Biostatistics (2022) **23**, 2, *pp.* 591–608
doi: 10.1093/biostatistics/kxaa044
Advance Access publication on November 6, 2020

Downloaded from https://academic.oup.com/biostatistics/article/23/2/591/5957418 by Univ of North Carolina at Chapel Hill Health Sci Lib user on 22 July 2022

Bayesian design of clinical trials using joint models for longitudinal and time-to-event data

JIAWEI XU, MATTHEW A. PSIODA, JOSEPH G. IBRAHIM*
Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA
ibrahim@bios.unc.edu

Summary

Joint models for longitudinal and time-to-event data are increasingly used for the analysis of clinical trial data. However, few methods have been proposed for designing clinical trials using these models. In this article, we develop a Bayesian clinical trial design methodology focused on evaluating the treatment's effect on the time-to-event endpoint using a flexible trajectory joint model. By incorporating the longitudinal outcome trajectory into the hazard model for the time-to-event endpoint, the joint modeling framework allows for non-proportional hazards (e.g., an increasing hazard ratio over time). Inference for the time-to-event endpoint is based on an average of a time-varying hazard ratio which can be decomposed according to the treatment's direct effect on the time-to-event endpoint and its indirect effect, mediated through the longitudinal outcome. We propose an approach for sample size determination for a trial such that the design has high power and a well-controlled type I error rate with both operating characteristics defined from a Bayesian perspective. We demonstrate the methodology by designing a breast cancer clinical trial with a primary time-to-event endpoint and where predictive longitudinal outcome measures are also collected periodically during follow-up.

Keywords: Bayesian design; Clinical trials; Joint models; Sampling prior.

1. Introduction

In clinical trials with time-to-event endpoints, often many biologic outcomes are measured longitudinally throughout the follow-up period (Brown and Ibrahim, 2003). The longitudinal outcomes, such as quality of life (QOL) measurements or immune response measures (e.g., CD4 counts), are typically measured intermittently at potentially different times with a potentially different number of measurements for each patient (Wulfsohn and Tsiatis, 1997). In many cases, the longitudinal data are predictive of time-to-event outcomes such as overall survival (OS), disease-free survival (DFS), or progression-free survival (PFS). Traditional approaches to time-to-event analysis in clinical trials have often ignored longitudinal outcome data when modeling the time-to-event distribution (e.g., Cox model) and thus failed to account for, or capitalize on, relationships between the two outcomes. Moreover, situations arise where the effect of treatment on the time-to-event outcome may be mediated by the longitudinal outcome (e.g., cancer immunotherapy (treatment) leads to immune system mobilization (mediator) which leads to improved

*To whom correspondence should be addressed.

© The Author 2020. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

time-to-event outcomes). In these cases, incorporating the longitudinal outcome trajectory into the time-to-event model may lead to a better understanding of treatment effects and may better account for deviations from the ubiquitous proportional hazards assumption. Approaches that jointly model longitudinal and time-to-event data offer possibilities for increased efficiency in the analysis of these types of data. One benefit of using joint models for the simultaneous analysis of longitudinal and time-to-event data is that joint models can produce more efficient estimates of treatment effects on both the longitudinal and time-to-event outcomes (Ibrahim *and others*, 2010). This greater efficiency will allow a smaller sample size and/or higher power in a trial. Chen *and others* (2011) also show that joint models can provide treatment estimates with less bias.

Early work on the use of joint models stemmed from clinical trials in acquired immune deficiency syndrome, where immunologic markers such as CD4 counts are measured frequently. See for example the work of De Gruttola and Tu (1994), Faucett and Thomas (1996), Wulfsohn and Tsiatis (1997), and Chi and Ibrahim (2007). Other applications of joint models include using QOL data in a cancer context. QOL data are typically collected via a questionnaire or assessed through monitoring adverse events during the follow-up period. Selected works on the use of joint models in this area include the works of Chi and Ibrahim (2006, 2007). Ibrahim *and others* (2004) and Brown and Ibrahim (2003) apply joint models to cancer-vaccine trials, where vaccines are intended to mobilize patients' immune response to destroy tumor cells. Measures of the immune response are often measured repeatedly to facilitate the study of the relationship between immune response and time-to-recurrence or death.

In this article, we develop a Bayesian clinical trial design framework using a joint model for longitudinal and time-to-event data. For ease of exposition, we focus on the design of a parallel two-group randomized, controlled trial. We assume the primary endpoint is a time-to-event endpoint (e.g., PFS) and that a longitudinal outcome (e.g., QOL) is measured repeatedly during the follow-up period, and that it potentially provides predictive or prognostic information about the time-to-event endpoint. We propose the use of a trajectory joint model that incorporates patient-specific random effects to account for patient-level heterogeneity in both the longitudinal and time-to-event outcomes. The joint model proposed allows the treatment to have both direct and indirect effects on the time-to-event endpoint. The direct effect is assumed to be multiplicative on the hazard for the time-to-event endpoint (i.e., consistent with the proportional hazard assumption), while the indirect effect is mediated through the longitudinal outcome. Thus, the indirect effect is characterized by the treatment's effect on the longitudinal outcome and the longitudinal outcome's effect on the time-to-event endpoint. When the treatment effect is entirely indirect, the longitudinal outcome can be considered a surrogate, as defined by Prentice (1989).

We develop a simulation-based approach whereby one can identify the necessary sample size required to obtain the desired level of Bayesian power while controlling a Bayesian type I error rate. Bayesian (i.e., average) type I error rate and power are defined with respect to sampling prior distributions which are based on the null and alternative hypotheses, respectively (Psioda and Ibrahim, 2018, 2019). For the special case where the sampling priors place a point-mass on a fixed value of the model parameters, which is our focus in this article, the Bayesian type I error rate and power for a design closely align with the frequentist versions. We evaluate the operating characteristics of designs based on a joint model incorporating patient-specific random effects to capture patient-level heterogeneity in outcomes, a simplified joint model that omits the random effects but otherwise incorporates the longitudinal outcome trajectory in the time-to-event model, a Cox proportional hazards regression model, and the log-rank test. Our results demonstrate that the random effects joint model outperforms the fixed effect joint model and the commonly used alternative methods, even in circumstances where the trajectory function is not perfectly specified.

The rest of this article is organized as follows: in Section 2, we introduce a trajectory joint model, define an average time-varying hazard ratio, and discuss its decomposition into direct and indirect treatment effect contributions. We also develop the study design and Bayesian sample size determination strategy. In Section 3, we provide a simulation study comparing our design based on the proposed joint model

Downloaded from https://academic.oup.com/biostatistics/article/23/2/591/5957418 by Univ of North Carolina at Chapel Hill Health Sci Lib user on 22 July 2022

to designs based on other commonly used methods (e.g., Cox model). We close the article with some discussion in Section 4.

2. Methods

2.1. *Trajectory joint models*

Let $y_i(t)$ be the longitudinal outcome at time t for patient i where $y_i(t) = \mu_i(t) + \epsilon_i(t)$ with $\epsilon_i(t) \sim N(0, \sigma^2)$. We refer to the conditional expectation of $y_i(t)$, denoted by $\mu_i(t)$, as the *longitudinal process* and consider a design model based on a trajectory function given as follows:

$$\begin{aligned}\mu_i(t) = E[y_i(t)|\boldsymbol{\theta}_i] &= \boldsymbol{g}(t)^T\boldsymbol{\theta}_i + \boldsymbol{X}_i(t)^T\boldsymbol{\gamma}\\ &= \boldsymbol{g}(t)^T\boldsymbol{\theta}_i + \boldsymbol{g}(t)^T\boldsymbol{\gamma}_t + x_i\boldsymbol{g}(t)^T\boldsymbol{\gamma}_x + \boldsymbol{z}_i^T\boldsymbol{\gamma}_z,\end{aligned} \tag{2.1}$$

where

- $\boldsymbol{g}(t)$ is a function of time t,
- $\boldsymbol{\theta}_i \sim N(\mathbf{0}, \Sigma_\theta)$ is a mean zero random effect with positive definite covariance matrix Σ_θ,
- $\boldsymbol{X}_i(t) = [\boldsymbol{g}(t)^T, x_i\boldsymbol{g}(t)^T, \boldsymbol{z}_i^T]^T$ represents the covariate process, with treatment indicator x_i and baseline covariate vector $\boldsymbol{z}_i$, and
- $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_t, \boldsymbol{\gamma}_x, \boldsymbol{\gamma}_z]$ is a vector of regression coefficients with $\boldsymbol{\gamma}_t$, $\boldsymbol{\gamma}_x$, and $\boldsymbol{\gamma}_z$ corresponding to $\boldsymbol{g}(t)$, $x_i\boldsymbol{g}(t)$, and $\boldsymbol{z}_i$, respectively.

On the basis of randomization and assuming that the longitudinal measurement at time $t = 0$ is pre-treatment, the main treatment effect can be set to 0 and excluded from the model. In our development, we keep the main effect simply for ease of exposition. Earlier works (Rizopoulos, 2010, 2016; Zhang *and others*, 2016) have included both treatment and time covariates in the model but not their interaction. Such a model corresponds to an instantaneous and constant difference between treatment groups over time, which may not be plausible in practice for the longitudinal outcome. Thus, we incorporate the interaction term between treatment and the time function to provide increased flexibility in how the impact of treatment can change over time.

For the time-to-event model, we assume a hazard function which takes form

$$\log\lambda_i(t) = \log\lambda_0(t) + \beta\left\{\boldsymbol{g}(t)^T\boldsymbol{\theta}_i + \boldsymbol{g}(t)^T\boldsymbol{\gamma}_t + x_i\boldsymbol{g}(t)^T\boldsymbol{\gamma}_x\right\} + x_i\alpha_x + \boldsymbol{z}_i^T\boldsymbol{\alpha}_z, \tag{2.2}$$

where β is an association parameter that controls the influence of the longitudinal process on the time-to-event distribution, α_x and $\boldsymbol{\alpha}_z$ are direct effects of the treatment and covariates on the time-to-event distribution, and $\boldsymbol{\lambda} = (\lambda_{01}, ..., \lambda_{0K})^T$ is a K-component piecewise constant baseline hazard associated with a fixed partition of the time axis. Specifically for the baseline hazard, we denote the $K-1$ change points by $L_1, ..., L_{K-1}$ which satisfy $L_0 = 0 < L_1 < ... < L_{K-1} < L_K = \infty$ and thus $\lambda_0(t) = \lambda_{0k}$ for $t \in [L_{k-1}, L_k)$. One can see that the model allows for deviations from the proportional hazards assumption on the treatment effect through the quantity $\beta\boldsymbol{g}(t)^T\boldsymbol{\gamma}_x$ in the hazard function when $\beta \neq 0$ and $\boldsymbol{\gamma}_x \neq 0$.

Zhang *and others* (2016) define a similar longitudinal process as (2.1) with $\boldsymbol{g}(t)$ being a polynomial vector of time. In contrast to (2.2), they only incorporate the random term $\boldsymbol{g}(t)^T\boldsymbol{\theta}_i$ into the hazard function, which accounts for patient-level heterogeneity in time-to-event outcomes but does not account for effects of the treatment on the hazard that are mediated by the longitudinal outcome (i.e., indirect effects). This

Downloaded from https://academic.oup.com/biostatistics/article/23/2/591/5957418 by Univ of North Carolina at Chapel Hill Health Sci Lib user on 22 July 2022

Downloaded from https://academic.oup.com/biostatistics/article/23/2/591/5957418 by Univ of North Carolina at Chapel Hill Health Sci Lib user on 22 July 2022

approach may be appropriate when the longitudinal outcome is prognostic of patient outcomes but not thought to be on the causal pathway for the time-to-event outcome. In our development of the joint model, we incorporate the longitudinal process $\mu_i(t)$ (excluding the baseline covariate component $z_i^T \gamma_z$) into the hazard function for the two-pronged purpose of capturing patient-level heterogeneity in time-to-event outcomes (i.e., overdispersion) and allowing for more accurate characterization of a treatment's effect on the time-to-event outcome in cases where mediation of the effect by the longitudinal outcome is plausible. In the case of a strong indirect treatment effect that does not change over time, survival curves should separate immediately. However, in cancer immunotherapy trials, a "late separation" is often observed (Zhang, 2017), which suggests a minimal treatment effect on the hazard early in the observation period. Thus, modeling a non-constant treatment effect through the longitudinal process (e.g., an effect that increases over a period of time) may be advantageous.

Let $\boldsymbol{\xi} = (\boldsymbol{\gamma}, \boldsymbol{\lambda}, \beta, \boldsymbol{\alpha}, \Sigma_\theta)$ denotes the complete set of fixed effect parameters and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n)$ denote the collection of random effect vectors for the set of n patients enrolled in the trial. We denote the observed data for the complete set of n patients by $\mathbf{D}$. Suppose patient $i = 1, ..., n$ has the longitudinal outcome measured m_i times, denoted by $t_{i1}, ..., t_{im_i}$. We let $y_{ij} = y_i(t_{ij})$ and $\mathbf{X}_{ij} = \boldsymbol{X}_i(t_{ij})$ denote the observed outcome and covariate process at time t_{ij}, respectively, for $j = 1, ..., m_i$. The complete data likelihood $L(\boldsymbol{\xi}, \boldsymbol{\theta}|\mathbf{D})$ is written as

$$L(\boldsymbol{\xi},\boldsymbol{\theta}|\mathbf{D}) = \prod_{i=1}^{n}\left[\prod_{j=1}^{m_i} f(y_{ij}|\mathbf{X}_{ij},\boldsymbol{\theta}_i,\boldsymbol{\gamma},\sigma^2)\right] f(s_i,\delta_i|\mathbf{z}_i,\boldsymbol{\theta}_i,\boldsymbol{\gamma},\boldsymbol{\lambda},\beta,\boldsymbol{\alpha}) f(\boldsymbol{\theta}_i|\Sigma_\theta), \tag{2.3}$$

where the density for the time-to-event endpoint takes the form

$$\begin{aligned} f(s_i,\delta_i|\mathbf{z}_i,\boldsymbol{\theta}_i,\boldsymbol{\gamma},\boldsymbol{\lambda},\beta,\boldsymbol{\alpha}) &= \left\{\lambda_0(s_i)\exp\left(\beta\mu_i(s_i) + x_i\alpha_x + \mathbf{z}_i^T\boldsymbol{\alpha}_z\right)\right\}^{\delta_i} \\ &\quad \times \exp\left\{-\int_0^{s_i} (\lambda_0(t)e^{\beta\mu_i(t)+x_i\alpha_x+z_i^T\boldsymbol{\alpha}_z})dt\right\}, \end{aligned}$$

where s_i is the observation time and δ_i is an indicator for whether an event is observed for the ith patient. Integrating over the random effects $\boldsymbol{\theta}_i$ for $i = 1, ..., n$ gives the observed data likelihood

$$L(\boldsymbol{\xi}|\mathbf{D}) = \prod_{i=1}^{n}\int_{-\infty}^{\infty}\left[\left\{\prod_{j=1}^{m_i} f(y_{ij}|\mathbf{X}_{ij},\boldsymbol{\theta}_i,\boldsymbol{\gamma},\sigma^2)\right\} f(s_i,\delta_i|\mathbf{z}_i,\boldsymbol{\theta}_i,\boldsymbol{\gamma},\boldsymbol{\lambda},\beta,\boldsymbol{\alpha}) f(\boldsymbol{\theta}_i|\Sigma_\theta)\right] d\boldsymbol{\theta}_i, \tag{2.4}$$

which cannot be computed in closed form.

2.2. *Piecewise linear time trajectory function*

As noted above, Zhang *and others* (2016) propose a model for the longitudinal process that takes $\boldsymbol{g}(t)$ to be a polynomial vector. Indeed, this is a common practice for these types of models (Chen *and others*, 2011; Crowther *and others*, 2013). For our approach, we assume a continuous, semiparametric, piecewise linear time trajectory function for $\boldsymbol{g}(t)$ which is constructed using a pre-specified number of segments with specified knot (or change point) locations. This approach is advantageous because it allows for a more flexible shape for the time trajectory. For example, a piecewise linear curve allows for the possibility that the trajectory levels off at some point in time whereas commonly used linear or quadratic polynomial trajectory functions cannot accommodate this behavior.

We assume a piecewise linear trajectory function with M components and $M - 1$ knots denoted by $k_1, ..., k_{M-1}$ which satisfy $k_0 = 0 < k_1 < \cdots < k_{M-1} < k_M = \infty$. Define the $M + 1$ dimensional vector

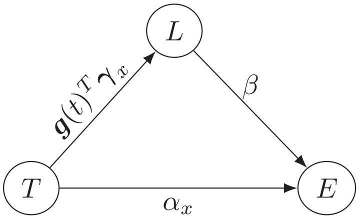

Fig. 1. Causal diagram for the treatment effect on the time-to-event outcome. T, treatment effect; L, longitudinal outcome; E, time-to-event outcome.

$\boldsymbol{g}(t)$ to have 1 as its first component and $f_m(t) = \max\{\min\{t, k_m\} - k_{m-1}, 0\}$ as its $(m+1)$th component for $m = 1, ..., M$ (i.e., $\boldsymbol{g}(t)^T = [1, f_1(t), \ldots, f_M(t)]$). It is easy to see based on this construction that $\boldsymbol{g}(t)$ is a continuous piecewise linear function of time. Based on this, one can see that the first components of $\boldsymbol{\theta}_i$, $\boldsymbol{\gamma}_t$, and $\boldsymbol{\gamma}_x$ correspond to intercept parameters and the remaining M components of each vector combine to determine each patient trajectory's slope over the M time intervals in the time axis partition.

2.3. *Direct and indirect treatment effects*

As mentioned above, the proposed joint model allows for the treatment to have both direct and indirect effects on the time-to-event outcome. We refer to α_x in (2.2) as the *direct* effect of treatment on the time-to-event endpoint. In the proposed joint model, we have assumed that the direct effect is consistent with a proportional hazards assumption. The *indirect* effect of treatment is $\beta\boldsymbol{g}(t)^T\boldsymbol{\gamma}_x$ which can be decomposed as the effect of treatment on the longitudinal outcome $\boldsymbol{g}(t)^T\boldsymbol{\gamma}_x$ multiplied by the effect of the longitudinal outcome on the time-to-event outcome β. Here, we note that the causal interpretation of the indirect treatment effect follows from the assumption that $\mu_i(t) = \boldsymbol{g}(t)^T\boldsymbol{\theta}_i + \boldsymbol{X}_i(t)^T\boldsymbol{\gamma}$ is the *true trajectory* for patient i with $\epsilon_i(t)$ corresponding to measurement error for the longitudinal outcome at time t.

Figure 1 provides a simple illustration of the causal diagram. There are several possibilities for how the treatment could impact the time-to-event distribution. The treatment could have only a direct effect ($\alpha_x \neq 0, \beta\boldsymbol{g}(t)^T\boldsymbol{\gamma}_x = 0$), only an indirect effect ($\alpha_x = 0, \beta\boldsymbol{g}(t)^T\boldsymbol{\gamma}_x \neq 0$), or both types of effects. The proposed formulation of the joint model is ideal for cases where both direct and indirect effects are plausible and where it is of interest to design a future trial with a goal of identifying their contribution to the total treatment effect.

2.4. *Study design*

We consider a study design where the goal is to demonstrate the superiority of an investigational treatment to control with respect to a time-to-event endpoint such as PFS. While it is of interest to make a global statement of the effectiveness of treatment, quantification of both direct and indirect effects is of interest. We assume that the longitudinal outcomes are measured at baseline and follow-up time points scheduled at regular intervals until some fixed time t_{traj}. Patients are followed for the time-to-event endpoint starting at baseline and up to time $t_{\max}$ corresponding to a random point in time when a specified number of events have accrued in the trial. Let $h_1(t)$ and $h_0(t)$ be the hazards for comparable patients in the treatment and control groups. Here, *comparable* implies that although the patients are treated differently, their random effects and baseline covariate vectors are equal.

We define the one-sided null and alternative hypotheses for superiority as: $H_0 : \phi(t_0) \geq 1$ versus $H_1 : \phi(t_0) < 1$, where $\phi(t_0) = G^{-1}\left(\int_0^{t_0} G\left(\frac{h_1(t)}{h_0(t)}\right)\Omega(t)\mathrm{d}t\right)$ for fixed t_0 with $t_{\text{traj}} \leq t_0 \leq t_{\max}$, $G(x)$ is a strictly increasing function, and $\Omega(t)$ is a non-negative weight function such that $\int_0^{t_0}\Omega(t)\mathrm{d}t = 1$. The quantity

Downloaded from https://academic.oup.com/biostatistics/article/23/2/591/5957418 by Univ of North Carolina at Chapel Hill Health Sci Lib user on 22 July 2022

$\phi(t_0)$ is thus derived using a very general framework for computing an average for a time-varying hazard ratio (Chen *and others*, 2015).

Chen *and others* propose and discuss choices for $G(x)$ such as $G(x) = x$, which gives an identity transformation such that $\phi(t_0) = \int_0^{t_0} \frac{h_1(t)}{h_0(t)} \Omega(t)\mathrm{d}t$, and $G(x) = \log(x)$. For this article, we consider $G(x) = \log(x)$ for several reasons. First, for this choice, when the log hazard and weight functions are piecewise linear functions in time, $\phi(t_0)$ has a closed form. Second, $\phi(t_0)$ is symmetric in $h_1(t)$ and $h_0(t)$ (i.e., $\phi(h_0/h_1) \cdot \phi(h_1/h_0) = 1$). Lastly, $\phi(t_0)$ can be decomposed into two components corresponding to the direct and indirect effects, respectively (Section 2.5). For more discussion on choices for $G(x)$, we refer the reader to Chen *and others* (2015).

We propose the weight function defined by $\Omega(t) \propto |\beta \boldsymbol{g}(t)^T \boldsymbol{\gamma}_x + \alpha_x|$, where $\beta \boldsymbol{g}(t)^T \boldsymbol{\gamma}_x + \alpha_x$ can be viewed as the difference in log hazards for two comparable subjects. Technically, $\Omega(t)$ is only a proper weight function when $\beta \boldsymbol{\gamma}_x$ and/or α_x is non-zero. In practice, being a function of unknown parameters, $\Omega(t)$ must be estimated from the data and so the estimated weight function will likely never be identically zero. Nonetheless, is it reassuring to add a small positive constant (e.g., $c_0 = 0.001$) to the weight function in software implementations to avoid any instability issues that might arise should such a pathological dataset arise. Such an approach was taken in our software implementation. Note that the weight function $\Omega(t)$ is directly incorporated into $\phi(t_0)$ to define the function of $\boldsymbol{\xi}$ that serves as the basis for inference. In other words, $\Omega(t)$ is not treated as fixed based on a plug-in type estimator.

2.5. *Decomposition of $\phi(t_0)$*

When $G(x) = \log(x)$, for the log hazard in (2.2) and for the arbitrary weight function $\Omega(t)$, it is straightfoward to show that

$$\phi(t_0) = \exp\left\{\int_0^{t_0} \alpha_x \Omega(t)\mathrm{d}t\right\} \exp\left\{\int_0^{t_0} \beta \boldsymbol{g}(t)^T \boldsymbol{\gamma}_x \Omega(t)\mathrm{d}t\right\} = e^{\alpha_x} \exp\left\{\int_0^{t_0} \beta \boldsymbol{g}(t)^T \boldsymbol{\gamma}_x \Omega(t)\mathrm{d}t\right\},$$

where $\phi_D = e^{\alpha_x}$ is the direct effect contribution to the average hazard ratio and $\phi_I = \exp\left\{\int_0^{t_0} \beta \boldsymbol{g}(t)^T \boldsymbol{\gamma}_x \Omega(t)\mathrm{d}t\right\}$ is the indirect effect contribution. Note that ϕ_D does not depend on time or the weight function $\Omega(t)$, whereas the value of ϕ_I depends on both. Thus, changing $\Omega(t)$ will only affect the indirect effect contribution ϕ_I to the average hazard ratio.

2.6. *Arbitrary $G(x)$ and $\Omega(t)$ functions*

In addition to the specific choices for $G(x)$ and $\Omega(t)$ discussed previously, one can approximate $\phi(t_0)$ for arbitrary choices using a trapezoidal approximation to the integral. Consider a partition of the interval $(0, t_0)$ into K intervals and let $0 = s_0 < s_1 < ... < s_K = t_0$ with $\Delta_k = s_k - t_{s-1}$ corresponding to the width of interval k. Then we have

$$\phi(t_0) \approx G^{-1}\left[\sum_{k=1}^{K} \frac{1}{2}\Delta_k \left\{\Omega(s_{k-1}) G\left(\frac{h_1(s_{k-1})}{h_0(s_{k-1})}\right) + \Omega(s_k) G\left(\frac{h_1(s_k)}{h_0(s_k)}\right)\right\}\right]. \tag{2.5}$$

Thus, assuming one can fit the model to estimate $\pi(\boldsymbol{\xi}|\mathbf{D})$, the posterior distribution $\pi(\phi(t_0)|\mathbf{D})$ can be readily estimated by applying the approximate transformation in (2.5) which can be made arbitrarily accurate by taking K to be large.

Downloaded from https://academic.oup.com/biostatistics/article/23/2/591/5957418 by Univ of North Carolina at Chapel Hill Health Sci Lib user on 22 July 2022

2.7. *Model estimation and posterior inference*

Even though the observed data likelihood $L(\boldsymbol{\xi}|\mathbf{D})$ in (2.4) cannot be computed in closed form, it is straightforward to estimate $\pi(\boldsymbol{\xi}|\mathbf{D})$ using Markov chain Monte Carlo (MCMC) methods to sample from $\pi(\boldsymbol{\xi}, \boldsymbol{\theta}|\mathbf{D})$, based on the complete data likelihood in (2.3). The samples for $\boldsymbol{\xi}$ can then be used to approximate $\pi(\boldsymbol{\xi}|\mathbf{D})$. Rizopoulos (2016) developed an R package, JMBayes, for fitting joint models using MCMC methods. That software provides a suite of commonly used priors and a flexible framework for modeling the baseline hazards. We refer the reader to the article by Rizopoulos (2016) for more specific details on the implementation.

Due to the substantial computational burden of using MCMC for large scale simulation studies, we use a posterior approximation for inference during design simulations. Using the observed data likelihood $L(\boldsymbol{\xi}|\mathbf{D})$, the posterior distribution for the fixed effects takes the form $\pi(\boldsymbol{\xi}|\mathbf{D}) \propto L(\boldsymbol{\xi}|\mathbf{D})\pi^{(f)}(\boldsymbol{\xi})$, where $\pi^{(f)}(\boldsymbol{\xi})$ is the fitting prior (Wang and Gelfand, 2002). When there is little prior information on $\boldsymbol{\xi}$, the fitting prior is generally specified to be non-informative, and can be an improper prior as long as $\pi(\boldsymbol{\xi}|\mathbf{D})$ is proper. We use standard software (e.g., The NLMIXED Procedure) to obtain an asymptotic posterior approximation for $\pi(\boldsymbol{\xi}|\mathbf{D})$ based on the Bayesian central limit theorem (Chen, 1985). Specifically, $\pi(\boldsymbol{\xi}|\mathbf{D}) \approx \text{Normal}(\hat{\boldsymbol{\xi}}, \hat{\Sigma}_{\boldsymbol{\xi}})$, where $\hat{\boldsymbol{\xi}}$ and $\hat{\Sigma}_{\boldsymbol{\xi}}$, respectively, are the maximum likelihood estimator (MLE) and approximate asymptotic covariance for the MLE, obtained by maximizing the approximate observed data likelihood obtained by integrating over the $\boldsymbol{\theta}_i$ using Gaussian Quadrature. When $\phi(t_0)$ has a closed form or is approximated by (2.5), application of the delta method (Doob, 1935) yields an approximate posterior for $\pi(\phi(t_0)|\mathbf{D}) \approx \text{Normal}(\hat{\phi}(t_0), \hat{\sigma}_{\phi}^2)$, where $\hat{\phi}(t_0)$ and $\hat{\sigma}_{\phi}^2$ are the MLE and estimated asymptotic variance of $\phi(t_0)$, respectively. It follows that $P(\phi(t_0) < 1|\mathbf{D}) \approx 1 - \Phi\left(\frac{\hat{\phi}(t_0)-1}{\hat{\sigma}_{\phi}}\right)$.

An appealing alternative approximation for $\pi(\phi(t_0)|\mathbf{D})$ for the case when $G(x)$ and $\Omega(t)$ do not yield a closed form for $\phi(t_0)$ can be obtained using Monte Carlo methods as follows. First, one must obtain samples $\boldsymbol{\xi}^{(1)}, ..., \boldsymbol{\xi}^{(M)}$ from $\pi(\boldsymbol{\xi}|\mathbf{D})$ based on the asymptotic approximation above. One can then compute $\phi^{(m)}(t_0) = \phi(t_0|\boldsymbol{\xi}^{(m)})$ using (2.5) to obtain the approximate posterior. Calculation of posterior probabilities is based on the fraction of samples meeting the desired criterion (i.e., $\phi(t_0) < 1$).

2.8. *Bayesian sample size determination*

The proposed method is designed to facilitate identification of the smallest sample size required for a trial, subject to Bayesian type I error rate and power requirements. Following Psioda and Ibrahim (2018, 2019), we define the Bayesian type I error rate and power using user-specified null and alternative sampling prior distributions for $\boldsymbol{\xi}$, respectively. In our context, the null sampling prior gives non-zero weight to values of $\boldsymbol{\xi}$ such that $\phi(t_0) \geq 1$ and the alternative sampling prior such that $\phi(t_0) < 1$. For this article, we only consider point-mass sampling priors such that $\pi_0^{(s)}(\boldsymbol{\xi}) = 1(\boldsymbol{\xi} = \boldsymbol{\xi}_0)$ and $\pi_1^{(s)}(\boldsymbol{\xi}) = 1(\boldsymbol{\xi} = \boldsymbol{\xi}_1)$, where the superscript (s) indicates that the prior is a sampling prior, the subscript h indicates whether the prior corresponds to a null $(h = 0)$ or alternative $(h = 1)$ sampling prior, and $1\{A\}$ is an indicator that A is true. In the case of point-mass sampling priors, the Bayesian type I error rate and power align with the frequentist versions.

Let $\alpha^{(s)}$ and $\beta^{(s)}$ denote the Bayesian type I and II error rates. Prespecify p_0 as the threshold for substantial evidence such that we reject the null hypothesis if $P(\phi(t_0) < 1|\mathbf{D}) \geq p_0$. For a fixed value of $\boldsymbol{\xi}$, the null hypothesis rejection rate is defined as $r(\boldsymbol{\xi}) = E[1\{P(\phi(t_0) < 1|\mathbf{D}) \geq p_0\}|\boldsymbol{\xi}]$, where the expectation is with respect to the distribution of $\mathbf{D}$ given $\boldsymbol{\xi}$. The Bayesian type I error rate and power are defined as $\alpha^{(s)} = E[r(\boldsymbol{\xi})|\pi_0^{(s)}]$ and $1 - \beta^{(s)} = E[r(\boldsymbol{\xi})|\pi_1^{(s)}]$, which, for non-degenerate sampling priors, are weighted averages of $r(\boldsymbol{\xi})$ with weights determined by $\pi_0^{(s)}(\boldsymbol{\xi})$ and $\pi_1^{(s)}(\boldsymbol{\xi})$, respectively.

Downloaded from https://academic.oup.com/biostatistics/article/23/2/591/5957418 by Univ of North Carolina at Chapel Hill Health Sci Lib user on 22 July 2022

2.9. *Simulation-based sample size determination*

We propose using simulations to identify the required number of events (effectively the sample size in time-to-event trials) such that the trial design has sufficiently high Bayesian power. The number of patients enrolled in the trial may be chosen to obtain a specified number of events in a specified interval of time *on average*. Let the sample size and number of events be given by n and v, respectively. We consider an approach that fixes the ratio $r = \frac{n}{v}$ but varies the number of events. If n_1 patients result in obtaining v_1 events in a specific time frame, then to obtain $v_2 \geq v_1$ events in the same time frame, one should increase n_2 proportionally. By fixing r, we ensure that the trial will complete in a reasonable period of time which may be desirable to limit extrapolation of trajectories (unless longitudinal outcome assessment continues for the duration of follow-up).

We want to determine the smallest v such that the Bayesian power for the design is at least $1 - \beta^{(s)}$. A simulation-based sample size determination procedure is given below:

S1. Let $v_1, ..., v_K$ denote the potential event totals at which the trial might be stopped.
S2. Initialize $k = 1$.
S3. Compute the Bayesian power $1 - \beta_k^{(s)}$ based on v_k.
S4. If $1 - \beta_k^{(s)} \geq 1 - \beta^{(s)}$ then set $v = v_k$ and stop; otherwise, increment k and return to S3.

Note that the approximate Bayesian type I error rate will be $\alpha^{(s)}$ when one takes $p_0 = 1 - \alpha^{(s)}$ for the case where point-mass null sampling priors are used (along with a non-informative fitting prior). Thus, for the identified choice of v, it will generally be the case that $\alpha^{(s)} \approx 1 - p_0$ and so specific efforts to control the Bayesian type I error rate at level $\alpha^{(s)}$ are not generally needed when p_0 is chosen in this way. Nonetheless, one can always compute the exact Bayesian type I error rate via simulation to ensure it is sufficiently close to the desired nominal level. The simulations studies presented in Section 3 illustrate that the property $\alpha^{(s)} \approx 1 - p_0$ indeed holds quite well.

Now we expound more on step S3 from the simple algorithm given above. Letting B be the number of simulation studies to be performed, to estimate the Bayesian power $1 - \beta_k^{(s)}$ associated with event total v_k, one does the following:

S3.1 Sample $\boldsymbol{\xi}^{(b)}$ from the alternative sampling prior $\pi_1^{(s)}(\boldsymbol{\xi})$.

S3.2 Simulate the *observed* data $\mathbf{D}^{(b)}$.

S3.3 Estimate the posterior distribution $\pi(\phi(t_0)|\mathbf{D})$ using an approach described in Section 2.7 and compute the null hypothesis rejection indicator

$$r^{(b)} = 1\{P(\phi(t_0) < 1|\mathbf{D}^{(b)}) \geq p_0\}.$$

S3.4 Approximate the Bayesian power:

$$1 - \beta_k^{(s)} \approx \frac{1}{B}\sum_{b=1}^{B} r^{(b)}.$$

See Appendix A of the Supplementary material available at *Biostatistics* online for an algorithm that can be used to generate the observed data $\mathbf{D}^{(b)}$ as required for step S3.2 above.

Downloaded from https://academic.oup.com/biostatistics/article/23/2/591/5957418 by Univ of North Carolina at Chapel Hill Health Sci Lib user on 22 July 2022

3. Example application: Bayesian clinical design for breast cancer

Our design methodology is motivated by a breast cancer trial undertaken by the International Breast Cancer Study Group (IBCSG) (IBCSG, 1996). The trial, IBCSG Trial VI, was conducted in pre-menopausal women with node-positive breast cancer to investigate the efficacy of different durations of adjuvant chemotherapy (3 versus 6 cycles of CMF—cyclophosphamide, methotrexate, and fluorouracil) and whether the reintroduction of CMF provided added benefit. Treatment strategies were evaluated with respect to the OS and DFS endpoints. During the study, four measures of QOL (appetite, mood, coping, physical well-being) were scheduled to be collected at baseline and every 3 months for up to 2 years (Hürny *and others*, 1992).

For our example application, we consider the design of a similar trial evaluating two treatments (e.g., whether or not chemotherapy was reintroduced) with respect to a primary PFS time-to-event endpoint. We consider one QOL measure (e.g., coping score) in this application. Similar to IBCSG Trial VI, we assume QOL scores are collected every 3 months starting at baseline for up to 2 years, and that these scores are approximately normally distributed (after appropriate transformation, e.g., a square root transformation as was done in Zhang *and others* (2016)). In the simulated trials, patients were randomized to the two treatments using a 1:1 allocation scheme. Patient accrual was simulated to be uniform over a 1-year period and censoring (i.e., dropout) was assumed to follow a mixture distribution whereby patients had probability $\rho = 0.05$ of dropping out of the trial early and, conditional on being a dropout, the time to dropout was simulated to be uniform over a 5-year period. All patients were administratively censored when the desired number of events for the simulated trial was reached. This resulted in approximately 66% of patients having a censored time-to-event outcome on average with the dominant type of censorship being administrative censoring (consistent with the IBCSG data). In Appendix B of the Supplementary material available at *Biostatistics* online, we explore the impact of increased rates of non-informative dropout and its impact on the Bayesian type I error rate and power. Therein, we compare design properties based on $\rho \in \{0.05, 0.10, 0.20\}$ and the results show that the design method we propose is robust even under a substantial degree of non-informative dropout. In IBCSG Trial VI, the number of positive lymph nodes identified for lymphadenectomy was prognostic of PFS and so we included a binary covariate ($>$ 3 versus $\leq$ 3 nodes) in our hypothetical trial simulations. The covariate was simulated such that approximately 50% of the subjects were in the more severe group.

We assumed a piecewise linear trajectory function with a longitudinal process for patient i given by $\mu_i(t) = \theta_i + \boldsymbol{g}(t)^T\boldsymbol{\gamma}_t + x_i\boldsymbol{g}(t)^T\boldsymbol{\gamma}_x + \gamma_z z_i$, where z_i is an indicator that the patient had $>$ 3 positive lymph nodes. In design simulations, we considered a four-component piecewise linear trajectory function with knots at 0.25, 0.75, and 1.25. For the baseline hazard, we considered a five-component piecewise constant function with knots at times 1.91, 2.43, 3.00, and 3.80. Knot placement was determined such that each component of the trajectory and hazard had approximately the same number of longitudinal measures and events, respectively. In fitted models, the knot placements were assumed to be known.

For the construction of $\phi(t_0)$, we took $t_0 = 5$ which was equal to the approximate expected duration of the trial and $G(x) = \log(x)$ leading to a closed form for $\phi(t_0)$ as described in Section 2.4. We considered various sampling priors for the treatment effect and association parameters. For the association parameter, we considered point-mass priors on $\beta \in \{-0.15, -0.30, -0.45\}$. For the direct treatment effect, we considered point-mass priors on $\alpha_x \in \{0.0, -0.2\}$. Due to the randomized nature of the trial, we only considered $\gamma_{x,0} = 0.0$. For the treatment effect on the longitudinal process, we considered point-mass priors on the four slope parameters $\left\{\gamma_{x,1}, ..., \gamma_{x,4}\right\}$ including $\{0.0, 0.0, 0.0, 0.0\}$ (no effect), $\{0.2, 0.2, 0.2, 0.2\}$ (favorable linear trajectory), and $\{0.4, 0.3, 0.2, 0.1\}$ (decreasing effectiveness over time). Appendix C of the Supplementary material available at *Biostatistics* online provides an explanation regarding how one can determine whether a combination for $\left(\beta, \alpha_x, \boldsymbol{\gamma}_x\right)$ combine to determine whether a given sampling prior corresponds to a null or alternative scenario. The nuisance parameter values were taken to equal the

Downloaded from https://academic.oup.com/biostatistics/article/23/2/591/5957418 by Univ of North Carolina at Chapel Hill Health Sci Lib user on 22 July 2022

Table 1. *Control group parameter estimates for IBCSG data.*

Parameter estimates				
Parameter description	Parameter	Posterior mode	SD	Approximate 95% credible region
Covariance parameter estimates				
SD for random intercept	Σ_θ	0.71	0.02	(0.68, 0.75)
Standard deviation	σ	0.66	0.01	(0.64, 0.67)
Longitudinal parameter estimates				
Intercept	$\gamma_{t,0}$	0.27	0.05	(0.17, 0.36)
Node group	γ_z	−0.03	0.05	(−0.14, 0.07)
Time trajectory($\gamma_{t,1}$-$\gamma_{t,4}$)	$\gamma_{t,1}$	−0.32	0.19	(−0.68, −0.05)
	$\gamma_{t,2}$	−0.72	0.10	(−0.93, −0.52)
	$\gamma_{t,3}$	−0.14	0.11	(−0.35, 0.07)
	$\gamma_{t,4}$	−0.22	0.28	(−0.77, 0.34)
Survival parameter estimates				
Node group	α_z	0.77	0.11	(0.55, 0.98)
	$\log \lambda_1$	−3.61	0.14	(−3.90, −3.33)
	$\log \lambda_2$	−2.22	0.15	(−2.51, −1.93)
Baseline hazard	$\log \lambda_3$	−2.25	0.15	(−2.55, −1.95)
	$\log \lambda_4$	−2.50	0.16	(−2.82, −2.18)
	$\log \lambda_5$	−2.70	0.18	(−3.05,−2.35)

Notes: Posterior quantities are computed based on a Laplace approximation to the posterior.

approximate posterior modes based on our analysis of the IBCSG data as shown in Table 1. To identify the desired number of events required to achieve Bayesian power equal to 0.8, we considered $\nu = 100$ to 400 in increments of 25. A total of 4000 simulated trials were performed to estimate the operating characteristics for each choice of sampling prior size was considered.

We evaluate the performance of the proposed method in two types of scenarios: one where the joint model is correctly specified and another where some type of misspecification exists. Section 3.1 compares the performance of the proposed method with other methods when the joint model is correctly specified. Sections 3.2 and 3.3 demonstrate the robustness of the proposed method in scenarios where the joint model trajectory function and the random effect structure are misspecified, respectively.

3.1. *Evaluation of the joint model when correctly specified*

We evaluated the performance of the proposed joint model against a *simplified* joint model that omits the random effects ($\theta_i = 0$), a Cox proportional hazards model, and the log-rank test. Note that, due to the correlated nature of longitudinal outcomes, patient-level heterogeneity will essentially always exist and so the simplified joint model should not be viewed as a competitor of the joint model which includes random effects. We merely included the simplified version to assess the impact of ignoring patient-level heterogeneity in an otherwise correctly specified model. Estimates of power based on the different methods are shown in Figures 2 and 3. Figure 2 presents power when there is no direct effect (i.e., $\alpha_x = 0$). Figure 3 presents power when there is only a direct effect (column 1) and both direct and indirect effects (columns 2 and 3). In Appendix D of the Supplementary material available at *Biostatistics* online, to help the reader appreciate how direct and indirect effects manifest differences in survival curves, we present and compare

Downloaded from https://academic.oup.com/biostatistics/article/23/2/591/5957418 by Univ of North Carolina at Chapel Hill Health Sci Lib user on 22 July 2022

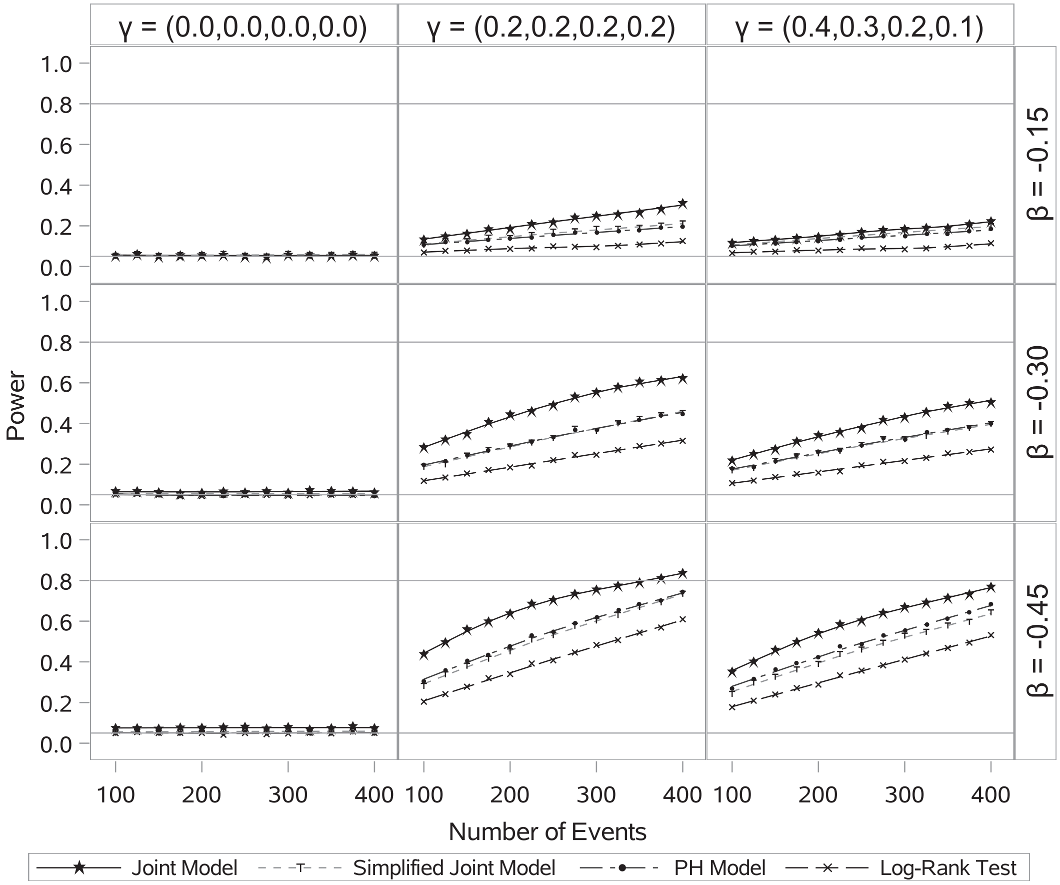

Fig. 2. Estimated power curves for no direct effect case (i.e., $\alpha_x = 0$). PH = proportional hazards.

curves for selected scenarios where there is only a direct effect, only an indirect effect, and where there are both types of effects.

All four analysis methods provide Bayesian type I error control at approximately the nominal level (i.e., $\alpha^{(s)} = 0.05$), regardless of the strength of the association parameter (Figure 2, column 1). A comparison of the power curves illustrates that the joint model always outperforms the standard log-rank test. When there is no indirect treatment effect (Figure 3, column 1), the power difference between the joint model and the proportional hazards model is modest, but increases as the association parameter increases in absolute value, reflecting the degraded performance of the proportional hazards model due to failing to capture patient heterogeneity. This suggests that power based on the joint model is as high as that based on the proportional hazards model even when the proportional hazards assumption holds. Columns 2 and 3 in both of Figures 2 and 3 illustrate that the joint model has significantly higher power than the proportional hazards model when there is a moderate or large indirect treatment effect.

Figure 3 (e.g., panel in row 2 and column 3) illustrates that the joint model that omits the random effect sometimes yields lower power compared to the proportional hazards model even though it is otherwise correctly specified. This phenomenon is not exhibited by the joint model that correctly accounts for the

Downloaded from https://academic.oup.com/biostatistics/article/23/2/591/5957418 by Univ of North Carolina at Chapel Hill Health Sci Lib user on 22 July 2022

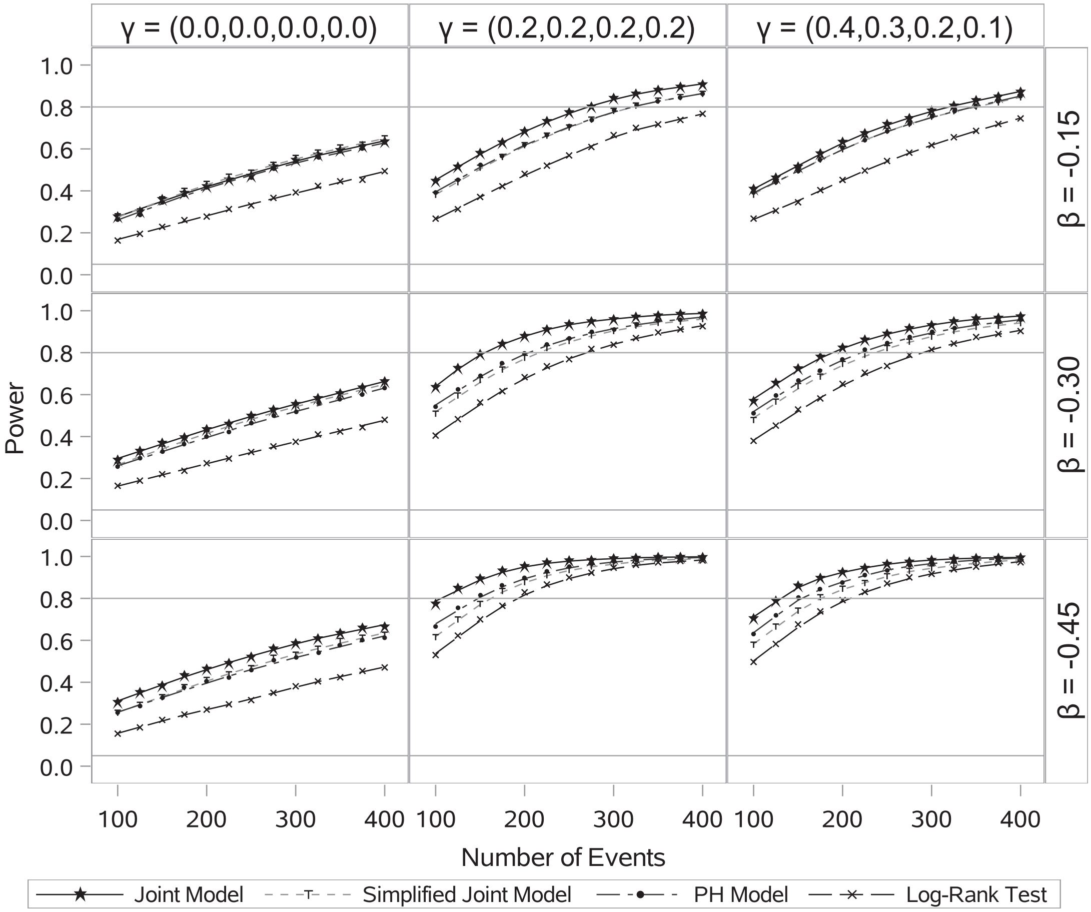

Fig. 3. Estimated power curves when the direct treatment effect $\alpha_x = -0.2$. PH = proportional hazards.

random effect. While these results demonstrate the importance of accounting for patient-level heterogeneity in the joint model, one should not construe them to imply the random effect structure in the joint model must be correct to gain efficiency through the use of that modeling technique. Indeed, the results we present in Section 3.3 illustrate the robustness of the proposed joint model when patient-level heterogeneity is accounted for in the analysis—but accounted for incorrectly. Performance of the joint model suffers unduly only when patient-level heterogeneity is ignored in the model (i.e., when independence between the longitudinal outcomes and time-to-event outcomes is assumed) or is not present in the data generation process. In Appendix E of the Supplementary material available at *Biostatistics* online, we explore the impact of patient-level heterogeneity on the performance of the joint model in more depth. Specifically, we generate data assuming varying different degrees of patient-level heterogeneity, including a case with no heterogeneity which is mainly included for pedagogical purposes. From the results of that Appendix of the Supplementary material available at *Biostatistics* online, one can see that the estimates from the joint model can be biased when there is no patient-level heterogeneity *or* when it is assumed to be the case (i.e., by fitting the simplified joint model) even when patient-level heterogeneity does exist. More specific details can be found in that Appendix of the Supplementary material available at *Biostatistics* online.

Downloaded from https://academic.oup.com/biostatistics/article/23/2/591/5957418 by Univ of North Carolina at Chapel Hill Health Sci Lib user on 22 July 2022

Downloaded from https://academic.oup.com/biostatistics/article/23/2/591/5957418 by Univ of North Carolina at Chapel Hill Health Sci Lib user on 22 July 2022

In this section, we present results based on a correctly specified joint model. We also considered the performance of the joint model in the presence of overfitting (i.e., fitting an overly complex joint model that contains the true model). In Appendix F of the Supplementary material available at *Biostatistics* online, we present simulations that illustrate power based on the joint model is quite robust to overfitting the baseline hazard, trajectory function, or both. As shown in the Appendix of the Supplementary material available at *Biostatistics* online, power does decline as the degree of overfitting increases. However, the decrease relative to the difference in power compared to the proportional hazards model is modest, suggesting that practitioners can be comfortable fitting a relatively complex joint model (i.e., one with 4 or more components in the trajectory) without concern (provided sufficient longitudinal outcomes are measured to support estimation of the trajectory).

Lastly, we note that modeling the baseline hazard using a piecewise constant function (Ibrahim *and others*, 2001) is well-established and provides a robust approximation to the underlying true baseline hazard. Nonetheless, a modification of our approach that more flexibly models the baseline hazard as a piecewise linear function (as we have done for the trajectory function) is straightforward and does not affect mathematical results presented earlier.

3.2. *Performance of the joint model under trajectory misspecification*

In this section, we present results from an investigation into the impact of trajectory function misspecification on the performance of the joint model. For this investigation, we compared power based on fitting a joint model using the correct trajectory function to power based on one that had a misspecified trajectory function. Figures 4 and 5 show the estimated Bayesian power curves based on fitting both models. The corresponding correct and average fitted trajectories for the two treatment groups are shown alongside the graphs presenting the operating characteristics. We considered two different scenarios for the trajectory functions: one scenario where the trajectory for the treated group increases initially but levels off over time and another scenario where the treated group trajectory increases initially and subsequently decreases. For simplicity, both scenarios incorporate a flat trajectory for the control group. The misspecified model approximated the correct six-component piecewise linear trajectory with a model that only included three-components. Figures 4 and 5 present the estimated power curves (column 1) under the two scenarios. The power estimates based on the misspecified models are nearly identical to those based on the true models for both cases. Figures 3 and 4 in Appendix G of the Supplementary material available at *Biostatistics* online present the estimated Bayesian type I error rate curves for scenarios where both treated and control group trajectories match the treated group trajectories in Figures 4 and 5, respectively. The nearly identical curves suggest that the misspecified model provides a well-controlled Bayesian type I error rate in the presence of model misspecification.

Recalling the discussion from Section 3.1 regarding the robustness of the joint model to overfitting, we offer practitioners the following advice. First, for design simulations, the use of a parsimonious trajectory function in order to compute power is reasonable. Using a parsimonious (but still realistic) trajectory function can greatly decrease the computational burden of large-scale simulation studies. Second, the analysis model used for real data analysis can make use of a more flexible trajectory function without great concern for the degradation of the design's operating characteristics.

3.3. *Performance of the joint model under random effect misspecification*

In Section 3.1, we considered a joint model that (correctly) included a random intercept to account for individual heterogeneity. In practice, the random effects structure needed to optimally model patient-level heterogeneity may be more complex (e.g., random slopes may be needed). In this section, we evaluate the performance of the proposed joint modeling framework in cases where only a random intercept is assumed, where random effects are omitted, and where a more complex random effects structure is

Downloaded from https://academic.oup.com/biostatistics/article/23/2/591/5957418 by Univ of North Carolina at Chapel Hill Health Sci Lib user on 22 July 2022

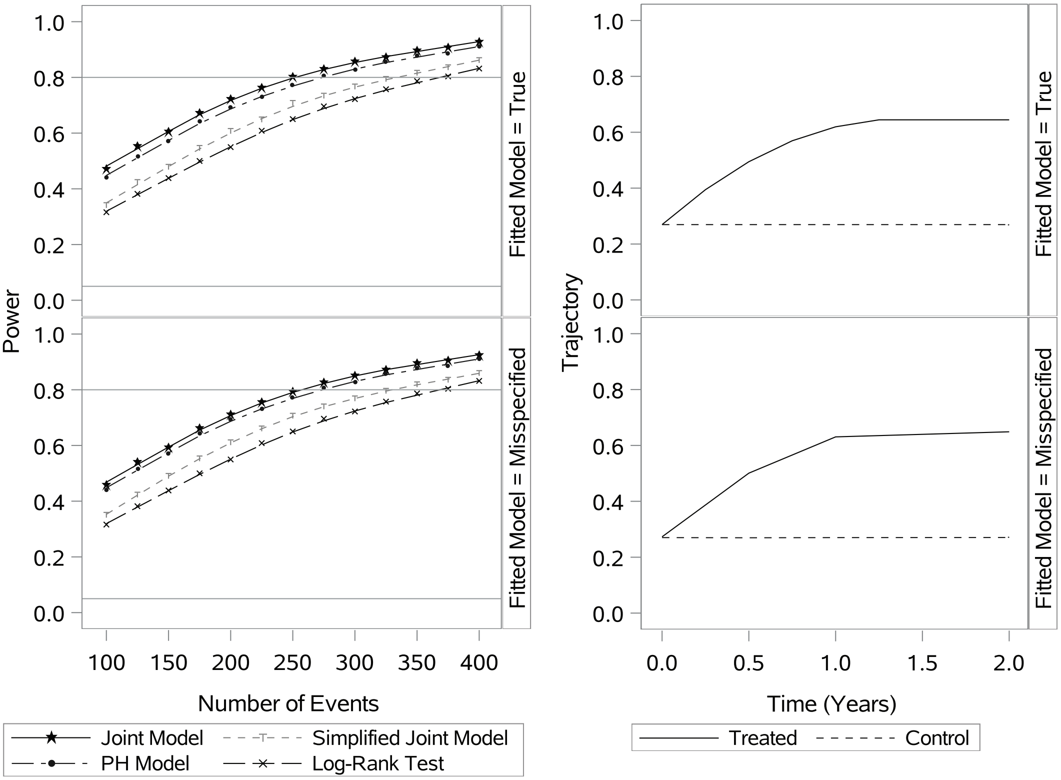

Fig. 4. Estimated power curves based on fitting the true and misspecified models. In this scenario, the trajectory function for the treated group increases initially but levels off over time.

correctly modeled. For this purpose, we generated the data assuming both a random intercept and random trajectory slope. Models with relatively simple random effect structures (e.g., only a random intercept or none) are misspecified. Figure 6 presents the estimated Bayesian type I error rate and power curves for scenarios where the random slope (common to each trajectory component) has $\omega = 0.25$ and $\omega = 0.50$ times the standard deviation of the random intercept. For simplicity, we assumed independence between the two random effects. The estimated type I error rate curves suggest that the misspecified models provide a well-controlled Bayesian type I error rate in the presence of random effects misspecification. The power estimates based on the misspecified model with only a random intercept are nearly identical to those based on the true models for both choices for ω whereas the model that entirely omits the random effects performs worse. In particular, the model that omits the random effects entirely has lower power compared to the proportional hazards model as previously observed. This underscores the point previously made that the joint model is robust to misspecification of the random effects structure provided one does not assume there is no patient-level heterogeneity.

4. Discussion

For the joint model used in the example application, we assumed that the longitudinal process maintains the same trajectory after the collection of the longitudinal measurements ceases. Thus, the trajectory is

Downloaded from https://academic.oup.com/biostatistics/article/23/2/591/5957418 by Univ of North Carolina at Chapel Hill Health Sci Lib user on 22 July 2022

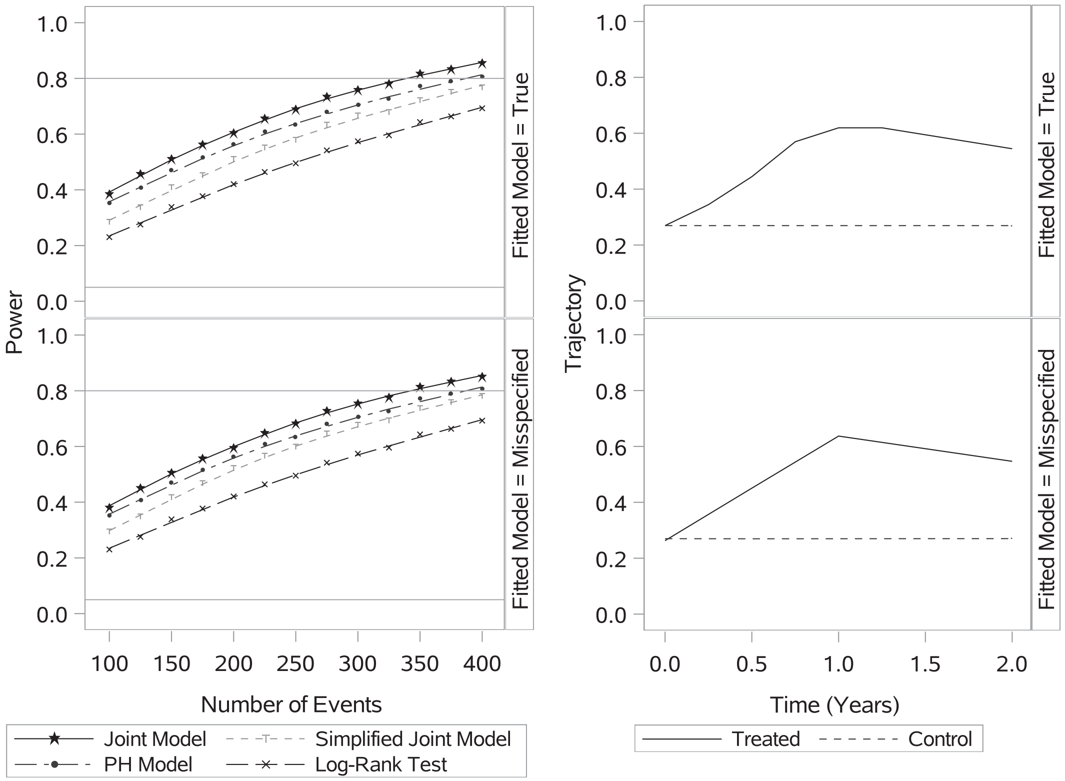

Fig. 5. Estimated power curves based on fitting the true and misspecified models. In this scenario, the trajectory function for the treated group increases initially and subsequently decreases.

effectively extrapolated from the point of the last measurement of the longitudinal outcome until the observation period ends. This may be problematic if the period of elapsed time is substantial. If data suggest a change in the trajectory may occur near the time when assessment of longitudinal outcome ends, it may be necessary to modify the hazard model to prevent erroneous extrapolation of the trajectory. *Zhang and others* (2016) provide some discussion on techniques for how the hazard function may be modified to address instances where the hazard ceases to follow the trajectory estimated based on the period of time over which the longitudinal outcome is measured. In an ideal setting, researchers will collect longitudinal outcomes over the period $[0, t_0]$, even if the frequency of collection lessens over time for logistics reasons. This will avoid having to extrapolate the trajectory altogether. Nonetheless, a key reason for modeling the time trajectory as a piecewise linear function instead of say, a linear or quadratic polynomial, is to allow the behavior of the trajectory near the period of time over which it must be extrapolated to be more strongly influenced by the recent longitudinal outcome measures.

When computing the average of the time-varying hazard ratio, we considered a function $\phi(t_0)$ using hazards for comparable treated and control patients having the same random effect values. An alternative formulation can be derived by averaging over the random effects. Note that $\phi\,(t_0)$ can be written more generally as $\phi\,(t_0, \boldsymbol{\xi}, \boldsymbol{\theta})$ to more explicitly indicate its dependence on $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$. One can define $\psi\,(t_0) = 1/n \sum_{i=1}^{n} \phi_i\,(t_0, \boldsymbol{\xi}, \boldsymbol{\theta}_i)$ to represent the average time-varying hazard ratio further averaged over the patient-specific random effects. The posterior distribution for $\psi\,(t_0)$ can be readily computed by using MCMC

Downloaded from https://academic.oup.com/biostatistics/article/23/2/591/5957418 by Univ of North Carolina at Chapel Hill Health Sci Lib user on 22 July 2022

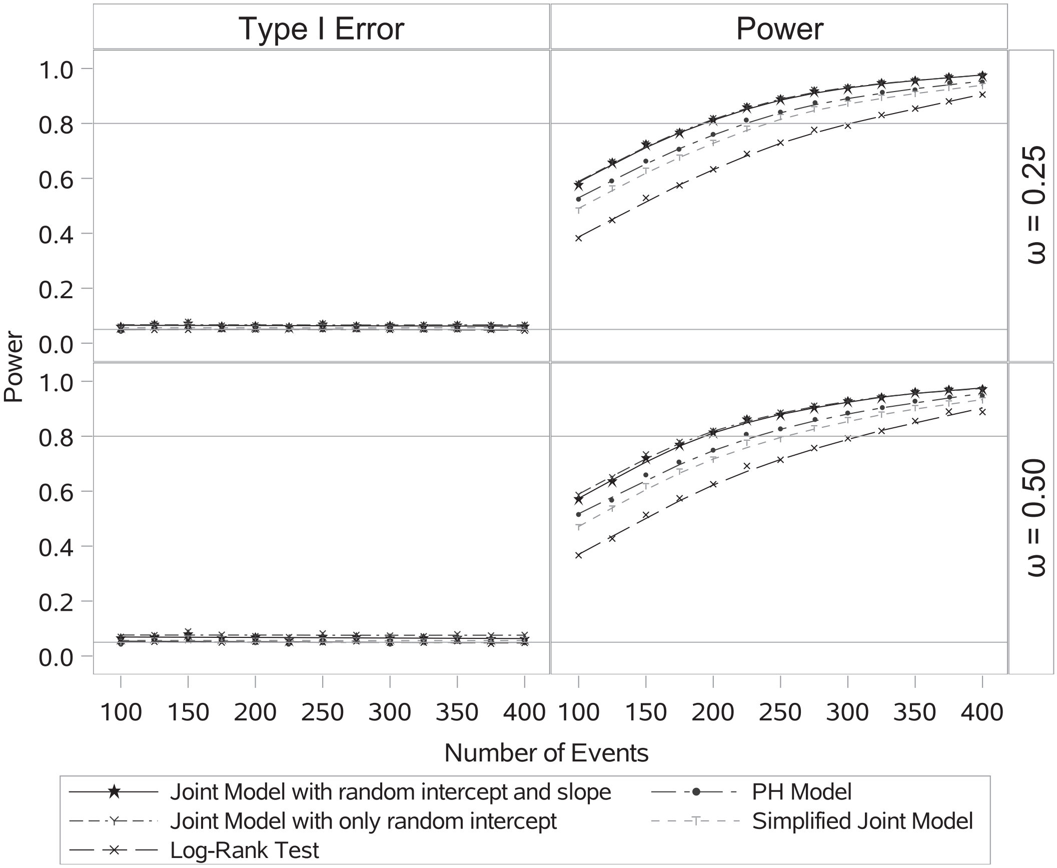

Fig. 6. Estimated type I error rate and power curves based on fitting the true and misspecified models. The joint model with a random intercept and slope is the true model. Both the joint model with only a random intercept and the simplified joint model are misspecified.

methods to obtain samples from the posterior distribution $\pi\left(\boldsymbol{\xi},\boldsymbol{\theta}\middle|\mathbf{D}\right)$ and then calculating $\psi\ (t_0)$ for each sampled value of $(\boldsymbol{\xi},\boldsymbol{\theta})$. The relative merits of performing inference using $\phi\ (t_0)$ versus $\psi\ (t_0)$ is a topic of future research for the authors.

For the sampling priors used in the example application, we used point-mass sampling prior distributions based on parameter estimates from an analysis of the IBCSG data. More generally, the Bayesian framework for power and type I error evaluation is applicable for non-degenerate sampling priors on the parameters as well. For a more extensive discussion on the use of non-degenerate sampling priors for computing Bayesian power and type I error rates, we refer the readers to the recent work of Psioda and Ibrahim (2018, 2019) and the references cited therein.

In the joint model setting, even for point-mass sampling priors, choosing the sampling priors can be challenging. The authors would give the following general advice. As was done with the IBCSG data in our example application, where such data are available, using existing data to inform choices for nuisance parameters is recommended even if the available data are not a perfect fit for the problem at hand. Where little information is available on covariate effect sizes, the associated covariates need not be considered in design simulations but can still be pre-specified for inclusion in the analysis model based on scientific rationale. If preliminary data are available regarding the treatment's effect on the longitudinal outcome (e.g., from phase I studies), these data can be used to construct a plausible trajectory model reducing the problem to having to hypothesize values for the association parameter β. In the absence of any data to inform the sampling priors, as with any sample size determination problem, one will need to explore a variety of sampling priors that cover a range of plausible "true" alternatives and ensure the chosen sample size is sufficient over that set of alternatives.

The design methodology developed in the article is for a single, continuous longitudinal outcome. It may be extended to two or more longitudinal outcomes (continuous and/or discrete), but the relationships between different longitudinal outcomes and the construction of a causal pathway need to be carefully considered. Extending the proposed framework to allow for multiple types of longitudinal outcomes is a topic for future research for the authors.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors wish to thank the editor, associate editor, and referees for helpful comments and suggestions, which have led to improvements of this article. A GitHub repository (https://github.com/JiaweiXu-code/Joint-Models) contains the SAS and R programs and other resources needed to reproduce the results presented in this article. All files (programs, datasets, logs, shell scripts, etc.) used in the article are provided at https://figshare.com.
Conflict of Interest: None declared.

Funding

This research was supported by National Institutes of Health (NIH) (#GM 70335 and P01CA142538), in part.

References

Brown, E. R. and Ibrahim, J. G. (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics* **59**, 221–228.

Chen, C.-F. (1985). On asymptotic normality of limiting density functions with Bayesian implications. *Journal of the Royal Statistical Society. Series B (Methodological)* **47**, 540–546.

Chen, L. M., Ibrahim, J. G. and Chu, H. (2011). Sample size and power determination in joint modeling of longitudinal and survival data. *Statistics in Medicine* **30**, 2295–2309.

Chen, Q., Zeng, D., Ibrahim, J. G., Chen, M.-H., Pan, Z. and Xue, X. (2015). Quantifying the average of the time-varying hazard ratio via a class of transformations. *Lifetime Data Analysis* **21**, 259–279.

Chi, Y.-Y. and Ibrahim, J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics* **62**, 432–445.

Chi, Y.-Y. and Ibrahim, J. G. (2007). Bayesian approaches to joint longitudinal and survival models accommodating both zero and nonzero cure fractions. *Statistica Sinica* **17**, 445–462.

Crowther, M. J., Abrams, K. R. and Lambert, P. C. (2013). Joint modeling of longitudinal and survival data. *The Stata Journal* **13**, 165–184.

Downloaded from https://academic.oup.com/biostatistics/article/23/2/591/5957418 by Univ of North Carolina at Chapel Hill Health Sci Lib user on 22 July 2022

De Gruttola, V. and Tu, X. M. (1994). Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics* **50**, 1003–1014.

Doob, J. L. (1935). The limiting distributions of certain statistics. *The Annals of Mathematical Statistics* **6**, 160–169.

Faucett, C. L. and Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine* **15**, 1663–1685.

Hürny, C., Bernhard, J., Gelber, R. D., Coates, A., Castiglione, M., Isley, M., Dreher, D., Peterson, H., Goldhirsch, A. and Senn, H.-J. (1992). Quality of life measures for patients receiving adjuvant therapy for breast cancer: an international trial. *European Journal of Cancer* **28**, 118–124.

IBCSG, International Breast Cancer Study Group. (1996). Duration and reintroduction of adjuvant chemotherapy for node-positive premenopausal breast cancer patients. *Journal of Clinical Oncology* **14**, 1885–1894.

Ibrahim, J. G., Chen, M.-H. and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer, New York, NY: Springer Science & Business Media.

Ibrahim, J. G., Chen, M.-H. and Sinha, D. (2004). Bayesian methods for joint modeling of longitudinal and survival data with application to cancer vaccine trials. *Statistica Sinica* **14**, 863–883.

Ibrahim, J. G., Chu, H. and Chen, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology* **28**, 2796–2801.

Prentice, Ross L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* **8**, 431–440.

Psioda, M. A. and Ibrahim, J. G. (2018). Bayesian design of a survival trial with a cured fraction using historical data. *Statistics in Medicine* **37**, 3814–3831.

Psioda, M. A. and Ibrahim, J. G. (2019). Bayesian clinical trial design using historical data that inform the treatment effect. *Biostatistics* **20**, 400–415.

Rizopoulos, D. (2010). JM: an R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software* **35**, 1–33.

Rizopoulos, D. (2016). The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. *Journal of Statistical Software* **72**, 1–46.

Wang, F. and Gelfand, A. E. (2002). A simulation based approach to sample size determination under a given model and for separating models. *Statistical Science* **17**, 193–208.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53**, 330–339.

Zhang, A. (2017). *Non-proportional Hazard in Cancer Immunotherapy*. https://sites.duke.edu/diss2017/files/2017/09/S5A_2017-09-08-DISS-NPH.pdf.

Zhang, D., Chen, M.-H., Ibrahim, J. G., Boye, M. E. and Shen, W. (2016). JMFit: a SAS macro for joint models of longitudinal and survival data. *Journal of Statistical Software* **71**, 1–24.

[*Received August 30, 2019; revised September 10, 2020; accepted for publication September 12, 2020*]

Downloaded from https://academic.oup.com/biostatistics/article/23/2/591/5957418 by Univ of North Carolina at Chapel Hill Health Sci Lib user on 22 July 2022