Borrowing from Historical Control Data in Cancer Drug Development: A Cautionary Tale and Practical Guidelines

Connor Jo Lewis*
Securian Financial Group, Inc., St. Paul, MN, USA
and
Somnath Sarkar
Flatiron, Inc., New York, USA
and
Jiawen Zhu
Roche-Genentech, New York, USA
and
Bradley P. Carlin
Counterpoint Statistical Consulting, Edina, MN, USA

May 22, 2018

Abstract

Some clinical trialists, especially those working in rare or pediatric disease, have suggested borrowing information from similar but already-completed clinical trials. This paper begins with a case study in which relying solely on historical control information would have erroneously resulted in concluding a significant treatment effect. We then attempt to catalog situations where borrowing historical information may or may not be advisable using a series of carefully designed simulation studies. We use an MCMC-driven Bayesian hierarchical parametric survival modeling approach to analyze data from a sponsor's colorectal cancer study. We also apply these same models to simulated data comparing the effective historical sample size, bias, 95% credible interval widths, and empirical coverage probabilities across the simulated

*This work reported in this article formed the basis for the first author's University of Minnesota Master's project paper, which was supervised by the last author. The authors are grateful to Dr. David Dejardin for helpful comments that significantly improved the manuscript. The work of the first and last authors was supported in part by a grant from Roche-Genentech. In addition, the work of the last author was partially supported by National Cancer Institute grant 1-R01-CA157458-01A1.

cases. We find that even after accounting for variations in study design, baseline characteristics, and standard-of-care improvement, our approach consistently identifies Bayesianly significant differences between the historical and concurrent controls under a range of priors on the degree of historical data borrowing. Our simulation studies are far from exhaustive, but inform the design of future trials. When the historical and current controls are not dissimilar, Bayesian methods can still moderate borrowing to a more appropriate level by adjusting for important covariates and adopting sensible priors.

Keywords: Bayesian analysis; Commensurate prior; Effective historical sample size; Hierarchical model.

1 Introduction

Borrowing information from previously completed trials is used extensively in medical device trials (U.S. Department of Health and Human Services, 2010), and is increasingly seen in drug trials in pediatrics (Gamalo-Siebers et al., 2017) and oncology (Gökbuget et al., 2016a,b). However, the practice of using historical information still faces resistance in clinical trials practice, particularly in Phase III. In general, trial sponsors and enrollees benefit in multiple ways from using Bayesian methods to borrow information from historical controls. The reductions in sample size, time, expense, and increased statistical power from borrowing from sufficiently similar data are obvious, but also important is the underlying ethical implications of reducing the number of participants assigned to the concurrent control arm, especially in late-stage cancer trials (U.S. Department of Health and Human Services, 2010; Viele et al., 2014). One difficulty facing researchers hoping to design a trial using Bayesian borrowing methods is understanding the *commensurability* of the auxiliary data and the trial data yet to be collected. How were the previous trials designed? Were similar patient populations utilized? Were the dosages and treatment schedules of the current drug the same? Has the standard of care evolved? If the historical information differs substantially from the concurrent, Bayesian methods may borrow strength from a biased source, leading to an inflated Type I error rate, as well as the possibility of needing to run a longer, more expensive trial in order to overcome the incommensurate prior data. Other obstacles in designing trials in a Bayesian framework include a lack of user-friendly Bayesian software and Bayesian-trained statisticians, but the landscape in these areas is changing quickly.

Deciding to use Bayesian methods for a late-phase clinical trial is sometimes difficult for a pharmaceutical company, because not only must the firm undertake the more demanding design calculations, it must also cope with regulatory guidance for Bayesian drug trials that is not yet well-developed. In the current environment, the Food and Drug Administration (FDA) has outlined the use of Bayesian methods for medical device trials, and is increasingly open to discussing such novel design approaches with biostatisticians in industry and academics. However, regulatory authorities are often concerned that the introduction of outside information could inflate the trial's Type I error rate or lead to bias in the estimated treatment effect (Pennello and Thompson, 2008), especially in late-phase (confirmatory) drug trials (Berry, 2006). In addition to the analysis issues accompanying the use of auxiliary data, regulators worry that some researchers might "cherry-pick" the historical data, including only information favorable to the company's interests. Quan et al. (2017) offer a recent practically-focused review the places the borrowing of strength from historical data in the broader context of integrated data analysis, with particular emphasis on the use of network meta-analysis and empirical Bayes to estimate the effects of various treatments.

In this paper, we begin with a "cautionary tale" wherein naive aggressive borrowing from historical controls turns out *not* to be advisable. We then go on to provide assistance in specifying hierarchical models that perform well regardless of whether auxiliary data borrowing is or is not advisable, creating modeling guidelines for companies and regulatory boards alike. Section 2 introduces our motivating data set, a real-world clinical trial setting from the field of metastatic colorectal cancer. Our Bayesian hierarchical parametric survival model, explanatory variables, prior distributions, and approach for calculating the effective historical sample size are outlined in Section 3. Next, the unfortunate results of applying overzealous borrowing methods to the motivating data set are presented in Section 4. In Section 5, we perform a simulation study to provide better guidelines for researchers as to when borrowing information from previously completed trials is advisable, comparing results using different commensurate priors for the group effect of the concurrent controls to those using standard vague priors. Credible interval widths, bias, empirical coverage percentages, and effective historical sample sizes are compared across scenarios favorable and unfavorable to borrowing. Finally, Section 6 summarizes our findings and suggests areas for future development in this area.

2 Description of Motivating Data Set

Our motivating real-world example uses sponsor-supplied individual patient-level data from three metastatic colorectal cancer (mCRC) trials run by Roche/Genentech. The historical control data were taken from two previously completed trials, named CRC1 and CRC2 in this article. A more recently collected data set, from a randomized trial named CRC3 in this article, compares the survival times using either duration of response (DoR) or progression-

	CRC1 (HC)	CRC2 (HC)	CRC3 (CC)	CRC3 (Drug A)
Count	350	62	62	63
ORR (%)	47.0	47.0	64.0	58.7
DoR: Median (mo)	8.3	9.9	11.1	10.8
PFS: Median (mo)	9.5	9.9	12.8	13.1

Table 1: Total counts, overall response rates (ORRs), and both DoR and PFS median survival times for each group of interest: HC-historical controls; CC-concurrent controls; and Drug A, the novel treatment of interest.

free survival (PFS) of patients receiving an additional Drug A compared to those receiving the current standard-of-care treatment, the concurrent controls. All patients in each of the trials have data when using PFS as the survival time, but to be considered for DoR, one first must achieve an overall response; thus, the DoR data are a subset of the PFS data. Table 1 shows summary statistics for the trials across the four treatment arms.

The median survival times for the historical control trials are substantially lower than those in both CRC3 arms. Moreover, the CRC3 concurrent controls and additional Drug A group median survival times do not seem to differ materially. The resulting conundrum is this: if the CRC3 Drug A survival times are compared to the historical controls alone, a significant treatment effect might emerge; however, within the context of the CRC3 trial, we see no evidence of a treatment effect. The sponsor was interested in evaluating in this small sample size setting of CRC3, when the truth is unknown, what benefit in precision can be achieved (without unacceptable increases in bias) by borrowing from similar, previously conducted trials using hierarchical Bayesian methods.

Using Bayesian survival modeling, we seek to compare group effects under the following two data groupings:

- A naive 2-arm approach that combines all controls and compares them with CRC3 patients receiving the additional treatment of Drug A (HC & CC vs. Drug A).

- A 3-arm approach where the group effects are compared between each combination of the historical controls, the concurrent controls, and the Drug A treatment arm of CRC3 (HC vs. CC vs. Drug A).

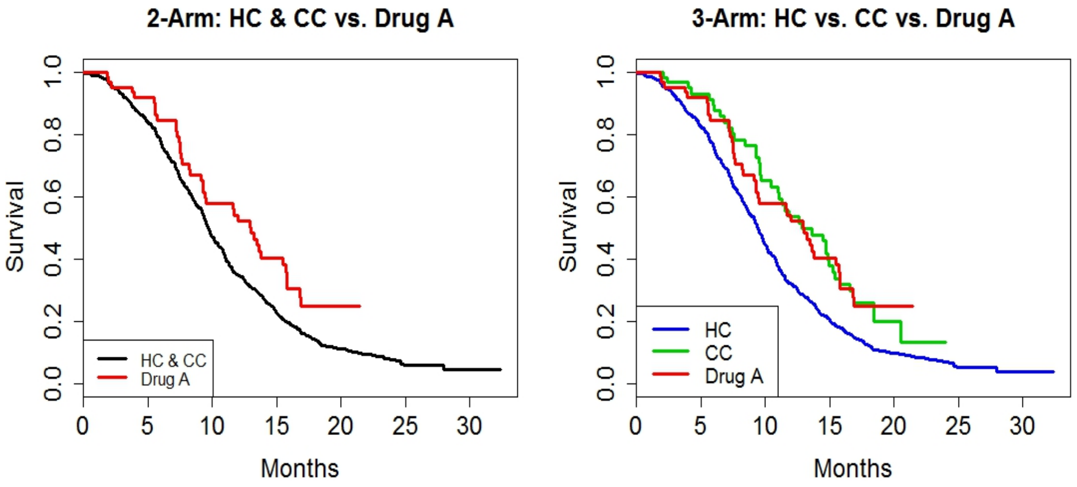

Figure 1: PFS: Kaplan-Meier plots for the naively-combined controls vs treatment (left) and the three separate treatment groups (right).

The smoothness and general shape of the Kaplan-Meier curves in Figure 1 supports the modeling assumption that the survival times follow a Weibull distribution. The Kaplan-Meier plots reiterate the problem faced in the analysis of the CRC3 data: the historical information in this example differs from that in the control arm of the CRC3 trial, but pooling all the control data masks this difference. The 3-arm plot shows that the survival curve for the historical controls is below the curve of both arms of the CRC3 trial, whereas the CRC3 trial curves are intertwined throughout the plot.

To check the robustness of our results and determine if we can properly explain the increase in survival time between the historical controls and CRC3 patients, analyses which include additional covariates are also of interest. The main additional covariate of interest is each patient's duration on oxaliplatin (OXA), a common chemotherapeutic agent. Figure 2 shows the distribution of OXA duration for each of the three groups. For historical controls, the distribution is more widely spread out and contains higher values than either the CRC3 group's distributions. In particular, the distribution of OXA duration for the historical control group is centered near eight months with a maximum of 15 months, whereas the range for both CRC3 groups is zero to five months with a median near four months (a reduction in OXA duration was mandated by the CRC3 protocol). Based on our current

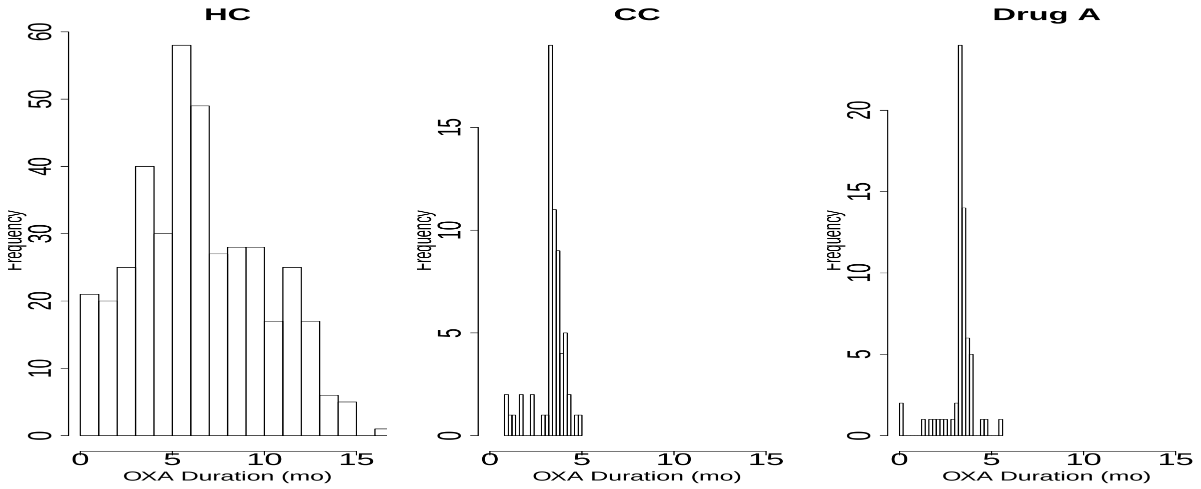

Figure 2: Histograms of the distribution of OXA duration for each treatment group using the PFS data.

understanding of chemotherapy, increased length (or intensity) of OXA treatment, as seen in the historical controls, would suggest longer survival times. However, as seen in Table 1 and Figure 1, the historical controls actually have *lower* survival times than both the concurrent controls and the Drug A recipients. This is another sign that the standard of care for these patients may have evolved since the historical studies were done.

To attempt to adjust for this evolution, a time component was also considered in our models. To do so, the added variable is time (months) since treatment began, counting back from the end of the 2016 calendar year. Note this differs from OXA duration, which only accounts for the length of time OXA was administered; the time variable instead refers to how long ago the patient initially started OXA. Additional demographic variables, namely, age, sex, and ECOG status (a baseline assessment of disease progression), are also potentially sensible additions to any statistical model.

3 Methods

An MCMC-driven Bayesian hierarchical parametric survival modeling approach (similar to that described in the "mice" example from the BUGS manual) was used to model these data (Lunn et al., 2000). For the most basic group comparisons (excluding, for the moment, the covariates referenced in Section 2), let $\boldsymbol{\beta}$ be a coefficient vector having one element for each comparison group, and $\mathbf{z_i}$ be the corresponding covariate vector containing indicator variables for each respective group for subject i. Let t_i be the time-to-event for subject i, a survival time assumed to follow a Weibull distribution with shape r and scale μ_i, i.e. having density function $f(t_i, \boldsymbol{z}_i) = re^{\boldsymbol{z}'_i\boldsymbol{\beta}}t_i^{r-1}exp(-e^{\boldsymbol{z}'_i\boldsymbol{\beta}}t_i^r)$. By setting $\mu_i = e^{\boldsymbol{z}'_i\boldsymbol{\beta}}$, the parameterization is $t_i \sim \text{Weibull}(r, \mu_i)$, $i = 1, \ldots, n$.

Censored observations instead follow a truncated Weibull distribution. All regression coefficients were assumed to have independent vague normal priors. For the shape parameter r, we used an $Exp(1)$ prior, which suggests values between 0 and 6, a range more than wide enough to contain the values that are plausible for our data. Our BUGS implementation used three parallel MCMC chains, each run for a 1,000-iteration burn-in period followed by a 20,000-iteration production run. This “base model” will be used in our data analysis and each of the data simulation cases in Section 5.

In our covariate-adjusted model, the continuous variable OXA duration, denoted by x_i, was added. Letting γ be a scalar representing the effect of OXA duration on survival times across the groups, the model's mean structure becomes $\mu_i = e^{\boldsymbol{z}'_i\boldsymbol{\beta}+x_i\gamma}$. We ran an additional model which allowed for the OXA duration effect to vary by group (i.e., an interaction effect). Though many hierarchical model choice criteria are now available, we used the Deviance Information Criterion (DIC) as implemented in BUGS to compare the models with and without this interaction (and indeed all the models throughout this paper). Models with smaller DIC values are preferred (Carlin and Louis, 2008). Here, the DIC of the model including the interaction effect (2653) did not differ meaningfully from that of the single γ model (2650). Thus, using a single γ to account for the OXA duration effect across all groups appears reasonable.

In hopes of better accounting for the improvement in standard-of-care, we briefly explored a model including both OXA duration and the time component. Finally, a “full model” was considered, including the demographic variables age, gender, and ECOG status, in addition to OXA duration and time. The only alteration to the two-covariate model is that $\boldsymbol{\gamma}$ is now a vector consisting of five entries and $\boldsymbol{x}_i$ is a vector having five elements ($\mu_i = e^{\boldsymbol{z}'_i\boldsymbol{\beta}+\boldsymbol{x}'_i\boldsymbol{\gamma}}$). All of these models were also refit using a *commensurate* *prior* (Hobbs et al., 2011) on the concurrent controls group effect. This prior assumes $\beta_{CC}|\beta_{HC} \sim N(\beta_{HC}, \tau)$, where τ is a crucial precision parameter that controls the degree of borrowing. Commensurate priors are cousins to power priors (Ibrahim and Chen, 2000), which can also be used to downweight the impact of the historical data, but are less convenient when the appropriate degree of borrowing is unknown *a priori*. Here, we either fix τ at a large value (say, 1000) that encourages borrowing, or else assign it a moderately informative hyperprior centered at the same value – say, a $Gamma(1, 0.001)$ distribution (Hobbs et al., 2011; Murray et al., 2014). Neuenschwander et al. (2016) broaden the commensurate prior notion of historical data borrowing to the use of *any* trial-external complementary data ("co-data"), including both control and treatment data, and from trials that are either completed or ongoing. Gamalo-Siebers et al. (2017) offer a review of methods for adaptive borrowing from auxiliary data, including both "static" methods that fix the degree of borrowing from the auxiliary data ahead of time (say, via a pre-chosen power prior), and "dynamic" methods that attempt to estimate this degree from the data. Note that a possibility here is first attempt to estimate the similarity of the two data sources and then determine the borrowing hyperparameters accordingly. While intutively appealing, such an empirical Bayes-style approach implicitly "uses the same data twice," and is thus invalid from a fully Bayesian point of view.

An important comparator across models and simulation cases is the effective historical sample size (EHSS), the effective number of historical patients borrowed, as it pertains to the treatment effect between Drug A and the concurrent controls, or some combination of the controls. For our present purposes, it suffices to use a straightforward, "univariate" calculation of EHSS that focuses entirely on the primary parameter of interest, namely

$$EHSS = N_{HC} \left(\frac{Prec(\beta_A - \beta_{CC} \mid all\ data)}{Prec(\beta_A - \beta_{CC} \mid CRC3\ data)} - 1 \right), \tag{1}$$

where Prec denotes the precision (inverse variance, $Prec = \frac{1}{sd^2}$) of the indicated quantity. Here we have selected the treatment effect, $\beta_A - \beta_{CC}$, though of course other choices could be contemplated. This definition is based on equation (3) of Hobbs et al. (2013), and is inspired by (but not identical to) previous versions mentioned in regulatory science by Pennello and Thompson (2008); the definition of Neuenschwander et al. (2016) is also similar. While conceptually simple, its reliance on estimated variance components can

	Mean	SD	95% CrI
$\beta_{HC\&CC}$	-4.23	0.18	(-4.58, -3.88)
β_A	-4.61	0.23	(-5.08, -4.15)
β_A - $\beta_{HC\&CC}$	-0.38	0.17	(-0.72, -0.06)

Table 2: Posterior PFS summaries for the two-arm model comparing all combined controls (n=474) and subjects receiving the additional Drug A treatment (n=63).

make it somewhat unstable. In our case, we used estimated posterior standard deviations taken from their respective BUGS output tables. Below, we evaluate biases as well as 95% credible interval widths and coverage rates for the estimated treatment effect differences when comparing the various methods via simulation.

4 Results for Motivating Data Set

This section shows the results when focusing on analyzing the PFS data (n=537), instead of the DoR subset (n=235), whose smaller sample size does not permit statistically significant findings. Throughout the remainder of this paper, since all analyses are Bayesian, we use the term "significant" to mean "Bayesianly significant," i.e., reflective of a 95% equal-tail credible interval that excludes 0.

Using the 2-arm and 3-arm models from Section 3 and the BUGS language, the Bayesian estimates for the group effects for the combinations of interest were computed along with their respective standard deviations and 95% credible intervals (CrI). The three versions of the 3-arm model included vague priors on the βs, a fixed τ prior, and a random τ prior. Table 2 gives the group effects and treatment effect for the naively-combined controls and Drug A under vague priors. All estimates are significantly negative, including the treatment effect difference between the groups, suggesting that receiving Drug A decreases the hazard rate compared to the naively-combined controls.

Table 2 shows the impacts that naively combining the controls can have on the significance and directionality of the treatment effect. In particular, the naive combination of the controls results in a Bayesianly significant treatment effect for the novel Drug A.

	$\bar{D}$	pD	DIC
Base model	2805	3.949	2809
OXA	2645	4.974	2650
OXA & time	2643	5.926	2649
Full model	2644	9.433	2653

Table 3: DIC values for various 3-arm models of interest when using the vague normal prior. The base model does not include any additional covariates; the full model adjusts for OXA duration, time, age, sex, and ECOG status.

The results comparing the naively-combined controls with the treatment group for models including additional covariates were broadly similar.

As discussed in Section 3, a variety of nested models were applied to the data. The DIC values for each model when comparing the three groups (HC, CC, and Drug A) separately are shown in Table 3. Quite obviously, the enormous decrease of 159 in DIC between the base model and the OXA duration-adjusted model suggests that a big improvement occurs when OXA duration is added to the model. However, once OXA duration is included, models with additional covariates beyond this do little to change the DIC values, meaning these models are equally acceptable statistically. Hence, our focus below will be on the model results and the changes in precision among just three models: the base model, the OXA duration-adjusted model, and the full model.

The posterior summaries for each 3-arm model comparing the three groups (HC, CC, Drug A) separately under vague priors are shown in Table 4. For the base model outlined in Section 3, significant group effect differences (-0.44) are seen for both CRC3 groups when compared to the historical controls. Since the estimated effect difference was significantly negative, both CRC3 groups' comparison with the historical controls behave as expected, decreasing the hazard rate. The most important result for the no-covariate model, because the treatment effect is the main question of interest, is that $\beta_A - \beta_{CC}$ is not significantly different from zero, which implies no significant PFS benefit from Drug A exists when comparing with the concurrent control arm of standard of care.

Next, we examine the effect individual patient-level OXA duration has on the treatment

	Base Model			OXA Duration			Full Model		
	Mean	SD	95% CrI	Mean	SD	95% CrI	Mean	SD	95% CrI
β_{HC}	-4.17	0.18	(-4.53, -3.82)	-4.36	0.18	(-4.72, -4.02)	-4.71	0.25	(-5.21, -4.22)
β_{CC}	-4.61	0.23	(-5.07, -4.16)	-5.51	0.25	(-6.01, -5.02)	-5.34	0.36	(-6.04, -4.63)
β_A	-4.61	0.23	(-5.08, -4.16)	-5.53	0.25	(-6.04, -5.04)	-5.36	0.36	(-6.06, -4.55)
γ_{OXA}	—	—	—	-0.18	0.02	(-0.21, -0.15)	-0.18	0.02	(-0.22, -0.15)
γ_{time}	—	—	—	—	—	—	0.01	<0.01	(-0.01, 0.01)
γ_{age}	—	—	—	—	—	—	0	<0.01	(-0.01, 0.01)
γ_{sex}	—	—	—	—	—	—	0.13	0.10	(-0.07, 0.33)
γ_{ECOG}	—	—	—	—	—	—	0.10	0.10	(-0.10, 0.30)
β_A-β_{CC}	-0.00	0.22	(-0.44, 0.44)	-0.03	0.22	(-0.47, 0.41)	-0.02	0.22	(-0.45, 0.42)
β_A-β_{HC}	-0.44	0.17	(-0.78, -0.12)	-1.17	0.18	(-1.54, -0.83)	-0.64	0.34	(-1.30, 0.03)
β_{CC}-β_{HC}	-0.44	0.16	(-0.77, -0.13)	-1.15	0.17	(-1.50, -0.81)	-0.62	0.33	(-1.26, 0.04)

Table 4: Posterior PFS summaries for the three-arm model using vague priors comparing the historical controls (n=412), concurrent controls (n=62), and the Drug A group (n=63) for the three main models of interest. The horizontal lines separate the treatment effects, covariate effects, and group effect differences for each model.

effect. Due to missing OXA duration data, 15 of the 537 observations were omitted from this analysis. Since the estimate for γ_{OXA} in Table 4 is significantly negative, an increase in the length of OXA duration is associated with a decrease in the hazard rate, when the treatment group is kept constant. The significance of all other results did not change in comparison with the no covariate model; i.e., a significant difference is still present between the historical controls and both concurrent controls and Drug A, while a treatment difference within the CRC3 trial does not exist. Similar results were found when replacing OXA duration with cumulative OXA dosage in the model.

Finally, the only notable change in the group effects under the vague-prior full model (rightmost section of Table 4) is that the differences in PFS between the historical controls and both concurrent controls and Drug A are no longer significant, albeit barely. Interestingly, the duration of OXA is the only covariate to have a significant effect (95% CrI

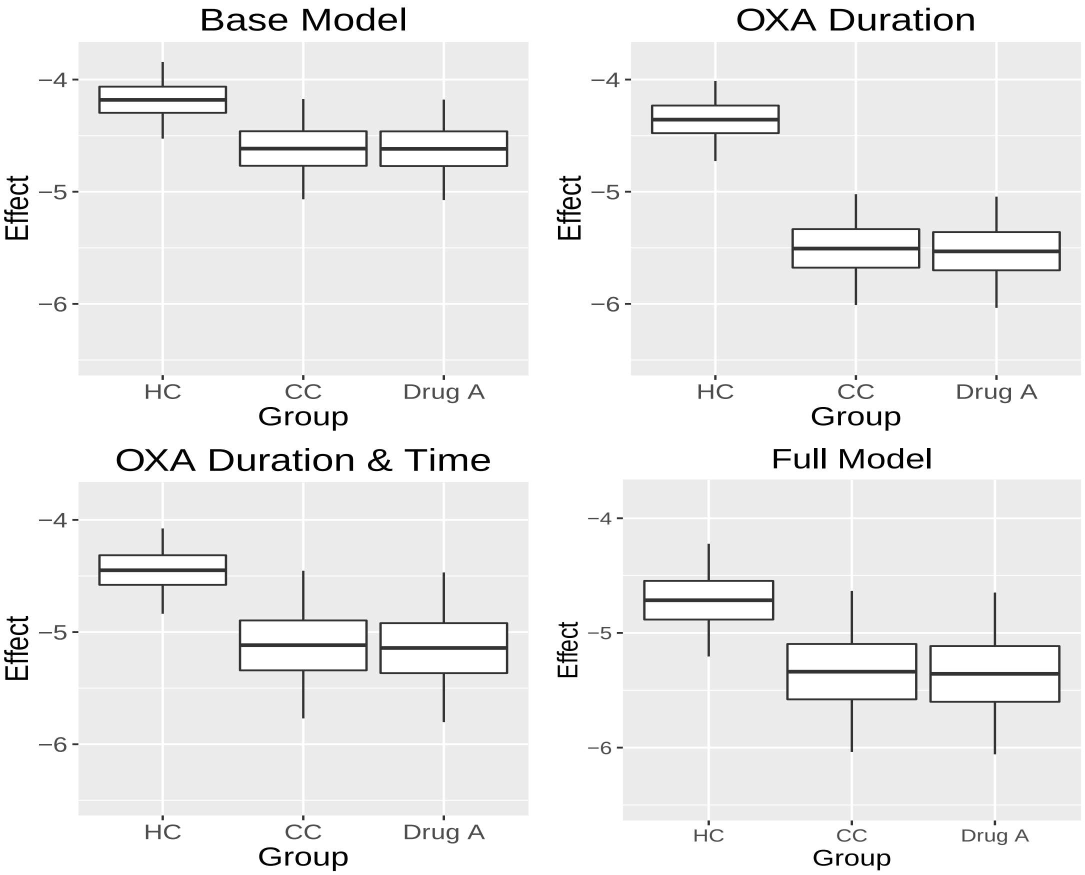

Figure 3: Posterior boxplots for the group effects (β_{HC}, β_{CC}, and β_A) for the four main models of interest: the base model which does not include any continuous covariates (upper left), a model including OXA duration (upper right), a model including both OXA duration and the time component (lower left), and a full model with these two plus three demographic variables (lower right).

(-0.22, -0.15)).

Figure 3 contains four boxplots comparing the group specific posteriors from the models described earlier in thise section, including the OXA duration and time model. First, none of the plots suggest a significant difference exists between the CRC3 trial arms (CC & Drug A). Some movement occurs in the group effects, but not within the CRC3 trial. The effects of both arms of the CRC3 trial moved further away from that of the historical controls for the OXA duration model (upper right). Also, the 95% credible intervals (indicated by the “whiskers” in the boxplots) for the models including covariates beyond OXA duration

	Base Model			OXA Duration			Full Model		
	Mean	SD	95% CrI	Mean	SD	95% CrI	Mean	SD	95% CrI
β_{HC}	-4.22	0.17	(-4.58, -3.89)	-4.40	0.18	(-4.77, -4.04)	-4.81	0.23	(-5.26, -4.36)
β_{CC}	-4.24	0.18	(-4.59, -3.91)	-4.45	0.19	(-4.82, -4.09)	-4.82	0.23	(-5.27, 4.36)
β_{A}	-4.61	0.23	(-5.06, -4.17)	-5.23	0.25	(-5.74, -4.75)	-4.96	0.30	(-5.55, -4.37)
γ_{OXA}	—	—	—	-0.13	0.02	(-0.16, -0.10)	-0.18	0.02	(-0.22, -0.15)
γ_{time}	—	—	—	—	—	—	0.01	<0.01	(0.01, 0.02)
γ_{age}	—	—	—	—	—	—	0	<0.01	(-0.01, 0.01)
γ_{sex}	—	—	—	—	—	—	0.14	0.10	(-0.06, 0.33)
γ_{ECOG}	—	—	—	—	—	—	0.10	0.10	(-0.10, 0.31)
β_A-β_{CC}	-0.37	0.17	(-0.71, -0.04)	-0.79	0.18	(-1.15, -0.45)	-0.14	0.21	(-0.56, 0.27)
β_A-β_{HC}	-0.38	0.17	(-0.73, -0.06)	-0.83	0.18	(-1.20, -0.50)	-0.14	0.21	(-0.57, 0.27)
β_{CC}-β_{HC}	-0.02	0.03	(-0.08, 0.04)	-0.05	0.03	(-0.11, 0.01)	-0.01	0.03	(-0.07, 0.06)

Table 5: Posterior PFS summaries comparing the three groups when using a commensurate prior ($\tau = 1000$) for the concurrent controls for the three main models of interest. The horizontal lines separate the treatment effects, covariate effects, and group effect differences for each model.

are larger than those for the smaller two models. A difference in the width of the credible intervals between the base model and OXA duration-adjusted model is not discernible from Figure 3, perhaps because, unlike some of the other full model covariates, OXA duration adds real explanatory power (and not mere "noise") to the model.

Tables 2 and 4 showcase the problematic nature of the analyses of the colorectal cancer data: if the CRC3 trial data are analyzed using vague priors, essentially separate from information garnered from previously completed trials, no statistical evidence of a treatment effect difference is found; however, significant treatment differences exist when comparing the Drug A group to either the historical controls alone or the naively-combined controls. In the latter setting, the historical controls account for 87% of the combined control group, and thereby drive the statistical difference with Drug A.

A significant difference between the historical controls and the concurrent controls is also

	Base Model			OXA Duration			Full Model		
	Mean	SD	95% CrI	Mean	SD	95% CrI	Mean	SD	95% CrI
β_{HC}	-4.12	0.18	(-4.55, -3.87)	-4.37	0.18	(-4.73, -4.01)	-4.84	0.24	(-5.31, -4.38)
β_{CC}	-4.31	0.21	(-4.75, -3.93)	-5.44	0.26	(-5.95, -4.94)	-4.87	0.25	(-5.36, -4.40)
β_{A}	-4.61	0.23	(-5.06, -4.16)	-5.53	0.26	(-6.03, -5.04)	-4.99	0.30	(-5.60, -4.42)
γ_{OXA}	—	—	—	-0.18	0.02	(-0.21, -0.14)	-0.18	0.02	(-0.22, -0.15)
γ_{time}	—	—	—	—	—	—	0.01	<0.01	(0.01, 0.02)
γ_{age}	—	—	—	—	—	—	0	<0.01	(-0.01, 0.01)
γ_{sex}	—	—	—	—	—	—	0.15	0.10	(-0.05, 0.34)
γ_{ECOG}	—	—	—	—	—	—	0.11	0.10	(-0.10, 0.31)
β_A-β_{CC}	-0.29	0.21	(-0.68, 0.14)	-0.09	0.23	(-0.53, 0.35)	-0.13	0.21	(-0.55, 0.28)
β_A-β_{HC}	-0.40	0.17	(-0.74, -0.07)	-1.16	0.18	(-1.53, -0.82)	-0.15	0.22	(-0.60, 0.27)
β_{CC}-β_{HC}	-0.10	0.13	(-0.47, 0.04)	-1.07	0.18	(-1.42, -0.73)	-0.02	0.09	(-0.25, 0.08)

Table 6: Posterior PFS summaries comparing the three groups when using a hyperprior of ($\tau \sim Gamma(1, 0.001)$) for the three main models of interest. The horizontal lines separate the treatment effects, covariate effects, and group effect differences for each model.

seen, raising the question as to whether or not it is reasonable to borrow any information from the previous trials at all. Another approach to this issue can be examined by placing various commensurate priors on β_{CC} with $\tau = 1000$ (a prior that actively encourages borrowing from the historical controls) or random $\tau \sim Gamma(1, 0.001)$. The fixed τ results are shown in Table 5. For the base and OXA duration-adjusted models, $\hat{\beta}_{CC}$ is shrinking towards $\hat{\beta}_{HC}$, leading to a now-significant treatment effect in which Drug A users have a lower hazard rate than the concurrent controls. On the other hand, the full model returns the more sensible non-significant treatment effect, suggesting again that proper covariate adjustment is a key part of the modeling process. The results of the fixed τ (fairly informative) commensurate prior shows the danger of giving too much weight to the historical data, especially when the data are not sufficiently similar and we do not properly adjust for covariates.

Finally, Table 6 shows that having a more flexible τ eliminates the significant treatment effect found when we fixed τ for the base and OXA-duration models, meaning we are not relying as heavily on the covariate adjustment to obtain a sensible result. The resulting τ estimates (SD) from the base, OXA-adjusted, and full models are 683 (900), 935 (983), and 947 (992), respectively. Since the estimated τ values and corresponding standard deviations for the models containing covariates are near the theoretical mean and standard deviation of 1000, those models do not do much to inform the amount of borrowing. The random τ model produces interval widths which are more comparable to but still smaller than the vague prior model, showing a modest gain in precision. Thus, the random τ model is more robust than the fixed τ model, yet still improves on the noninformative prior model.

The EHSS for these three models helps shape the picture. We calculated the standard deviation for the treatment effect when only using information from the CRC3 trial, $\widehat{SD}(\beta_A - \beta_{CC}|\ CRC3\ data)$, for each model (base, OXA, and full). The precision in the denominator of equation (1) is computed using this value, and the standard deviations needed to compute the precisions in the numerator are taken from Tables 4, 5, and 6. The EHSS for the non-informative prior models are, as expected, all near zero (0, 3, 19), whereas the EHSS for the base, OXA duration, and full models are 287, 236, and 63, respectively under the fixed τ commensurate prior and 76, 0, and 75 under the random τ prior. Thus, the random τ model uses enough information from the historical controls to improve upon the precision of the vague models, but not as much as the fixed τ models. Happily and intuitively, this leads to the most sensble conclusion of a non-significant treatment effect.

A referee has asked about the utter lack of borrowing in the OXA duration model under the random τ prior, which surprised us as well. To check this, we redid these calculations under a couple of other priors, including a $Gamma(10, .01)$ (less vague, but still centered at 1000) and a $Gamma(10, .1)$ (same shape, but now centered near 100). These produced EHSS values of 204 and 47, respectively, indicating that borrowing can occur under random τ for the OXA duration model, but only if the prior offers a bit of guidance.

5 Simulation Study

5.1 Simulation Methods

A simulation study was conducted to examine what impacts changing the true treatment effects could have on the proper amount of information to borrow from the historical controls, and the resulting precision, bias, and interval coverage probabilities of the group differences. The main point of alteration was the true survival rate for the concurrent controls, which was allowed to vary between that of the historical controls and the Drug A group as estimated from Table 4. Specifically, except for the "null" Case 1, β_{CC} is the only β value that changes throughout the cases, varying between $\hat{\beta}_{HC} = -4.17$ and $\hat{\beta}_{A} = -4.61$, the base model values estimated from our data. The cases we consider are:

- Case 1: The "null case" where all group effects were considered equal ($\beta_{HC} = \beta_{CC} = \beta_{A} = -4.17$).
- Case 2: The historical and concurrent controls are equivalent, $\beta_{CC} = \beta_{HC} = -4.17$, but $\beta_{A} = -4.61$.
- Cases 3,4,5: β_{CC} takes on the values of -4.28, -4.39, and -4.50, respectively, with $\beta_{HC} = -4.17$ and $\beta_{A} = -4.61$ fixed.
- Case 6: All β's fixed at essentially their Table 4 values; we use $\beta_{HC} = -4.17$, and $\beta_{CC} = \beta_{A} = -4.61$.

For each case, we used the observed group totals of $N_{HC} = 412$, $N_{CC} = 62$, and $N_{A} = 63$, and generated 1,000 artificial data sets using the corresponding true β coefficients. Since each group had a different observed percentage of censored data (HC: 18%, CC: 32%, Drug A: 62%), censoring times were simulated for each group to mimic this. Specifically, the failure times follow a Weibull distribution with mean function $\mu_i = e^{\boldsymbol{z_i \beta}}$, using the corresponding β values for each case and shape parameter r from the output of the colorectal cancer analysis. We assume that censoring times independently follow a normal distribution with parameters $tcen$ and σ^2, which again were manually adjusted to achieve the desired proportions of censored individuals in each group. Thus, for $k = 1 \ldots 1000$, we generate the individual failure times and censoring times as

$$t_{ijk} \sim Weibull(r, \mu_i) \text{ and } c_{ijk} \sim N(tcen_j, \sigma_j^2),$$

for $i = 1 \ldots 537$ and where $j(i)$ represents the treatment group for individual i (HC, CC, or Drug A). If $t_{ijk} > c_{ijk}$, we replace the t_{ijk} value with a missing value (NA), signifying a censored survival time. We let $\sigma_j^2 = 10$ for all j, and reverse engineered sensible $tcen_j$ values of 18, 17, and 10 for the historical controls, concurrent controls, and Drug A, respectively.

Next, for each data set we apply the three versions of the 3-arm base model. The most simple of the 3-arm base models places a vague prior on each β. We then consider the simple fixed τ commensurate prior on β_{CC}, namely $\beta_{CC}|\beta_{HC} \sim N(\beta_{HC}, \tau)$ where τ is fixed at 1000, a value that is fairly informative but still appropriate given the scale of our data. Finally, we explore the use of a $Gamma(1, 0.001)$ hyperprior on τ, having mean and standard deviation both equal to 1000, a model that allows the data to influence the estimated value of τ. The resulting estimates are then used to calculate the average EHSS, the average biases of the treatment differences, and the average widths and empirical coverages of the resulting 95% CrIs of the treatment differences. We also compare the percentage of significant treatment effect results across the simulation cases and priors.

We also repeated the comparison process for the full model, instead of the base model, using generated data sets where our mean function is now $\mu_i = e^{\boldsymbol{z_i \beta} + \boldsymbol{x_i \gamma}}$ and $N_{HC} = 397$, again as in the colorectal cancer data. By incorporating all of the covariates in data simulation, we chose new $tcen_j$ values of 32, 30, and 17 to obtain the desired proportion of censored individuals for the historical controls, concurrent controls, and Drug A, respectively. We use the observed data values for OXA duration, time, age, sex, and ECOG status, and let the true covariate effect vector be $\boldsymbol{\gamma} = (-0.2, 0, 0, 0, 0)$. These outputs are compared across the various priors, as well as with the corresponding base model outputs.

5.2 Simulation Results

Table 7 shows the pertinent simulation results for the primary treatment effect ($\beta_A - \beta_{CC}$) for each prior using the base model. As expected, the vague prior model (V) does not borrow much information from the historical controls, which leads to minimal bias

	EHSS			Bias			Interval Width			Coverage		
Case	V	F	R	V	F	R	V	F	R	V	F	R
1	3	229	197	-0.005	-0.005	-0.005	0.88	0.71	0.73	93.8%	94.9%	95.3%
2	3	201	173	-0.012	-0.012	-0.012	0.92	0.76	0.78	94.5%	95.1%	95.2%
3	1	202	170	-0.01	-0.103	-0.092	0.92	0.76	0.78	94.4%	91.9%	93.0%
4	3	211	150	-0.003	-0.194	-0.165	0.93	0.76	0.81	94.9%	81.0%	86.5%
5	3	219	106	-0.007	-0.284	-0.217	0.93	0.76	0.84	94.3%	67.1%	80.9%
6	7	231	51	-0.008	-0.370	-0.237	0.94	0.76	0.91	94.1%	49.5%	79%

Table 7: The simulation results for the 3-arm base model comparing the EHSS, precision, bias, and coverage percentages of the treatment effect ($\beta_A - \beta_{CC}$) for each variation of the prior (V=vague, F=fixed τ, R=random τ).

and empirical coverage percentages near the desired 95%. The fixed τ model (F) which encourages borrowing did just that, borrowing more from the historical controls than either of the other priors. This results in narrower intervals in Case 1, but also an increasing amount of bias as the true value of β_{CC} moves further away from β_{HC} (i.e., working through Cases 2 to 6). The empirical coverage percentages for the treatment effect show that as the concurrent simulated data become more different from the historical simulated data, the 95% BCrI widths remain stable but the intervals do a progressively poorer job of covering the true treatment effect.

Turning to the random τ model (R), its EHSS behaves as we would expect, decreasing as the true β_{CC} value moves further away from the true β_{HC} value. While still borrowing some information from the historical controls (with intervals nearly as narrow as those of model F), the model R has less bias for corresponding cases than model F, and smaller interval widths than model V. We can see that by wisely selecting the hyperprior on τ, we gain precision over method V, as evidenced by the 95% interval widths, particularly in Cases 1 and 2. The method R coverage percentages also withstand the unfavorable Cases 5 and 6 much better than those of method F.

The simulation results for each case when adjusting for all possible covariates are shown in Table 8. Differences in bias, interval width and coverage percentages are smaller between

	EHSS			Bias			Interval Width			Coverage		
Case	V	F	R	V	F	R	V	F	R	V	F	R
1	24	91	88	-0.009	-0.008	-0.008	0.89	0.83	0.83	95.3%	95.0%	94.6%
2	27	89	87	-0.023	-0.021	-0.021	0.94	0.88	0.88	94.8%	94.1%	94.2%
3	26	93	88	-0.02	-0.051	-0.05	0.95	0.88	0.89	95.1%	93.8%	93.6%
4	24	99	93	-0.009	-0.078	-0.075	0.96	0.88	0.89	96%	94%	93.8%
5	25	104	97	-0.006	-0.109	-0.103	0.97	0.89	0.89	95.5%	92.3%	92.9%
6	24	111	100	-0.004	-0.141	-0.131	0.98	0.89	0.90	95.5%	89.3%	90.2%

Table 8: The simulation results for the 3-arm full model comparing the EHSS, precision, bias, and coverage percentages of the treatment effect ($\beta_A - \beta_{CC}$) for each variation of the prior (V=vague, F=fixed τ, R=random τ).

the three variations of priors when using the full model, which is not unexpected. In the analysis of the observed data (Tables 4, 5, and 6), the estimates, standard deviations, and intervals of the treatment effects did not vary much across the various priors for the full model. The F and R models still show an increase in bias as β_{CC} moves towards β_{HC} and a slow decrease in interval coverage percentages. The interval widths for the vague model are larger than those for the fixed and random τ models, since fewer historical controls are being borrowed, as evident from the EHSSs. In comparison with the base model simulations in Table 7, the full model has improved performance, but at the cost of precision, with corresponding interval widths being larger on average.

Table 9 contains the percentages of 95% CrIs that return a significant treatment effect ($\beta_A - \beta_{CC}$) for each case, model, and prior, with associated Monte Carlo standard errors in parentheses. All priors and models do recover close to the nominal Type I error rate (5%) in the null Case 1. More power exists under the F and R priors in Cases 2-5, which is expected due to the borrowing. Case 6 represents a (relatively high) Type I error, because the null is, again, true ($\beta_A = \beta_{CC}$). The corresponding power (Cases 2-5) and Type I error (Case 6) rates for the base model compared to the full model are consistently higher for the F and R priors. This is also expected, since we have shown that the full model borrows less information from the historical controls. Thus, Table 9 suggests that covariate adjustment

	Base Model			Full Model		
Case	V	F	R	V	F	R
1	6.2 (0.76)	5.1 (0.70)	4.7 (0.67)	4.7 (0.8)	5.0 (0.82)	5.4 (0.86)
2	49.6 (1.58)	66.7 (1.49)	65.3 (1.51)	47.4 (1.9)	52.8 (1.9)	52.6 (1.9)
3	29.3 (1.44)	63.1 (1.53)	58.0 (1.56)	32.6 (1.75)	40.3 (1.83)	39.6 (1.82)
4	17.5 (1.20)	59.3 (1.55)	47.3 (1.58)	14.7 (1.51)	28.4 (1.92)	27.1 (1.90)
5	8.7 (0.89)	56.1 (1.57)	35.8 (1.52)	7.5 (1.11)	17.0 (1.59)	15.5 (1.53)
6	5.9 (0.74)	56.4 (1.57)	21 (1.29)	4.4 (0.87)	10.7 (1.31)	9.8 (1.26)

Table 9: Percent of significant treatment effects ($\beta_A - \beta_{CC}$) (MC standard error) for each simulation study case by model and prior (i.e., credible interval did not contain zero; Type I error in Cases 1 and 6, power in Cases 2-5); V=vague, F=fixed τ, and R=random τ.

can help compensate for misspecification elsewhere in the model.

6 Discussion, Alternatives, and Conclusions

6.1 Discussion

Our motivating data set provides a real-world setting in which either designing the CRC3 trial as a single active treatment arm study and relying solely on historical control data, or implementing a two-arm study and borrowing too much from these same historical controls, would have resulted in a false positive estimate of the treatment effect. Clearly, the driving force behind the significant difference between Drug A and the combined controls is mostly attributable to the large number of historical controls (412) compared to concurrent controls (62). This is particularly worrisome because the historical controls and the concurrent controls group effects remain significantly different from each other throughout the analysis. When a hyperprior for τ is carefully selected, each of the models (base, OXA-adjusted, and full) all more sensibly return a non-significant treatment effect. The same cannot be said for the fixed τ models, where the shrinkage of the $\hat{\beta}_{CC}$ estimate towards $\hat{\beta}_{HC}$ causes the treatment effect to be significant for the base and OXA-duration model; however, the full model is able to overcome the shrinkage and remains non-significant. Thus, a combination of carefully selected priors and hyperpriors as well as proper covariate adjustment can lead to sensible results even in Bayesian hierarchical modeling settings where the historical data are not similar to the concurrent.

We note that our data analysis ignores possible correlation between the OXA duration covariate and the survival endpoint, which if nothing else muddies the causal interpretation of this covariate. A traditional fix here might be to binarize the OXA covariate (say, to "long duration" and "short duration"), or to fit a model including an OXA-by-treatment interaction term. For our data, this latter approach revealed that there is no reason to believe an interaction effect between OXA and treatment exists. In any case, this is a common problem in oncology trials that our methods must be tuned to combat.

Our simulation study examined one null case and five non-null cases of varying β_{CC} values, ranging between $\hat{\beta}_A$ and $\hat{\beta}_{HC}$ from the initial analysis of the colorectal data. We used the base model with non-informative priors on each of the βs, one with a commensurate prior on β_{CC} having $\tau = 1000$, and one placing a hyperprior on τ. By applying the fixed commensurate prior to the concurrent controls we saw an increase in the amount of bias and a decrease in the coverage percentages of the treatment effect as the β_{CC} value moved further away from β_{HC}. The same phenomenon occurred for the hyperprior model, but each corresponding case had less bias and better coverage than the fixed τ model. The random τ model also had smaller interval widths than the non-informative prior model. Thus, the hyperprior model increases the level of precision relative to the vague model, yet protects against bias better than the fixed τ model.

In particular, when the historical controls are sufficiently similar to the concurrent data (Cases 1 and 2, and perhaps 3), comparing the simulation using the random τ model to the vague model show 15% and 6.4% gains in precision (interval width) from borrowing in Case 2 for the base model and full model, respectively. Neither of the models for Case 2 show an increase in bias for the pro-borrowing priors, and coverage rates remain at the desired 95% level, confirming bias is not being sacrificed for this increase in precision.

By using the full model instead of the base model, the differences in bias, interval width, and coverage percentages between the fixed and random τ models are greatly attenuated. The full model simulations had less bias and better coverage percentages for each prior compared to the base model, but larger interval widths. These differences can also be seen in the Type I error and power rates for the two different models. The full model has lower power rates for corresponding cases (2-5) and priors (F and R) than the base model, but these rates do not differ significantly between the F and R priors in the full model as they did in the base model. Thus, adjusting for the correct covariates can improve model performance and allow for flexibility when selecting the prior, at some cost in precision.

6.2 Summary

When is it acceptable to borrow from historical trials? Is it reasonable to design a single-arm clinical trial and compare the treatment effect to solely historical controls or naively borrow information from the historical trials? In situations where a particular unmet medical need exists and randomizing additional patients to control is infeasible or unethical, a single arm study could be the only feasible design. If there is a high likelihood that the historical controls are similar to the current single-arm intervention patients in all important respects, and if we can be reasonably confident that here has been no drift in the control rate since the historical study was conducted, then this approach may suffice. The emergence of master protocols in oncology (Renfro and Sargent, 2017) will certainly be of help here, especially in drug development for rare diseases and some cancers. In more general cases, significant recent effort has focused on the selection of "synthetic control" patients from historical populations arising from similar protocols that are matched in some way to the current single-arm intervention patients, where the matching might be determined by the patients' baseline characteristics (Berry et al., 2017). However, even here it can be a strong assumption that there has been no drift in the control rate since the historical studies were conducted, even in the presence of propenity score or other similar statistical adjustment. As indicated by the analysis of our motivating data set and our simulation study, such approaches can somtimes be misleading and inaccurate. In our setting, running a two-arm study, choosing a flexible borrowing model, and including proper covariate information were all crucial to reaching a sensible conclusion.

In a broader sense, understanding the comparison between historical controls and concurrent controls is crucial in designing a clinical trial using a Bayesian framework. The historical and concurrent data must be sufficiently similar, and if not, there must be a proper way to account for evolving standard of care, differences in the patient population, or variations in the study designs. Bayesian methods can help with making sure defensible results are still obtained, even under modestly informative commensurate (pro-borrowing) priors. Simulations to analyze the impact of the priors and of the data in various situations can help better understand the connections between the historical and current data, and should be employed in future cases of this approach with historical data.

References

Berry, D. (2006), 'Bayesian clinical trials', *Nature Reviews Drug Discovery* **5**, 27–36.

Berry, D., Elashoff, M., Blotner, S., Davi, R., Beineke, P., Chandler, M., Lee, D., Chen, L. and Sarkar, S. (2017), 'Creating a synthetic control arm from previous clinical trials: Application to establishing early end points as indicators of overall survival in acute myeloid leukemia (AML)', *Journal of Clinical Onocology* **35**(15), 7021–7021.

Carlin, B. and Louis, T. (2008), *Bayesian Methods for Data Analysis, Third Edition*, Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis.
URL: *https://books.google.com/books?id=EcjBrQEACAAJ*

Chen, N., Carlin, B. and Hobbs, B. (2018), 'Web-based statistical tools for the analysis and design of clinical trials that incorporate historical controls', *Computational Statistics and Data Analysis* **125**(to appear).

Gamalo-Siebers, M., Savic, J., Basu, C., Zhao, X., Gopalakrishnan, M., Gao, A., Baygani, S., Thompson, L., Xia, A., Price, K., Tiwara, R. and Carlin, B. P. (2017), 'Statistical modeling for Bayesian extrapolation of adult clinical trial information in pediatric drug evaluation', *Pharmaceutical Statistics* **16**, 232–249.

Gökbuget, N., Dombret, H., Ribera, J., Fielding, A., Advani, A., Bassan, R., Chia, V., Doubek, M., Giebel, S., Hoelzer, D., Ifrah, N., Katz, A., Kelsh, M., Martinelli, G., Morgades, M., O'Brien, S., Rowe, J., Stieglmaier, J., Wadleigh, M. and Kantarjian, H. (2016a), ‘International reference analysis of outcomes in adults with b-precursor ph-negative relapsed/refractory acute lymphoblastic leukemia’, *Haematologica* **101**(12), 1524–1533.

Gökbuget, N., Kelsh, M., Chia, V., Advani, A., Bassan, R., Dombret, H., Doubek, M., Fielding, A., Giebel, S., Haddad, V., Hoelzer, D., Holland, C., Ifrah, N., Katz, A., Maniar, T., Martinelli, G., Morgades, M., O’Brien, S., Ribera, J., Rowe, J., Stein, A., Topp, M., Wadleigh, M. and Kantarjian, H. (2016b), ‘Blinatumomab vs historical standard therapy of adult relapsed/refractory acute lymphoblastic leukemia’, *Blood Cancer Journal* **6**(9), e473.

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

Hobbs, B., Carlin, B. and Sargent, D. (2013), ‘Adaptive adjustment of the randomization ratio using historical control data’, *Clinical Trials* **10**(3), 430–440.

Hobbs, B., Sargent, D. and Carlin, B. (2012), ‘Commensurate prior for incorporating historical information in clinical trials using general and generalized linear models’, *Bayesian Analysis* **7**(3), 639–674.

Ibrahim, J. and Chen, M.-H. (2000), ‘Power prior distributions for regression models’, *Statistical Science* **15**, 46–60.

Lunn, D., Thomas, A., Best, N. and Spiegelhalter, D. (2000), ‘WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility’, *Statistics and Computing* **10**, 325–337.

Murray, T. A., Hobbs, B., Lystig, T. and Carlin, B. (2014), ‘Semiparametric Bayesian commensurate survival model for post-market medical device surveillance with non-exchangeable historical data’, *Biometrics* **70**, 185–191.

Neuenschwander, B., Roychoudhury, S. and Schmidli, H. (2016), ‘On the use of co-data in clinical trials’, *Statistics in Biopharmaceutical Research* **8**(3), 345–354.

Pennello, G. and Thompson, L. (2008), ‘Experience with reviewing Bayesian medical device trials’, *Journal of Biopharmaceutical Statistics* **18**(1), 81–115.

Quan, H., Zhang, B., Chuang-Stein, C., Jones, B. and on behalf of the EFSPI Integrated Data Analysis Efficacy Working Group (2017), ‘Integrated data analysis for assessing treatment effect through combining information from all sources’, *Statistics in Biopharmaceutical Research* **9**(1), 52–64.

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

U.S. Department of Health and Human Services (2010), ‘Guidance for the use of Bayesian statistics in medical device clinical trials’.

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J., Kinnersley, N., Lindborg, S., Micallef, S., Roychoudhury, S. and Thompson, L. (2014), ‘Use of historical control data for assessing treatment effects in clinical trials’, *Pharmaceutical Statistics* **13**(1), 41–54.

Appendix

JAGS code

The following is the `JAGS` code for the full model. To switch to `OpenBUGS`, the first two lines for each group can be replaced with `t[i] ∼ dweib(r, mu[i])C(t.cen[i],)`. The historical controls are represented by beta[1]; the concurrent controls, beta[2]; and Drug A, beta[3]. The OXA duration effect is gamma[1], time is gamma[2], age is gamma[3], sex is gamma[4], and ECOG status is gamma[5].

```
model{ # The basic structure of this model has been taken from
   the ”MICE” example in BUGS:
```

```
for(i in 1 : N1) { #where N1 = the number of historical
   control patients
    is.censored[i] ~ dinterval(t[i], t.cen[i])
    t[i] ~ dweib(r, mu[i]) #the event times are assumed
       to follow a Weibull distribution
    mu[i]<- exp(beta[1]+gamma[1]*(x[i]- mean(x[]))+ gamma
       [2]*(y[i]-mean(y[]))+gamma[3]*(age[i]-mean(age[]))
      +gamma[4]*sex[i]+gamma[5]*ecog[i])
    #median survival times for each patient
        median[i] <- pow(log(2) * exp(-(beta[1]+gamma
           [1]*(x[i]- mean(x[]))+ gamma[2]*(y[i]-mean(y
           []))+gamma[3]*(age[i]- mean(age[]))+gamma[4]*
           sex[i]+gamma[5]*ecog[i])), 1/r)
                }
for(i in (N1+1): (N2+N1)) { #N2 is the number of
   concurrent controls
        is.censored[i] ~ dinterval(t[i], t.cen[i])
        t[i] ~ dweib(r, mu[i])
        mu[i]<- exp(beta[2]+gamma[1]*(x[i]- mean(x[]))+
           gamma[2]*(y[i]-mean(y[]))+gamma[3]*(age[i]-
           mean(age[]))+gamma[4]*sex[i]+gamma[5]*ecog[i])
median[i] <- pow(log(2) * exp(-(beta[2]+gamma[1]*(x[i]-
   mean(x[]))+ gamma[2]*(y[i]-mean(y[]))+gamma[3]*(age[i]
    mean(age[]))+gamma[4]*sex[i]+gamma[5]*ecog[i])), 1/r)
        }
for(i in (N1+N2+1):(N1+N2+N3)) { #N3 is the number of
   patients receiving the new treatment, drug A
        is.censored[i] ~ dinterval(t[i], t.cen[i])
        t[i] ~ dweib(r, mu[i])
```

```
            mu[i] <- exp(beta[3]+gamma[1]*(x[i]- mean(x[]))+
              gamma[2]*(y[i]-mean(y[]))+gamma[3]*(age[i]-
              mean(age[]))+gamma[4]*sex[i]+gamma[5]*ecog[i])
    median[i] <- pow(log(2) * exp(-(beta[3]+gamma[1]*(x[i]-
      mean(x[]))+ gamma[2]*(y[i]-mean(y[]))+gamma[3]*(age[i
      ]- mean(age[]))+gamma[4]*sex[i]+gamma[5]*ecog[i])), 1/
      r)
                    }
    for(j in 1:5){gamma[j] ~ dnorm(0,0.001)}
    for(j in 1:3){beta[j] ~ dnorm(0.0, 0.001)}
#When using the commensurate prior, the beta priors change to
   :
#beta[1] ~ dnorm(0.0, 0.001)
    #beta[3] ~ dnorm(0.0, 0.001)
    #beta[2] ~ dnorm(beta[1], tau)
    #tau <- 1000 #tau ~ dgamma(1, 0.001)
r ~ dexp(1)
            #The treatment effects:
                trt.cc <- beta[3] - beta[2]
                trt.hc <- beta[3] - beta[1]
                cc.hc <- beta[2] - beta[1]
    }
```

Online Alternative: The SMEEACT Calculator

While our JAGS implementation above is fairly straightforward, a free, even-easier-to-use alternative is offered by an online hierarchical Bayesian software tool developed by staff at the M.D. Anderson Cancer Center, titled SMEEACT (an odd grant-related acronym not worth explaining here; see http://research.mdacc.tmc.edu/SmeeactWeb/ASurv.aspx) and the forthcoming related paper by Chen et al. (2018). The interface allows users to test the effectiveness of a new treatment by using information from both concurrent and previous studies in a commensurate prior formulation (Hobbs et al., 2012, 2013). The user can upload his or her own response data (though not covariate data) and then input values for four parameters that control algorithm performance: parameters that determine the degree of borrowing from the historical controls (i.e., the parameters of the hyperprior on τ), the required minimum number of events within each AIC-optimal time-axis partition, and the total number of future patients to be allocated in the current study (a feature useful only for ongoing trials). The software actually replaces our Gamma hyperprior on τ with a "spike-and-slab" distribution that spreads p_0 of its mass uniformly across the interval $(0.1, S_u)$, and places the remaining $1 - p_0$ in a single large "spike" at $K = 5000 > S_u$. Typically S_u is taken to be much smaller than K, encouraging a fairly dichotomous decision between borrowing (the spike) and no borrowing (the slab). That is, Hobbs et al. (2011, 2012) found that encouraging the procedure to lean toward either full borrowing or no borrowing (as this hyperprior does) led to slightly better operating characteristics. However, this hyperprior can be tricky to tune, unlike the simpler, more standard gamma forms used earlier in this paper. The SMEEACT dynamic borrowing algorithm uses a piecewise constant hazard model to analyze the given data, and outputs a series of plots and a table of results, which, in particular, contains the posterior summary for the treatment effect of interest and the effective historical sample size.

We set the input values for the four SMEEACT parameters as follows: the minimum number of events in each interval was set to 10, the upper slab parameter S_u was set to 200, the slab prior probability p_0 was set to 0.5, and the number of subjects left to be allocated was arbitrarily set to 100 (this last value is of no consequence to us, since the clinical trials have been completed). These values correspond to a spike-and-slab hyperprior that encourages an "all-or-nothing" decision on the amount of historical strength to borrow. Figure 4 shows the graphical results of using the SMEEACT interface on our 3-arm PFS data. The rightmost plot shows the time-to-event credible bands for the control and treatment groups having significant overlap, supplying additional evidence supporting the lack of treatment effect for Drug A. The plots indicate only minimal similarity between the historical (CRC1 and CRC2) and concurrent (CRC3) controls, similar to our previous results. SMEEACT also outputs a table (not shown) containing the treatment effect estimate, credible inter-

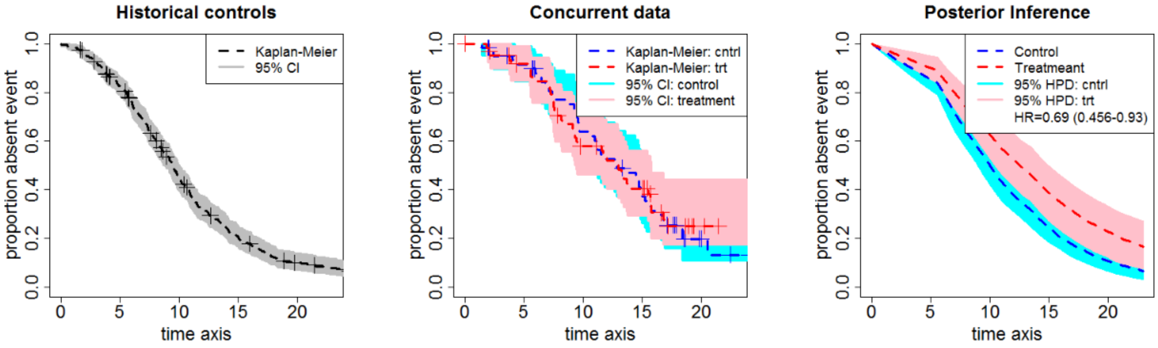

Figure 4: SMEEACT interface plots for the PFS data.

val, and allocation ratios, among other information. Included in this table is an effective historical sample size estimate of 56, or about 14% of the 412 historical controls available. This is fairly close to the EHSS of 76 that we got from the base model in Section 4 using the gamma commensurate prior for τ.