab frequentist properties example

This subsection presents frequentist properties for the spike-and-slab hyperprior specifications described in Table 2 of Section 4.2.2. Fig. A.3 elucidates the bias–variance trade-offs that result from the spike-and-slab hyperprior formulations. For example, Hyperprior 2 offers more gain in MSE in the absence of bias than all other comparators. However, bias accrues more rapidly in this case. Hyperprior 1 offers perhaps the most desirable bias–variance trade-off, with improved MSE with respect to no borrowing over the entire domain of historical bias.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2018.05.002.

References

Anderson, G., LeBlanc, M., Liu, P.Y., Crowley, J., 2011. On use of covariates in randomization and analysis of clinical trials. In: Crowley, J., Hoering, A. (Eds.), Handbook of Statistics in Clinical Oncology, third ed. Chapman and Hall/CRC Press, Boca Raton, FL.

Berry, D.A., 2006. Bayesian clinical trials. Nat. Rev. Drug Discov. 5, 27–36.

Browne, W.J., Draper, D., 2006. A Comparison of Bayesian and likelihood-based methods for fitting multilevel models. Bayesian Anal. 1 (3), 473–514.

Carlin, B.P., Louis, T.A., 2009. Bayesian Methods for Data Analysis, third ed. Chapman and Hall/CRC Press, Boca Raton, FL.

Chen, M.-H., Ibrahim, J.G., Lam, P., Yu, A., Zhang, Y., 2011. Bayesian design of noninferiority trials for medical devices using historical data. Biometrics 67 (3), 1163–1170.

Cheung, Y.K., Inoue, L.Y.T., Wathen, J.K., Thall, P.F., 2006. Continuous Bayesian adaptive randomization based on event times with covariates. Stat. Med. 25, 55–70.

Committee on Cancer Clinical Trials and the NCI Cooperative Group Program Board on Health Care Services, 2010. Institute of Medicine: A National Cancer Clinical Trials System for the 21st Century: Reinvigorating the NCI Cooperative Group Program. In: Nass, S.J., Moses, H.L., Mendelsohn, J. (Eds.), National Academies Press.

Daniels, M.J., 1999. A prior for the variance in hierarchical models. Canad. J. Statist. 27 (3), 567–578.

De Santis, F., 2006. Power priors and their use in clinical trials. Amer. Statist. 60 (2), 122–129.

Doi, S., Barendregt, J., Mozurkewich, E., 2011. Meta-analysis of heterogeneous clinical trials: An empirical example. Contemp. Clin. Trials 32 (2), 288–298.

French, J.L., Thomas, N., Wang, C., 2012. Using historical data with Bayesian methods in early clinical trial monitoring. Stat. Biopharmaceutical Res. 4 (4), 384–394.

Friedman, L., Furberg, C., DeMets, D.L., 1998. Fundamentals of Clinical Trials, third ed. Springer-Verlag, New York.

Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models. Bayesian Anal. 1 (3), 515–534.

Goodman, S.N., Sladky, J.T., 2005. A Bayesian approach to randomized controlled trials in children utilizing information from adults: The case of Guillain-Barre. Clinical Trials 2 (4), 305–310.

Hobbs, B.P., Carlin, B.P., Mandrekar, S.J., Sargent, D.J., 2011. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. Biometrics 67 (3), 1047–1056.

Hobbs, B.P., Carlin, B.P., Sargent, D.J., 2013. Adaptive adjustment of the randomization ratio using historical control data. Clin. Trials 10, 430–440.

Hobbs, B.P., Chen, N., Lee, J.J., 2018. Controlled multi-arm platform design using predictive probability. Stat. Methods Med. Res. 27 (1), 65–78.

Hobbs, B.P., Sargent, D.J., Carlin, B.P., 2012. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. Bayesian Anal. 7, 639–674.

Hu, F., Rosenberger, W., 2006. The Theory of Response-Adaptive Randomization in Clinical Trials. John Wiley and Sons, New York.

Huang, X., Ning, J., Li, Y., Estey, E., Issa, J.-P., Berry, D.A., 2009. Using short-term response information to facilitate adaptive randomization for survival clinical trials. Stat. Med. 28, 1680–1689.

Ibrahim, J.G., Chen, M.-H., 2000. Power prior distributions for regression models. Statist. Sci. 15 (1), 46–60.

Kaizer, A.M., Hobbs, B.P., Koopmeiners, J.S., 2018. A multisource adaptive platform design for testing sequential combinatorial therapeutic strategies. Biometrics.

Kaizer, A.M., Koopmeiners, J.S., Hobbs, B.P., 2018. Bayesian hierarchical modeling based on multisource exchangeability. Biostatistics 19 (2), 169–184.

Kass, R.E., Natarajan, R., 2006. A default conjugate prior for variance components in generalized linear mixed models. Bayesian Anal. 1 (3), 535–542.

Korn, E.L., Freidlin, B., 2011. Outcome-adpative randomization: is it useful?. J. Clin. Oncol. 29, 771–776.

Lee, J.J., Gu, X., Liu, S., 2010. Bayesian adaptive randomization designs for targeted agent development. Clin. Trials 7, 584–597.

Mitchell, T.J., Beauchamp, J.J., 1988. Bayesian variable selection in linear regression. J. Amer. Statist. Assoc. 83, 1023–1032.

Morita, S., Thall, P.F., Müller, P., 2008. Determining the effective sample size of a parametric prior. Biometrics 64, 595–602.

Morita, S., Thall, P.F., Müller, P., 2012. Prior effective sample size in conditionally independent hierarchical models. Bayesian Anal. 7, 591–614.

Murray, T.A., Hobbs, B.P., Lystig, T.C., Carlin, B.P., 2014. Semiparametric bayesian commensurate survival model for post-market medical device surveillance with non-exchangeable historical data. Biometrics 70 (1), 185–191.

Natarajan, R., Kass, R.E., 2000. Reference Bayesian methods for generalized linear mixed models. J. Amer. Statist. Assoc. 95 (449), 227–237.

Neelon, B., O'Malley, A.J., 2010. Bayesian analysis using power priors with application to pediatric quality of care. J. Biometrics Biostat. 1 (103).

Neuenschwander, B., Capkun-Niggli, G., Branson, M., Spiegelhalter, D.J., 2010. Summarizing historic information on controls in clinical trials. Clin. Trials 7, 5–18.

Pennello, G., Thompson, L., 2008. Experience with reviewing bayesian medical device trials. J. Biopharm. Statist. 18, 81–115.

Pocock, S.J., 1976. The combination of randomized and historical controls in clinical trials. J. Chronic Dis. 29, 175–188.

Pocock, S.J., Simon, R., 1975. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics 31, 103–115.

Rietbergen, C., Klugkist, I., Janssen, K., Moons, K., Hoijtink, H., 2011. Incorporation of historical data in the analysis of randomized therapeutic trials. Contemp. Clin. Trials 32, 848–855.

Smith, T.C., Spiegelhalter, D.J., Thomas, A., 1995. Bayesian approaches to random-effects meta-analysis: A Comparative study. Stat. Med. 14 (24), 2685–2699.

Spiegelhalter, D.J., 2001. Bayesian methods for cluster randomized trials with continuous responses. Stat. Med. 20, 435–452.

Sverdlov, O., Tymofyeyevb, Y., Wong, W.K., 2011. Optimal response-adaptive randomized designs for multi-armed survival trials. Stat. Med. 30, 2890–2910.

Thall, P., Fox, P., Wathen, J., 2015. Statistical controversies in clinical research: Scientific and ethical problems with adaptive randomization in comparative clinical trials. Ann. Oncol. 26, 1621–1628.

Whitehead, J., Valdés-Márquez, E., Johnson, P., Graham, G., 2008. Bayesian sample size for exploratory clinical trials incorporating historical data. Stat. Med. 27 (13), 2307–2327.
Williams, R.J., Tse, T., DiPiazza, K., Zarin, D.A., 2015. Terminated trials in the clinicaltrials.gov results database: Evaluation of availability of primary outcome data and reasons for termination. PLoS One 10 (5).
Yin, G., Chen, N., Lee, J.J., 2012. Phase II trial design with Bayesian adaptive randomization and predictive probability. J. R. Stat. Soc. Ser. C. Appl. Stat. 61, 219–235.