Commensurate Priors for Incorporating Historical Information in Clinical Trials Using General and Generalized Linear Models

Brian P. Hobbs[*], Daniel J. Sargent[†] and Bradley P. Carlin[‡]

Abstract. Assessing between-study variability in the context of conventional random-effects meta-analysis is notoriously difficult when incorporating data from only a small number of historical studies. In order to borrow strength, historical and current data are often assumed to be fully homogeneous, but this can have drastic consequences for power and Type I error if the historical information is biased. In this paper, we propose empirical and fully Bayesian modifications of the commensurate prior model (Hobbs et al. 2011) extending Pocock (1976), and evaluate their frequentist and Bayesian properties for incorporating patient-level historical data using general and generalized linear mixed regression models. Our proposed commensurate prior models lead to preposterior admissible estimators that facilitate alternative bias-variance trade-offs than those offered by pre-existing methodologies for incorporating historical data from a small number of historical studies. We also provide a sample analysis of a colon cancer trial comparing time-to-disease progression using a Weibull regression model.

Keywords: clinical trials, historical controls, meta-analysis, Bayesian analysis, survival analysis, correlated data

1 Introduction

1.1 Background

Clinical trials are not designed without consideration of earlier results from similar studies. Prior distributions derived from *historical data*, data from previous studies in similar populations, can be used prospectively to provide increased precision of parameter estimates. Our understanding of the "standard care" group in a trial can almost always be augmented by information derived from previous investigations. In a seminal article, Pocock (1976) considers incorporating historical control data into clinical trial analysis given that it satisfies six "acceptability" conditions. Conventionally, acceptable evidence from multiple trials is synthesized using random-effects meta-analyses (Spiegelhalter et al. 2004, p.268). Such borrowing of strength to assess "population averaged" effects in the full comparative evaluation of a new treatment has long been encouraged by the Center for Devices and Radiological Health (CDRH) at the U.S. Food and Drug Administration (FDA); see http://www.fda.gov/cdrh/osb/guidance/1601.html.

---

[*]Department of Biostatistics, M.D. Anderson Cancer Center, Houston, TX, bphobbs@mdanderson.org

[†]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, sargent.daniel@mayo.edu

[‡]Division of Biostatistics, University of Minnesota, Minneapolis, MN, brad@biostat.umn.edu

However, for the case of just a few historical studies, this approach is overly sensitive to the hyperprior distribution on the variance parameter that controls the amount of cross-study borrowing. Furthermore, with only one historical study, assessing the uncertainty of the between-study variability is difficult (Spiegelhalter 2001; Gelman 2006). Therefore, implementing the conventional meta-analytic approach to borrow strength from only one historical study requires informative prior distributions that may have drastic consequences for power and Type I error.

In this paper we propose empirical and fully Bayesian modifications to the "commensurate prior" approach (Hobbs et al. 2011) and extend the method to regression analysis using general and generalized linear regression models, in the context of two successive clinical trials. Throughout the paper we assume that the current trial compares a novel intervention to a previously studied control therapy that was used in the first trial, and thus historical data are available *only* for the control group. Furthermore, commensurate priors are constructed to inform about fixed regression effect parameters.

The goal of our proposed methodology is to formulate Bayesian hierarchical models that facilitate more desirable bias-variance trade-offs than those offered by pre-existing methodologies for incorporating historical data from a small number of historical studies. Sutton and Abrams (2001) consider empirical Bayesian methods in meta-analysis. Alternative solutions include "robust" Cauchy priors (Fúquene et al. 2009), meta-analytic-predictive methods (Neuenschwander et al. 2010), and power priors (Ibrahim and Chen 2000; Neelon and O'Malley 2010). Proper implementation of power prior models for non-Gaussian data requires formidable numerical computation that may prohibit their use in practice for clinical trial design. Therefore, we do not consider the commensurate power prior methodology proposed by Hobbs et al. (2011).

1.2   Connection to meta-analysis

Before developing our method, we briefly discuss the conventional random-effects meta-analytic approach for incorporating historical data. Let $\boldsymbol{y}$ denote a vector of i.i.d. responses of length n from patients enrolled in a current trial, such that $y_i \overset{iid}{\sim} N(\mu + d_i\lambda, \sigma^2)$ where d_i is an indicator of novel treatment. Suppose that we have patient-level data for patients assigned to the current control arm from H historical trials. Let $\boldsymbol{y}_{0,h}$ denote response vectors of length $n_{0,h}$ for the historical data, $\boldsymbol{y}_{0,h} \sim N(\mu_{0,h}, \sigma_{0,h}^2)$, where $h = 1, ..., H$. Suppose that current trial's objective is to compare the novel treatment to the previously studied control, and thus the posterior distribution of λ is of primary interest for treatment evaluation. Historical data is incorporated for the purpose of facilitating more precise estimates of μ and λ.

The conventional random-effects meta-analytic approach for borrowing strength from the historical data (see e.g. Spiegelhalter et al. 2004, p.268) assumes that $\mu_{0,1}$, ..., $\mu_{0,H}$, and μ are exchangeable:

$$\mu_{0,1}, ..., \mu_{0,H}, \mu \sim N(\xi, \eta^2).$$

The model allows for both between-study *heterogeneity* and within-study variability.

Parameters ξ and η^2 characterize the population mean and between-study variance, respectively. The estimate of ξ is a weighted average of the observed historical and current study effects, with weights $1/(\sigma^2 + \eta^2), 1/(\sigma_{0,1}^2 + \eta^2), ..., 1/(\sigma_{0,H}^2 + \eta^2)$.

The estimates of the $\mu_{0,h}$ and μ are "shrunk" toward ξ by an amount depending on the relative between-study and within-study variances. Following Spiegelhalter et al. (2004, p.94), $B = \sigma^2/(\sigma^2 + \eta^2)$ controls the amount of shrinkage of the estimate of μ towards ξ. Relatively small values of η^2 suggest that the data provide little evidence for heterogeneity with respect to the effect of control among the trial populations. This results in more borrowing of strength from the historical data. Fixing $\eta^2 = 0$ induces a model that assumes "full homogeneity."

Often the data provides sufficient information to estimate location parameters μ, $\mu_{0,h}$, ξ, and within-study variances $\sigma_{0,h}^2$ and σ^2 using common noninformative prior distributions, $h = 1, ..., H$. Table 1 lists common noninformative and "weakly informative" prior distributions for η^2 suggested by Spiegelhalter et al. (2004, p.170), Gelman (2006), and Daniels (1999).

Table 1: Common priors for η^2

prior	form
uniform variance	$p(\eta^2) = U(0, a)$
inverse gamma	$p(\eta^2) = \Gamma^{-1}(\epsilon, \epsilon)$
uniform standard deviation	$p(\eta) = U(0, \sqrt{a})$
half-Cauchy	$p(\eta) \propto (\eta^2 + b)^{-1}$
uniform shrinkage	$p(\eta^2) \propto \sigma^2/\{(\sigma^2 + \eta^2)^2\}, \sigma_{0,h}^2 = \sigma^2$, for all h

The first option for $p(\eta^2)$ considered by Spiegelhalter et al. (2004) is a uniform prior distribution with a relatively large range ($a = 100$). Gelman (2006) does not recommend it because it tends to unduly favor higher values, resulting in excessive heterogeneity. Spiegelhalter et al. (2004) also consider the inverse gamma prior for η^2 with both hyperparameters small, say $\epsilon = 0.001$. This prior distribution is often used because it is proper and conditionally conjugate. However, it is sharply peaked near zero and thus induces strong prior preference for homogeneity. As detailed in Gelman (2006), inferences with this prior are sensitive to the choice of ϵ for datasets in which homogeneity is feasible, and in the limit ($\epsilon \to 0$) results in an improper posterior density.

Both authors consider a uniform density on η, which is equivalent to $p(\eta^2) \propto 1/\eta$ (Gelman 2006). Assuming a uniform prior on the scale of η facilitates more homogeneity. However, for small H this prior tends overestimate heterogeneity. Gelman (2006)
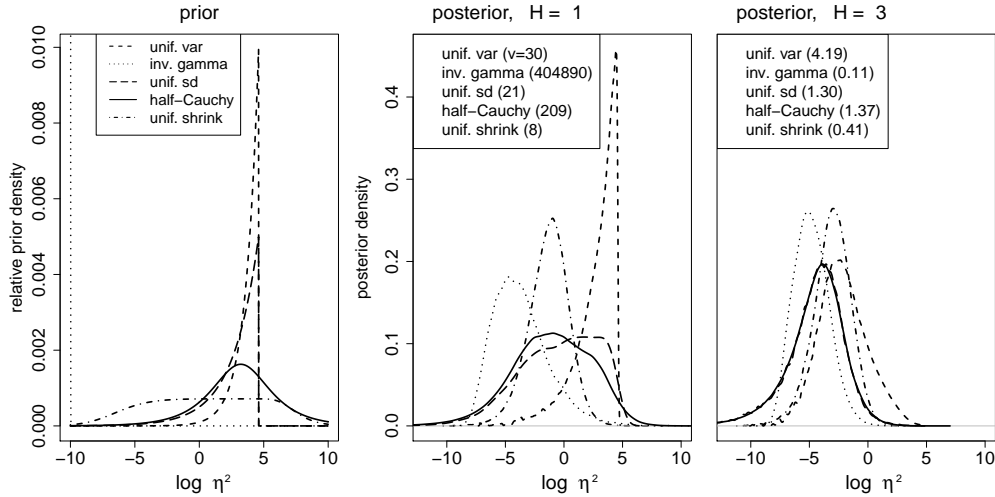
Figure 1: Prior distribution for $\log(\eta^2)$ and posterior distributions for $\log(\eta^2)$ for truly unbiased historical data: $\Delta_1^{tr} = ... = \Delta_H^{tr} = 0$, for $H = 1$ and $= 3$ historical studies. Values in parentheses are approximate posterior standard deviations on the scale η^2.

proposes a "weakly informative" half-Cauchy prior distribution on η with scale parameter, b, as a sensible compromise between the inverse gamma and uniform priors. For large values of b, (e.g. 25), this family of prior distributions has better behavior near 0, compared to the inverse gamma family; gentle slopes in the tails constrain the posterior away from large values and allow the data to dominate.

Daniels (1999) derives properties of the proper uniform shrinkage prior, which is equivalent to assuming a $U(0,1)$ prior on the shrinkage parameter, B. One attractive property of "uniform shrinkage" is that the density is maximized at zero, but less sharply peaked compared to the inverse gamma family. However, in this context, it requires *identical* within-study variances $\sigma_{0,h}^2 = \sigma^2$, for $h = 1, ..., H$, which is an undesirable assumption.

Denote the difference between the current and hth historical intercept, or *unknown bias* (Pocock 1976), by $\Delta_h = \mu - \mu_{0,h}$, and let $\boldsymbol{\Delta} = (\mu - \mu_{0,1}, ..., \mu - \mu_{0,H})$. Denote the parameter vector $\boldsymbol{\theta} = (\mu, \lambda, \boldsymbol{\Delta}, \sigma^2, \sigma_{0,1}^2, ..., \sigma_{0,H}^2)$, let $\mathcal{D} = (\boldsymbol{y}, \boldsymbol{y}_0)$ denote the collection of current and historical response data, where $\boldsymbol{y}_0 = (\boldsymbol{y}_{0,1}, ..., \boldsymbol{y}_{0,H})$, and let $L(\mathcal{D}|\boldsymbol{\theta})$ denote the joint likelihood. Figure 1 illustrates the propriety of each of the aforementioned prior distributions for η^2 in this context on the scale of $\log(\eta^2)$. The left plot portrays the relative prior densities, $p\{\log(\eta^2)\} \propto \eta^2 p(\eta^2)$. The plot illustrates each prior's relative proclivity for homogeneity. In this context, assuming η^2 is uniform

over the interval $[0, 100]$ induces a preference for heterogeneity, while the excessive tail behavior of the inverse gamma prior, which is characterized by a vertical dotted line on the far left side, induces a preference for homogeneity. The other three alternatives induce varying compromises. The center and right plots of Figure 1 contain marginal posterior distributions for $\log(\eta^2)$ for truly unbiased historical data for $H = 1$ and $= 3$ historical trials. Specifically, the plots depict

$$q\{\log(\eta^2)|\boldsymbol{\theta}^{tr}\} \propto \int_{\mathcal{D}} \left\{ \int_{\boldsymbol{\theta}} \eta^2 p(\eta^2, \boldsymbol{\theta}) L(\mathcal{D}|\boldsymbol{\theta}) d\boldsymbol{\theta} \right\} L(\mathcal{D}|\boldsymbol{\theta}^{tr}) d\mathcal{D},$$

where $\boldsymbol{\theta}^{tr}$ contains fixed $\boldsymbol{\Delta}^{tr} = \mathbf{0}$. We also set true parameters $\lambda^{tr} = 0$ and $\sigma^{tr} = \sigma_{0,1}^{tr} = ... = \sigma_{0,H}^{tr} = 1$; fixed historical sample sizes $n_{0,h} = 60$, and current sample size $n = 180$, and assumed equal allocation of patients to treatment and control in the current trial. The log-transformation of η^2 facilitates comprehensive characterization of the distributions over a vast portion of the parameter space. However, given the asymmetry of the parameter space (i.e., in this context homogeneity is realized for values < -3) it is difficult to assess uncertainty on the log-scale. Thus, legends in the center and right plots of Figure 1 provide approximate posterior standard deviations on the scale of η^2 in parentheses.

Figure 1 illustrates the primary drawback to using meta-analysis to incorporate historical data from one historical study. Looking from the left to center plots reveals that relatively little Bayesian updating of $p\{\log(\eta^2)\}$ has occurred despite the fact that the historical data is unbiased. The uniform prior on η^2 is relatively unchanged, while the inverse gamma prior results in a posterior that covers a wide range of homogeneity, yet it is still very diffuse given that its posterior standard deviation on the η^2 scale is greater than $400,000$. The center plot reaffirms the aforementioned authors' preferences for the uniform standard deviation, half-Cauchy, or uniform shrinkage priors, which facilitate sensible compromises. Yet, even these posteriors are still diffuse, facilitating little borrowing of strength from the historical data in this setting. The right plot illustrates that for $H = 3$, a relative "convergence" to the preference for homogeneity begins to emerge, although the priors still systematically influence the relative degree of homogeneity.

The remainder of the paper proceeds as follows. Section 2 introduces our proposed commensurate prior models and evaluates their frequentist properties for Gaussian data. Section 3 introduces general linear and general linear mixed models for Gaussian data. Then in Section 4 we expand the method to include non-Gaussian responses for generalized linear and generalized linear mixed models. Section 5 offers an illustrative time-to-event analysis that demonstrates the benefit of our proposed method, while Section 6 evaluates the frequentist and Bayesian operating characteristics of our method using simulation. Finally, Section 7 concludes, discusses our findings, and suggests avenues for further research.

2   Commensurate prior models

Hobbs et al. (2011) consider the simple case involving incorporation of data from one historical trial into the analysis of a single-arm trial. The authors define the location commensurate prior for μ to be the product of the historical likelihood and a normal prior on μ with mean μ_0 and precision or "commensurability parameter" τ. The general formulation follows from Pocock (1976), who suggested that historical parameters are biased representations of their concurrent counterparts. Pocock (1976) also suggested that models for incorporating historical information must account for unknown bias $\Delta = \mu - \mu_0$ in the historical data. The commensurate prior is essentially a structural prior distribution that describes the extent to which a parameter in a new trial varies about the analogous parameters in a set of historical trials when the direction of the bias is unknown. The approach assumes that the current analysis should borrow strength from the historical data in the absence of evidence for heterogeneity. Thus, lack of evidence for large absolute bias, $|\Delta|$, relative to the data's informativeness, implies commensurability. A one-to-one relationship exists between the commensurability parameter and the between-study variance parameter η^2 for the random-effects meta-analytic models discussed in the previous section for the case of one historical study. Pocock (1976) proposes repeated analysis under several fixed values of $1/\tau$, while Hobbs et al. (2011) propose a fully Bayesian approach that assumes a diffuse uniform prior distribution on $\log(\tau)$.

There are two issues with the pre-existing formulation. First, the diffuse prior of Hobbs et al. (2011) on $\log(\tau)$ (an attempt at objectivity) is actually quite informative, in that it strongly favors either full homogeneity or heterogeneity (on the scale of η, nearly 83% of the *a priori* probability is placed on values less than 0.05 or larger than 10, the effective range for substantial to very little shrinkage). Second, the historical likelihood should perhaps more properly be considered *data* instead of a component of the prior. Nevertheless, Hobbs et al. (2011) demonstrate that analysis using commensurate priors may lead to more powerful procedures than an analysis that ignores the historical data, even when Type I error is controlled at 0.05. Henceforth, we proceed with two modifications of the preceding commensurate prior methodology. We consider the historical data to be a part of the likelihood, and propose new empirical and fully Bayesian modifications for estimating τ from the data. Our proposed commensurate prior models lead to preposterior admissible estimators that facilitate alternative bias-variance trade-offs from those offered by pre-existing methodologies for incorporating historical data from a small number of historical studies. The proposed methodology provides the most gains for the case of only one historical study. Before applying this methodology in practice we must assess frequentist properties in the context of other important factors, such as potential disparities in sample size among the historical and current studies.

2.1 One historical study

As demonstrated in Subsection 1.2, in a random-effects meta-analysis it is difficult to estimate η^2 if the number of historical studies is small. As illustrated in Figure 1, for only one historical trial in the above context, the data provide very little information about η^2. Yet, if evidence regarding the efficacy and safety profile of the current control arm derives from a single study because data is exceptionally expensive, the patient population is sparse, or the therapy is unusually hazardous, then it may be highly desirable to facilitate more borrowing of strength from the existing evidence. This is especially true if the trial implements adaptive decision rules that are designed to minimize the number of patients that are exposed to the inferior treatment. Consequently, in practice inference often proceeds under the assumption of full homogeneity, $\eta^2 = 0$, which violates Pocock's (1976) proposition that one must allow for unknown bias in the historical controls. Assuming homogeneity yields designs with undesirable frequentist operating characteristics.

Let $\boldsymbol{\theta}$ denote the parameter vector $\boldsymbol{\theta} = (\lambda, \mu, \mu_0, \sigma^2, \sigma_0^2)$. For $H = 1$, the joint posterior distribution under the commensurate prior model is proportional to

$$q(\boldsymbol{\theta}|\tau, \boldsymbol{y}, \boldsymbol{y}_0) \propto N(\mu|\,\mu_0, 1/\tau)p(\mu_0)p(\sigma, \sigma_0) \prod_{j=1}^{n_0} N(y_{0j}|\,\mu_0, \sigma_0^2) \prod_{i=1}^{n} N(y_i|\,\mu + d_i\lambda, \sigma^2). \quad (1)$$

Throughout this paper we refer to $\mu \sim N(\mu|\,\mu_0, 1/\tau)$ as the *commensurate prior*, and to $p(\mu_0)$ as the *initial prior*, since it characterizes information before the historical data was observed. The commensurate prior assumes that μ is a non-systematically biased representation of μ_0. For one historical study a one-to-one relationship exists between τ and the meta-analytic between-study variance, η^2: $\tau = 1/(2\eta^2)$. Larger values of τ indicate increased commensurability, and induce increased borrowing of strength from the historical data.

Define $n_d = \sum_{i=1}^{n} d_i$, $\bar{y}_d = \sum_{i=1}^{n} d_i y_i/n_d$, $s_0^2 = \sum_{j=1}^{n_0}(y_{0,j}-\bar{y}_0)^2$, $s^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2$, $\hat{\mu} = \frac{n\bar{y}-n_d\bar{y}_d}{n-n_d}$, $\hat{\lambda}_\mu = \sum_{i=1}^{n} d_i(y_i-\mu)/n_d$, $\hat{\mu}_{0,\mu} = \left(\frac{\bar{y}_0}{\sigma_0^2/n_0} + \frac{\mu}{1/\tau}\right)/\left(\frac{n_0}{\sigma_0^2}+\tau\right)$, and $\hat{\mu}_\tau = \left\{\frac{\hat{\mu}}{\sigma^2/(n-n_d)} + \frac{\bar{y}_0}{\sigma_0^2/n_0+1/\tau}\right\}/\left(\frac{n-n_d}{\sigma^2}+\frac{1}{\sigma_0^2/n_0+1/\tau}\right)$. If we assume a flat prior for μ_0 (no initial information), reference priors for the within-study variances, $p(\sigma^2, \sigma_0^2) = 1/(\sigma^2\sigma_0^2)$, the joint posterior in (1) can be written as

$$q(\boldsymbol{\theta}\,|\,\tau, \boldsymbol{y}, \boldsymbol{y}_0) \propto N\left(\lambda\,\Big|\,\hat{\lambda}_\mu, \sigma^2/n_d\right) N\left\{\mu_0\,\Big|\,\hat{\mu}_{0,\mu}, \left(\frac{n_0}{\sigma_0^2}+\tau\right)^{-1}\right\}$$

$$\times N\left[\mu\,\Big|\,\hat{\mu}_\tau, \left\{\frac{n-n_d}{\sigma^2}+\left(\frac{\sigma_0^2}{n_0}+\frac{1}{\tau}\right)^{-1}\right\}^{-1}\right] N\left(\hat{\mu}-\bar{y}_0\,\Big|\,0, \frac{\sigma^2}{n-n_d}+\frac{\sigma_0^2}{n_0}+\frac{1}{\tau}\right)$$

$$\times \Gamma^{-1}\left(\sigma_0^2\,\Big|\,\frac{n_0-3}{2}, s_0^2/2\right)\Gamma^{-1}\left\{\sigma^2\,\Big|\,n/2-2, \frac{s^2-1/(1/n-1/n_d)(\bar{y}-\bar{y}_d)^2}{2}\right\},$$

$$(2)$$

where Γ^{-1} denotes the inverse gamma distribution. We can reduce dimensionality in the numerical marginalization of $\boldsymbol{\theta}$ by replacing the product of the initial prior and the historical likelihood with a normal approximation; Section 4 discusses this approach in generality for generalized linear models. Subsection 2.3 proposes fully Bayesian as well as empirical approaches for estimating τ.

2.2   Multiple historical studies

For $H > 1$ historical studies, we propose an extension of the commensurate prior model in the previous subsection that assumes homogeneity among the historical studies by formulating the commensurate prior for μ conditional on the historical population mean. Usually data from only a few historical studies will be available. If the historical studies are markedly heterogenous, one may need to consider whether the historical data satisfies Pocock's "acceptability" conditions (Pocock 1976), or fix the degree of heterogeneity to acknowledge the conflict.

Following our previous notation, the parameter vector contains one parameter for the historical population mean, μ_0: so that $\boldsymbol{\theta} = (\lambda, \mu, \sigma^2, \mu_0, \sigma_{0,1}^2, ..., \sigma_{0,H}^2)$, where $\mu_{0,1} = \cdots = \mu_{0,H} = \mu_0$. Denote the hth historical within-study variance divided by the sample size by $\omega_{0,h} = \sigma_{0,h}^2/n_{0,h}$. The relationship between τ and the meta-analytic between-study variance η^2 is more complicated than before. For $H > 1$, τ^{-1} characterizes the meta-analytic between-study variability, plus the difference between the summed variability among the sample means $\bar{y}$ and the $\bar{y}_{0,h}$s, and the population mean (previously ξ) when heterogeneity is estimated η^2 versus when full homogeneity is assumed,

$$\tau = \left[\eta^2 + \left\{ 1 \Big/ \left(\frac{\sigma^2}{n - n_d} + \eta^2 \right) + \sum_{h=1}^{H} 1/(\omega_{0,h} + \eta^2) \right\}^{-1} \right.$$
$$\left. - \left(\frac{n - n_d}{\sigma^2} + \sum_{h=1}^{H} 1/\omega_{0,h} \right)^{-1} \right]^{-1}.$$

Moreover, constraining the H historical means, $\mu_{0,h}$, to be equal to each other but perhaps not to the current mean, μ inserts an asymmetry into the model that is not present in the usual exchangeability model. As in the previous section, the data typically provide sufficient information to estimate parameters λ, μ_0, σ, and $\sigma_{0,h}$ (given no initial information) under common noninformative prior distributions, $h = 1, ..., H$. The joint posterior distribution is proportional to

$$q(\boldsymbol{\theta}|\tau, \boldsymbol{y}, \boldsymbol{y}_0) \propto N(\mu|\, \mu_0, 1/\tau) p(\mu_0) p(\lambda, \sigma, \sigma_{0,1}, ..., \sigma_{0,H_0})$$
$$\times \prod_{h=1}^{H} \prod_{j=1}^{n_{0,h}} N(y_{0,hj}|\, \mu_0, \sigma_{0,h}^2) \prod_{i=1}^{n} N(y_i|\, \mu + d_i\lambda, \sigma^2). \quad (3)$$

The joint posterior distribution follows from (2) by replacing σ_0^2/n_0 with $v_0 = (\sum_{h=1}^{H} 1/\omega_{0,h})^{-1}$, $\bar{y}_0$ with $v_0 \left(\sum_{h=1}^{H} \bar{y}_{0,h}/\omega_{0,h} \right)$, and $\Gamma^{-1}\left(\sigma_0^2 \,\middle|\, \frac{n_0-3}{2}, s_0^2/2 \right)$ with

$\prod_{h=1}^{H} \Gamma^{-1}\left(\sigma_{0,h}^2 \,\Big|\, \frac{n_{0,h}-3}{2}, s_{0,h}^2/2\right)$, where $s_{0,h}^2 = \sum_{j=1}^{n_{0,h}}(y_{0,h,j} - \bar{y}_{0,h})^2$.

2.3   Estimation of τ

In this subsection we propose empirical and fully Bayesian methods for estimating τ, and evaluate frequentist properties of the corresponding models for estimating λ in the context of our proposed commensurate prior model.

Empirical Bayesian

In this subsection we consider parametric empirical Bayesian (EB) estimation (see e.g. Morris 1983; Kass and Steffey 1989; Carlin and Louis 2009) in the context of our proposed commensurate prior models. Define $\hat{\mu}_0 = v_0\left(\sum_{h=1}^{H} \bar{y}_{0,h}/\omega_{0,h}\right)$. For known σ^2 and v_0, the marginal distribution of the current and historical data given hyperparameter τ, $m(\boldsymbol{y}, \boldsymbol{y}_0|\tau) = \int_{\boldsymbol{\theta}} q(\boldsymbol{\theta}|\tau, \boldsymbol{y}, \boldsymbol{y}_0)d\boldsymbol{\theta}$, follows from (2) and (3) as

$$m(\boldsymbol{y}, \boldsymbol{y}_0|\tau) \propto N\left\{\hat{\mu} - \hat{\mu}_0 \,\Big|\, 0, \frac{\sigma^2}{n - n_d} + v_0 + \frac{1}{\tau}\right\}. \tag{4}$$

EB inference for $\boldsymbol{\theta}$ proceeds by replacing the scalar hyperparameter τ in (2) and (3) with its marginal maximum likelihood estimate (MMLE). Larger values of τ indicate an increasing lack of empirical evidence for heterogeneity, and lead to increased borrowing of strength from the historical data. Figure 2 contains marginal distributions of the data in (4) for three observed values of $|\hat{\Delta}|$, where $\hat{\Delta} = \hat{\mu} - \hat{\mu}_0$ under the scenario that produced Figure 1 for one historical study. For sufficiently small $|\hat{\Delta}|$, $m(\boldsymbol{y}, \boldsymbol{y}_0|\tau)$ is monotonically increasing as a function of τ, evident in the left plot corresponding to the case when $\hat{\Delta} = 0$. It follows that $\arg\max_{\tau>0} m(\boldsymbol{y}, \boldsymbol{y}_0|\tau) = \infty$, if $\hat{\Delta}^2 < \frac{\sigma^2}{n-n_d} + v_0$. The center and right plots reveal that larger values of $|\hat{\Delta}|$, yield more peaked, unimodal functions.

Let $\nu = \tau^{-1}$. We propose fixing ν^* at the value that maximizes the marginal density of the data in (4), restricted to a pre-specified interval capturing the effective range of borrowing of strength. This leads to the following estimate for ν:

$$\nu^* = \arg\max_{\nu \in [l_\nu, u_\nu]} \{m(\boldsymbol{y}, \boldsymbol{y}_0 \mid 1/\nu)\} = \max\left[\min\left\{\hat{\Delta}^2 - \frac{\sigma^2}{n - n_d} - v_0, \ u_\nu\right\}, \ l_\nu\right], \tag{5}$$

where $0 < l_\nu < u_\nu$. Bounding the MMLE precludes full homogeneity when evidence for heterogeneity is not strong. That is, decreasing (increasing) l_ν will induce more (less) borrowing of strength from the historical data when unrestricted maximization results in infinite $1/\nu^*$. Restricting $\nu^* \in [2(0.05^2), 2(10^2)]$, which corresponds to $\eta \in [0.05, 10]$ for one historical study, usually captures the effective range of borrowing of strength. However, these limits should be selected via formal evaluation of the induced frequentist operating characteristics and bias-variance trade-offs in context.

Let $\mathcal{D} = (\boldsymbol{y}, \boldsymbol{y}_0)$ denote the current and historical data, and $\boldsymbol{\theta}^{tr} = (\lambda^{tr}, \ \Delta^{tr},$
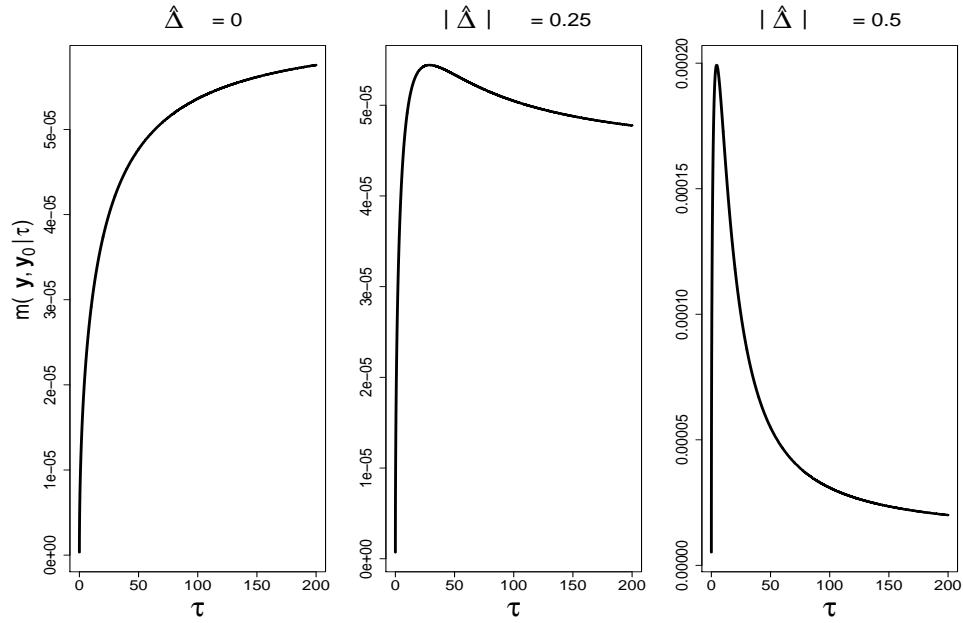
Figure 2: Marginal density of $\boldsymbol{y}, \boldsymbol{y}_0 \mid \tau$ (4) as a function of τ for three values of $|\hat{\Delta}|$, where $\hat{\Delta} = \hat{\mu} - \hat{\mu}_0$ and $\sigma^2 = 1$, $v_0 = 1/n_0$, $n_0 = 60$, $n = 180$, $n_d = 90$.

μ^{tr}, $(\sigma^{tr})^2$, $v_0^{tr})$ denote a set of fixed parameters characterizing a true state of the model. Define $Z^2 = \left(\hat{\Delta}^2\right) / \left(\frac{\sigma^2}{n - n_d} + v_0\right)$. Given $\boldsymbol{\theta}^{tr}$, Z^2 follows a non-central χ^2 distribution, $Z^2 | \boldsymbol{\theta}^{tr} \sim \chi_1^2 \left\{ (\Delta^{tr})^2 / \left(\frac{(\sigma^{tr})^2}{n - n_d} + v_0^{tr} \right) \right\}$. Let $F_{Z^2 | \boldsymbol{\theta}^{tr}}(\cdot)$ denote its cumulative density function. The probability that ν^* is fixed at l_ν in the EB inference of $\boldsymbol{\theta}$ is $F_{Z^2 | \boldsymbol{\theta}^{tr}} \left(\frac{l_\nu}{\frac{(\sigma^{tr})^2}{n - n_d} + v_0^{tr}} + 1 \right)$.

Various maximization techniques can be used to estimate ν^* for the intractable case when variances are unknown, including Markov chain Monte Carlo (MCMC) methods considered by Geyer and Thompson (1992) and Doucet et al. (2002). EB inference typically "underestimates" variability in $\boldsymbol{\theta}$, since posterior uncertainty in ν^* is unacknowledged in the analysis. However, in the following sections we demonstrate that our EB procedure has several desirable properties when compared to conventional random effects meta-analytic models.

Fully Bayesian

The EB procedure yields approximate full homogeneity when evidence for heterogeneity is not strong (although this can be adjusted via careful selection of the lower bound, l_ν). However, even for this scenario approximate full homogeneity may not be

warranted. In this subsection we discuss fully Bayesian estimation in the context of our proposed commensurate prior models. By accounting for prior uncertainty when estimating τ, the fully Bayesian approach takes full account of uncertainty in the parameter estimates. Specifically, we consider two families of priors for τ, a conditionally conjugate gamma distribution, as well as a variant of the "spike and slab" distribution introduced by Mitchell and Beauchamp (1988) for variable selection.

For known σ^2 and v_0, the marginal posterior distribution of $\tau | \boldsymbol{y}, \boldsymbol{y}_0$ follows from (2) and (3) as

$$q(\tau | \boldsymbol{y}, \boldsymbol{y}_0) \propto \tau N \left\{ \hat{\Delta} \, \bigg| \, 0, \, \frac{\sigma^2}{n - n_d} + v_0 + \frac{1}{\tau} \right\} p(\tau). \tag{6}$$

Assuming a flat prior on τ clearly leads to an improper posterior, e.g. $\lim_{\tau \to \infty} q(\tau | \boldsymbol{y}, \boldsymbol{y}_0)$ $\left(\sqrt{a + 1/\tau} \right) / \tau = \exp \left\{ -0.5 \hat{\Delta}^2 / a \right\}$ and for $a > 0$, $\int_0^\infty \tau \left(\sqrt{a + 1/\tau} \right)^{-1} d\tau$ is divergent.

A default choice for $p(\tau)$ is the gamma family of distributions, $p(\tau) = \Gamma(c\tilde{\tau}, c)$, since, from (1) and (3), it leads to the following conjugate full conditional posterior distribution,

$$q(\tau | \boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{y}_0) \propto \Gamma \left(c\tilde{\tau} + 1/2, \, \Delta^2/2 + c \right). \tag{7}$$

For this parameterization $\tilde{\tau} > 0$ can be thought of as a prior guess at τ, while scalar $c > 0$ represents degree of confidence, with a smaller value corresponding to weaker prior belief.

We also propose an alternative prior for τ that derives from the aptly named "spike and slab" distribution (Mitchell and Beauchamp 1988). As the nomenclature suggests, the distribution is locally uniform between two limits, $0 \leq \mathcal{S}_l < \mathcal{S}_u$ except for a bit of probability mass concentrated at point $\mathcal{K} > \mathcal{S}_u$. Let p_0 denote the prior probability that $\mathcal{S}_l \leq \tau \leq \mathcal{S}_u$. Then, formally, the prior assumes

$$Pr(\tau < \mathcal{S}_l) = 0,$$
$$Pr(\tau < u) = p_0 \left\{ (u - \mathcal{S}_l) / (\mathcal{S}_u - \mathcal{S}_l) \right\}, \, \mathcal{S}_l \leq u \leq \mathcal{S}_u$$
$$\text{and } Pr(\tau > \mathcal{S}_u) = Pr(\tau = \mathcal{K}) = 1 - p_0. \tag{8}$$

The reason that the spike and slab approach is appropriate in this context derives from the fact that sufficiently small values of $\hat{\Delta}$ result in marginal densities of the data (4) with nearly flat, gradually decreasing slopes as functions of τ, for sufficiently large τ. This is illustrated in the left plot of Figure 2. Therefore, given little evidence for heterogeneity, the marginalized likelihood prefers a large value for τ, but is virtually flat over a vast portion of the parameter space, providing little information to distinguish among values. This suggests a sensible approach may be to choose one carefully selected large value of τ (a "spike"), that characterizes commensurability. We demonstrate in the following sections that, when properly calibrated, this prior yields desirable frequentist properties. While the fully Bayesian methods are at less risk of dramatically overestimating τ, all of our methods are somewhat subjective and can therefore produce poor results if the model is incorrectly specified (the usual Bayesian "good model"

assumption), thus making model checking an important component of this approach in practice.

Point estimation of λ

In this subsection we evaluate the proposed empirical and fully Bayesian commensurate prior models for estimating λ under squared error loss (SEL) and compare results to the full homogeneity model and the associated "no borrowing" model that ignores the historical data completely. Recall that we have assumed that the current trial's objective is to compare a novel treatment to the previously studied control therapy. Thus, posterior inference on the novel treatment effect parameter, λ, is of primary interest. Let $\mathcal{D} = (\boldsymbol{y}, \boldsymbol{y}_0)$, and define

$$V_\tau = \left[\left\{ \left(\frac{n}{n_d} \right)^2 \left(\frac{\sigma^2}{n} + v_0 + \frac{1}{\tau} \right) \right\}^{-1} + \frac{n_d(1 - n_d/n)}{\sigma^2} \right]^{-1}, \tag{9}$$

and $\hat{\lambda}_\tau = \left\{ \frac{\bar{y} - \hat{\mu}_0}{\left(\frac{n}{n_d} \right) \left(\frac{\sigma^2}{n} + v_0 + \frac{1}{\tau} \right)} + \frac{\bar{y}_d - \bar{y}}{\sigma^2/n_d} \right\} / V_\tau^{-1}. \tag{10}$

For known σ^2 and v_0, it follows from (2) and (3) that the marginal posterior distribution of $\lambda, \tau | \mathcal{D}$ can be represented as the following product:

$$q(\lambda, \tau | \boldsymbol{y}, \boldsymbol{y}_0) \propto \tau q(\lambda | \boldsymbol{y}, \boldsymbol{y}_0, \tau) m(\boldsymbol{y}, \boldsymbol{y}_0 | \tau) p(\tau)$$

$$\propto \tau N \left(\lambda \,\middle|\, \hat{\lambda}_\tau, V_\tau \right) N \left\{ \hat{\Delta} \,\middle|\, 0, \frac{\sigma^2}{n - n_d} + v_0 + \frac{1}{\tau} \right\} p(\tau). \tag{11}$$

If the historical data are ignored, the marginal posterior for $\lambda | \boldsymbol{y}$ follows as

$$q(\lambda | \boldsymbol{y}) \propto N \left[(\bar{y}_d - \bar{y}) / (1 - n_d/n) \,,\, \sigma^2 / \{ n_d (1 - n_d/n) \} \right]. \tag{12}$$

Let $\boldsymbol{\theta}^{tr} = (\lambda^{tr}, (\sigma^{tr})^2, \Delta^{tr}, v_0^{tr})$ denote a set of fixed parameters. The preposterior risk under squared error loss (Carlin and Louis 2009, p. 433), conditional on $\boldsymbol{\theta}^{tr}$ is

$$E_{\mathcal{D}|\boldsymbol{\theta}^{tr}} \left[\left\{ E_{\lambda|\mathcal{D}}(\lambda) - \lambda^{tr} \right\}^2 \right]. \tag{13}$$

Under fully Bayesian inference, the marginal posterior expectation of $\lambda | \mathcal{D}$ follows as $E_{\lambda|\mathcal{D}}(\lambda) = \int \hat{\lambda}_\tau q(\tau | \mathcal{D}) d\tau$. EB inference sets $E_{\lambda|\mathcal{D}}(\lambda) = \hat{\lambda}_{1/\nu^*}$, where ν^* is defined in (5). The no borrowing and homogeneity models result in $E_{\lambda|\mathcal{D}}(\lambda) = \hat{\lambda}_0$ and $\hat{\lambda}_\infty$, respectively.

Preposterior risk (13) can be simulated using the following relationships derived from the conditional likelihood of $\mathcal{D}|\boldsymbol{\theta}^{tr}$,

$$\bar{y} - \hat{\mu}_0 \sim N \left\{ \Delta^{tr} + \frac{n_d}{n} \lambda^{tr}, \frac{\sigma_{tr}^2}{n} + v_0^{tr} \right\}, \tag{14}$$

and $\bar{y}_d - \bar{y} \sim N \left\{ \left(1 - \frac{n_d}{n} \right) \lambda^{tr}, \sigma_{tr}^2 \left(\frac{1}{n_d} - \frac{1}{n} \right) \right\}. \tag{15}$
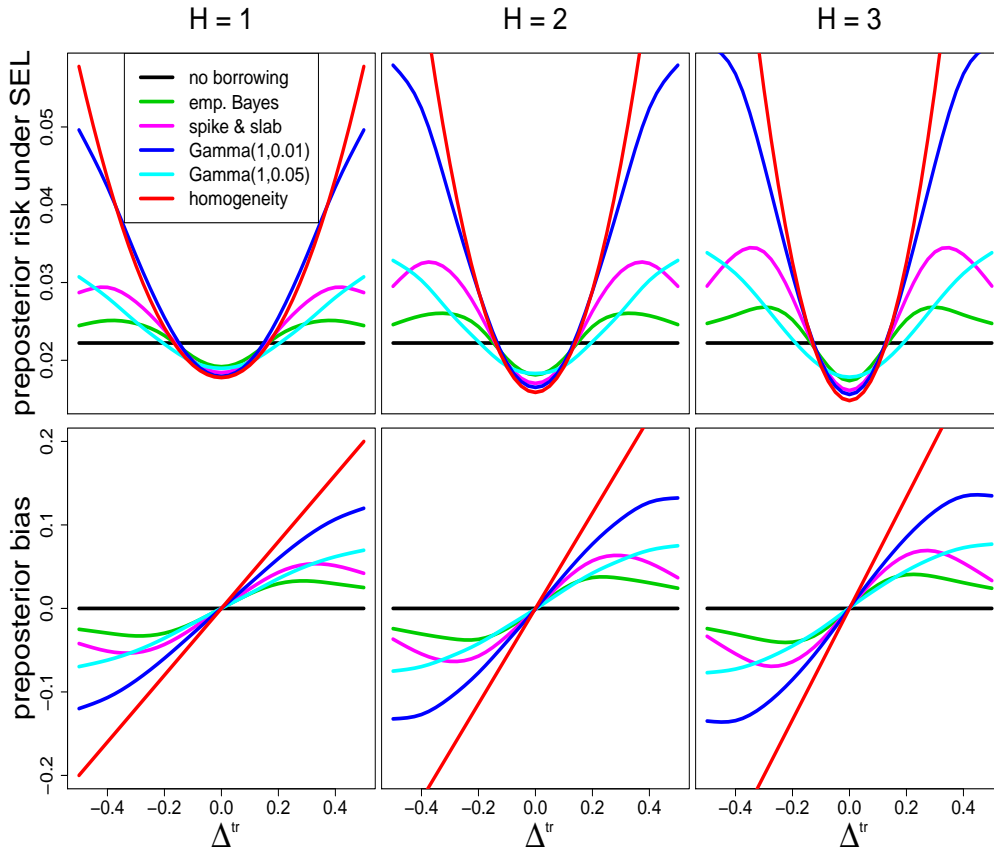
Figure 3: Preposterior risk under squared error loss and bias as functions of Δ^{tr} for $H = 1, 2$, and 3 historical studies.

Figure 3 contains preposterior risk under SEL and bias as functions of Δ^{tr} resulting from inference under no borrowing, full homogeneity, as well as empirical and fully Bayesian commensurate prior models for $H = 1$, 2, and 3 historical studies under the same scenario that produced Figure 1: $n = 180$, $n_d = 90$, $n_{0,h} = 60$, $(\sigma^{tr})^2 = 1$ and $v_0^{tr} = \left(\sum_{h=1}^{H} n_{0,h} \right)^{-1}$. Results are shown for an EB commensurate model that restricts $\nu^* \in [2(0.05^2), 2(10^2)] = [0.005, 200]$, a fully Bayesian model that assumes a spike and slab prior for τ with hyperparameters, $\mathcal{S}_l = 0.005$, $\mathcal{S}_u = 2$, $\mathcal{K} = 200$, and $p_0 = 0.99$, as well as two fully Bayesian models assuming gamma priors for τ with hyperparameters $c = \tilde{\tau}^{-1}$ and $\tilde{\tau} = 100$, and 20, respectively.

Under no borrowing the posterior mean of $\lambda | \mathcal{D}$ is unbiased. Therefore, preposterior risk under SEL for the no borrowing inference follows from (12) as $(\sigma^{tr})^2 / \{n_d (1 - n_d/n)\}$, and thus is constant as a function of Δ^{tr}. The plots reveal that all models obtain preposterior biases of zero when $\Delta^{tr} = 0$. Assuming homogeneity a priori offers maximal

variance reduction in this context, and thus is associated with the largest reductions in preposterior risk for Δ^{tr} near zero. Therefore, the resulting estimator is preposterior admissible. However, homogeneity leads to prohibitively biased estimators with sharply, monotonically increasing preposterior risk for $|\Delta^{tr}| > 0$, rapidly trumping the gains in variance reduction. The top row of plots in Figure 3 reveal that preposterior risk for homogeneity exceeds no borrowing in this context for $H = 1, 2$, and 3 when $|\Delta^{tr}| > 0.2$, > 0.16, and > 0.1, respectively.

In contrast to the no borrowing and homogeneity models, the commensurate prior models offer preposterior admissible estimators, with alternative bias variance trade-offs that facilitate more borrowing of strength for Δ^{tr} near zero and less borrowing for large values of $|\Delta^{tr}|$. Numerous alternative bias variance trade-offs are attainable via adjustment to the model hyperparameters.

2.4   Comparison to meta-analysis

This subsection compares frequentist properties of our proposed commensurate prior models (2) and (3) with results for the meta-analysis models in Section 1.2 for the scenario that produced Figure 1 for the case when sampling level variances are unknown. Tables 2 and 3 augment Figure 3 and synthesize the relative bias-variance trade-offs when using the corresponding posterior expectation of $\lambda|\mathcal{D}$ as an estimator. Table 2 contains the percent change from no borrowing in preposterior risk under SEL given $\boldsymbol{\theta}^{tr}$ for the five meta-analysis models, our proposed EB and fully Bayesian commensurate prior models, and the model that assumes full homogeneity. Negative values indicate reductions in preposterior risk. Table 3 contains the corresponding preposterior bias. Results are shown for $H = 1, 2$, and 3 historical studies and fixed true values of Δ^{tr} indicating various degrees of historical bias. Results are shown for fixed true standard deviation values $\sigma^{tr} = \sigma_{0,1}^{tr} = ... = \sigma_{0,H}^{tr} = 1$. The spike and slab prior uses the same hyperparameters that were proposed in the previous subsection with the exception of p_0 which we now adjust to 0.7. If the historical data is ignored, the marginal posterior for $\lambda|\boldsymbol{y}$ now follows as

$$q(\lambda|\boldsymbol{y}) \propto t \left\{ n - 2, (\bar{y}_d - \bar{y})/(1 - n_d/n), \frac{s^2 - 1/(1/n - 1/n_d)(\bar{y} - \bar{y}_d)^2}{n_d(n-2)(1 - n_d/n)} \right\}. \qquad (16)$$

As in the previous subsection, preposterior risk under SEL given $\boldsymbol{\theta}^{tr}$ follows as $(\sigma^{tr})^2$ $/ \left\{ n_d(1 - \frac{n_d}{n}) \right\}$. Note that marginal posteriors for $\lambda|\mathcal{D}$ corresponding to the commensurate prior models discussed in the previous subsection are now intractable due to required marginalization of the sampling variances.

First we consider the case when $\Delta^{tr} = 0$. For the case of one historical study, Table 3 shows that all estimators are unbiased, yet Table 2 reveals that the meta-analysis estimators correspond to only slight reductions in preposterior risk under SEL, with the exception of the inverse gamma prior. In contrast, the commensurate models facilitate relatively large reductions in preposterior risk compared to the first four meta-analytic models. As in Figure 3, the gamma prior facilitates the largest reductions in

Table 2: Percent change from no borrowing in preposterior risk under squared error loss for the specified model. For this scenario the result under no borrowing is equal to $(\sigma^{tr})^2/\left\{n_d(1 - \frac{n_d}{n})\right\} = 0.0222$.

	$\underline{\Delta^{tr} = 0}$			$\underline{\Delta^{tr} = 0.25}$			$\underline{\Delta^{tr} = 0.5}$		
	$H = 1$	2	3	$H = 1$	2	3	$H = 1$	2	3
unif. var.	-1	-3	-11	-1	-1	-1	0	1	4
unif. shrink	-4	-9	-14	-2	-2	-1	0	3	7
half-Cauchy	-4	-14	-21	-1	1	3	1	8	14
unif. sd	-3	-14	-21	-1	0	5	1	7	16
inv. gamma	-12	-20	-25	-1	4	9	8	20	31
emp. Bayes	-13	-17	-22	7	11	20	8	9	16
spike & slab	-13	-17	-22	1	5	11	9	10	11
Gamma(1,0.01)	-16	-22	-24	0	5	8	25	35	38
homogeneity	-19	-28	-32	25	61	86	152	337	475

Table 3: Preposterior bias for the specified model: $E_{\mathcal{D}|\boldsymbol{\theta}^{tr}}\left\{E_{\lambda|\mathcal{D}}(\lambda)\right\} - \lambda^{tr}$

	$\underline{\Delta^{tr} = 0}$			$\underline{\Delta^{tr} = 0.25}$			$\underline{\Delta^{tr} = 0.5}$		
	$H = 1$	2	3	$H = 1$	2	3	$H = 1$	2	3
unif. var.	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.02
unif. shrink	0.00	0.00	0.00	0.01	0.02	0.03	0.01	0.02	0.03
half-Cauchy	0.00	0.00	0.00	0.01	0.03	0.04	0.01	0.03	0.04
unif. sd	0.00	0.00	0.00	0.01	0.03	0.05	0.01	0.02	0.04
inv. gamma	0.00	0.00	0.00	0.03	0.05	0.07	0.03	0.05	0.07
emp. Bayes	0.00	0.00	0.00	0.03	0.04	0.05	0.02	0.02	0.03
spike & slab	0.00	0.00	0.00	0.03	0.03	0.04	0.02	0.02	0.02
Gamma(1,0.01)	0.00	0.00	0.00	0.05	0.06	0.07	0.07	0.08	0.08
homogeneity	0.00	0.00	0.00	0.10	0.14	0.17	0.19	0.28	0.34

preposterior risk, nearly approaching the reduction produced by highly subjective, full homogeneity.

When $H > 1$, inference using the uniform standard deviation and half-Cauchy priors in the meta-analysis framework provides considerably more reduction in preposterior risk than for $H = 1$. The inverse gamma prior facilitates even more reduction in preposterior risk. The uniform variance and uniform shrinkage priors offer considerably less reduction in preposterior risk. The commensurate prior models still facilitate more borrowing of strength than the first four models for meta-analysis leading to estimators with more reduction in preposterior risk, but less reduction than that obtained by full homogeneity.

When $0 < \Delta^{tr} \le 0.5$, the tables suggest that preposterior risk and bias are non-

decreasing in H for all models. The highly subjective homogeneity model yields highly biased estimators corresponding to substantial increases in preposterior risk. The gamma model estimator is most biased among the commensurate prior estimators, which leads to large increases in preposterior risk for $\Delta^{tr} = 0.5$. The spike and slab model estimator provides perhaps the best overall bias-variance trade-off among the commensurate prior estimators, given that it is least biased and results in relatively smaller increases in preposterior risk. Moreover, it provides equal or less bias than the half-Cauchy meta-analysis prior estimator when $H > 1$. Among the meta-analysis models, only the inverse gamma prior facilitates meaningful variance reduction from borrowing of strength for the case of one historical study. However, when $\Delta^{tr} = 0.5$ the inverse gamma meta-analysis prior results in an increase in relative preposterior risk that is nearly two and three times larger than that for the spike and slab commensurate prior estimator for $H = 2$ and $= 3$, respectively.

3   General linear models

In this section we introduce general linear and general linear mixed commensurate prior models for Gaussian response data in the context of two successive clinical trials. In addition, we assume that both trials identically measure $p - 1$ covariates representing fixed effects which are to be incorporated into the analysis. As before, we assume that the second (current) trial compares a novel intervention to a previously studied control therapy that was used in the first trial, and thus historical data are available *only* for the control group. Furthermore, commensurate priors are constructed to inform about fixed regression effect parameters.

3.1   Fixed effect models

Assume $\boldsymbol{y}_0$ is a vector of n_0 responses from patients in the historical study of an intervention that is to be used as a control in a current trial testing a newly developed intervention for which no reliable prior data exist. Let $\boldsymbol{y}$ be the vector of n responses from subjects in both the treatment and control arms of the current trial. Suppose that both trials are designed to identically measure $p - 1$ covariates of interest. Let X_0 be an $n_0 \times p$ design matrix and X be an $n \times p$ design matrix, both of full column rank p, such that the first columns of X_0 and X are vectors of 1s corresponding to intercepts.

Suppose $\boldsymbol{y}_0 \sim N_{n_0}(X_0\boldsymbol{\beta_0}, \sigma_0^2)$ and $\boldsymbol{y} \sim N_n(X\boldsymbol{\beta} + \boldsymbol{d}\lambda, \sigma^2)$ where λ is the (scalar) treatment effect and $\boldsymbol{d}$ is an $n \times 1$ vector of $0 - 1$ indicator variables for the new treatment. Let y_i, X_i, d_i represent data corresponding to the ith subject in the current trial, $i = 1, ..., n$. Commensurability in the linear model depends upon similarity among the intercepts *and* covariate effects. Yet, the strength of empirical evidence for heterogeneity among the current and historical data may vary across covariates. We formulate the commensurate linear model by replacing τ in Section 2 with a vector $\boldsymbol{\tau} = (\tau_1, ..., \tau_p)$ containing a commensurability parameter, τ_g, for each associated pair of parameters in β_g and β_{0g}. The commensurate priors follow as $p(\beta_g|\beta_{0g}) \propto N\left(\beta_g|\beta_{0g}, \tau_g^{-1}\right)$, for

$g = 1, ..., p$, and the β_gs are assumed *a priori* independent.

Let $\boldsymbol{\theta} = (\lambda, \boldsymbol{\beta}, \boldsymbol{\beta}_0, \sigma^2, \sigma_0^2)$ and $\boldsymbol{\Delta} = \boldsymbol{\beta} - \boldsymbol{\beta}_0$, let $diag\{\boldsymbol{u}\}$ denote the diagonal matrix consisting of the elements of vector $\boldsymbol{u}$, and let I_a denote the identity matrix of dimension a. The joint posterior distribution of $\boldsymbol{\theta}|\boldsymbol{\tau}, \boldsymbol{y}, \boldsymbol{y}_0$ follows as proportional to

$$p(\lambda, \sigma, \sigma_0) N_{n_0}(\boldsymbol{y}_0|X_0\boldsymbol{\beta}_0, \sigma_0^2 I_{n_0}) N_n(\boldsymbol{y}|X\boldsymbol{\beta} + \boldsymbol{d}\lambda, \sigma^2 I_n) \prod_{g=1}^{p} N\left(\beta_g|\beta_{0g}, \tau_g^{-1}\right) p(\beta_{0g}). \quad (17)$$

Let $\hat{\boldsymbol{\beta}}_\lambda = (X^T X)^{-1} X^T (\boldsymbol{y} - \boldsymbol{d}\lambda)$ and $\hat{\boldsymbol{\beta}}_0 = (X_0^T X_0)^{-1} X_0^T \boldsymbol{y}_0$, and $\hat{\lambda}_\beta = \sum_{i=1}^{n} d_i(y_i - X_i\boldsymbol{\beta})/n_d$. Under a flat prior, the full conditional posterior for λ is proportional to $N\left(\hat{\lambda}_\beta, \sigma^2/n_d\right)$. Let $V_{\boldsymbol{\tau}}$ denote the precision matrix that results from averaging the commensurate prior over the historical likelihood:

$$V_{\boldsymbol{\tau}} = diag\{\boldsymbol{\tau}\} - diag\{\boldsymbol{\tau}\} \left(X_0^T X_0/\sigma_0^2 + diag\{\boldsymbol{\tau}\}\right)^{-1} diag\{\boldsymbol{\tau}\}. \quad (18)$$

Thus, assuming a flat initial prior and marginalizing $\boldsymbol{\beta}_0$ results in a conditional posterior for $\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{y}_0, \lambda, \sigma^2, \sigma_0^2, \boldsymbol{\tau}$ proportional to

$$N_p\left\{\boldsymbol{\beta}\,\middle|\,\left(V_{\boldsymbol{\tau}} + \frac{X^T X}{\sigma^2}\right)^{-1}\left(V_{\boldsymbol{\tau}}\hat{\boldsymbol{\beta}}_0 + \frac{X^T X}{\sigma^2}\hat{\boldsymbol{\beta}}_\lambda\right), \left(V_{\boldsymbol{\tau}} + \frac{X^T X}{\sigma^2}\right)^{-1}\right\}. \quad (19)$$

Notice that the full conditional posterior mean for λ, $\hat{\lambda}_\beta$, is a function of residuals ($\boldsymbol{y} - X\boldsymbol{\beta}$), whereas the conditional posterior mean of $\boldsymbol{\beta}$ in (19) is an average of the historical and concurrent data relative to the estimated commensurability parameter vector $\boldsymbol{\tau}$. As $\boldsymbol{\tau}$ approaches $\boldsymbol{0}$, the posterior for $\boldsymbol{\beta}$ converges to a normal density with mean $\hat{\boldsymbol{\beta}}_\lambda$ and variance $\left(\frac{X^T X}{\sigma^2}\right)^{-1}$, recovering the standard result from linear regression that ignores all of the historical data. In this case, $\hat{\lambda}_\beta$ also converges to the "no borrowing" estimate of the treatment effect. Moreover, as τ_g approaches infinity, for all $g = 1, ..., p$, precision $V_{\boldsymbol{\tau}} \to X_0^T X_0/\sigma_0^2$, fully incorporating the historical data, recovering full homogeneity. Full conditional posteriors for β_0, σ^2, and σ_0^2 under noninformative priors are provided in Appendix A.

3.2 Mixed models

In this subsection, we extend the model to include random effects. We begin with a familiar and useful one-way ANOVA model. Then following McCulloch and Searle (2001, p.156), we give the linear mixed model for general variance-covariance structures between and within levels of the random components.

One-way random effects model

Following the notation of Browne and Draper (2006), suppose $y_{0jk} = \mu_0 + u_{0k} + \epsilon_{0jk}$, where $\epsilon_{0jk} \stackrel{iid}{\sim} N(0, \sigma_{\epsilon 0}^2)$ for $k = 1, ..., n_{0j}$, $j = 1, ..., m_0$, and $u_{0j} \stackrel{iid}{\sim} N(0, \sigma_{u0}^2)$. Variance component $\sigma_{\epsilon 0}^2$ represents the conditional variance of $y_{0jk}|u_{0j}$, while the marginal

variance of y_{0jk} follows as $\sigma_{u0}^2 + \sigma_{\epsilon0}^2$. Note that observations from different subjects are assumed to be uncorrelated. The model for responses in the current trial follows $y_{ig} = \mu + u_i + d_i\lambda + \epsilon_{ig}$, for $g = 1, ..., n_i$; $i = 1, ..., m$, where $\epsilon_{ig} \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$ and $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$. As before, $d_i = 1$ indicates treatment and $d_i = 0$ corresponds to the standard of care for the ith subject in the current trial. Therefore, fixed effects μ and λ represent the intercept and treatment effect for a patient receiving the new intervention. Given no initial information, the commensurate prior for μ is proportional to a normal distribution with mean μ_0 and precision τ.

Following the recommendations of Gelman (2006), we use independent noninformative uniform priors on σ_{u0} and the current random-effects standard deviation, for large m (≥ 5), which is equivalent to a product of inverse-χ^2 densities with -1 degrees of freedom, $p(\sigma_{u0}^2, \sigma_u^2) \propto \frac{1}{\sigma_{u0}\sigma_u}$. For small m (< 5), we use the half-Cauchy prior discussed in Section 1.2.

Let $\boldsymbol{\theta}$ denote the parameter vector. The joint posterior distribution for $\boldsymbol{\theta}|\tau, \boldsymbol{y}, \boldsymbol{y}_0$ is proportional to

$$\prod_{j=1}^{m_0} N_{n_{0j}}\left(\boldsymbol{y_{0j}} \,\Big|\, \mu_0 1_{n_{0j}}, \sigma_{u0}^2 J_{n_{0j}} + \sigma_{\epsilon0}^2 I_{n_{0j}}\right) \prod_{i=1}^{m} N_{n_i}\left(\boldsymbol{y_i} \,\Big|\, \mu 1_{n_i} + d_i\lambda, \sigma_u^2 J_{n_i} + \sigma_\epsilon^2 I_{n_i}\right)$$

$$\times N\left(\mu \,\Big|\, \mu_0, 1/\tau\right) / \left(\sigma_{u0}\sigma_u\sigma_{\epsilon0}\sigma_\epsilon\right), \quad (20)$$

where 1_u is a $1 \times u$ column vector of 1s, and J_u is a $u \times u$ matrix of 1s. The Gibbs sampler is implemented by sampling the latent variables u_{0j} and u_i.

To ease the subsequent algebra required to marginalize (20), note that $\sigma_{\epsilon0}^2$ times the inverted marginal estimated historical covariance for all observations in subject j, $\sigma_{\epsilon0}^2 \left(\sigma_{u0}^2 J_{n_{0j}} + \sigma_{\epsilon0}^2 I_{n_{0j}}\right)^{-1}$, is equal to $\left(I_{n_{0j}} - \frac{\sigma_{u0}^2}{\sigma_{\epsilon0}^2 + n_{0j}\sigma_{u0}^2} J_{n_{0j}}\right)$, which has $n_{0j} - 1$ eigenvalues equaling 1 and one non-unit eigenvalue equal to $1 - r_{0j}^{-1}$, where $r_{0j} = \frac{\sigma_{\epsilon0}^2}{n_{0j}\sigma_{u0}^2} + 1$. Similarly, $\left(I_{n_i} - \frac{\sigma_u^2}{\sigma_\epsilon^2 + n_i\sigma_u^2} J_{n_i}\right)$ has $n_i - 1$ eigenvalues equaling 1 and one non-unit eigenvalue equal to $1 - r_i^{-1}$, where $r_i = \frac{\sigma_\epsilon^2}{n_i\sigma_u^2} + 1$. Let

$$M_\tau = \left\{\frac{\sum_{j=1}^{m_0}\sum_{k=1}^{n_{0j}} y_{0jk}(1 - r_{0j}^{-1})}{\sum_{j=1}^{m_0} n_{0j}(1 - r_{0j}^{-1})}\right\} \Big/ \left\{\frac{1}{\tau} + \frac{\sigma_{\epsilon0}^2}{\sum_{j=1}^{m_0} n_{0j}(1 - r_{0j}^{-1})}\right\}$$

$$+ \left\{\frac{\sum_{i=1}^{m}\sum_{g=1}^{n_i}(y_{ig} - d_i\lambda)(1 - r_i^{-1})}{\sum_{i=1}^{m} n_i(1 - r_i^{-1})}\right\} \Big/ \left\{\frac{\sigma_\epsilon^2}{\sum_{j=1}^{m} n_i(1 - r_i^{-1})}\right\},$$

and
$$\Sigma_\tau = \left\{\frac{1}{\tau} + \frac{\sigma_{\epsilon0}^2}{\sum_{j=1}^{m_0} n_{0j}(1 - r_{0j}^{-1})}\right\}^{-1} + \left\{\frac{\sigma_\epsilon^2}{\sum_{j=1}^{m} n_i(1 - r_i^{-1})}\right\}^{-1}. \quad (21)$$

After marginalizing μ_0, the conditional posterior for $\mu|\lambda, u_0, u, \sigma_{\epsilon0}^2, \sigma_\epsilon^2, \tau$ is proportional to $N(\Sigma_\tau^{-1} M_\tau, \Sigma_\tau^{-1})$. Full conditional posteriors for the remaining parameters can be found in Appendix B.

Several alternative prior specifications for the correlation structure may be more natural for incorporating prior information on the variance components. For example, we could formulate our prior opinion about the model smoothness by specifying a prior on the variance ratio, $\sigma_\epsilon^2/\sigma_u^2$ or the degrees of freedom it induces (Hodges and Sargent 2001). For discussion about degrees of freedom and how they can be used to sensibly determine variance component priors, as well as the general marginal posterior for the variance ratio, see Reich and Hodges (2008), and Cui et al. (2010).

Linear mixed model

The one-way random effects model presented above is a special case of a linear mixed model for which between-subject observations are independent, all within subject observations have identical covariance (compound symmetry within groups), and there are no fixed regression effects (only intercepts). In this subsection we extend to the general linear mixed model.

Following the notation presented above, denote $n_0 = \sum_{j=1}^{m_0} n_{0j}$ and $n = \sum_{i=1}^{m} n_i$. Suppose X_0 and X are $n_0 \times p$ and $n \times p$ design matrices such that the first columns contain vectors of 1s corresponding to the intercepts, $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$ are vectors of identically measured regression coefficients of length p representing fixed covariate effects, and $\boldsymbol{d}$ is an $n \times 1$ new intervention indicator. Furthermore, let $\boldsymbol{u}_0, \boldsymbol{u}$ and Z_0, Z denote $m_0 \times 1$ and $m \times 1$ random effects vectors and their respective $n_0 \times m_0$ and $n \times m$ design matrices for the historical and current data.

Adopting the notation of McCulloch and Searle (2001, p.156), we formulate the general linear mixed model by first assuming normally distributed random effects with covariances D_0 and D, $\boldsymbol{u}_0 \sim N(0, D_0)$ and $\boldsymbol{u} \sim N(0, D)$. Models for the historical and concurrent responses are $\boldsymbol{y}_0 = X_0\boldsymbol{\beta}_0 + Z_0\boldsymbol{u}_0 + \boldsymbol{\epsilon}_0$ and $\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{u} + \boldsymbol{d}\lambda + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}_0 \sim N_{n_0}(0, R_0)$ and $\boldsymbol{\epsilon} \sim N_n(0, R)$; R_0 and R represent the conditional covariances of $\boldsymbol{y}_0|\boldsymbol{u}_0$ and $\boldsymbol{y}|\boldsymbol{u}$. The marginal covariances for $\boldsymbol{y}_0$ and $\boldsymbol{y}$ are $\Sigma_0 = Z_0 D_0 Z_0^T + R_0$ and $\Sigma = ZDZ^T + R$, respectively. The precision matrix $V_{\boldsymbol{\tau}}$ in (18) follows as

$$V_{\boldsymbol{\tau}} = diag\{\boldsymbol{\tau}\} - diag\{\boldsymbol{\tau}\} \left(X_0^T \Sigma_0 X_0 + diag\{\boldsymbol{\tau}\}\right)^{-1} diag\{\boldsymbol{\tau}\}. \tag{22}$$

The conditional posterior distribution for $\boldsymbol{\beta}|\lambda, D_0, R_0, D, R, \boldsymbol{\tau}$ is proportional to

$$N_p \left\{\boldsymbol{\beta} \,\middle|\, \left(V_{\boldsymbol{\tau}} + X^T \Sigma X\right)^{-1} \left(V_{\boldsymbol{\tau}}\tilde{\boldsymbol{\beta}}_0 + X^T \Sigma X \tilde{\boldsymbol{\beta}}_\lambda\right), \left(V_{\boldsymbol{\tau}} + X^T \Sigma X\right)^{-1}\right\}, \tag{23}$$

where $\tilde{\boldsymbol{\beta}}_\lambda = (X^T \Sigma X)^{-1} X^T (\boldsymbol{y} - \boldsymbol{d}\lambda)$, and $\tilde{\boldsymbol{\beta}}_0 = (X_0^T \Sigma_0 X_0)^{-1} X_0^T \boldsymbol{y}_0$, are the usual integrated least squares estimates. Treating the random effects as latent variables and adopting conjugate Wishart priors for D^{-1} and R^{-1} and their historical counterparts, namely $D^{-1} \sim W\left(\phi\tilde{D}, \phi\right)$ and $R^{-1} \sim W\left(\rho\tilde{R}, \rho\right)$, posterior inference may proceed via the Gibbs sampler. The remaining full conditional posteriors can be found in Appendix C.

If we assume that observations are uncorrelated across subjects, the covariance structures simplify to $D_0 = \sigma_{u0}^2 I_{m_0}$, $D = \sigma_u^2 I_m$, $R_0 = \sigma_{\epsilon0}^2 I_{n_0}$, and $R = \sigma_\epsilon^2 I_n$. The

resulting model extends the one-way random effects model above to incorporate fixed covariate effects. Using the same priors as before, the full conditional posteriors for σ_u^2 and σ_ϵ^2 follow as $q(\sigma_\epsilon^2|\lambda, \boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{y}_0, \boldsymbol{y}) \propto \Gamma^{-1}\left\{(n-1)/2, \parallel \boldsymbol{y} - X\boldsymbol{\beta} - \boldsymbol{d}\lambda - Z\boldsymbol{u} \parallel^2 /2\right\}$ and $q(\sigma_u^2|\boldsymbol{u}, \boldsymbol{y}_0, \boldsymbol{y}) \propto \Gamma^{-1}\left\{(m-1)/2, \sum_{j=1}^m u_j^2/2\right\}$. See Kass and Natarajan (2006) for an empirical Bayes approach using an inverted Wishart prior on Σ for general covariance structures and design matrices.

4 Generalized linear models

In this section we extend the methodology to incorporate generalized linear models for non-Gaussian error distributions. The methodology is generalized to include data from exponential families assuming Gaussian approximations of the product of the initial prior and historical likelihood using the Bayesian Central Limit Theorem (see e.g. Carlin and Louis 2009, p.108). For a flat initial prior, the approximation takes mean equal to the historical MLE and variance equal to the inverted observed Fisher information matrix. These approximations are used frequently in Bayesian analysis for data assumed to follow from exponential families (Spiegelhalter et al. 2004, p.23; Gelman et al. 2004, p.101). While computational methods can handle the full model, the approximations facilitate dimension reduction for the numerical problem of estimating the crucial MMLE of $\boldsymbol{\tau}$ when using EB inference, since $\boldsymbol{\beta}_0$ is readily marginalized analytically.

We first present the general method for fixed effects models, and then discuss logistic regression models for binary outcomes and a Weibull regression model. We then extend the general method to incorporate random effects, and illustrate in two important specific cases: a probit regression model for binary outcomes, and a Poisson regression model for count data. The Weibull regression model is used to analyze patient-level data from two successive colorectal cancer trials in Section 5 for both the approximate and full commensurate prior models.

4.1 Fixed effect models

Let $\boldsymbol{y}_0$ and $\boldsymbol{y}$ denote column vectors of length n_0 and n consisting of independent measurements from a distribution that is a member of the exponential family, $f_Y(y)$. That is, we suppose $y_{0j} \overset{indep}{\sim} f_{Y_{0j}}(y_{0j})$ and $y_i \overset{indep}{\sim} f_{Y_i}(y_i)$ such that the log-likelihoods are of form

$$\log f_{Y_0}(y_{0j}) = \sum_{j=1}^{n_0}\left\{y_{0j}\gamma_{0j} - b_0(\gamma_{0j})\right\}/\nu_0^2 - \sum_{j=1}^{n_0} c_0(y_{0j}, \nu_0) \qquad (24)$$

$$\text{and } \log f_Y(y_i) = \sum_{i=1}^{n}\left\{y_i\gamma_i - b(\gamma_i)\right\}/\nu^2 - \sum_{i=1}^{n} c(y_i, \nu), \qquad (25)$$

for $j = 1, ..., n_0$ and $i = 1, ..., n$ (McCulloch and Searle 2001, p.139). Following the notation of Section 3, let $g(\boldsymbol{\mu}_0) = \boldsymbol{X}_0\boldsymbol{\beta}_0$ and $g(\boldsymbol{\mu}) = X\boldsymbol{\beta} + \boldsymbol{d}\lambda$, for known "link" function $g()$ where $E[\boldsymbol{y}_0] = \boldsymbol{\mu}_0$ and $E[\boldsymbol{y}] = \boldsymbol{\mu}$.

Using the Bayesian Central Limit Theorem, we replace the product of the initial prior and historical likelihood, $f_{Y_0}(\boldsymbol{y}_0|\boldsymbol{\beta}_0)p(\boldsymbol{\beta}_0)$, with an asymptotic normal approximation, $\hat{p}(\boldsymbol{\beta}_0|\boldsymbol{y}_0)$. This density is an approximate sequential Bayesian update of the initial prior for $\boldsymbol{\beta}_0$. For a flat initial prior, the approximation takes mean equal to the historical MLE (computed numerically via Newton-Raphson or Fisher scoring) and variance equal to the inverted observed Fisher information matrix, $\hat{p}(\boldsymbol{\beta}_0|\boldsymbol{y}_0) = N_p\left\{\hat{\boldsymbol{\beta}}_0, \left(X_0^T\hat{W}_0 X_0\right)^{-1}\right\}$, where $\hat{W}_0 = W_0(\hat{\boldsymbol{\mu}}_0)$ is an $n_0 \times n_0$ diagonal matrix having $jj-$element

$$W_{0jj}(\mu_0) = \left\{\nu^2 v(\boldsymbol{\mu}_0)\left(\frac{\partial g(\boldsymbol{\mu}_0)}{\partial \boldsymbol{\mu}_0}\right)^2\right\}^{-1}, \tag{26}$$

where $v(\mu_{0j}) = var(y_{0j})/\nu^2 = \partial^2 b(\gamma_{0j})/\partial\gamma_{0j}^2$, for $j = 1, ..., n_0$ (McCulloch and Searle 2001, p.141). Following Subsection 3.1, the commensurate prior for $\boldsymbol{\beta}$ is proportional to $N_p(\boldsymbol{\beta}_0, diag\{\boldsymbol{\tau}\}^{-1})$. Let $\boldsymbol{\theta}$ denote the general parameter vector, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\beta}_0, \nu, \nu_0, \lambda)$. Assuming a flat prior for λ, the approximate joint posterior distribution of $\boldsymbol{\theta}|\boldsymbol{\tau}, \boldsymbol{y}, \boldsymbol{y}_0$ is proportional to

$$N_p(\boldsymbol{\beta}|\boldsymbol{\beta}_0, diag\{\boldsymbol{\tau}\}^{-1})N_p\left\{\boldsymbol{\beta}_0|\hat{\boldsymbol{\beta}}_0, \left(X_0^T\hat{W}_0 X_0\right)^{-1}\right\}\prod_{i=1}^n f_Y(y_i|\boldsymbol{\beta}, \lambda). \tag{27}$$

After marginalizing over $\boldsymbol{\beta}_0$, the precision matrix in (18) follows as,

$$V_{\boldsymbol{\tau}} = diag\{\boldsymbol{\tau}\} - diag\{\boldsymbol{\tau}\}\left(X_0^T\hat{W}_0 X_0 + diag\{\boldsymbol{\tau}\}\right)^{-1} diag\{\boldsymbol{\tau}\}, \tag{28}$$

and the posterior distribution of $\boldsymbol{\beta}, \lambda|\tau, \boldsymbol{y}, \boldsymbol{y}_0$ is proportional to

$$N_p(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}_0, V_{\boldsymbol{\tau}}^{-1})\prod_{i=1}^n f_Y(y_i|\boldsymbol{\beta}, \lambda). \tag{29}$$

Non-Gaussian data results in intractable non-conjugate full conditional distributions, therefore posterior inference requires alternative MCMC sampling methods (see e.g. Carlin and Louis 2009), such as the Metropolis algorithm.

Binary response

Let $\boldsymbol{y}_0$ and $\boldsymbol{y}$ denote the historical and current data such that $y_{0j} \sim Ber\{\pi_0(X_{0j})\}$, $\pi_0(X_{0j}) \in [0,1]$, for $j = 1, ..., n_0$, and $y_i \sim Ber\{\pi(X_i, d_i)\}$, $\pi(X_i, d_i) \in [0,1]$, for $i = 1, ..., n$. The logistic link function transforms the expectations of y_0 and y such that $\log\left\{\frac{\pi_0(X_0)}{1-\pi_0(X_0)}\right\} = X_0\boldsymbol{\beta}_0$ and $\log\left\{\frac{\pi(X,d)}{1-\pi(X,d)}\right\} = X\boldsymbol{\beta} + \boldsymbol{d}\lambda$. The diagonal elements of $\hat{W}_0$ in (27) and (28) now consist of the estimated historical sampling variance, $\hat{W}_{0jj} = \hat{\pi}_0(X_{0j})\{1 - \hat{\pi}_0(X_{0j})\}$, where $\hat{\pi}_0(X_{0j}) = \left(1 + e^{-X_{0j}\hat{\boldsymbol{\beta}}_0}\right)^{-1}$. Assuming a flat prior for λ, the posterior distribution of $\boldsymbol{\beta}, \lambda|\boldsymbol{\tau}, \boldsymbol{y}, \boldsymbol{y}_0$ is proportional to

$$N_p(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}_0, V_{\boldsymbol{\tau}}^{-1})\prod_{j=1}^n \pi(X_j)^{y_j}\{1 - \pi(X_j)\}^{1-y_j}.$$

Sampling proceeds by Metropolis, though switching to a probit link function can lead to closed form full conditionals (Albert and Chib 1993). Fúquene, Cook, and Pericchi (2009) propose an approach using a robust Cauchy prior for univariate logistic models.

Time-to-event response

Following the notation of Kalbfleisch and Prentice (2002, p.52), data for the historical and current trials consist of triples $(t_{0j}, \delta_{0j}, X_{0j})$ for $j = 1, ..., n_0$ and (t_i, δ_i, X_i) for $i = 1, ..., n$. Here, $t_{0j}, t_i > 0$ are the observed, possibly censored, failure times; δ_{0j}, δ_i are noncensoring indicators (0 if censored, 1 if failure); and X_{0j} and X_i are row vectors of p covariates associated with historical subject j and current subject i. Let $\tilde{t}_{0j}$ and $\tilde{t}_i$ be the underlying uncensored failure times, with corresponding densities $f(t_{0j})$ and $f(t_i)$. Denote the survival functions for the jth historical and ith current individuals by $Pr(\tilde{t}_i > t) = F(t)$ and $Pr(\tilde{t}_{0j} > t_0) = F(t_0)$.

Log-linear models are commonly used for analyzing time-to-event data. Suppose $\boldsymbol{y}_0 = \log(\boldsymbol{t}_0) = X_0\boldsymbol{\beta}_0 + \sigma_0\boldsymbol{e}_0$ and $\boldsymbol{y} = \log(\boldsymbol{t}) = X\boldsymbol{\beta} + \boldsymbol{d}\lambda + \sigma\boldsymbol{e}$ where $\boldsymbol{e}_0 = (\boldsymbol{y}_0 - X_0\boldsymbol{\beta}_0)/\sigma_0$ and $\boldsymbol{e} = (\boldsymbol{y} - X\boldsymbol{\beta} - \boldsymbol{d}\lambda)/\sigma$. Assuming that censoring times are conditionally independent of each other and of the independent failure times given X_0 and X (*noninformative censoring*), the historical and current data likelihoods follow as $L_0(\boldsymbol{\beta}_0, \sigma_0|\boldsymbol{y}_0) = \prod_{i=j}^{n_0} \left\{\frac{1}{\sigma_0}f(e_{0j})\right\}^{\delta_{0j}} F(e_{0j})^{1-\delta_{0j}}$ and $L(\boldsymbol{\beta}, \sigma|\boldsymbol{y}) = \prod_{i=j}^{n} \left\{\frac{1}{\sigma}f(e_i)\right\}^{\delta_i} F(e_i)^{1-\delta_i}$. Assuming a flat prior for λ, the commensurate prior model follows from (27) where ML estimates for $\boldsymbol{\beta}_0$ and σ_0 are computed numerically via Newton-Raphson or Fisher-Scoring; see Kalbfleisch and Prentice (2002, p.66–69). For discussion about censoring mechanisms see Kalbfleisch and Prentice (2002, p.193) or Klein and Moeschberger (2003, p.63).

Weibull regression arises when $\boldsymbol{e}_0$ and $\boldsymbol{e}$ are assumed to follow the extreme value distribution, $f(u) = \exp\{u - exp(u)\}$. This results in a parametric regression model which has both a proportional hazards and an accelerated failure-time representation. The hazard function, $h(t) = -d\log F(t)/dt$, is monotone decreasing for shape parameter $\sigma > 1$, increasing for $\sigma < 1$, constant for $\sigma = 1$. We assume commensurate priors for both the regression coefficients, $\boldsymbol{\beta}$, and log transformation of the shape parameter, σ. Let $\boldsymbol{\zeta}_0 = \left\{\boldsymbol{\beta}_0^T, \log(\sigma_0)\right\}^T$ and $\boldsymbol{\zeta} = \left\{\boldsymbol{\beta}^T, \log(\sigma)\right\}^T$ denote column vectors of length $p + 1$. Let $\boldsymbol{e}_0$, $\boldsymbol{\delta}_0$, and $\exp(\boldsymbol{e}_0) - \boldsymbol{\delta}_0$ denote vectors of length n_0 such that the jth element is equal to e_{0j}, δ_{0j}, and $\exp(e_{0j}) - \delta_{0j}$, $j = 1, ..., n_0$. ML equations for the historical coefficients follow as $\frac{\partial \log L_0}{\partial \boldsymbol{\beta}_0} = \frac{1}{\sigma_0}X_0^T \{\exp(\boldsymbol{e}_0) - \boldsymbol{\delta}_0\}$ and $\frac{\partial \log L_0}{\partial \log(\sigma_0)} = \sum_{j=1}^{n_0} \{e_{0j}\exp(e_{0j}) - \delta_{0j}(e_{0j} + 1)\}$. Let $\hat{e}_{0j} = (y_{0j} - X_{0j}\hat{\boldsymbol{\beta}}_0)/\hat{\sigma}_0$. The observed Fisher information matrix, $\hat{\Psi}_0(\hat{\boldsymbol{\zeta}}_0)$, follows as

$$\hat{\Psi}_0(\hat{\boldsymbol{\zeta}}_0) = \begin{bmatrix} X_0^T \hat{W}_0 X_0 & X_0^T \hat{E}_0 \\ \hat{E}_0^T X_0 & \sum_{j=1}^{n_0} \hat{e}_{0j} \{\delta_{0j} - (\hat{e}_{0j} + 1)\exp(\hat{e}_{0j})\} \end{bmatrix}, \qquad (30)$$

where $\hat{\boldsymbol{E}}_0$ is the vector of length n_0 containing elements $\hat{E}_{0j} = (\hat{e}_{0j} + 1)\exp\{\hat{e}_{0j} - \log(\hat{\sigma}_0)\} - \delta_{0j}/\hat{\sigma}_0$, and the diagonal elements of $\hat{W}_0$ are $\hat{W}_{0jj} = \frac{1}{\hat{\sigma}_0^2}\exp(\hat{e}_{0j})$, for $j = $

$1, ..., n_0$ (see Breslow and Clayton 1993). Assuming no initial prior information for $\boldsymbol{\zeta}_0$ and that the prior on λ is flat, the posterior distribution for $\boldsymbol{\zeta}, \lambda | \tau, \boldsymbol{y}, \boldsymbol{y}_0$ follows directly from (29) by replacing $X_0^T \hat{W}_0 X_0$ in (27) and (28) with $\hat{\Psi}_0(\hat{\boldsymbol{\zeta}}_0)$. Note that the exponential model is a special case where $\sigma_0 = \sigma = 1$, since this leads to probability density functions for $\boldsymbol{t}_0$ and $\boldsymbol{t}$ following $f(\boldsymbol{t}_0) = \frac{1}{\boldsymbol{\mu}_0} \exp(\boldsymbol{t}_0 / \boldsymbol{\mu}_0)$ and $f(\boldsymbol{t}) = \frac{1}{\boldsymbol{\mu}} \exp(\boldsymbol{t} / \boldsymbol{\mu})$, where $\boldsymbol{\mu}_0 = \exp(X_0 \boldsymbol{\beta}_0)$ and $\boldsymbol{\mu} = \exp(X \boldsymbol{\beta} + \boldsymbol{d} \lambda)$; $\boldsymbol{t}_0 / \boldsymbol{\mu}_0$ is a vector of length n_0 with jth element equal to $t_{0j} / \exp(X_{0j} \boldsymbol{\beta}_0)$, and similarly, $\boldsymbol{t} / \boldsymbol{\mu}$ is a vector of length n with ith element equal to $t_i / \exp(X_i \boldsymbol{\beta} + d_i \lambda)$.

4.2 Mixed models

As with the general linear models in Section 3, we now extend our generalized linear model to the mixed model case. Following the notation of Subsection 3.2, let $\boldsymbol{y}_{0j}$ and $\boldsymbol{y}_i$ denote response vectors of lengths n_{0j} and n_i consisting of conditionally independent measurements given random effects u_{0j} and u_i, $j = 1, ..., m_0$; $i = 1, ..., m$, where $\boldsymbol{u}_0$ and $\boldsymbol{u}$ are vectors of length m_0 and m. For generalized linear mixed models, we assume the conditional distributions of $\boldsymbol{y}_0$ given $\boldsymbol{u}_0$ and $\boldsymbol{y}$ given $\boldsymbol{u}$ have p.d.f.s from the exponential family, $\boldsymbol{y}_{0j} | \boldsymbol{u}_0 \overset{indep}{\sim} f_{Y_0}(\boldsymbol{y}_{0j} | \boldsymbol{u}_0)$ and $\boldsymbol{y}_i | \boldsymbol{u} \overset{indep}{\sim} f_Y(\boldsymbol{y}_i | \boldsymbol{u})$, where $\log f_{Y_0}(\boldsymbol{y}_0 | \boldsymbol{u}_0)$ and $\log f_Y(\boldsymbol{y}_i | \boldsymbol{u})$ follow (24) and (25). In addition, assume that the random effects follow distributions $\boldsymbol{u}_0 \sim f_{U_0}(\boldsymbol{u}_0)$ and $\boldsymbol{u} \sim f_U(\boldsymbol{u})$. The marginal likelihoods follow as $f_{Y_0}(\boldsymbol{y}_{0j}) = \int \prod_{j=1}^{m_0} f_{Y_0}(\boldsymbol{y}_{0j} | \boldsymbol{u}_0) f_{U_0}(\boldsymbol{u}_0) d\boldsymbol{u}_0$ and $f_Y(\boldsymbol{y}_i) = \int \prod_{i=1}^{m} f_Y(\boldsymbol{y}_i | \boldsymbol{u}) f_U(\boldsymbol{u}) d\boldsymbol{u}$.

Assuming that the conditional means are $E(\boldsymbol{y}_0 | \boldsymbol{u}_0) = \boldsymbol{\mu}_0$ and $E(\boldsymbol{y} | \boldsymbol{u}) = \boldsymbol{\mu}$, let $g(\boldsymbol{\mu}_0) = X_0 \boldsymbol{\beta}_0 + Z_0 \boldsymbol{u}_0$ and $g(\boldsymbol{\mu}) = X \boldsymbol{\beta} + \boldsymbol{d} \lambda + Z \boldsymbol{u}$, where $g()$ is a known link function (note that $Z_0, Z, X_0, X, \boldsymbol{d}$ and parameters D, D_0 are defined in Subsection 3.2). Following Subsection 4.1 for normally distributed random effects, $\boldsymbol{u}_0 \sim N(\boldsymbol{0}, D_0)$ and $\boldsymbol{u} \sim N(\boldsymbol{0}, D)$, the posterior distribution of $\boldsymbol{\beta}, \lambda | \tau, \boldsymbol{y}, \boldsymbol{y}_0$, follows from (29) with $\hat{W}_0$ in (27) and (28) replaced by $(\hat{W}_0^{-1} + Z_0 \hat{D}_0 Z_0^T)^{-1}$. See McCulloch (1997), McCulloch and Searle (2001, p.263), or Breslow and Clayton (1993) for algorithms for computing ML estimates for fixed effects and prediction of random effects.

Binary response

Suppose that the historical responses follow $y_{0jk} | u_{0j} \sim Ber\{\pi_0(X_{0j})\}$, where y_{0jk} denotes observation k for the jth patient, $k = 1, ..., n_{0j}$, $j = 1, ..., m_0$. The current responses follow $y_{ig} | u_i \sim Ber\{\pi(X_{ig}, d_i)\}$, where y_{ig} denotes observation g for the ith patient, $g = 1, ..., n_i$, $i = 1, ..., m$. Since the logit link function was discussed in Section 4.1, we will consider the probit link. Given $\boldsymbol{u}_0 \sim N(\boldsymbol{0}, D_0)$, the probit uses the standard normal c.d.f., $\Phi()$, to transform the means of $\boldsymbol{y}_0$ and $\boldsymbol{y}$ such that $\pi_0(X_0, Z_0) = \Phi(\boldsymbol{\beta}_0 X_0 + Z_0 \boldsymbol{u}_0)$ and $\pi(X, \boldsymbol{d}, Z) = \Phi(X \boldsymbol{\beta} + \boldsymbol{d} \lambda + Z \boldsymbol{u})$. The posterior distribution of $\boldsymbol{\beta}, \lambda | \boldsymbol{\tau}, \boldsymbol{y}, \boldsymbol{y}_0$ follows from (29) by replacing $\hat{W}_0$ in (27) and (28) with

$(\hat{W}_0^{-1} + Z_0\hat{D}_0Z_0^T)^{-1}$, and setting

$$\hat{W}_0 = diag\left\{ \frac{\phi(X_0\hat{\boldsymbol{\beta}}_0 + Z_0\hat{\boldsymbol{u}}_0)^2}{\Phi(X_0\hat{\boldsymbol{\beta}}_0 + Z_0\hat{\boldsymbol{u}}_0)\left\{1 - \Phi(X_0\hat{\boldsymbol{\beta}}_0 + Z_0\hat{\boldsymbol{u}}_0)\right\}} \right\},$$

where $\phi()$ is the standard normal p.d.f. (McCulloch and Searle 2001, p.136).

Count response

Suppose that the historical responses follow $y_{0jk}|u_{0j} \overset{indep}{\sim} Poisson(\mu_{0jk})$, where $\log(\mu_{0jk}) = X_{0jk}\boldsymbol{\beta}_0 + u_{0j}$ and $u_{0j} \overset{i.i.d.}{\sim} N(0, \sigma_{u_0}^2)$. Here, y_{0jk} denotes the kth count observed for the jth patient, where $k = 1, ..., n_{0j}$ and $j = 1, ..., m_0$. Similarly, let y_{ig} denote the gth count observed for the ith patient, where $y_{ig}|u_i \overset{indep}{\sim} Poisson(\mu_{ig})$, $\log(\mu_{ig}) = X_{ig}\boldsymbol{\beta} + u_g$, and $u_g \overset{i.i.d.}{\sim} N(0, \sigma_u^2)$, for $g = 1, ..., n_i$ and $i = 1, ..., m$. The posterior distribution of $\boldsymbol{\beta}, \lambda|\boldsymbol{\tau}, \boldsymbol{y}, \boldsymbol{y}_0$ follows from (29) by replacing $\hat{W}_0$ in (27) and (28) with $(\hat{W}_0^{-1} + \hat{\sigma}_{u0}^2 Z_0 Z_0^T)^{-1}$, and setting

$$\hat{W}_0 = diag\left\{\exp(X_0\hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{u}}_0)\right\}.$$

5   Case study: analysis of successive colon cancer trials

In this section, we illustrate our method using data from two successive randomized controlled colorectal cancer clinical trials originally reported by Saltz et al. (2000) and Goldberg et al. (2004), respectively. The initial trial randomized $N_0 = 683$ 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 progression-free survival and overall survival than both Irinotecan alone and 5FU/LV and became the standard of care treatment (Saltz et al. 2000).

The subsequent trial compared three new (at the time) drug combinations in $N = 795$ 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.

Both trials recorded two bi-dimensional measurements on each tumor for each patient at regular cycles. The Saltz trial measured patients every 6 weeks for the first 24 weeks and every 12 weeks thereafter until death or disease progression, while the Goldberg trial measured every 6 weeks for the first 42 weeks, or until death or disease progression.
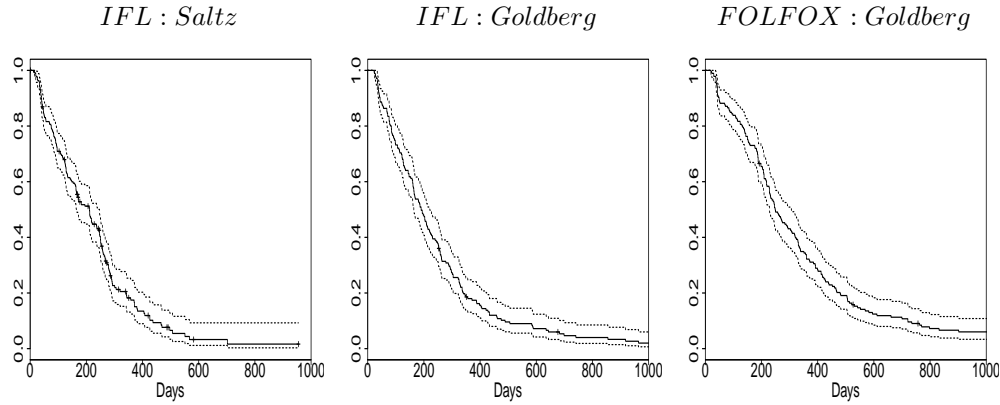
Figure 4: Separate Kaplan-Meier curves for time to disease progression corresponding to subjects on IFL in the Saltz trial (left), IFL in the Goldberg trial (center), and FOLFOX in the Goldberg trial.

We computed the sum of the longest diameter in cm ("ld sum") for up to 9 tumors for each patient at baseline. In both trials, disease progression was defined as a 25% or greater increase in measurable tumor or the appearance of new lesions.

This section offers a fixed effects time-to-event analysis using the Weibull regression model presented in Subsection 4.1 to compare disease progression among the FOLFOX and IFL regimens. The historical data consists of the IFL treatment arm from the initial study, while the current data consists of patients randomized to IFL or FOLFOX in the subsequent 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. The model incorporates baseline ld sum as a predictor.

We restricted our analysis to patients that had measurable tumors and observed baseline covariates bringing the total sample size to 586: 224 historical and 362 current observations. Among the current patients, 176 are controls (IFL) and 186 are patients treated with the new regimen (FOLFOX). Figure 4 contains Kaplan-Meier estimated time to progression curves for subjects on each treatment regimen in both trials. The plots suggest that the time to progression experience for subjects on IFL was similar in both the Saltz (left panel) and Goldberg trials (center), and that FOLFOX (right) is associated with somewhat prolonged time-to-progression.

Following the log-linear model notation of Subsection 4.1 for progression times t_0 and t, let $y_0 = X_0\beta_0 + \sigma_0 e_0$ and $y = X\beta + d\lambda + \sigma e$, where $y_0 = \log(t_0)$ and $y = \log(t)$; here, X_0 and X are $n_0 \times 2$ and $n \times 2$ design matrices with columns corresponding to (1, ld sum at baseline), and d is the FOLFOX indicator. Thus, the β_0 and β parameters contain intercepts as well as a regression coefficient corresponding to the baseline covariate,

Table 4: Weibull model fits to colorectal cancer data $n_0 = 224$, $n = 362$.

	Separate analyses				Pooled analysis	
	Historical		Current			
	est	sd	est	sd	est	sd
Intercept	5.503	0.058	5.555	0.067	5.533	0.045
BL ld sum	-0.043	0.051	-0.115	0.045	-0.092	0.034
FOLFOX	–	–	0.417	0.092	0.453	0.077
$\log(\sigma)$	-0.291	0.060	-0.153	0.039	-0.186	0.033

while $\exp(\lambda)$ represents the acceleration factor associated with FOLFOX. Note that since $F(t|d = 1) = F\left(te^\lambda|d = 0\right)$ for all t, a negative value is indicative of *decreased* survival. Exploratory data analysis on the covariate and age at baseline suggested that the trials enrolled patients from comparable populations. The first, second, and third quartiles for baseline tumor sum are $5, 8.5, 12.8$ in the Saltz trial and $4.7, 7.9, 12.7$ in the Goldberg trial; for age they are $54, 62, 69$ in the Saltz trial and $53, 61, 69$ in the Goldberg trial.

Table 4 summarizes results from separate and pooled classical linear regression fits to the historical $(\boldsymbol{t}_0, X_0)$ and current $(\boldsymbol{t}, X, d)$ data. The "current" values constitute a "no borrowing" analysis. Results from the current data alone suggest that the estimated acceleration factor corresponding to FOLFOX is highly significant at the 0.05 level. Point estimates (posterior means) and posterior standard deviations corresponding to both the Gaussian approximation and full commensurate prior models are provided in Table 5. The fully Bayesian spike and slab model uses the same hyperparameters for $p(\tau)$ that were used to produce Tables 2 and 3 in Section 2.4. For the EB model $\tau_j = 1/\nu_j^*$, for $j = 1, 2$, and 3 denote the MMLEs of the commensurability parameters corresponding to the intercept, regression coefficient for ld sum at baseline, and log-shape parameters, respectively. The ν_j^*s were maximized over a restriction to the interval $(0.005, 200)$, the same interval used in Section 2.3. Table 5 reveals that posterior inferences using the Gaussian approximated and full MCMC models provide congruous results in this context. Posterior distributions for the τs are highly right-skewed, with large standard deviations inducing compromises among the evidence supplied by the historical and current data.

Incorporating the Saltz data into the analysis of the Goldberg trial using our proposed method leads to more precise parameter estimates (i.e., reductions to the posterior standard deviation for the FOLFOX effect of nearly 9%, 8%, and 5% for the EB, spike and slab, and gamma models, respectively). The two fully Bayesian commensurate prior analyses offer relatively similar amounts of borrowing of strength from the historical data. Moreover, these models provide considerably less borrowing of strength than that provide by pooling, which facilitates a 16% reduction in posterior standard deviation for estimating the FOLFOX effect. The EB procedure leads to more precise, naive estimates, and thus more borrowing of strength overall when compared to the fully

Table 5: Commensurate prior Weibull model fits to colorectal cancer data $n_0 = 224$, $n = 362$. Gaussian approximate model (top), full model (bottom).

	Approximate					
	EB		spike & slab		Gamma$(1, 0.01)$	
	est	sd	est	sd	est	sd
Intercept	5.540	0.054	5.544	0.058	5.546	0.059
BL ld sum	-0.100	0.040	-0.103	0.041	-0.105	0.042
FOLFOX	0.436	0.084	0.434	0.085	0.432	0.087
$\log(\sigma)$	-0.152	0.038	-0.159	0.039	-0.158	0.038
τ_1	200	–	153.7	84.0	123.4	106.3
τ_2	200	–	157.1	81.8	124.2	106.2
τ_3	40.0	–	121.6	97.2	98.7	90.7

	Full Model					
	EB		spike & slab		Gamma$(1, 0.01)$	
	est	sd	est	sd	est	sd
Intercept	5.541	0.054	5.547	0.058	5.546	0.058
BL ld sum	-0.100	0.040	-0.103	0.042	-0.105	0.042
FOLFOX	0.435	0.085	0.431	0.086	0.432	0.085
$\log(\sigma)$	-0.152	0.038	-0.158	0.038	-0.158	0.038
τ_1	200	–	153.2	84.4	124.8	107.3
τ_2	200	–	153.8	83.9	123.8	106.7
τ_3	40.0	–	126.1	96.1	102.5	93.7

Bayesian results. However, EB inference actually leads to less borrowing of strength from the historical data for estimating $\log(\sigma)$.

Note that our commensurate prior models estimate time-to-progression to be nearly $\exp(0.43)$ times larger on average in the FOLFOX group, and since the posteriors for $\log(\sigma)$ are less than 0, the hazard rates are increasing slightly over time. This finding is consistent with those of Goldberg et al. (2004), who determined FOLFOX to have superior time-to-progression and response rate compared to IFL.

6   Simulation study

In this section we use simulation to evaluate the Bayesian and frequentist operating characteristics of our proposed empirical Bayesian commensurate prior model for the challenging case of one historical study. Figure 5 plots coverage and width of the 95% highest posterior density (HPD) intervals for λ by $\Delta^{tr} = \mu^{tr} - \mu_0^{tr}$ for the Gaussian (left) and exponential (right) time-to-event models described in Sections 3 and 4. Results for Gaussian data are compared for the no borrowing, half-Cauchy meta-analytic, EB
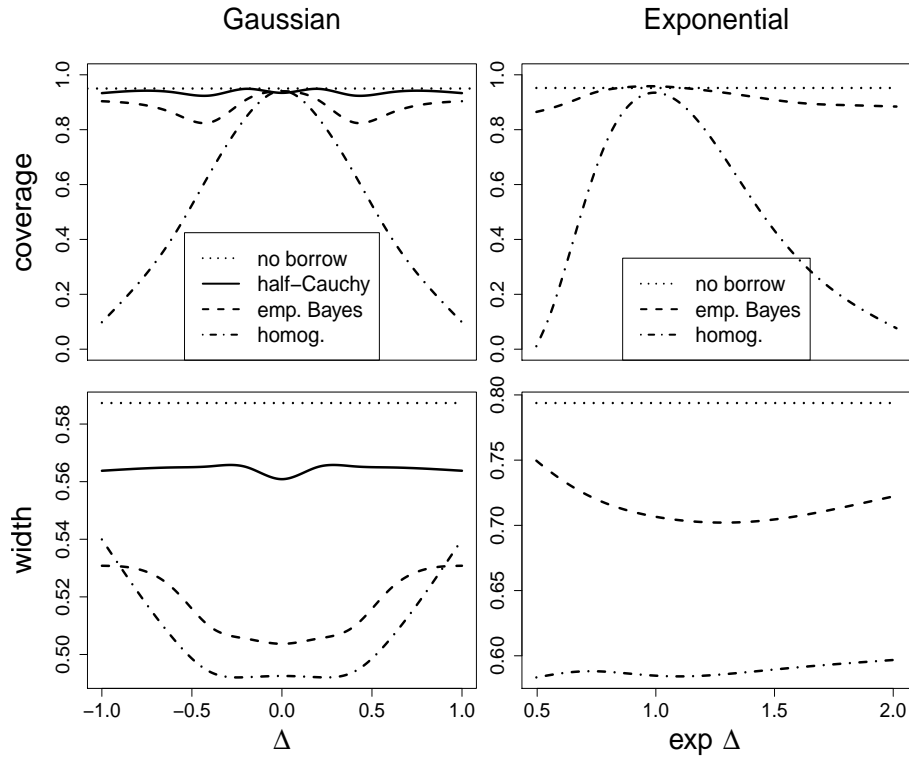
Figure 5: Coverage and width of 95% HPD intervals for λ by Δ^{tr} or $\exp(\Delta^{tr})$ for Gaussian and exponential data for one historical study.

commensurate, and homogeneity models under simulated data for $n_0 = 90$ historical patients, equal allocation of $n = 180$ current patients, and fixed true parameters, $\mu^{tr} = 0$, and $\sigma^{tr} = \sigma_0^{tr} = 1$. Results for the exponential model are compared for the no borrowing, EB commensurate, and homogeneity models simulated for $\mu^{tr} = 2$, $n_0 = 200$, and equal allocation of $n = 100$ current patients.

Figure 5 reveals that all four of the Gaussian models (left) provide 95% highest posterior density coverage and achieve their minimum interval widths when the historical data are unbiased, $\Delta^{tr} = 0$. As $|\Delta^{tr}| > 0$, we see that the interval widths increase. While coverage for the relatively conservative half-Cauchy models deviates little from the no borrowing model, the highly subjective full homogeneity model provides increasingly poorer coverage. The EB commensurate prior model offers a sensible compromise that offers more precision given lack of strong empirical evidence for heterogeneity, yet protects against poor coverage provided that the historical data is estimated to be highly biased. Results for the exponential model given in the right column of Figure 5 follow the same general trends. A referee encourages us to remind the reader that coverage probabilities of EB intervals are sometimes too small (Carlin and Louis 2009, Chapter

5); we would expect our fully Bayesian spike and slab or gamma prior models would perform slightly better in this regard.

Next we present a sensitivity analysis for the four Gaussian models. Following Freedman et al. (1984), suppose the parameter space of λ^{tr} is partitioned into three intervals characterizing three true states of nature: $\lambda^{tr} < -\delta$ implies failure, $-\delta \leq \lambda^{tr} \leq \delta$ implies equivalence, and $\lambda^{tr} > \delta$ implies efficacy. Suppose that the current trial analysis evaluates decision rules using posterior tail densities such that

$$
\begin{aligned}
q(\lambda < -\delta|\boldsymbol{y}, \boldsymbol{y}_0) &> 0.95, \quad \text{failure} \\
q(\lambda \in [-\delta, \delta]|\boldsymbol{y}, \boldsymbol{y}_0) &> 0.90, \quad \text{equivalence} \\
q(\lambda > \delta|\boldsymbol{y}, \boldsymbol{y}_0) &> 0.95, \quad \text{efficacy} \\
\text{otherwise,} & \qquad \text{inconclusive.}
\end{aligned}
\tag{31}
$$

We approximated the probability surfaces of the four stopping rules in (31) using simulation for a myriad of true values of $\Delta^{tr} = \mu^{tr} - \mu_0^{tr}$ and λ^{tr} for no borrowing, half-Cauchy meta-analytic, EB commensurate, and homogeneity models. We formulated a metric that facilitates comprehensive assessment of the four models by synthesizing the amount of probability allocated to the correct and incorrect decision spaces. Let $\mathcal{D} = (\boldsymbol{y}, \boldsymbol{y}_0)$ denote the collection of current and historical response data, and let $L(\mathcal{D}|\boldsymbol{\theta})$ denote the joint likelihood, where $\boldsymbol{\theta} = (\mu, \lambda, \Delta, \sigma^2, \sigma_0^2)$. Let $I(a)$ denote the indicator function of the event a, and let $\phi(\cdot|\boldsymbol{\theta}^{tr})$ denote the probability of a decision rule given a fixed set of parameters, $\boldsymbol{\theta}^{tr}$ i.e.

$$
\phi(\lambda < -\delta|\boldsymbol{\theta}^{tr}) = \int I\{q(\lambda < -\delta|\mathcal{D}) > 0.95\} L(\mathcal{D}|\boldsymbol{\theta}^{tr}) d\mathcal{D}.
\tag{32}
$$

For fixed true values σ^{tr}, σ_0^{tr}, and μ_0^{tr} the metric of comparison follows as

$$
\begin{aligned}
M(c) =& \int_{-\infty}^{-\delta} \int_{-\infty}^{\infty} \phi(\lambda < -\delta|\Delta^{tr}, \lambda^{tr}) p(\Delta^{tr}) p(\lambda^{tr}) d\Delta^{tr} d\lambda^{tr} \\
&+ \int_{-\delta}^{\delta} \int_{-\infty}^{\infty} \phi(\lambda \in [-\delta, \delta]|\Delta^{tr}, \lambda^{tr}) p(\Delta^{tr}) p(\lambda^{tr}) d\Delta^{tr} d\lambda^{tr} \\
&+ \int_{\delta}^{\infty} \int_{-\infty}^{\infty} \phi(\lambda > \delta|\Delta^{tr}, \lambda^{tr}) p(\Delta^{tr}) p(\lambda^{tr}) d\Delta^{tr} d\lambda^{tr} \\
&- \left\{ \int_{-\infty}^{-\delta} \int_{-\infty}^{\infty} \phi(\lambda > -\delta|\Delta^{tr}, \lambda^{tr}) p(\Delta^{tr}) p(\lambda^{tr}) d\Delta^{tr} d\lambda^{tr} \right. \\
&+ \int_{-\delta}^{\delta} \int_{-\infty}^{\infty} \phi(\lambda < -\delta \cup \lambda > \delta|\Delta^{tr}, \lambda^{tr}) p(\Delta^{tr}) p(\lambda^{tr}) d\Delta^{tr} d\lambda^{tr} \\
&+ \int_{\delta}^{\infty} \int_{-\infty}^{\infty} \phi(\lambda < \delta|\Delta^{tr}, \lambda^{tr}) p(\Delta^{tr}) p(\lambda^{tr}) d\Delta^{tr} d\lambda^{tr} \\
&+ c \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(\text{inconclusive}|\Delta^{tr}, \lambda^{tr}) p(\Delta^{tr}) p(\lambda^{tr}) d\Delta^{tr} d\lambda^{tr} \left. \right\}
\end{aligned}
\tag{33}
$$

Table 6: Simulation results for $M(c)$ and its components in (33) for $p(\Delta^{tr}) = N(0, s^2)$ and $c = 1/2$.

	$M(c)$	$+Eff$	$+Fail$	$+Equi$	$-Eff$	$-Fail$	$-Equi$	$-Inc$
				s=0				
no borrow	0.19	0.13	0.13	0.20	0.00	0.00	0.01	0.26
half-Cauchy	0.20	0.13	0.14	0.20	0.00	0.00	0.01	0.26
emp. Bayes	0.29	0.14	0.15	0.24	0.00	0.00	0.01	0.23
homog.	0.32	0.15	0.15	0.25	0.00	0.00	0.01	0.22
				s=1				
no borrow	0.19	0.13	0.13	0.20	0.00	0.00	0.01	0.26
half-Cauchy	0.21	0.13	0.14	0.20	0.00	0.00	0.01	0.25
emp. Bayes	0.24	0.14	0.15	0.21	0.00	0.00	0.02	0.24
homog.	0.14	0.14	0.15	0.12	0.00	0.00	0.07	0.21
				s=30				
no borrow	0.19	0.13	0.13	0.20	0.00	0.00	0.01	0.26
half-Cauchy	0.21	0.13	0.14	0.20	0.00	0.00	0.01	0.25
emp. Bayes	0.23	0.14	0.14	0.21	0.00	0.00	0.02	0.24
homog.	0.07	0.14	0.14	0.08	0.00	0.00	0.07	0.21

where $c \in [0, 1]$ is a positive scalar that weights the relative importance of an inconclusive result. The metric $M(c) \in [-1, 1]$, a generalized version of expected $0 - 1$ loss, is evaluated for $c = 1/2$ and $\delta = 0.33$ under prior $p(\Delta^{tr}) = N(0, s^2)$, for three values of s: $s = 0, 1, 30$, characterizing likely degrees of bias. All results are computed for $p(\lambda^{tr}) = N(0, 1/4)$. Larger values of $M(c)$ indicate overall more desirable probability allocation among the four decision rules. The prior distributions on Δ^{tr} and λ^{tr} essentially weight the relative importance of different subsets of the decision rule probability surfaces.

Table 6 summarizes the simulation results for $M(c)$ in (33). The top portion of the table contains results for the unrealistic case of *a priori* absolute certainty that the historical data is unbiased. In this scenario the highly subjective full homogeneity model exceeds our proposed EB commensurate model by 0.03. This occurs largely because borrowing of strength in this context yields unbiased results for λ. For the more likely cases when the *a priori* assessment of historical bias is more uncertain in the middle and bottom portions of Table 6, the EB commensurate model provides the best overall probability allocation as measured by $M(c)$, while the highly subjective full homogeneity model provides the least desirable results. These results are sensitive to alternative weights we might subjectively assign to the outcomes we are penalizing.

7   Discussion

In this paper, we provided empirical and fully Bayesian modifications of commensurate prior formulations (Hobbs et al. 2011), and extended the method to facilitate linear and generalized linear models. The proposed models are shown to lead to pre-posterior admissible estimators that facilitate alternative bias-variance trade-offs than those offered by pre-existing methodologies for incorporating historical data from a small number of historical studies. The method was also used to analyze data from two recent studies in colorectal cancer.

Future work looks toward extending the methodology to include time-dependent covariates, smoothed hazards, multiple events, Bayesian semi-parametric Cox models as e.g. Ibrahim, Chen, and Sinha (2001, p.47), and non-normal formulations of the commensurate prior itself. We are also currently pursuing the use of commensurate priors with adaptive randomization that allows the sample size or allocation ratio in the ongoing trial to be altered if this is warranted. For example, if historical and concurrent controls emerge as commensurate, we might assign fewer patients to the control group, thus enhancing the efficiency of the ongoing trial by imposing "information" balance.

References

Albert, J. H. and Chib, S. (1993). "Bayesian analysis of binary and polychotomous response data." *Journal of the American Statistical Association*, 88: 669–679. 660

Breslow, N. E. and Clayton, D. G. (1993). "Approximate inference in generalized linear mixed models." *Journal of the American Statistical Association*, 88: 9–25. 661

Browne, W. J. and Draper, D. (2006). "A comparison of Bayesian and likelihood-based methods for fitting multilevel models." *Bayesian Analysis*, 1: 473–514. 655

Carlin, B. P. and Louis, T. A. (2009). *Bayesian Methods for Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC Press, 3rd edition. 647, 650, 658, 659, 666

Cui, Y., Hodges, J. S., Kong, X., and Carlin, B. P. (2010). "Partitioning degrees of freedom in hierarchical and other richly-parameterised models." *Technometrics*, 52: 124–136. 657

Daniels, M. J. (1999). "A prior for the variance in hierarchical models." *The Canadian Journal of Statistics*, 27: 567–578. 641, 642

Doucet, A., Godsill, S. J., and Robert, C. P. (2002). "Marginal maximum a posteriori estimation using Markov chain Monte Carlo." *Statistics and Computing*, 12: 77–84. 648

Freedman, L. S., Lowe, D., and Macaskill, P. (1984). "Stopping rules for clinical trials incorporating clinical opinion." *Biometrics*, 40: 575–586. 667

Fúquene, J. A., Cook, J. D., and Pericchi, L. R. (2009). "A case for robust Bayesian priors with applications to clinical trials." *Bayesian Analysis*, 4: 817–846. 640, 660

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*, 1: 515–534. 640, 641, 656

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC Press, 2nd edition. 658

Geyer, C. J. and Thompson, E. A. (1992). "Constrained Monte Carlo maximum likelihood for dependent data." *Journal of the Royal Statistical Society. Series B*, 54: 657–699. 648

Goldberg, R. M., Sargent, D. J., Morton, R. F., Fuchs, C. S., Ramanathan, R. K., Williamson, S. K., Findlay, B. P., Pitot, H. C., and Alberts, S. R. (2004). "A randomized controlled trial of fluorouracil plus leucovorin, irinotecan, and oxaliplatin combinations in patients with previously untreated metastatic colorectal cancer." *Journal of Clinical Oncology*, 22: 23–30. 662, 665, 673

Hobbs, B. P., Carlin, B. P., Mandekar, S. J., and Sargent, D. J. (2011). "Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials." *Biometrics*, 67: 1047–1056. 639, 640, 644, 669

Hodges, J. S. and Sargent, D. J. (2001). "Counting degrees of freedom in hierarchical and other richly-parameterised models." *Biometrika*, 88: 367–379. 657

Ibrahim, J. G. and Chen, M.-H. (2000). "Power prior distributions for regression models." *Statistical Science*, 15: 46–60. 640

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. New York: Springer-Verlag. 669

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. New York: Wiley, 2nd edition. 660

Kass, R. E. and Natarajan, R. (2006). "A default conjugate prior for variance components in generalized linear mixed models." *Bayesian Analysis*, 1: 535–542. 658

Kass, R. E. and Steffey, D. (1989). "Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models)." *Journal of the American Statistical Association*, 84: 717–726. 647

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer-Verlag, 2nd edition. 660

McCulloch, C. E. (1997). "Maximum likelihood algorithms for generalized linear mixed models." *Journal of the American Statistical Association*, 92: 162–170. 661

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. New York: Wiley. 655, 657, 658, 659, 661, 662

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian variable selection in linear regression." *Journal of the American Statistical Association*, 83: 1023–1032. 649

Morris, C. N. (1983). "Parametric empirical Bayes inference: theory and applications." *Journal of the American Statistical Association*, 78: 47–55. 647

Neelon, B. and O'Malley, A. J. (2010). "Bayesian analysis using power priors with application to pediatric quality of care." *Journal of Biometrics and Biostatistics*, 1:103. doi: 10.4172/2155-6180.1000103. 640

Neuenschwander, B., Capkun-Niggli, G., Branson, M., and Spiegelhalter, D. J. (2010). "Summarizing historical information on controls in clinical trials." *Clinical Trials*, 7: 5–18. 640

Pocock, S. J. (1976). "The combination of randomized and historical controls in clinical trials." *Journal of Chronic Diseases*, 29: 175–188. 639, 642, 644, 646

Reich, B. J. and Hodges, J. S. (2008). "Identification of the variance components in the general two-variance linear model." *Journal of Statistical Planning and Inference*, 138: 1592–1604. 657

Saltz, L. B., Cox, J. V., Blanke, C., et al. and the Irinoteca Study Group (2000). "Irinotecan plus fluorouracil and leucovorin for metastatic colorectal cancer." *The New England Journal of Medicine*, 343: 905–914. 662, 673

Spiegelhalter, D. J. (2001). "Bayesian methods for cluster randomized trials with continuous responses." *Statistics in Medicine*, 20: 435–452. 640

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. London: Wiley. 639, 640, 641, 658

Sutton, A. J. and Abrams, K. R. (2001). "Bayesian methods in meta-analysis and evidence synthesis." *Statistical Methods in Medical Research*, 10: 277–303. 640

Appendix A

Full conditional posterior distributions for $\boldsymbol{\beta}_0$, σ, and σ_0 assuming noninformative priors for the commensurate prior linear model presented in Subsection 3.1 are as follows:

$q(\boldsymbol{\beta_0}|\boldsymbol{y}_0, \boldsymbol{\beta}, \sigma_0^2, \boldsymbol{\tau}) \propto$
$N\left[\left(X_0^T X_0/\sigma_0^2 + diag\{\boldsymbol{\tau}\}\right)^{-1}\left(X_0^T X_0 \hat{\boldsymbol{\beta}}_0/\sigma_0^2 + diag\{\boldsymbol{\tau}\}\boldsymbol{\beta}\right), \left(X_0^T X_0/\sigma_0^2 + diag\{\boldsymbol{\tau}\}\right)^{-1}\right],$
$q(\sigma^2|\boldsymbol{y}, \boldsymbol{\beta}, \lambda) \propto \Gamma^{-1}\left\{(n-1)/2, \parallel \boldsymbol{y} - X\boldsymbol{\beta} - \boldsymbol{d}\lambda \parallel^2 /2\right\}, \text{ and}$
$q(\sigma_0^2|\boldsymbol{y}_0, \boldsymbol{\beta}_0) \propto \Gamma^{-1}\left\{(n_0-1)/2, \parallel \boldsymbol{y}_0 - X_0\boldsymbol{\beta}_0 \parallel^2 /2\right\}.$

Appendix B

Here we present full conditional posterior distributions corresponding to the commensurate prior one-way random effects model presented in Subsection 3.2. Let $\hat{\mu}_{0,\mu} =$

$$\left[\left\{\frac{\sum_{j=1}^{m_0}\sum_{k=1}^{n_{0j}} y_{0jk}(1-r_{0j}^{-1})}{\sum_{j=1}^{m_0} n_{0j}(1-r_{0j}^{-1})}\right\}\bigg/\left\{\frac{\sigma_{\epsilon0}^2}{\sum_{j=1}^{m_0} n_{0j}(1-r_{0j}^{-1})}\right\}+\tau\mu\right]\bigg/\left(\frac{\sum_{j=1}^{m_0} n_{0j}(1-r_{0j}^{-1})}{\sigma_{\epsilon0}^2}+\tau\right) \text{ and } \Sigma_{0,\tau}$$

$= \frac{\sum_{j=1}^{m_0} n_{0j}(1-r_{0j}^{-1})}{\sigma_{\epsilon0}^2}+\tau.$ Full conditional posteriors for μ_0, λ, $\boldsymbol{u}$, $\boldsymbol{u}_0$, σ_ϵ^2, $\sigma_{\epsilon0}^2$, σ_u^2, and σ_{u0}^2 follow as:

$$q(\mu_0|\boldsymbol{y}_0,\mu,u_0,\sigma_{\epsilon0}^2,\sigma_{u0}^2,\tau) \propto N\left(\hat{\mu}_{0,\mu},\,\Sigma_{0,\tau}^{-1}\right),$$

$$q(\lambda|\boldsymbol{y},\boldsymbol{u},\mu,\sigma_\epsilon^2,\sigma_u^2) \propto N\left(\frac{\sum_{i=1}^m d_i\sum_{g=1}^{n_i}\{y_{ig}-n_i(\mu+u_i)\}}{\sum_{i=1}^m n_i d_i},\,\frac{\sigma_\epsilon^2}{\sum_{i=1}^m n_i d_i}\right),$$

$$q(\boldsymbol{u}|\boldsymbol{y}_0,\lambda,\mu,\sigma_\epsilon^2,\sigma_u^2) \propto \prod_{i=1}^m N\left\{\frac{\sigma_u^2\sum_{g=1}^{n_i}\{y_{ig}-(\mu+d_i\lambda)\}}{n_i\sigma_u^2+\sigma_\epsilon^2},\,\left(\frac{n_i}{\sigma_\epsilon^2}+\frac{1}{\sigma_u^2}\right)^{-1}\right\},$$

$$q(\boldsymbol{u}_0|\boldsymbol{y}_0,\mu_0,\sigma_{\epsilon0}^2,\sigma_{u0}^2) \propto \prod_{j=1}^{m_0} N\left\{\frac{\sigma_{u0}^2\sum_{k=1}^{n_{0j}}(y_{0jk}-\mu_0)}{n_{0j}\sigma_{u0}^2+\sigma_{\epsilon0}^2},\,\left(\frac{n_{0j}}{\sigma_{\epsilon0}^2}+\frac{1}{\sigma_{u0}^2}\right)^{-1}\right\},$$

$$q(\sigma_\epsilon^2|\boldsymbol{y},\lambda,\mu,\boldsymbol{u}) \propto \Gamma^{-1}\left(\frac{n-1}{2},\,\frac{\sum_{i=1}^m\sum_{g=1}^{n_i}(y_{ig}-\mu-u_i-d_i\lambda)^2}{2}\right),$$

$$q(\sigma_{\epsilon0}^2|\boldsymbol{y}_0,\boldsymbol{u}_0,\mu_0) \propto \Gamma^{-1}\left(\frac{n_0-1}{2},\,\frac{\sum_{j=1}^{m_0}\sum_{k=1}^{n_{0j}}(y_{0jk}-\mu_0-u_{0j})^2}{2}\right),$$

$$q(\sigma_u^2|\boldsymbol{y},\boldsymbol{u}) \propto \Gamma^{-1}\left(\frac{m-1}{2},\,\frac{\sum_{i=1}^m u_i^2}{2}\right), \text{ and}$$

$$q(\sigma_{u0}^2|\boldsymbol{y}_0,\boldsymbol{u}_0) \propto \Gamma^{-1}\left(\frac{m_0-1}{2},\,\frac{\sum_{j=1}^m u_{0j}^2}{2}\right).$$

Appendix C

Here we present full conditional posteriors for the general commensurate prior linear mixed model presented in Subsection 3.2. The full conditional posteriors for λ, $\boldsymbol{\beta}_0$, $\boldsymbol{u}$, $\boldsymbol{u}_0$, D^{-1}, D_0^{-1}, R^{-1}, and R_0^{-1} follow as:

$$q(\lambda|\boldsymbol{\beta},\boldsymbol{u},R,\boldsymbol{y}) \propto N\left\{\left(\boldsymbol{d}^T R^{-1}\boldsymbol{d}\right)^{-1}\boldsymbol{d}^T R^{-1}(\boldsymbol{y}-X\boldsymbol{\beta}-Z\boldsymbol{u}),\,\left(\boldsymbol{d}^T R^{-1}\boldsymbol{d}\right)^{-1}\right\},$$

$$q(\boldsymbol{\beta}_0|\boldsymbol{y}_0,\lambda,\boldsymbol{\beta},R_0,\boldsymbol{\tau}) \propto$$

$$N_p\left[\left(X_0^T\Sigma_0 X_0+diag\{\boldsymbol{\tau}\}\right)^{-1}\left(X_0^T\Sigma_0 X_0\tilde{\boldsymbol{\beta}}_0+diag\{\boldsymbol{\tau}\}\boldsymbol{\beta}\right),\,\left(X_0^T\Sigma_0 X_0+diag\{\boldsymbol{\tau}\}\right)^{-1}\right]$$

$$q(\boldsymbol{u}|\lambda,\boldsymbol{\beta},D,R,\boldsymbol{y}) \propto$$

$$N_m\left\{\left(D^{-1}+Z^T R^{-1}Z\right)^{-1}Z^T R^{-1}(\boldsymbol{y}-X\boldsymbol{\beta}-\boldsymbol{d}\lambda),\,\left(D^{-1}+Z^T R^{-1}Z\right)^{-1}\right\},$$

$$q(\boldsymbol{u}_0|\boldsymbol{\beta}_0, D_0, R_0, \boldsymbol{y}_0) \propto$$

$$N_{m0}\left\{\left(D_0^{-1} + Z_0^T R_0^{-1} Z_0\right)^{-1} Z_0^T R_0^{-1}(\boldsymbol{y}_0 - X_0\boldsymbol{\beta}_0),\ \left(D_0^{-1} + Z_0^T R_0^{-1} Z_0\right)^{-1}\right\},$$

$$q(D^{-1}|\lambda, \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{y}) \propto W\left\{\left(\boldsymbol{u}\boldsymbol{u}^T + \phi\tilde{D}\right), m + \phi\right\},$$

$$q(D_0^{-1}|\boldsymbol{\beta}_0, \boldsymbol{u}_0, \boldsymbol{y}_0) \propto W\left\{\left(\boldsymbol{u}_0\boldsymbol{u}_0^T + \phi_0\tilde{D}_0\right), m_0 + \phi_0\right\},$$

$$q(R^{-1}|\lambda, \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{y}) \propto W\left[\left\{\|\ \boldsymbol{y} - X\boldsymbol{\beta} - \boldsymbol{d}\lambda - Z\boldsymbol{u}\ \|^2 + \rho\tilde{R}\right\}, n + \rho\right],\ \text{and}$$

$$q(R_0^{-1}|\boldsymbol{\beta}_0, \boldsymbol{u}_0, \boldsymbol{y}_0) \propto W\left[\left\{\|\ \boldsymbol{y}_0 - X_0\boldsymbol{\beta}_0 - Z_0\boldsymbol{u}_0\ \|^2 + \rho_0\tilde{R}_0\right\}, n_0 + \rho_0\right],\ \text{where}\ \|\ .\ \|\ \text{de-}$$
notes the inner product.

Acknowledgments