Randomization in the presence of historical controls

Brian P. Hobbs\textsuperscript{a}, Bradley P. Carlin\textsuperscript{b}, and Daniel J. Sargent\textsuperscript{c}

\textsuperscript{a}Department of Biostatistics, M.D. Anderson Cancer Center, Houston, TX, 77030, USA

\textsuperscript{b}Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota, 55455, USA

\textsuperscript{c}Department of Health Sciences Research, Mayo Clinic, Rochester, MN, 55905, USA

word count 5391

\footnotesize{The work of the first author was supported in part by the University of Texas M.D. Anderson’s Cancer Center Support Grant NIH P30 CA016672. The work of the second author was supported in part by National Cancer Institute grant 2-R01-CA095955-05A2. The work of the third author was supported in part by NCI grant CA25224. Author for correspondence: Brian P. Hobbs, Department of Biostatistics, M.D. Anderson Cancer Center, Houston, TX, 77030, USA. E-mail: bphobbs@mdanderson.org}
Abstract

**Background:** Prospective trial design often occurs in the presence of “acceptable” [1] historical control data. Typically this data is only utilized for treatment comparison in a posteriori retrospective analysis to estimate population-averaged effects in a random-effects meta-analysis.

**Purpose:** We propose and investigate an adaptive trial design in the context of an actual randomized controlled colorectal cancer trial. This trial, originally reported by Goldberg et al. [2], succeeded a similar trial reported by Saltz et al. [3], and used a control therapy identical to that tested (and found beneficial) in the Saltz trial.

**Methods:** The proposed trial implements an adaptive randomization procedure for allocating patients aimed at balancing total information (concurrent and historical) among the study arms. This is accomplished by assigning more patients to receive the novel therapy in the absence of evidence for heterogeneity among the concurrent and historical controls. Allocation probabilities adapt as a function of the effective number of historical controls (EHC) characterizing relative informativeness defined in the context of a semiparametric piecewise exponential model for evaluating time-to-disease progression. Commensurate priors [4] are utilized to assess historical and concurrent heterogeneity at two interim analyses and to borrow strength from the historical data in the final analysis. The adaptive trial’s frequentist properties are simulated using the actual patient-level historical data from the Saltz trial and the actual enrollment dates for patients enrolled into the Goldberg trial.

**Results:** Assessing concurrent and historical heterogeneity at interim analyses and balancing total information with the adaptive randomization procedure leads to trials that on average assign more new patients to the novel treatment when the historical controls are unbiased or slightly biased compared to the concurrent controls. Biased historical controls lead to trials that on average allocate patients equally among the treatment arms. Using the proposed commensurate prior model to borrow strength from the historical data, after balancing total information with the adaptive randomization procedure, produces an estimator of the treatment effect that dominates the no-borrowing estimator in terms of preposterior risk under squared error loss.
**Limitations:** Adaptive randomization methods in general are sensitive to population drift and more suitable for trials that initiate with gradual enrollment. Balancing information among study arms in a time-to-event analyses is difficult in the presence of informative right-censoring.

**Conclusions:** The proposed design could prove important in trials that follow recent evaluations of a control therapy. Efficient use of the historical controls is especially important in contexts where reliance on pre-existing information is unavoidable because the control therapy is exceptionally hazardous, expensive, or the patient population is sparse.

Keywords: adaptive designs, Bayesian analysis, historical controls

1 **Introduction**

Adaptive designs of clinical trials facilitate mid-trial modifications based on interim information from internal or external sources. Recently methods have been proposed to facilitate adaptivity for prospective modification of many trial features including randomization [5, 6]. Friedman, Furberg, and DeMets [7] broadly refer to randomization methods that facilitate mid-trial adjustments to the allocation ratios as *adaptive*. Two types of adaptive randomization (AR) procedures are commonly implemented in clinical trials. *Baseline AR* designs are used to balance the study arms with respect to prognostic factors that are available at baseline [8, 9]. By contrast, *response-adaptive* or *outcome-adaptive* designs were developed for the purpose of assigning more patients to more effective or safer treatment regimens based on interim data from an ongoing trial [10, 11, 12, 13, 14]. While Korn and Freidlin [15] have questioned the usefulness of designs that use outcome-adaptive randomization for this purpose, the potential remains to benefit from adaptive designs that assign more new patients to learn about newer, less studied therapies in the presence of historical controls that satisfy Pocock’s [1] “acceptability” criteria. Such designs promise to enhance efficiency when implementing controlled clinical trials that follow recent evaluations of a control therapy by facilitating more precise estimates of the treatment effect. However, highly subjective a priori
assumptions of homogeneity among historical and concurrent controls may lead to poor frequentist operating characteristics and highly biased results.

In this article we propose an adaptive analog of an actual randomized controlled colorectal cancer trial originally reported by Goldberg et al. [2] that succeeded a similar trial reported by Saltz et al. [3]. The Goldberg trial used as its control arm an identical treatment to that found superior in the Saltz trial. The proposed adaptive design is based on a randomization method that adapts as a function of the relative informativeness of the historical data for evaluating the endpoint of interest (in this case time-to-disease progression). Actual patient-level data from the Saltz trial (historical) and enrollment dates from the Goldberg (current) are used to simulate the proposed adaptive design’s frequentist properties. Historical data is formally incorporated into the Bayesian analysis of Goldberg trial using commensurate priors [16, 4] in the context of a semiparametric piecewise exponential model [17].

In the absence of strong evidence for heterogeneity among the historical and accumulating current controls, the proposed adaptive design assigns more new patients to learn more about the safety and efficacy of the novel therapies, thus attempting to impose balance between concurrent and historical information. After an initial stage of equal allocation, the allocation probability is adjusted as a function of the effective number of historical controls (EHC). EHC is defined as a function of a measure of the degree to which estimates of model parameters are “shrunk” toward their historical counterparts.

Designs such as this promise to be important in contexts where reliance on pre-existing information is unavoidable, as occurs when the control therapy is exceptionally hazardous or expensive, or when the patient population is sparse. There are many sub-areas of oncology that could benefit from applications of these methodologies, including rare subtypes of sarcomas, rare progressions of renal cell carcinoma such as advanced sarcomatoid, and pediatric brain tumors such as choroid...
plexus carcinoma, medulloblastoma, and pontine glioma, to mention a few.

This article proceeds as follows. Section 2 describes the colorectal cancer clinical trials that motivated the research. Section 3 introduces the concept of effective historical sample size in the context of Bayesian analysis with a hierarchical model. Section 4 introduces the probability models used to evaluate time-to-disease progression in the proposed, adaptive trial. In Section 5 we formulate our proposed alternative, adaptive design. Section 6 discusses the adaptive design’s frequentist properties. Finally, Section 7 concludes and suggests avenues for further development.

2 Colon cancer trials

The proposed adaptive design is motivated by two successive randomized controlled colorectal cancer clinical trials originally reported by [2] and [3]. The initial trial [3] randomized 683 patients with previously untreated metastatic colorectal cancer between May 1996 and May 1998 to one of three regimens: Irinotecan alone; Irinotecan and bolus Fluorouracil plus Leucovorin (IFL); or a regimen of Fluorouracil and Leucovorin (5FU/LV) (“standard therapy”). IFL resulted in significantly longer time-to-disease progression and overall survival than both Irinotecan alone and 5FU/LV, and thus became the standard of care therapy. The subsequent trial [2] compared two new (at the time) drug combinations in 795 patients with previously untreated metastatic colorectal cancer, randomized between May 1999 and April 2001. Patients in the first drug group received the current “standard therapy,” the IFL regimen identical to that used in the historical study. The second group received Oxaliplatin and infused Fluorouracil plus Leucovorin (abbreviated FOLFOX), while the third group received Irinotecan and Oxaliplatin (abbreviated IROX); both of these latter two regimens were new as of the beginning of the second trial.

Historical control data from the Saltz trial appears to satisfy Pocock’s acceptability criteria [1]. Less than one year elapsed in the time between the last patient enrolling into the Saltz trial and the
first patient enrolling into the Goldberg trial. Both trials used identically defined IFL therapies, both trials used similar inclusion and exclusion criteria, and both used identical criteria to assess the endpoint of interest, time-to-disease progression (TTP).

Our proposed adaptive trial compares TTP between the FOLFOX and IFL regimens. The historical controls consist of patients randomized to IFL in the Saltz study; the current data consists of patients randomized to IFL or FOLFOX in the Goldberg trial. For simplicity, we omit data from the Irinotecan alone and 5FU/LV arms in the Saltz study, and the IROX arm in the Goldberg study. We consider only patients that had measurable tumors at baseline, bringing the total sample size to 643: 224 historical and 419 current observations. Among the current patients, there are 208 controls (IFL) and 211 patients treated with the new regimen (FOLFOX).

3 Effective historical sample size

Balancing information among treatment arms as data accrues in an adaptive, controlled trial, requires interim assessment of the relative informativeness of the historical data. Before formulating our adaptive trial design we briefly discuss the concept of effective historical sample size (EHSS) in the context of Bayesian analysis with a hierarchical model, a concept that will be used in Section 5 to adjust randomization allocation probabilities among study arms for new patients enrolled thereafter. Our work is related to that of Morita et al. [18], who considered effective sample size of parametric prior distributions for non-hierarchical models.

Let $y_0$ denote the vector of outcomes for $n_0$ patients assigned to the control therapy in the historical trial. Similarly, let $y$ denote the vector of outcomes for $n$ patients in the current trial. Let $\theta_0$ and $\theta$ denote analogous model parameters and $L(y|\theta)$ and $L(y_0|\theta_0)$ denote the likelihood functions corresponding to the historical and current data, respectively, where $\theta$ is the parameter of interest. Let $p(\theta)$ denote a suitable non-informative prior distribution for $\theta$. If the historical
information is ignored, inference proceeds with respect to the posterior distribution of \( \theta | y \): 

\[
q^* (\theta | y) \propto p(\theta) L(y | \theta).
\] (1)

We refer to (1) as the “reference” model.

Now consider borrowing strength from the historical data in the context of a hierarchical model. We may model a conditional relationship between \( \theta \) and \( \theta_0 \) by assuming a conditional prior distribution, \( p(\theta | \theta_0, \eta) \), that is dependent upon a hyperparameter, \( \eta \), controlling the amount of cross-study borrowing. Inference proceeds with respect to the posterior distribution of \( \theta | y_0, y, \eta \) induced by the hierarchical model:

\[
q (\theta | y_0, y, \eta) \propto L(y | \theta) \int_{\theta_0} p(\theta | \theta_0, \eta) p(\theta_0) L(y_0 | \theta_0) d\theta_0.
\] (2)

Model (2) facilitates borrowing of strength via joint modeling of the historical and current data. Thus, we refer to (2) as the “joint” model.

Let \( P^* (y) \) denote the posterior precision of \( \theta | y \) corresponding to inference under the reference model, \( P^* (y) = \left[ E_{\theta | y} \{\theta - E_{\theta | y}(\theta)\}^2 \right]^{-1} \). Let \( P (y_0, y, \eta) \) denote the posterior precision of \( \theta | y_0, y, \eta \) corresponding to inference under the joint model, \( P (y_0, y, \eta) = \left[ E_{\theta | y_0, y, \eta} \{\theta - E_{\theta | y_0, y, \eta}(\theta)\}^2 \right]^{-1} \). If the relationship among sample size and precision is reasonably linear under the reference model, then the rate of precision per patient may be approximated by \( P^* (y) / n \). Therefore, the effective sample size of the joint model’s posterior is approximately \( n \frac{P (y_0, y, \eta)}{P^* (y)} \), suggesting that a sensible functional relationship between joint posterior precision and EHSS follows as,

\[
EHSS \approx n \left\{ \frac{P (y_0, y, \eta)}{P^* (y)} - 1 \right\}.
\] (3)
In this formulation, EHSS is approximately the effective sample size of the joint posterior minus the current sample size $n$. If the joint model leads to little gain in precision, then the EHSS is small. In contrast, relatively large gains in precision will produce large values of EHSS. Appendix A illustrates the concept of EHSS in the context of a random-effects meta-analytic joint model.

4 Probability models

Pocock [1] proposed Bayesian hierarchical models for borrowing of strength from historical controls for time-to-event data in the context of parametric exponential models. He suggested that inference should proceed with fixed “magnitudes” of the historical bias, noting that one may wish to repeat the analysis with several alternative values expressing varying degrees of trust in the historical controls. For the case of one historical study, the general commensurate prior approach [4] extends the approach of Pocock and proposes a set of prior distributions for estimating the magnitude of historical bias from the observed data in the context of fully hierarchical and empirical Bayesian analysis for parametric inference with exponential families.

4.1 Semiparametric piecewise exponential model

Let $y$ and $c$ denote vectors containing time-to-event variables and right-censoring indicators for $n$ patients, respectively. We use a piecewise exponential model that assumes constant baseline hazards within finite partitions of the time axis. Such a flexible semiparametric model accommodates numerous shapes of the baseline hazard over the partition intervals. Following the notation of Ibrahim, Chen, and Sinha [17, p.48], we partition the time axis into $J + 1$ intervals, $(0, s_1], (s_1, s_2], \ldots, (s_{J-1}, s_J], (s_J, \infty)$ where $0 < s_1 < s_2 < \ldots < s_J < \infty$, and $s_J$ denotes the end of patient follow-up, the maximum possible time that a patient could be followed during the trial. For $y$ in the $j$th interval, $s_{j-1} < y \leq s_j$, the model assumes a constant baseline hazard $\lambda_j > 0$, \
for $j = 1, ..., J$. Suppose $x_i$ is a vector of $p$ patient-specific baseline covariates, where $i = 1, ..., n$ indexes patient, and $\beta = (\beta_1, ..., \beta_p)$ is the corresponding vector of regression coefficients. Let $\lambda$ denote the vector of baseline hazards, $\lambda = (\lambda_1, ..., \lambda_J)$. Within the $j$th time partition interval, the hazard for the $i$th patient is assumed to be $h(\lambda_j, x_i, \beta) = \lambda_j \exp(x_i \beta)$. Let $\theta$ denote the vector of model parameters $\theta = (\lambda, \beta)$. The $i$th patient’s full contribution to the likelihood follows as,

$$L(y_i | \theta) = \prod_{j=1}^{J+1} \{\lambda_j \exp(x_i \beta)\}^{I(s_{j-1} < y_i \leq s_j)(1 - c_i)}$$

$$\times \exp \left[ -I(s_{j-1} < y_i \leq s_j) \left\{ \lambda_j(y_i - s_{j-1}) + \sum_{l=1}^{j-1} \lambda_l(s_l - s_{l-1}) \right\} \exp(x_i \beta) \right].$$

(4)

The likelihood in (4) assumes that the ratio of hazards for two individuals is constant over time, and thus represents a proportional hazards model.

4.2 Commensurate prior model

In this subsection we formulate commensurate prior distributions [4] to model the scenario presented by the motivating successive colon cancer trials, where historical data are available only for the control group. Let $y_0$ and $c_0$ denote vectors of length $n_0$ containing time-to-event variables and right-censoring indicators for patients assigned to the current control therapy in the historical study. Similarly, let $y$ and $c$ denote vectors containing time-to-event variables and right-censoring indicators for $n$ patients in the current trial. Furthermore, let $d_i$ denote a 0-1 scalar indicator of the novel treatment for the $i$th patient in the current trial, $i = 1, ..., n$. In this context, the hazard for a historical patient in the $j$th time partition interval is identical to the $j$th historical baseline hazard, $h(\lambda_{0,j}) = \lambda_{0,j}$. For the $i$th patient in the current trial, the $j$th baseline hazard, $\lambda_j$, is modified by novel treatment status, $d_i$: $h(\lambda_j, d_i, \xi) = \lambda_j \exp(d_i \xi)$, where $\xi$ is the log acceleration factor corresponding to the novel treatment. Both historical and current baseline hazard parameters
are defined with respect to a single partition of the time axis characterized by boundary points

\[ 0 < s_1 < s_2 < ... < s_J < \infty. \]

To borrow strength from the historical controls, the commensurate prior approach can be applied by assuming a normal prior for the \( j \)th log baseline hazard, \( \log(\lambda_j) \), centered at its historical counterpart, \( \log(\lambda_{0,j}) \), with precision \( \tau_j \): \( \log(\lambda_j) \sim N\{\log(\lambda_{0,j}), 1/\tau_j\} \). Given no additional prior information about the historical baseline hazards, inference may proceed with a non-informative prior distribution, \( p\{\log(\lambda_{0,j})\} \), for each of the \( \log(\lambda_{0,j}) \). In addition, we assume a non-informative Gaussian prior distribution, \( p(\xi) \), for the novel treatment effect parameter, \( \xi \). Let \( D \) and \( D_0 \) denote the current and historical data: \( D = (y, d, c) \) and \( D_0 = (y_0, c_0) \), respectively. Let \( \theta \) denote the parameter vector \( \theta = \{\xi, \log(\lambda_1), ..., \log(\lambda_J)\} \), let \( \theta_0 \) denote the parameter vector \( \theta_0 = \{\log(\lambda_{0,1}), ..., \log(\lambda_{0,J})\} \), and let \( \tau \) denote vector \( \tau = (\tau_1, ..., \tau_J) \). Following from (4), the joint posterior distribution of \( \theta | \tau \) is proportional to

\[
q(\theta | \tau, D, D_0) \propto p(\xi) \prod_{i=1}^n L(y_i | \theta) \int_{\theta_0} \prod_{j=1}^J N\{\log(\lambda_j) \mid \log(\lambda_{0,j}), 1/\tau_j\} p\{\log(\lambda_{0,j})\} \prod_{k} L(y_{0,k} | \theta_0) d\theta_0
\]

Following Hobbs et al. [4], we assume “spike and slab” prior distributions [19] for the \( \tau_j \)s. This distribution is locally uniform between two limits, \( 0 \leq S_l < S_u \) except for a bit of probability mass concentrated at a point \( K > S_u \):

\[
Pr(\tau_j < u) = p_0 \{ (u - S_l) / (S_u - S_l) \}, \ S_l \leq u \leq S_u \ \text{and} \ \ Pr(\tau_j > S_u) = 1 - p_0.
\]

The additional hyperparameter, \( p_0 \), denotes the prior probability that \( S_l \leq \tau_j \leq S_u \). The aforementioned authors [4] show that this weakly informative prior, when properly calibrated, leads to desirable bias-variance trade-offs in the context of commensurate prior models for exponential families. The results of our simulation study discussed in Section 6 illustrate the advantages of the
spike and slab commensurate prior model in this context. Markov chain Monte Carlo (MCMC) methods can be used to sample from the joint posterior (5).

5 Adaptive trial design

In this section we formulate our proposed alternative design of the Goldberg trial following completion of the Saltz trial, which includes both non-adaptive and adaptive phases. During the initial, non-adaptive phase, new patients are randomized equally between the control (IFL) and novel (FOLFOX) treatments until a targeted number of events selected to ensure that sufficient information has accrued to assess heterogeneity among the concurrent and historical controls. Thereafter, the treatment allocation probability is adjusted at two interim analyses as a function of the effective number of historical controls, and the numbers of observed events for each regimen. We use the piecewise exponential commensurate prior model (5) to jointly model the concurrent and historical data at each interim analysis. EHC follows from EHSS (3) defined in the context of the commensurate prior model. Given large EHC, the adaptive design will randomize more new patients to the newer, less studied therapy.

5.1 Historical data

We use likelihood inference based on (4) to select the time axis partition. The Akaike information criterion optimal partition contains two intervals with boundary point at \( s = 243 \) days. Thus, our piecewise exponential model contains two parameters characterizing baseline hazards within intervals \([0, 243)\) and \([243, \max(Y_0)]\), respectively. Figure 1 contains the associated TTP curves. The dashed line represents the Kaplan-Meier curve, while dotted lines correspond to associated 95% log-transformed [20, see e.g. p.105] pointwise confidence intervals. Results for the piecewise exponential analysis are plotted with a blue, solid line. Posterior summaries corresponding to the
piecewise exponential analysis are provided in Table 1. The posterior mean of median survival is approximately 205 days.

Table 1: Posterior means and standard deviations derived from the piecewise exponential analysis of time-to-disease progression for patients randomized to IFL in the Saltz trial, $n_0 = 224$.

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>log baseline hazard 1</td>
<td>-5.689</td>
<td>0.094</td>
</tr>
<tr>
<td>log baseline hazard 2</td>
<td>-4.858</td>
<td>0.133</td>
</tr>
</tbody>
</table>

5.2 New trial

The Goldberg trial actually enrolled a total of $n = 419$ patients with measurable tumors assigned to either IFL (control) or FOLFOX (novel treatment) regimens. Figure 2 plots actual enrollment over time. Each enrollment is characterized by a dot representing the trial’s cumulative sample size
(y-axis) in calendar time (x-axis). The first patient was enrolled on May 20, 1999, the last patient was enroll on April 25, 2001. We use the Goldberg et al. [2] article’s submission date as the time of final analysis, which occurred approximately 28 months after close of enrollment.

Our proposed trial consists of two phases. During the first, non-adaptive phase, patients are randomized 1:1 to treatment regimens until $O_1$ events are observed. An initial interim analysis to assess historical and concurrent control heterogeneity and compute EHC will occur at that time. An adaptive randomization procedure (a function of EHC) will be used to assign new patients to treatment arms thereafter. EHC will be re-assessed at a second interim analysis after a total of $O_2$ events, with the final analysis to occur 51 months after the trial’s initiation. $O_1$ needs to be large enough so that sufficient information to assess control heterogeneity has accrued, yet small enough to justify the adaptive treatment allocation. In practice, $O_1$ and $O_2$ should be selected in the context of a comprehensive assessment of the trial’s operating characteristics in the presence of the historical data and expected patient enrollment.

5.2.1 Data structure

Let $T = 1552$ denote the time of the trial’s final analysis in days. Let $t$ denote “trial time,” $0 < t < T$, and let $e_i$, $0 < e_i < 706$, represent the time at which the $i$th patient enrolls. Let $Y_i(t)$ denote the value of the $i$th patient’s TTP process at time $t$. Note that $Y_i(t)$ is recorded from $e_i$, so that $0 < Y_i(t) < T - e_i$. Let $Y(t) = \{Y_1(t), ..., Y_n(t)\}$ denote the vector of values of the TTP processes for the $n(t)$ patients enrolled prior to time $t$, $n(t) = \sum_{i=1}^{n} I(e_i < t)$, where $I$ is the indicator function. Let $c_i(t)$ denote an 0-1 indicator of the $i$th patient’s right-censoring status at time $t$, and let $c(t) = \{c_1(t), ..., c_n(t)\}$ denote the vector of values of censoring processes for the $n(t)$ patients enrolled prior to time $t$. Finally, let $d_i(t)$ denote an 0-1 treatment indicator for the $i$th patient, with 0 indicating IFL and 1 FOLFOX.
Figure 2: Enrollment over time for patients randomized to the FOLFOX or IFL regimens in the Goldberg trial. The y-axis represents cumulative sample size. The x-axis represents calendar time.

5.2.2 Probability models

The time axis partition for our piecewise exponential model contains two baseline hazard parameters, $\lambda_1$ and $\lambda_2$, corresponding to intervals $[0, 243)$ and $[243, T]$ (based on the AIC-optimal partition for the historical data). The reference model consists of the Bayesian piecewise exponential model corresponding to (5) that ignores the historical data. Following the notation of Section 4, let $\xi$ denote the FOLFOX effect, and denote the current and historical parameter vectors by $\theta$ and $\theta_0$, where $\theta = \{\xi, \log(\lambda_1), \log(\lambda_2)\}$, and $\theta_0 = \{\log(\lambda_{0,1}), \log(\lambda_{0,2})\}$, respectively. Denote concurrent data at time $t$ by $D(t) = \{Y(t), c(t), d(t)\}$ and historical data by $D_0 = (Y_0, c_0)$.

At the two interim analyses the joint commensurate prior model is formulated to borrow strength from the historical data for estimating only the initial baseline hazard, $\lambda_1$, due to the expected paucity of events observed within second interval. Thus a non-informative Gaussian prior distribution (with small precision $w = 0.0001$) is assumed for $\log(\lambda_2)$. In addition, non-informative
Gaussian prior distributions are assumed for the FOLFOX effect, $\xi$, and the historical log baseline hazards, $\log(\lambda_{0,1})$ and $\log(\lambda_{0,2})$. Following from (4), (5), and (6), the posterior distribution of $\theta | D(t), D_0$ for interim analysis under the joint model at trial time $t$ follows as

$$q\left\{ \theta \mid D(t), D_0 \right\} \propto N(\xi \mid 0, 1/w) N \left\{ \log(\lambda_2) \mid 0, 1/w \right\} \prod_{i=1}^{n(t)} L \left\{ Y_i(t) \mid \theta \right\} \times \int_{\tau_1} \int_{\theta_0} N \left\{ \log(\lambda_1) \mid \log(\lambda_{0,1}), 1/\tau_1 \right\} \prod_{j=1}^{2} N \left\{ \log(\lambda_{0,j}) \mid 0, 1/w \right\} p(\tau_1) \prod_{k} L(Y_{0,k} | \theta_0) d\theta_0 d\tau_1. \quad (7)$$

The posterior distribution of $\theta | D(t)$ for interim analysis at time $t$ under the reference model is

$$q^* \left\{ \theta \mid D(t) \right\} \propto N(\xi \mid 0, 1/w) \prod_{j=1}^{2} N \left\{ \log(\lambda_j) \mid 0, 1/w \right\} \prod_{i=1}^{n(t)} L \left\{ Y_i(t) \mid \theta \right\}. \quad (8)$$

At the time of the final analysis, $T$ (51 months after initiation), a sufficient number of events will have occurred within both intervals of the time axis partition. Therefore, the joint model used in the final analysis will borrow strength from the historical data for estimating both baseline hazards. Let $\tau = (\tau_1, \tau_2)$ denote a vector of the commensurability hyperparameters. The posterior distribution of $\theta | D(T), D_0$ follows by modifying (7) to

$$q \left\{ \theta \mid D(T), D_0 \right\} \propto N(\xi \mid 0, 1/w) \prod_{i=1}^{n} L \left\{ Y_i(T) \mid \theta \right\} \times \int_{\tau} \int_{\theta_0} \prod_{j=1}^{2} N \left\{ \log(\lambda_j) \mid \log(\lambda_{0,j}), 1/\tau_j \right\} N \left\{ \log(\lambda_{0,j}) \mid 0, 1/w \right\} p(\tau) \prod_{k} L(Y_{0,k} | \theta_0) d\theta_0 d\tau. \quad (9)$$

### 5.2.3 Effective number of historical controls and adaptive randomization

We compute the effective number of historical controls at each interim analysis, with use thereafter to guide the randomization procedure. In Appendix B we investigate the linear association between number of events and posterior precision for $\log(\lambda_1) | D(t)$ in the context of the refer-
ence model (8) in the presence of the historical data. Following from EHSS (3), EHC in this context maps the joint model’s relative gain in precision for estimating the initial baseline hazard, $\lambda_1$, to a number of additional events observed for control. The $h$th interim assessment of EHC will occur at the trial time $t_h = \arg\min_t \left[ \sum_{i=1}^{n(t)}\{1 - c_i(t)\} = O_h \right]$, where $h = 1, 2$. Let $O^{IFL}_j(t)$ denote the number of events observed among concurrent controls (patients assigned to the IFL regimen in the current trial) in the $j$th partition interval, $j = 1, 2$, at trial time $t$, e.g., $O^{IFL}_1(t) = \sum_{i=1}^{n(t)}(1 - d_i)\{1 - c_i(t)\} I \{Y_i(t) < 243\}$. Following Section 3, let $P\{\mathcal{D}(t_h)\}$ denote the posterior precision of $\log(\lambda_1) | \mathcal{D}(t_h)$ at the $h$th interim analysis under the reference model (8). Similarly, let $P\{\mathcal{D}(t_h), \mathcal{D}_0\}$ denote the posterior precision of $\log(\lambda_1) | \mathcal{D}(t_h), \mathcal{D}_0$ at the $h$th interim analysis under the joint model (7). The effective number of historical control events at the $h$th interim analysis follows from (3) as

$$EHC_h = O^{IFL}_1(t_h) \left[ \frac{P\{\mathcal{D}(t_h), \mathcal{D}_0\}}{P^*\{\mathcal{D}(t_h)\}} - 1 \right].$$

Patients will be allocated equally among control (IFL) and novel treatment (FOLFOX) until $O_1$ events are observed. After the $h$th interim analysis, we seek to balance total information among the treatment regimens. Following the previous subsection let $O^{FOX}_j(t)$ denote the number of events observed among patients assigned to FOLFOX in the $j$th partition interval, $j = 1, 2$, at trial time $t$. In addition, let $O^{FOX}(t)$ and $O^{IFL}(t)$ denote the total number of events observed for patients assigned to FOLFOX and IFL regimens at time $t$, respectively: $O^{FOX}(t) = \sum_{j=1}^{2} O^{FOX}_j(t)$ and $O^{IFL}(t) = \sum_{j=1}^{2} O^{IFL}_j(t)$. Following the $h$th interim analysis, the probability that the $(i + 1)$st patient is randomized to novel treatment (FOLFOX) is defined as the proportion of total
information (observed and effective number of events) accrued for the control (IFL) regimen:

\[
\pi_{i+1} = \frac{O^{IFL}(e_{i+1}) + EHC_h}{O^{FOX}(e_{i+1}) + O^{IFL}(e_{i+1}) + EHC_h}.
\] (11)

6 Simulations

We used simulations to evaluate the frequentist properties of the proposed alternative, adaptive design of the Goldberg trial. The simulation uses the actual patient-level data from \( n_0 = 224 \) historical controls assigned to IFL in the Saltz trial. Patient accrual in the simulated trials follows the actual enrollment dates (shown in Figure 2) for \( n = 419 \) patients randomized to FOLFOX or IFL during the Goldberg trial. Frequentist properties are considered under varying magnitudes of historical bias characterized by median time-to-progression.

Following Subsection 5.2, each simulated trial proceeds as follows. Patients are randomized 1:1 to IFL and FOLFOX until \( O_1 = 80 \) total events are observed at time \( t_1 \). Thereafter, the probability of assigning a new patient to FOLFOX (11) adjusts as a function of EHC and numbers of observed concurrent events. EHC is updated at a second interim analysis after a total of \( O_2 = 120 \) events are observed at time \( t_2 \). At the \( h \)th interim analysis, \( h = 1, 2 \), \( EHC_h \) is computed from (10) using reference and joint posterior distributions of the initial log baseline hazard, \( \log(\lambda_1) | \mathcal{D}(t_h) \) and \( \log(\lambda_1) | \mathcal{D}(t_h), \mathcal{D}_0 \), derived from models (8) and (7), respectively. A final analysis occurs at time \( T \), 51 months following initial enrollment using the joint model in (9). The joint commensurate prior models (7) and (9) assume spike and slab priors (6) for \( \tau_j \), \( j = 1, 2 \), with the following hyperparameters recommended by [4]: \( S_l = 0.005, S_u = 2, K = 200, \) and \( p_0 = 0.6 \).

Posterior inference on the FOLFOX treatment effect parameter, \( \xi \), is of primary interest. Results are compared to fixed 1:1 allocation designs with analyses under the associated reference ("no borrowing") and "pooled" models, the latter of which simply pools the historical and current
controls. We let \( \theta^{tr} = \{\xi^{tr}, \log(\lambda_1^{tr}), \log(\lambda_2^{tr})\} \) denote a set of fixed parameters characterizing a true state of the model. Progression times for current patients are generated from \( \theta^{tr} \) under various adjustments of the baseline hazards from their historical estimates in Table 1. Without out loss of generality, we fix \( \xi^{tr} = 0 \). In addition, we fix \( \lambda_2^{tr} = \exp(-4.858) \) and set \( \lambda_1^{tr} = \exp(-5.689 + \Delta^{tr}) \), where \( \Delta^{tr} \) parameterizes historical bias on the log baseline hazard scale for estimating \( \lambda_1 \). Historical bias for estimating \( \xi \) in this context is more naturally interpretable on the median survival scale. Let \( \Lambda^{tr}(\theta^{tr}) \) denote median TTP, i.e., \( \int_0^{\Lambda^{tr}(\theta^{tr})} L(y|\theta^{tr}) = 0.5 \) from (4). Recall that inference for the historical data in Subsection 5.1 estimated median TTP for \( \mathcal{D}_0 \) to be nearly 205.

The preposterior risk under squared error loss [21, p. 433], at the final analysis of the adaptive
trial conditional on the true parameters, $\theta^{tr}$, and the observed historical data, $D_0$, follows as

$$E_{D(T)|\theta^{tr}, D_0} \left[ \left( E_{\xi|D,D_0}(\xi) - \xi^{tr} \right)^2 \right]. \quad (12)$$

Figure 3 presents preposterior risk under SEL and bias for estimating $\xi$ as functions of true median TTP, $\Lambda^{tr}(\theta^{tr})$ (x-axis). Results are shown for inference under no borrowing, pooling, and final analysis of our proposed adaptive trial. A total of 1500 replicated trials were simulated for each of 11 scenarios for true values of $\Delta^{tr}$ inducing values $\Lambda^{tr}(\theta^{tr})$ ranging from 100 to 300. The historical controls are truly unbiased for $\Delta^{tr} = 0$ which yields $\Lambda^{tr}(\theta^{tr}) \approx 205$ on the scale of median TTP.

Under no borrowing, $D_0$ is ignored. Thus, the posterior mean of $\xi|D(T)$ is unbiased and its preposterior risk is constant as a function of $\Lambda^{tr}(\theta^{tr})$. Pooling the historical and concurrent controls offers maximal variance reduction, and thus is associated with the largest reductions in preposterior risk for $\Lambda^{tr}(\theta^{tr})$ near 205. Therefore, the pooled estimator is preposterior admissible. However, pooling leads to prohibitively biased estimators with sharply, monotonically increasing preposterior risk for $\Delta^{tr} > 0$. The adaptive trial offers a preposterior admissible estimator of $\xi$, that approaches the gains in preposterior risk obtained by pooling when $\Delta^{tr}$ near zero, yet minimizes bias for large values of $|\Delta^{tr}|$. Overall, the simulation suggests that the adaptive trial facilitates an estimator that dominates that of no borrowing for preposterior risk under SEL, while also providing a much more desirable bias-variance trade-off than naive pooling.

Our simulation study also evaluates EHC at both interim analyses and total number of patients (out of $n = 419$) assigned to FOLFOX in our proposed adaptive version of the Goldberg trial as functions of $\Lambda^{tr}(\theta^{tr})$. The simulation results strongly suggest that the adaptive trial facilitates efficient use of the historical controls. As shown in Figure 4, truly minimally biased historical controls are associated with increases in EHC at the interim analyses facilitated by the joint commensurate
prior model. Thus, when the historical information is minimally biased (i.e., median TTP close to 205), the proposed design properly adjusts for the relative information imbalance (favoring IFL) by assigning more new patients to the novel FOLFOX therapy. Biased historical controls induce negligible interim EHC and thus, on average results in enrolling patients 1:1 to the two study arms.

7 Discussion

In this article we proposed an adaptive trial design that implements a randomization procedure for allocating patients aimed at balancing total information (concurrent and historical) among the
study arms by assigning more patients to receive the novel therapy in the absence of evidence for heterogeneity among the concurrent and historical controls. We have only considered computing EHC in the context of a single concurrent control parameter. In the case of multiple parameters, we might define EHC as a function of the relative gain in the sum of precisions.

Adaptive randomization methods are sensitive to population drift in general, and more suitable for trials that initiate with gradual enrollment. Furthermore, the success of clinical investigation requires that trials are properly designed and meticulously implemented such that data can be promptly updated to allow interim statistical monitoring. Our adaptive design was shown to produce desirable results without continuous monitoring, which is challenging in practice.

The proposed design promises to enhance efficiency in the evaluation of new therapies involving controlled clinical trials that follow recent evaluation of the control therapy. The proposed adaptive trial was considered in a “non-sequential” setting involving a single treatment comparison at the end of the trial. That is, the design, while sequential in assessing evidence for heterogeneity among the historical and concurrent controls, did not incorporate decision rules for early stopping for efficacy or futility. Extending the proposed adaptive trial to allow for earlier stopping is an area for potential future development. Indeed, other areas for future development are many, since they include any setting where we wish to adjust trial enrollment adaptively based on an updated estimate of how much strength may be sensibly borrowed from external data sources. For instance, when historical information is also available on the new treatment, we may have two commensurability calculations, each having their own impact on the randomization ratio (11). Alternatively, in a one-arm device safety trial we may use commensurate priors to decide when to stop accrual based on the predictive probability of a favorable end result given the current estimate of the total effective sample size.
References


Here we briefly illustrate the concept of effective historical sample size for Gaussian data in the context of a conventional random-effects meta-analytic joint model. Again, let \( y = \{y_1, \ldots, y_n\} \) denote a vector of i.i.d. responses of length \( n \) for patients enrolled in the current trial such that \( y_i \sim N(\mu, \sigma^2), \ i = 1, \ldots, n \). Assuming known variance, \( \sigma^2 \), under a flat non-informative prior for \( \mu \), the reference posterior for \( \mu | y \) follows as \( N(\mu | \bar{y}, \sigma^2/n) \), where \( \bar{y} \) denotes the observed sample mean. Suppose we have patient-level data for patients assigned to the control therapy from \( H \) historical trials. Let \( y_{0,h} \) denote response vectors of length \( n_{0,h} \) for the \( h \)th historical trial, \( y_{0,h} \overset{iid}{\sim} N(\mu_{0,h}, \sigma^2_{0,h}) \), where \( h = 1, \ldots, H \). The conventional random-effects meta-analytic approach for borrowing strength from the historical data [see e.g. 22, p.268] assumes that \( \mu_{0,1}, \ldots, \mu_{0,H} \), and \( \mu \) are exchangeable: \( \mu_{0,1}, \ldots, \mu_{0,H}, \mu \overset{iid}{\sim} N(\xi, \eta^2) \). The model allows for both between-study
heterogeneity and within-study variability. Parameters $\xi$ and $\eta^2$ characterize the population mean and between-study variance, respectively. Assume a flat prior for $\xi$ and let $v$ and $v_{0,h}$ denote $v = \sigma^2/n$ and $v_{0,h} = \sigma_{0,h}^2/n_{0,h}$, respectively. Conditional on known variance components, the corresponding posterior distribution under the joint modeling of the concurrent and historical data follows as $q(\mu|y_0, y, \eta^2, v, v_{0,1}, \ldots, v_{0,H}) \propto N(VM, V)$, where

$$V = \left\{ \frac{\sum_{h=1}^H (\eta^2 + v_{0,h})^{-1} + 1}{1 + \eta^2 \sum_{h=1}^H (\eta^2 + v_{0,h})^{-1}} \right\}^{-1}, \quad \text{and} \quad M = \frac{\sum_{h=1}^H \bar{y}_{0,h} / (\eta^2 + v_{0,h})}{1 + \eta^2 \sum_{h=1}^H (\eta^2 + v_{0,h})^{-1}} + \frac{\bar{y}}{v}. \quad (13)$$

Therefore, $P^*(y) = 1/v, P(y_0, y(t), \eta^2, v, v_{0,1}, \ldots, v_{0,H}) = V^{-1}$, and thus from (3) EHSS is constant as a function of the current sample size $n : EHSS = \frac{\sigma^2 \sum_{h=1}^H (\eta^2 + v_{0,h})^{-1}}{1 + \eta^2 \sum_{h=1}^H (\eta^2 + v_{0,h})^{-1}}$.

**Appendix B**

Here we investigate the association between number of events and posterior precision for $\log(\lambda_1) | \mathcal{D}(t)$ in the context of the reference model (8). Specifically, we simulated 7,000 datasets from a piecewise exponential probability distribution. For each simulated dataset, $\mathcal{D}$, the sample size, $\tilde{n}$, and censored probability, $\pi^{cens}$, were randomly generated from uniform distributions covering broad ranges of the sample spaces of relevance: $\tilde{n} \sim Uniform(50, 600)$ and $\pi^{cens} \sim Uniform(0.1, 0.8)$. Time-to-event observations were generated for each of the $\tilde{n}$ patients from $\theta^{tr}$ containing baseline hazards fixed at their corresponding historical estimates in Table 1, $\log(\lambda_1^{tr}) = -5.689$ and $\log(\lambda_2^{tr}) = -4.858$, and treatment effect $\xi^{tr} = 0$. For each dataset we recorded the number of events observed in the initial partition interval, and fit the reference model (8) to obtain the associated posterior precision of $\log(\lambda_1) | \mathcal{D}$. The results in Figure 5 reveal a strong linear association (Pearson correlation of 0.996) among posterior precision and number of observed events for the considered portion of the parameter space.
Figure 5: Scatterplot of simulation results. Each dot depicts the number of events (y-axis) as a function of posterior precision of $\log(\lambda_1)|D(t)$ (x-axis) under the reference model (8).