Computational Statistics and Data Analysis 53 (2009) 609–620

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A random walk approach for quantifying uncertainty in group sequential survival trials

Daniel L. Gillen *

Department of Statistics, University of California, Irvine, CA 92697-1250, United States

ARTICLE INFO

Article history:
Available online 20 August 2008

ABSTRACT

The development of group sequential methods has produced multiple criteria that are used to guide the decision of whether a clinical trial should be stopped early given the data observed at the time of an interim analysis. However, the potential for time-varying treatment effects should be considered when monitoring survival endpoints. In order to quantify uncertainty in future treatment effects it is necessary to consider future alternatives which might reasonably be observed conditional upon data collected up to the time of an interim analysis. A method of imputation of future alternatives using a random walk approach that incorporates a Bayesian conditional hazards model and splits the prior distribution for model parameters across regions of sampled and unsampled support is proposed. By providing this flexibility, noninformative priors can be used over regions of sampled data while providing structure to model parameters over future time intervals. The result is that inference over areas of sampled support remains consistent with commonly used frequentist statistics while a rich class of predictive distributions of treatment effect over the maximal duration of a trial are generated to assess potential treatment effects which may be plausibly observed if the trial were to continue. Selected operating characteristics of the proposed method are investigated via simulation and the approach is applied to survival data stemming from trial 002 of the Community Programs for Clinical Research on AIDS (CPCRA) study.

1. Introduction

Right-censored survival endpoints are common in many clinical trials and due to ethical considerations, it has become standard for researchers to periodically analyze accruing trial data. Such testing is typically carried out using a group sequential framework (cf. Jennison and Turnbull (2000)). The ultimate goal of group sequential methodology is to provide researchers with sufficient confidence, via probabilistic statements, in favor of a decision regarding the efficacy, futility or harm of an experimental treatment as soon as possible. However, when a longitudinal or survival endpoint is of interest one can consider whether or not a potential treatment by time interaction could result in a different decision if the trial were allowed to progress to a longer duration.

As an example of a time-varying treatment effect on survival, Abrams et al. (1994) report results from trial 002 the Community Programs for Clinical Research on AIDS (CPCRA) study, a comparative trial of Didanosine (DDI) or Zalcitabine (DDC) after treatment with Zidovudine in patients with human immunodeficiency virus (HIV) infection. The CPCRA study was a multicenter, randomized open-label trial designed to test whether DDC was non-inferior to DDI with respect to the primary endpoint of progression-free survival. Planned under a proportional hazards model the study protocol specified that

* Corresponding address: Department of Statistics, 2226 Donald Bren Hall, University of California, Irvine, CA 92697-1250, United States. Tel.: +1 949 824 9862; fax: +1 949 824 9863.
E-mail address: dgillen@uci.edu.

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.08.018

DDC would be judged non-inferior to DDI if one could rule out that the DDC/DDI hazard ratio was less than 1.25 (Fleming et al., 1995). As is commonly done in non-inferiority trials, statistical evidence for non-inferiority was based upon the upper limit of a 95% confidence interval for the DDC/DDI hazard ratio for progression events (or death). To obtain sufficient information on the time to disease progression, the study protocol specified that patient followup was to be continued until at least 243 patients experienced disease progression or death. The protocol also specified that the study's data safety monitoring committee (DSMC) would conduct a total of 4 analyses (3 interim analyses and one final analysis) scheduled after each recruitment of 25% of the protocol-specified 243 progression events. The Lan-DeMets error spending implementation of the O'Brien–Fleming guideline was employed by the DSMC to formulate repeated confidence intervals, allowing the DSMC to consider recommendations for early termination of the trial.

At the first interim analysis, the early study results strongly favored DDI. Patients treated with DDI had experienced substantially fewer AIDS/death events than those on DDC (19 vs. 39 events; DDC/DDI relative risk, 2.1; nominal 95% CI: 1.20–3.66; repeated 95% CI: 0.80–5.70) and fewer deaths (6 vs. 12) (Demets et al., 1995). The DSMC, charged with making a recommendation regarding discontinuation of the trial on the grounds of futility, considered the possibility that future treatment effects may differ from those estimated at this early analysis. Relying primarily on the guidance of the O'Brien–Fleming stopping rule the DSMC recommended continuation of the trial (Ellenberg et al., 2002). Observed treatment effects turned around shortly after the first interim analysis and the study ultimately continued to the planned maximal sample size where the hypothesis of inferiority was rejected, establishing non-inferiority of DDC relative to DDI. With a mean followup of 15 months, 152 of 237 DDC patients and 157 of 230 DDI patients had experienced AIDS/death events (DDC/DDI relative risk, .94; nominal 95% CI: 0.75–1.18), whereas death had occurred for 100 DDI patients and only 88 DDC patients (DDC/DDI relative risk, .78; nominal 95% CI: 0.58–1.04).

Although the DSMC made the recommendation to continue the CPCRA trial in the face of a negative estimated treatment effect at early analyses, their decision to do so was primarily based upon confidence intervals for the hazard ratio using only data observed at the time of an interim analysis. However, in the case of a nonproportional hazards treatment effect consideration of confidence intervals using data obtained under truncated support (at an interim analysis) is not sufficient for quantifying likely values of the treatment effect under full support. Thus one might not only consider the variability of the results obtained at an interim analysis, but could also account for uncertainty in whether or not the estimator based upon truncated support will consistently estimate the parameter which corresponds to the originally planned full support of the trial.

In the current paper, we address the quantification of uncertainty when decisions regarding trial continuation are to be made at interim analyses. When such decisions are to be made at times where the full support of survival has yet to be sampled it may be necessary for one to consider a sufficient class of plausible treatment effects, or alternatives, which might arise if the trial were to continue. Naturally, the derivation of a class of alternatives could depend upon prior beliefs derived from previous scientific experiments as well as data observed up to the time of the interim analysis. Such a process can be formalized in a Bayesian framework where a prior distribution is placed upon the treatment effect over the planned maximal duration of the trial. In this case, sampling from the posterior distribution based upon the prior and updated with data obtained at an interim analysis allows one to quantify the variability of a treatment effect over the maximal support of the study (cf. Follmann and Albert (1999) and Rosner (2005)). Previous work in this area has considered a class of parametric models to quantify uncertainty in treatment effects that might arise under various alternatives (cf. Rosner (2005)). However, restricting attention to a single parametric family does not consider all plausible scenarios that may arise, and thus an investigation of a more general construction of alternatives is appealing. In Section 2, we consider the use of a Bayesian conditional hazards model for nonparametrically estimating the survival distribution of each comparison group using observed data (up to the time of an interim analysis) and propose the use of a random walk process for generating predictive survival distributions past the time of maximal followup. The proposed methodology generates a richer class of potential alternatives than parametric methods previously proposed by focusing on the local behavior of model parameters while conditioning upon the posterior distribution of survival over intervals of observed data. In Section 3 we present a simulation study to describe the operating characteristics of the suggested procedure under various proportional and nonproportional hazards alternatives. We apply the random walk framework to the CPCRA data in Section 4 and compare the results to those which were obtained in the actual monitoring of the trial. Section 5 concludes with a further discussion of the issues that might be considered in the design and monitoring of group sequential survival trials.

2. Procedure for quantifying uncertainty in future survival differences

2.1. Estimation of the survival distribution over observed support

Many authors have considered Bayesian approaches to estimation of survival distributions (cf. Ibrahim et al. (2001)). Recently, Follmann and Albert (1999) and Rosner (2005) proposed Bayesian nonparametric models based on the Dirichlet distribution to analyze survival rates in clinical trials. In this section we describe a Bayesian conditional hazards model similar to that proposed by McKeague and Tighiouart (2000) for estimating survival. Our goal is to allow for a flexible parameterization of the survival distribution which will incorporate random changes in the hazard for each group. By modeling the survival experience of each treatment group *separately* we avoid placing strong assumptions regarding the relative treatment effect such as that made in a proportional hazards framework.

We consider modeling the hazard function for each comparison group via a piecewise constant hazards model with random split times for defining the piecewise intervals. To this end, let X_i denote the observed survival time of subject i and let $\tau_{\max} = \max\{X_i;\, i = 1, \ldots, n\}$ denote the maximum observed survival time where n represents the total number of patients accrued to the group at the time of the analysis. Consider a partition of the observed study time scale $[0, \tau_{\max}]$ into K mutually exclusive intervals, defined by split times $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_K, \tau_{K+1}\}$ where $\tau_1 \equiv 0$ and $\tau_{K+1} \equiv \tau_{\max}$. For fixed K and $\boldsymbol{\tau}$, we assume the log-hazard function, $\lambda(\cdot)$, to be piecewise constant as follows:

$$\lambda(t) = \sum_{k=1}^{K} 1_{[\tau_k < t \leq \tau_{k+1}]} \lambda_k, \tag{1}$$

where $\exp\{\lambda_k\}$ is the height of the hazard function on the kth study time interval $(\tau_k, \tau_{k+1}]$.

For fixed $\tau_{\max}$, we follow McKeague and Tighiouart (2000) and assume the jump times $\tau_2, \tau_3, \ldots$ form a time-homogeneous Poisson process on $[0, \tau_{\max}]$ with rate α. That is, the number of split times is taken to be distributed according to a Poisson distribution, with mean α, and that their positions on $[0, \tau_{\max}]$ are uniformly distributed. Conditional on the partition $\boldsymbol{\tau}$, the K log-hazard heights are assumed to be distributed according to a K-dimensional multivariate Normal distribution:

$$\boldsymbol{\lambda} \sim \text{MVN}_K(\boldsymbol{\mu},\ \sigma^2 \boldsymbol{\Sigma}), \tag{2}$$

where $\boldsymbol{\mu}$ a K-vector denoting the overall trend (assumed constant across time and parameterized by a single trend μ) in the log-hazard heights and $\sigma^2 > 0$ is the overall variability. The $K \times K$ correlation matrix $\boldsymbol{\Sigma}$ is specified to enforce structure on the log-hazard heights across time. The resulting smoothing may be viewed in the context of (one-dimensional) spatial models, for which the specification of $\boldsymbol{\Sigma}$ has received much attention (see, for example, Wakefield et al. (2000)). Following Besag and Kooperberg (1995) we specify the joint distribution of $\boldsymbol{\lambda}$ via a Gaussian conditional autoregression. For $k = 1, \ldots, K$, let $\boldsymbol{\lambda}_{-k}$ denote the vector $\boldsymbol{\lambda}$ with λ_k removed and assume $\lambda_k | \boldsymbol{\lambda}_{-k} \sim \text{Normal}(\nu_k, \sigma_k^2)$ with the mean for the height of the kth interval, conditional on the remaining intervals, given by $\nu_k = \mu + \sum_{j \neq k} W_{kj}(\lambda_j - \mu)$. Thus ν_k is the overall (marginal) trend plus a weighted sum of the remaining interval-specific deviations (from the overall trend). The influence of a given interval is assumed to be a function of its width $\Delta_k = \tau_{k+1} - \tau_k$. One possible weighting scheme is to set

$$W_{k(k-1)} = \frac{(\Delta_{k-1} + \Delta_k)c}{\Delta_{k-1} + 2\Delta_k + \Delta_{k+1}}, \tag{3}$$

$$W_{k(k+1)} = \frac{(\Delta_k + \Delta_{k+1})c}{\Delta_{k-1} + 2\Delta_k + \Delta_{k+1}}, \tag{4}$$

where Δ_0 and Δ_{K+1} are defined to equal zero, and $c \in [0, 1]$ dictates the extent of dependence (and hence smoothing). All other weights are set to zero, so that only adjacent intervals have any influence, resulting in a nearest neighbor smoothing scheme. McKeague and Tighiouart (2000) use a similar scheme, although one key difference is that we define the boundary intervals as only having a single neighbor while they identified the endpoint intervals as being neighbors. Finally the conditional variance is given as $\sigma_k^2 = \sigma^2 Q_k$ where

$$Q_k = \frac{2}{\Delta_{k-1} + 2\Delta_k + \Delta_{k+1}}. \tag{5}$$

Given (3)–(5), the full joint specification may be recovered by noting that $\boldsymbol{\Sigma} = (\boldsymbol{I} - \boldsymbol{W})^{-1}\boldsymbol{Q}$ where $\boldsymbol{I}$ is the K-dimensional identity matrix, $\boldsymbol{W}$ is a $K \times K$ matrix with elements W_{jk} and $\boldsymbol{Q}$ is a K-dimensional diagonal matrix with the kth entry given by Q_k. Propriety of the matrix $\boldsymbol{\Sigma}$ depends on its symmetry and positive-definitiveness which may be verified using conditions set out by Besag and Kooperberg (1995). Specifically, for $j, k = 1, \ldots, K$, the elements of $\boldsymbol{W}$ and $\boldsymbol{Q}$ must satisfy $W_{jk}Q_k = W_{kj}Q_j$ and $\sum_{k=1}^{K} W_{jk} \leq 1$, with at least one strict inequality for the latter. We note that given the specific choices of $\boldsymbol{W}$ and $\boldsymbol{Q}$, the conditions are satisfied for all $c \in [0, 1]$.

To directly incorporate uncertainty regarding the second-stage hyperparameters μ, σ, and c, we consider a Bayesian hierarchical approach. For the marginal mean μ, we adopt a flat prior on the real line. For the variance component σ^2, a standard approach is to parameterize the model in terms of the precision and adopt a conjugate Gamma prior. Finally, for the spatial hyperparameter, c we assume a uniform prior between 0 and 1.

For fixed K, the posterior distribution is taken to be the combination of the likelihood for the data and the two prior stages as follows:

$$p(K, \boldsymbol{\lambda}, \mu, \sigma^2) = L(\boldsymbol{\tau}, \boldsymbol{\lambda}) \times \text{MVN}_K(\boldsymbol{\lambda} |\, \mu, \sigma^2 \boldsymbol{\Sigma}) \times \text{Poisson}(N |\, \alpha) \times\ \text{Gamma}(\sigma^{-2} | a, b), \tag{6}$$

where $N = K - 1$ is the number of split times defining the hazard process, and $L(\boldsymbol{\tau}, \boldsymbol{\lambda})$ denotes the survival likelihood that is proportional to

$$\prod_{i=1}^{n} \exp\{\lambda(X_i)\}^{\delta_i} \prod_{i=1}^{n} \exp\left\{-\int_0^{X_i} e^{\lambda(s)} ds\right\} = \exp\left\{\sum_{k=1}^{K} D_k \lambda_k - \int_0^{\tau_{\max}} \left[\sum_{i=1}^{n} I(X_i \geq s) e^{\lambda(s)}\right] ds\right\},$$

where D_k denotes the number of observed failures in interval k.

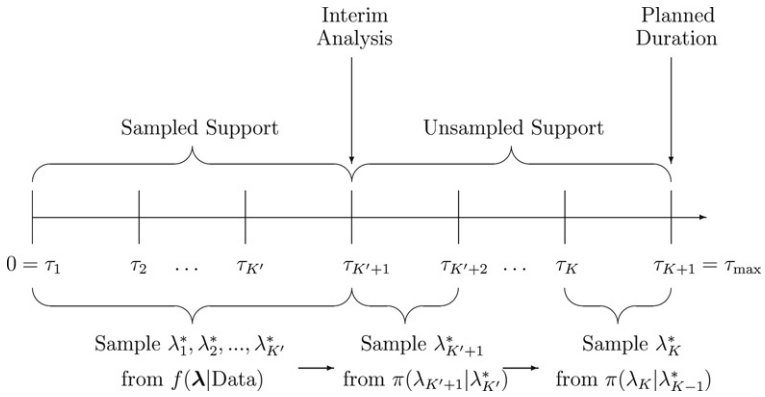


Fig. 1. Graphical depiction of the random walk process. $f(\boldsymbol{\lambda}|\text{Data})$ denotes the posterior density of $\boldsymbol{\lambda}$ using data obtained up to the time of the interim analysis and $\pi(\lambda_k|\lambda_{k-1}^*)$ denotes the assumed distribution for sampling λ_k^* conditional upon λ_{k-1}^* in the random walk.

Because the number of split times is a random variable, the dimension of the parameter space of interest is also random. Thus standard sampling algorithms are not of use in this case. To overcome this, it is possible to use a mixture of Metropolis–Hastings (Hastings, 1970) and Metropolis–Hastings–Green (Green, 1995) algorithms to sample from the posterior distribution given in (6). The details of this algorithm are provided in the Appendix.

To apply the conditional hazards model in the setting of group sequential testing, it is possible to specify the prior distribution of $\boldsymbol{\lambda}$ over the entire maximal study time, taking $\tau_{\max} = T$, where T is the originally planned duration of the trial. An initial number and placement splits, $0 = \tau_1 < \cdots < \tau_{K+1} = T$, can then be chosen over the interval $(0, T]$ and samples from the posterior distribution of $\boldsymbol{\lambda}$ can be obtained. In this case, the posterior results in the prior distribution being updated with observed data over areas of sampled support while the influence of such data quickly diffuses over areas of unsampled support, eventually leaving only the prior distribution to be sampled from. Upon sampling from the posterior distribution of $\boldsymbol{\lambda}$, summaries of the posterior distribution of any statistic that is a function of the hazard for each group such as the logrank statistic, the difference in mean survival restricted to T years, or differences in time-specific survival estimates can easily be computed relative the planned maximal duration of the trial.

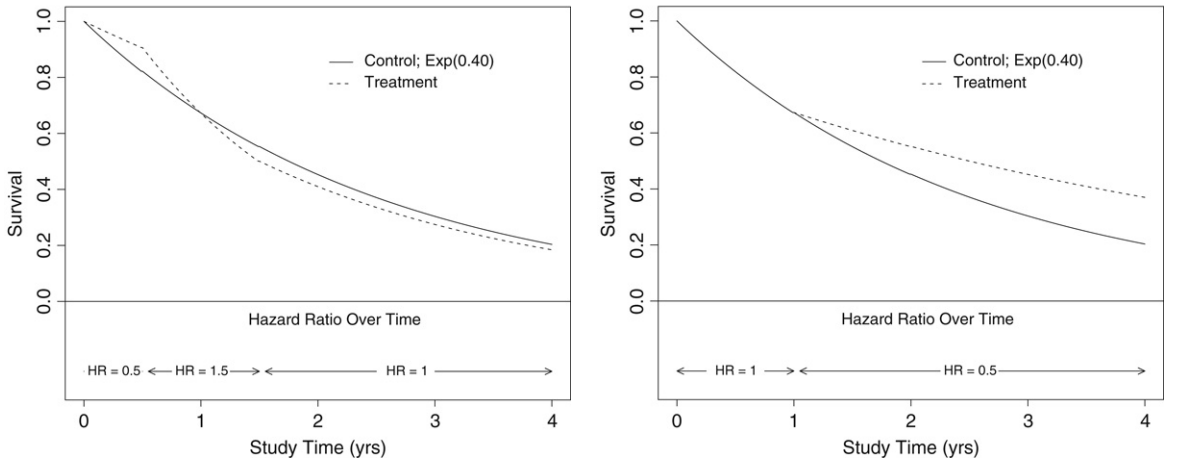
2.2. A random walk approach to obtaining predictive survival distributions

The fully Bayesian approach which places a single prior distribution over the planned maximal duration of the trial is a natural way for quantifying future uncertainty. However, this parameterization of the model does suffer some drawbacks. Specifically, our goal is to use a noninformative prior over all areas where data is available so that prior information is easily overwhelmed when data is obtained, thus yielding equivalent inference to that obtained via a frequentist analysis. However, by using a single prior, we are required to be at least somewhat informative in order to provide structure to potential alternatives occurring over areas of unsampled support. Sampling from the full posterior distribution in this case allows the prior distribution used over areas of unsampled support to influence intervals where data have been observed. This could lead to a more informative prior in these observed areas than may be desired.

While the fully Bayesian model presented in Section 2.1 could be used to obtain predictive survival distributions, the use of a single prior over the entire support distribution may be too constraining. Thus an alternative is to consider the two areas separately. Here we propose to model areas with observed data using a noninformative prior while sampling future alternatives from a prior distribution that depends upon the posterior distribution of $\boldsymbol{\lambda}$ over observed support and an assumed correlation structure. Although the incorporation of separate priors that change from one interim analysis to the next may be viewed as incoherent if one wished to report Bayesian inference over the entire support of the distribution, our goal focuses on the development of predictive distributions for guiding decisions regarding early trial termination and not for final inference which should be based solely on observed data. To this end we propose to obtain predictive distributions via a first-order random walk process as depicted in Fig. 1. This alternative approach focuses on the local behavior of the hazard for each group over a small interval of time. In this case, the location and variability of predicted hazards over intervals of unsampled support is controlled by specification of the distribution $\pi(\cdot)$.

To obtain predicted survival estimates via the random walk approach, we institute the following algorithm:

(1) Fit the nonparametric conditional hazards model discussed in Section 2.1 to data obtained up to the time of an interim analysis, denoted $\tau_{K'+1}$.
(2) Specify split times $\tau_{K'+2} < \cdots < \tau_{K+1} = \tau_{\max}$ defining time intervals past the observed support where hazards are to be drawn in order to form predicted survival curves. Here $\tau_{\max}$ denotes the maximal planned study duration.
(3) Sample $\lambda_1^*, \ldots, \lambda_{K'}^*$ from the posterior distribution of the log-hazards over intervals of sampled support.



(a) Early beneficial effect.

(b) Late beneficial effect.

Fig. 2. Depiction of the nonproportional hazards framework used in the simulation study.

(4) For $k = K' + 1, \ldots, K$ sequentially sample λ_k^* from $\pi(\lambda_k|\lambda_{k-1}^*)$, where $\pi(\lambda_k|\lambda_{k-1}^*)$ is a distribution centered at the log-hazard sampled in interval $k-1$, λ_{k-1}^*.
(5) Repeat steps (3) and (4) B times to obtain B predicted survival curves.

Choosing weakly informative priors over intervals of unsampled support results in the observed data overwhelming the prior distributions in intervals immediately following the end of observation (e.g. $(\tau_{k+1}, \tau_{k+2}]$ and $T(\tau_{k+2}, \tau_{k+3}]$ in Fig. 1). Conversely, late occurring intervals where data have yet to be gathered are quickly overwhelmed by the prior chosen, an effect similar to that of the single prior approach presented in Section 2.1 while still allowing for noninformative priors over time periods where data have been obtained.

One choice for $\pi(\cdot)$ is to assume a normal distribution such that $\lambda_k|\lambda_{k-1} \sim \mathcal{N}(\lambda_{k-1}, \sigma_k^2)$, where $\sigma_k^2 = \widehat{\text{Var}}(\lambda_{k-1}) \times (1 - \xi_{\Delta_k}^2)$ with $\widehat{\text{Var}}(\lambda_{k-1}) = \text{Var}(\lambda_{k-1}^{(1)}, \ldots, \lambda_{k-1}^{(B)})$, the empirical variance calculated over B draws from the posterior distribution of λ_{k-1}. In this context, ξ_{Δ_k} represents the assumed 'correlation' between log hazards a distance of $\Delta_k = \tau_k - \tau_{k-1}$ apart. More intuitively, $2\sigma_k$ can be roughly interpreted as a bound on the change in λ_k over a period of length Δ_k. Under this specification for $\pi(\cdot)$, the covariance structure imposed by the random walk process can be seen as a special case of the model described in Section 2.1 where $W_{k(k-1)} \to 1$ and $W_{k(k+1)} \to 0$ over time intervals where data have yet to be observed. However, by changing the balance of the weights prior assumptions about the variance of the log hazards are less restrictive under the random walk model and the spacing between jump times does not influence (dependency between) the height of jumps as it does in the model given in Section 2.1.

As in the fully Bayesian approach, after drawing B sample paths via the random walk one obtains a sample from the predictive distribution of the log hazard for each treatment group conditional upon the prior specification. This implies that summaries of the predictive distribution of any statistic that is a contrast of functionals defined by the hazard of the comparison groups can be calculated via this approach. Software implementing all of the proposed methods using the R language is available from the author.

3. Simulation study

In this section we present a simulation study to characterize the utility of the random walk approach described in Section 2.2. We consider the setting of a two-arm clinical trial ($n = 500$ per arm) designed to last a maximum of four years, with analyses performed at 1, 2, 3, and 4 years following the start of accrual. We further assume uniform accrual over 3 years with 1 additional year of followup. The performance of the random walk approach is considered under four scenarios: (1) Under the null hypothesis of no difference in survival for all $t > 0$; (2) Under a proportional hazards alternative with constant hazard ratio of 0.75; (3) Under an early beneficial treatment effect that reverses over time as depicted in Fig. 2(a) which assumes a hazard ratio of 0.5 during the first 6 months of treatment, 1.5 for the next 18 months of treatment, and 1 over the remainder of followup; and (4) a late occurring treatment effect as depicted in Fig. 2(b) which assumes no difference in the hazards over the first year of treatment and a constant hazard ratio of 0.5 over the remainder of followup. In each case, baseline survival was assumed to follow an Exponential distribution with hazard parameter 0.4.

For the simulation study, we focus on two functionals commonly used to compare survival in clinical settings, but again note that any statistic that is a functional of the underlying hazard for each comparison group can be summarized via the predictive distributions obtained using the proposed methods. Briefly, let T_{ik} and C_{ik} denote the survival and censoring time of individual i, $i = 1, \ldots, m_k$, belonging to group k, $k = 0, 1$, where T_{ik} and C_{ik} are assumed to be independent. Further,

Table 1
Simulation results for the hazard ratio over a maximal followup of 4 years assuming $N = 500$ subjects per treatment arm

Estimation procedure	Alternative							
	Null ($S_0 = S_1$)		Prop. haz.		Early beneficial effect		Late beneficial effect	
	Med. est.	Cov. pr.	Med. est.	Cov. pr.	Med. est.	Cov. pr.	Med. est.	Cov. pr.
Observed data (Cox est.)								
Analysis 1	1.019	0.960	0.784	0.940	0.723	0.539	0.931	0.877
Analysis 2	0.999	0.955	0.782	0.911	0.943	0.767	0.803	0.940
Analysis 3	1.000	0.948	0.804	0.931	1.064	0.947	0.758	0.934
Analysis 4	1.000	0.949	0.833	0.945	1.080	0.957	0.779	0.956
Bayesian (4 yr prior)								
Analysis 1	1.104	0.761	0.901	0.795	0.785	0.484	1.102	0.691
Analysis 2	1.017	0.892	0.778	0.863	1.108	0.844	0.801	0.878
Analysis 3	1.006	0.938	0.815	0.925	1.121	0.925	0.763	0.925
Analysis 4	1.007	0.943	0.838	0.933	1.088	0.954	0.783	0.948
RW ($\xi_{0.5} = 0.65$)								
Analysis 1	1.106	0.980	0.886	0.986	0.789	0.889	1.081	0.960
Analysis 2	1.016	0.970	0.773	0.937	1.096	0.952	0.789	0.979
Analysis 3	1.007	0.939	0.814	0.938	1.127	0.937	0.762	0.937
Analysis 4	1.007	0.943	0.838	0.933	1.088	0.954	0.783	0.948
RW ($\xi_{0.5} = 0.80$)								
Analysis 1	1.100	0.969	0.882	0.973	0.789	0.835	1.077	0.947
Analysis 2	1.016	0.958	0.774	0.926	1.092	0.948	0.790	0.964
Analysis 3	1.007	0.938	0.814	0.935	1.127	0.938	0.762	0.942
Analysis 4	1.007	0.943	0.838	0.933	1.088	0.954	0.783	0.948
RW ($\xi_{0.5} = 0.95$)								
Analysis 1	1.092	0.931	0.879	0.942	0.789	0.743	1.075	0.889
Analysis 2	1.015	0.938	0.776	0.910	1.084	0.920	0.791	0.938
Analysis 3	1.007	0.937	0.815	0.934	1.126	0.937	0.762	0.930
Analysis 4	1.007	0.943	0.838	0.933	1.088	0.954	0.783	0.948

The 4 year hazard ratio consistently estimated by the Cox model was considered to be the target of interest in calculating coverage probabilities. This is 1.00, 0.75, 1.08, and 0.78 under the null, proportional hazards, early beneficial effect, and late beneficial effect alternatives, respectively. Results are based on 5000 simulated datasets.

define $X_{ik} = \min(T_{ik}, C_{ik})$ to be the observed time for individual i in group k and let $\delta_{ik} = I(X_{ik} = T_{ik})$ denote the indicator that the actual survival time is observed on the ith individual in group k. Finally, let $N_k(t) = \sum_i^{m_k} I(X_{ik} \leq t, \delta_{ik} = 1)$ denote the number of events observed in group k occurring up to time t and $Y_k(t) = \sum_i^{m_k} I(X_{ik} \geq t)$ denote the number of patients at risk in group k at time t. The logrank statistic (Mantel, 1966) is defined as

$$\mathcal{U} = \left(\frac{M_1 + M_0}{M_1 M_0}\right)^{1/2} \int_0^\infty \left\{\frac{Y_1(t)Y_0(t)}{Y_1(t) + Y_0(t)}\right\} \left\{\frac{\mathrm{d}N_1(t)}{Y_1(t)} - \frac{\mathrm{d}N_0(t)}{Y_0(t)}\right\}, \tag{7}$$

where M_k denotes the number of patients initially at risk in group k, $k = 0, 1$. Noting that a consistent estimator of the hazard for group k at time t, $h_k(t)$, is given by $\hat{h}_k(t) = \mathrm{d}N_k(t)/Y_k(t)$, it is easy to see that the logrank statistic given in (7) is simply the sum, over all observed failure times, of the weighted difference in estimated hazard functions between the two groups, i.e. $\mathcal{U} = \left(\frac{M_1+M_0}{M_1M_0}\right)^{1/2} \sum_{t\in\mathcal{F}} w(t)\left[\hat{h}_1(t) - \hat{h}_0(t)\right]$, with $w(t) = \{Y_1(t)Y_0(t)/[Y_1(t) + Y_0(t)]\}$. For interpretability, we standardize results pertaining to the logrank statistic to the hazard ratio scale.

We also consider the difference in 4 year restricted mean survival, defined as $\mu_r(4) = \int_0^4 \left[\hat{S}_1(t) - \hat{S}_0(t)\right] \mathrm{d}t$, and interpretable as the expected years of life saved per individual over a total followup of 4 years, comparing group 1 to group 0.

For each treatment effect scenario, Table 1 presents simulated results computed at each interim analysis using four different methods: a standard proportional hazards analysis using the Cox model which bases inference and estimation only on data observed up to the time of the interim analysis, the Bayesian conditional hazards model with a single 4 year prior distribution placed over maximal support of the trial, and three random walk analyses assuming the correlation between log hazards a distance of 6 month apart, $\xi_{0.5}$, to be 0.65, 0.80, and 0.95, yielding a range of uncertainty in the stationarity of hazards over unsampled support. In all cases, prior distributions for the conditional hazards model were defined as in Section 2.1 and it was assumed that $\alpha = 3$ and $\sigma^{-2} \sim \mathit{Gamma}(0.5, .01)$, a distribution which places 95% of the central mass for σ^2 between 0.06 and 4.51. Under each alternative, 5000 simulations were conducted and the median estimate of the hazard ratio under each estimation procedure is reported. In addition, the empirical coverage probability corresponding to 95% confidence intervals (for the Cox proportional hazards analysis) and 95% credible intervals (for the Bayesian and random walk analyses) is reported. We note that in each case our goal is to estimate the parameter consistently

Table 2
Simulation results for the difference in 4 year restricted mean survival assuming $N = 500$ subjects per treatment arm

Estimation procedure	Alternative							
	Null ($S_0 = S_1$)		Prop. haz.		Early beneficial effect		Late beneficial effect	
	Med. est.	Cov. pr.	Med. est.	Cov. pr.	Med. est.	Cov. pr.	Med. est.	Cov. pr.
Observed data (Integrated Kaplan–Meier est.)								
Analysis 1	0.001	0.953	0.038	0.000	0.046	0.075	0.016	0.000
Analysis 2	0.004	0.950	0.113	0.424	0.017	0.699	0.109	0.086
Analysis 3	0.004	0.947	0.179	0.893	−0.028	0.918	0.215	0.832
Analysis 4	0.004	0.949	0.225	0.937	−0.061	0.959	0.288	0.957
Bayesian (4 yr prior)								
Analysis 1	0.004	0.767	0.242	0.793	0.259	0.530	0.120	0.679
Analysis 2	0.011	0.893	0.290	0.866	−0.050	0.862	0.301	0.868
Analysis 3	0.006	0.936	0.249	0.923	−0.080	0.922	0.326	0.925
Analysis 4	0.003	0.945	0.225	0.935	−0.064	0.952	0.303	0.949
RW ($\xi_{0.5} = 0.65$)								
Analysis 1	0.007	0.987	0.240	0.991	0.237	0.930	0.125	0.959
Analysis 2	0.011	0.983	0.288	0.974	−0.040	0.977	0.305	0.984
Analysis 3	0.006	0.949	0.246	0.949	−0.077	0.944	0.322	0.950
Analysis 4	0.003	0.945	0.225	0.935	−0.064	0.952	0.303	0.949
RW ($\xi_{0.5} = 0.80$)								
Analysis 1	0.006	0.974	0.244	0.983	0.246	0.896	0.129	0.951
Analysis 2	0.010	0.979	0.290	0.959	−0.035	0.971	0.308	0.976
Analysis 3	0.006	0.945	0.246	0.946	−0.076	0.938	0.323	0.949
Analysis 4	0.003	0.945	0.225	0.935	−0.064	0.952	0.303	0.949
RW ($\xi_{0.5} = 0.95$)								
Analysis 1	0.006	0.944	0.250	0.946	0.257	0.796	0.135	0.886
Analysis 2	0.011	0.949	0.292	0.920	−0.029	0.930	0.312	0.948
Analysis 3	0.006	0.941	0.246	0.936	−0.074	0.930	0.324	0.937
Analysis 4	0.003	0.945	0.225	0.935	−0.064	0.952	0.303	0.949

The true difference in 4 year restricted mean survival is 0.00, 0.34, −0.07, and 0.29 under the null, proportional hazards, early beneficial effect, and late beneficial effect alternatives, respectively. Results are based on 5000 simulated datasets.

estimated by the Cox model over the maximal followup of 4 years, hence coverage probabilities are computed accordingly. For this study, the true 4 year parameter values were 1.00, 0.75, 1.08, and 0.78 under the null, proportional hazards, early beneficial, and late beneficial alternatives, respectively. From Table 1 it is clear that all methods yield approximately unbiased parameter estimates and correct coverage probabilities under the null hypothesis and proportional hazards alternatives, with exception of the Bayesian single prior at the first two analyses where much of the information in this estimate is gleaned from the prior distribution. However, under the early and late beneficial treatment effect alternatives, basing inference only on data observed up to the time of an interim analysis yields inconsistent estimates of the 4 year hazard ratio and coverage probabilities as low as 0.539. Similar, though not as extreme, results are obtained using a single prior for the conditional hazards model. However, when using the random walk approach, coverage probabilities remain high (0.743 to 0.986, depending upon the degree of correlation imposed) at all analyses, illustrating how uncertainty in future results can be characterized by this method. As expected, one can also see that coverage probabilities for the random walk approach increase as the assumed correlation between hazards is decreased (indicating less certainty in stationarity of the hazard).

Table 2 yields analogous results for the difference in 4 year restricted mean survival. In the presented scenarios, the true difference in 4 year restricted mean survival is 0.00, 0.34, −0.07, and 0.29 under the null, proportional hazards, early beneficial, and late beneficial alternatives, respectively. Results for restricted mean survival are consistent with those presented for the hazard ratio, but further highlight problems with the standard proportional hazards analysis that only uses data observed up to the time of the interim analysis. Excluding results under the null hypothesis, 95% confidence intervals obtained from the standard proportional hazards analysis at 1 year of maximal followup obtained coverage probabilities for the true difference in 4 year restricted mean survival ranging from 0 to 0.075. In contrast, by using the random walk approach to project survival and account for uncertainty in future survival differences, credible intervals computed using data up to year 1 obtained coverage probabilities for the true difference in 4 year restricted mean survival ranging from 0.796 (for an early treatment effect and assuming the strongest correlation structure presented) to 0.991 (for a proportional hazards treatment effect and assuming the weakest correlation structure presented).

4. Application to trial 002 of the CPCRA study

Trial 002 of the Community Programs for Clinical Research on AIDS (CPCRA) study was a non-inferiority trial designed to compare Zalcitabine (DDC) to Didanosine (DDI) after treatment with Zidovudine in patients with human immunodeficiency

Table 3
Application of the random walk approach to data resulting from trial 002 of the CPCRA

Statistic	Analysis				
	8/29/1991 ($N = 287$; 58 events)	11/8/1991 ($N = 403$; 105 events)	2/13/1992 ($N = 467$; 220 events)	8/21/1992 ($N = 467$; 299 events)	9/20/1992 ($N = 467$; 309 events)
Hazard ratio					
Observed data					
Cox estimate	2.100	1.235	1.124	0.962	0.943
Unadjusted 95% CI	(1.200, 3.660)	(0.842, 1.811)	(0.863, 1.464)	(0.767, 1.207)	(0.755, 1.179)
Repeated CI (Pocock)	(1.074, 4.107)	(0.777, 1.966)	(0.817, 1.547)	(0.731, 1.263)	–
Repeated CI (OBF)	(0.825, 5.347)	(0.704, 2.169)	(0.820, 1.542)	(0.761, 1.215)	–
Random walk					
$\xi_{30} = 0.65$	1.933 (0.906, 5.228)	1.103 (0.543, 2.415)	1.099 (0.802, 1.496)	0.971 (0.757, 1.220)	0.951 (0.755, 1.168)
$\xi_{30} = 0.80$	2.004 (0.995, 5.086)	1.077 (0.544, 2.313)	1.099 (0.793, 1.484)	0.971 (0.754, 1.223)	0.951 (0.755, 1.168)
$\xi_{30} = 0.95$	2.276 (1.292, 4.468)	1.093 (0.619, 2.004)	1.098 (0.805, 1.433)	0.971 (0.758, 1.209)	0.951 (0.755, 1.168)
Difference in restricted mean survival					
Observed data					
Cox estimate	−0.090	−0.077	−0.070	−0.003	0.000
Unadjusted 95%	(−0.144, −0.036)	(−0.139, −0.015)	(−0.153, 0.013)	(−0.116, 0.110)	(−0.115, 0.115)
Repeated 95% CI (Pocock)	(−0.155, −0.025)	(−0.151, −0.003)	(−0.169, 0.029)	(−0.139, 0.133)	–
Repeated 95% CI (OBF)	(−0.181, 0.001)	(−0.167, 0.013)	(−0.168, 0.028)	(−0.119, 0.113)	–
Random walk					
$\xi_{30} = 0.65$	−0.056 (−0.517, 0.500)	−0.003 (−0.028, 0.009)	−0.005 (−0.028, 0.009)	0.000 (−0.020, 0.023)	0.003 (−0.014, 0.024)
$\xi_{30} = 0.80$	−0.332 (−0.888, 0.093)	−0.055 (−0.411, 0.297)	−0.050 (−0.243, 0.151)	0.017 (−0.115, 0.163)	0.041 (−0.088, 0.175)
$\xi_{30} = 0.95$	−0.559 (−0.891, −0.055)	−0.048 (−0.358, 0.303)	−0.046 (−0.226, 0.140)	0.018 (−0.118, 0.165)	0.041 (−0.088, 0.175)

Analysis results are presented assuming scientific interest lies in testing the hazard ratio and the difference in restricted mean survival. For each statistic, an analysis using only data observed up to the time of an interim analysis is given as well as the results from projecting survival via the random walk. For the analyses using only data observed up to the time of each interim analysis, three confidence intervals are presented for comparison to the random walk results: (1) an unadjusted confidence interval not accounting for group sequential testing, (2) a repeated confidence interval assuming a Pocock (1977) boundary was used to monitor the trial data, and (3) a repeated confidence interval accounting for the O'Brien and Fleming (1979) boundary (OBF) actually used to monitor the CPCRA study. The random walk analysis projects to a maximal followup of 2 years. Median posterior estimates with corresponding 95% credible intervals are presented assuming hazards 30 days apart have pairwise correlations of 0.65, 0.80, and 0.95.

virus (HIV) infection (see Abrams et al. (1994) for further details). Briefly, the primary endpoint of the trial was progression-free survival, with the maximum duration of the trial expected to last approximately 2 years in order to achieve adequate precision to establish non-inferiority of DDC relative to DDI. Planned under a proportional hazards model the study protocol specified that DDC would be judged non-inferior to DDI if one could rule out that the DDC/DDI hazard ratio was less than 1.25. Due to ethical concerns, the trial was monitored by an independent DSMC. The protocol also specified that the study's data safety monitoring committee (DSMC) would conduct a total of 4 analyses (3 interim analyses and one final analysis) scheduled after each recruitment of 25% of the protocol-specified 243 progression events. Ultimately all four formal interim analyses were conducted and the study rejected the null hypothesis of inferiority, establishing equivalence of DDC relative to DDI with respect to progression-free survival (the dates of the actual analyses and the total number of accrued patients and observed events (progression or death) are presented in Table 3). The published report for the study also included a final analysis performed one month after the last DSMC meeting (22 months after the start of enrollment) that incorporated all overrunning data available at that time. While the analysis dates listed in Table 3 reflect the actual dates that analyses were performed by the DSMC, due to a lag in data reporting the number of events reported here is greater than those that were available for each analysis during the actual monitoring of the study. As noted in the introduction, results stemming from the initial interim analyses suggested early inferiority of DDC when compared to DDI. However, these analyses did not account for uncertainty in the primary endpoint of disease-free survival over 2 years. To address this we consider application of the random walk for quantifying such uncertainty and compare these results to those obtained by the standard proportional hazards analysis that was actually conducted.

To illustrate how the methods proposed here could have been used as an additional tool to guide the DSMC's decision regarding continuation of the trial we present the results of the random walk under three correlation structures. In practice one could elucidate the proposed correlation structures by utilizing the scientific expertise of the DSMC members (potentially combined with prior information) to hypothesize bounds for how much the risk of the primary event could reasonably change over a relatively short period of time (say 30 days). For illustration, we consider three scenarios indicating weak, modest, and strong belief in the stationarity of the hazard over time by assuming the within group correlation between log hazards over 30 days to be 0.65 (weak), 0.80 (modest), and 0.95 (strong). Noting from Section 2.2 that an upper bound on the relative change in the hazard over 30 days is given by $\exp\left\{\widehat{\mathrm{Var}}(\lambda) \times (1 - \xi_{30}^2)\right\}$ and that the empirical variance of the

estimated log hazard over 30 days was calculated to be approximately 0.058 in the CPRC trial, these correlations translate to bounds on the relative change in the hazard of 3.4%, 2.1% and 0.6% over a 30 day time period. Non-informative priors were used over periods of observed support. Thus while the computational methods are complex, researchers need only consider plausible local changes in the risk of an event to characterize uncertainty in potential long-term effects.

Table 3 presents results from the CPCRA data for both the hazard ratio and the difference in restricted mean survival. For comparison with the random walk approach we include repeated confidence intervals adjusted for monitoring under two group sequential designs: a Pocock (1977) boundary and an O'Brien and Fleming (1979) boundary. We note that the O'Brien–Fleming stopping rule was actually used in the trial but have included the Pocock adjusted confidence intervals for illustration purposes. At the first analysis the standard proportional hazards and random walk analyses all yield point estimates for the hazard ratio that indicate inferiority of DDC relative to DDI, though credible intervals corresponding to the random walk under weak and modest beliefs in stationarity are very wide, indicating high uncertainty in what the hazard ratio at 2 years of followup may be. From a practical standpoint this informs investigators that if decisions are truly to be made about 2 year survival effects, little information has been gathered. Interestingly, we note that the Pocock adjusted confidence interval still excludes a hazard ratio of 1 (indicating harm at the first analysis) illustrating the anti-conservative nature of this stopping boundary, while the O'Brien–Fleming adjusted confidence interval is even wider than the random walk credible interval assuming little stationarity in the group-specific hazard over time. By examining the random walk results we can gauge the relative conservativeness of the two monitoring plans by relating them to the resulting inference that would be obtained under plausible local changes in the risk of an event.

Point estimates resulting from the random walk analysis at the second interim analysis are consistently lower than the estimate obtained from the analysis based solely on the observed data at that time. This stems from the random walk's emphasis on late occurring trends in the data that indicate a sharper declining survival curve in the DDI group at the time of the analysis. However, despite this change in the point estimates, credible intervals for the 2 year comparison still remain wide. By the fourth and fifth analyses results stemming from the random walk begin to coincide with the standard proportional hazards analysis. This is because nearly all of the planned 2 years of followup had been sampled by these times (21 and 22 months, respectively).

5. Discussion

The development of group sequential methods has produced multiple criteria that are used to guide the decision of whether a trial should be stopped early given the data observed. Examples of such criteria include estimates of treatment effect at the time of analysis and measures of stochastic curtailment such as conditional power. However, as currently implemented, these criteria typically assume time-invariant treatment effects and are defined for settings where the average effect of treatment up to the interim analysis is the same as that which would be observed if the trial continued on to maximum duration.

Even though researchers typically design studies with a maximal followup in mind, e.g. a 5 year cure-rate or years of life saved over 10 years in childhood cancer, it is still necessary from an ethical perspective to periodically test survival differences as data are being accrued. In clinical trials where one might reasonably hypothesize a time-varying treatment effect, several statistical issues could be addressed at both the trial design and monitoring stages. These include the choice of test statistic that will be used for comparing treatment groups, the rate at which statistical information of the test statistic grows in relation to the observed numbers of events, and quantification of uncertainty when trial continuation decisions are to be made at interim analyses. A common choice of test statistic for monitoring survival data under nonproportional hazards is a weighted version of the logrank statistic. Gillen and Emerson discuss the transitivity (Gillen and Emerson, 2007), information growth (Gillen and Emerson, 2005a), and inference (Gillen and Emerson, 2005b) corresponding to the $G^{\rho,\gamma}$ family of weighted logrank statistics (Fleming and Harrington, 1991) under group sequential testing.

A direct result of interim testing is the truncation of followup, limiting our ability to estimate treatment effects defined over a fixed maximal support. Although the use of weighted logrank statistics can, in some cases, increase statistical power under nonproportional hazards alternatives, when used in a group sequential setting these statistics do not address the potential for variation in future treatment effects. Therefore at the time of an interim analysis it is desirable to draw upon observed data and prior information for guiding the decision of early trial termination. By combining these two sources of information decisions can be based on predictive distributions from a probability model, however the construction of such predictive distributions should allow for plausible alternatives to be represented given the data observed while incorporating scientific knowledge about the degree to which treatment effects might change if the study were to continue. Ultimately final inference regarding trial results should only be based upon observed data, however the use of such predictive distributions can play a critical role in guiding a study sponsor's decision for early trial termination and may also be useful for the implementation of adaptive design methods that condition on observed data (cf. Chow and Chang (2007) and Müller and Schäfer (2004)). Indeed, in the context of the CPCRA trial discussed here, the final number of events analyzed was far greater than the originally planned 243 events stated in the study protocol. While this increase appears to have been the result of lagged data being included in the analyses presented, the random walk approach could have also been used to guide the decision to extend the trial to a greater sample size. In this case, since the number of events for the study would have been influenced by the treatment effect observed at previous interim analyses then appropriate adaptive methods would need to be employed in order to control the type I error rate of the study.

The methodology proposed here quantifies uncertainty in future survival differences by emphasizing late occurring trends in the observed data and focusing on local behavior of the hazard over small intervals of time. We have demonstrated the utility of the random walk approach using data from trial 002 of the CPCRA. However, numerous examples of time-varying effects in clinical trials exist. Recently, Peters et al. (2005) reported the results of a trial comparing survival following surgery among breast cancer patients randomized to high- or low-dose chemotherapy. Because high-dose chemotherapy is known to be associated with mortality shortly following surgery, the investigators hypothesized that any survival benefit gained in the high-dose arm would not manifest until late in the trial. Study results later verified this, showing a slight increase in the risk of mortality soon after randomization for subjects randomized to the high-dose group, but proving the high-dose regime superior with respect to 5 year survival. In another high-profile trial, the Women's Health Initiative observed that the hazard ratio associated with estrogen+medroxyprogesterone substantially decreased with time since initiation of therapy when considering coronary artery disease and venous thrombotic events (Prentice et al., 2005). In both of these cases, the random walk approach described here could provide study investigators with a useful tool for monitoring trial results and making decisions on whether or not the trial should be continued.

Although not addressed here, similar methods could also be used in the design, evaluation, and monitoring of longitudinal studies, since the potential for time-varying treatment effects in these settings forces one to consider future alternative which might arise following an interim analysis. We leave this topic open for future investigation.

Acknowledgements

The author is grateful to the anonymous referees for the comments and suggestions which have considerably clarified and improved this paper.

Appendix

Here we describe the Markov Chain Monte Carlo scheme for sampling from the posterior distribution of $\boldsymbol{\lambda}$ in the conditional hazards model. Due to the changing dimension of the parameter space caused by allowing the number and timing of splits in the piecewise constant hazards model to be random, samples from the posterior distribution of $\boldsymbol{\lambda}$ are obtained via mixture of Metropolis–Hastings and Metropolis–Hastings–Green steps. As we model the survival distribution for each group separately we describe the algorithm for only one of the groups. In addition, we present the algorithm with random hyperparameters noting that it is trial to hold these parameters fixed if so desired. Samples from the posterior distribution of the second-stage parameters are obtained via a Metropolis-within-Gibbs sampler.

There are four potential types of moves through the Markov Chain: (H) sampling of a new log hazard; (S) sampling second-stage hyperparameters; (B) the birth of a new split time; and (D) the death of an existing split time. According to Green (1995) we assign the probability of selecting one of the above moves as follows. Suppose $N \sim \text{Poisson}(\alpha)$ and let $K+1$ denote the current number of split times in the Markov Chain. Define

$$p_B = \rho \ \min\left\{1, \frac{P(N=K+2)}{P(N=K+1)}\right\} = \rho \min\left\{1, \frac{\alpha}{K+2}\right\},$$
$$p_D = \rho \ \min\left\{1, \frac{P(N=K)}{P(N=K+1)}\right\} = \rho \min\left\{1, \frac{K+1}{\alpha}\right\},$$

where ρ is chosen such that $p_B + p_D < \phi$. Here the parameter ϕ is a tuning parameter which dictates the proportion of time a move of type B or D is executed. Green (1995) takes $\phi = 0.90$ and this value was also used in our presented examples. The probability each of the two remaining moves is then specified by $p_H = p_S = [1 - (p_B + p_D)]/2$.

Move of type H: Sample k uniformly from $1, \ldots, K$. Next, sample $V \sim \text{Unif}(-\delta, \delta)$ where δ denotes a specified sampling parameter. λ_k in then updated as $\lambda_k^* = \lambda_k + V$. Define $S_k(\boldsymbol{\tau}) = \sum_{i=1}^n \max\{0, \min\{(X_i - \tau_k), \tau_{k+1} - \tau_k\}\}$. Then likelihood ratio for λ_k^* vs. λ_k is computed as $\text{lr} = \exp\left\{(\lambda_k^* - \lambda_k)D_k + (e^{\lambda_k} - e^{\lambda_k^*})S_k(\boldsymbol{\tau})\right\}$, where D_k denotes the number of deaths occurring in interval k, and the prior ratio for λ_k^* vs. λ_k is computed as $\text{pr} = \frac{\text{MVN}_K(\boldsymbol{\lambda}^*|\mu, \sigma^2\boldsymbol{\Sigma})}{\text{MVN}_K(\boldsymbol{\lambda}|\mu, \sigma^2\boldsymbol{\Sigma})}$. By symmetry of the proposal distribution the proposal ratio is 1, thus we accept λ_k^* with probability $\min\{1, \text{lr} \times \text{pr}\}$.

Move of type B : Sample $\tau^* \sim$ Uniform$(0, \tau_{K+1})$ and determine $k \in \{2, \ldots, K+1\}$ such that $\tau_{k-1} < \tau^* < \tau_k$. If $\tau^* = \tau_j$ for some j then resample τ^*. Based upon τ^* we relabel the split times as $(\tau_{k-1}, \tau_k) \to (\tau_{k-1}, \tau^*, \tau_k) \to (\tau_{k-1}^*, \tau_k^*, \tau_{k+1}^*)$, splitting interval $k-1$ into two adjacent intervals. The new split time similarly induces two new log-hazard values, λ_{k-1}^*, and λ_k^*. To obtain the new values, draw $U \sim \text{Unif}(0, 1)$. Similar to Green (1995) we compute λ_{k-1}^*, and λ_k^* via a convex perturbation of λ_{k-1} given by

$$\lambda_{k-1}^* = \lambda_{k-1} - \frac{\Delta_k^*}{\Delta_{k-1}} \log\left(\frac{1-U}{U}\right) \quad \text{and} \quad \lambda_k^* = \lambda_{k-1} + \frac{\Delta_{k-1}^*}{\Delta_{k-1}} \log\left(\frac{1-U}{U}\right),$$

where $\Delta_{k-1} = \tau_k - \tau_{k-1}$, $\Delta_{k-1}^* = \tau_k^* - \tau_{k-1}^*$, and $\Delta_k^* = \tau_{k+1}^* - \tau_k^*$. Thus the proposed parameter value resulting from the addition of τ^* is $\boldsymbol{\lambda}^* = (\lambda_1, \ldots, \lambda_{k-2}, \lambda_{k-1}^*, \lambda_k^*, \lambda_k, \ldots, \lambda_m)$, a vector of length $m+1$. The covariance matrix $\sigma^2\boldsymbol{\Sigma}$ is similarly

updated based upon the new split times to obtain $\sigma^2\boldsymbol{\Sigma}^*$. The resulting likelihood ratio is then given by

$$\text{lr} = \exp\left\{(\lambda^*_{k-1} - \lambda_{k-1})D^*_{k-1} + (\lambda^*_k - \lambda_{k-1})D^*_k - \left[\exp\{\lambda^*_{k-1}\}S_{k-1}(\boldsymbol{\tau}^*) + \exp\{\lambda^*_k\}S_k(\boldsymbol{\tau}^*)\right] + \exp\{\lambda_{k-1}\}S_{k-1}(\boldsymbol{\tau})\right\}$$

where D^*_{k-1} and D^*_k denote the number of deaths occurring in interval $(\tau^*_{k-1}, \tau^*_k]$ and $(\tau^*_k, \tau^*_{k+1}]$, respectively, and $D_{k-1} = D^*_{k-1} + D^*_k$. The prior ratio is given by $\text{pr} = \frac{\text{Poisson}(K+2|\alpha)}{\text{Poisson}(K+1|\alpha)} \frac{\text{MVN}_{K+1}(\boldsymbol{\lambda}^*|\mu,\sigma^2\boldsymbol{\Sigma}^*)}{\text{MVN}_K(\boldsymbol{\lambda}|\mu,\sigma^2\boldsymbol{\Sigma})}$, and the corresponding proposal ratio is equal to $\text{prop} = \frac{p_D^{(K+2)}\tau_{K+1}}{p_B^{(K+1)}(K+2)} = \frac{\rho\min\{1,\frac{K+2}{\alpha}\}}{\rho\min\{1,\frac{\alpha}{K+2}\}} \times \frac{\tau_{K+1}}{K+2} = \frac{\tau_{K+1}}{\alpha}$.

In order to take into account the changing size of the parameter space it is necessary to account for the jacobian of the transformation from λ_{k-1} to $(\lambda^*_{k-1}, \lambda^*_k)$. In this case, the jacobian of the transformation via U is given by

$$\text{J} = \begin{vmatrix} \frac{\text{d}\lambda^*_{k-1}}{\text{d}\lambda_{k-1}} & \frac{\text{d}\lambda^*_{k-1}}{\text{d}U} \\ \frac{\text{d}\lambda^*_k}{\text{d}\lambda_{k-1}} & \frac{\text{d}\lambda^*_k}{\text{d}U} \end{vmatrix} = \begin{vmatrix} 1 & \frac{\Delta^*_k}{\Delta_{k-1}}\frac{1}{U(1-U)} \\ 1 & -\frac{\Delta^*_{k-1}}{\Delta_{k-1}}\frac{1}{U(1-U)} \end{vmatrix} = \frac{1}{U(1-U)},$$

since $\Delta^*_{k-1} + \Delta^*_k = \Delta_{k-1}$.

Based upon the above calculations, we accept the new split time with probability $\min\{1, \text{lr} \times \text{pr} \times \text{prop} \times \text{J}\}$.

Move of type D: Sample k uniformly from $\{2, \ldots, K\}$ and relabel the split times as $(\tau_{k-1}, \tau_k, \tau_{k+1}) \rightarrow (\tau_{k-1}, \tau_{k+1}) \rightarrow (\tau^*_{k-1}, \tau^*_k)$, i.e. the intervals $(\tau_{k-1}, \tau_k]$ and $(\tau_k, \tau_{k+1}]$ are combined to form the single interval $(\tau^*_{k-1}, \tau^*_k]$. The induced value for λ^*_{k-1} in the combined interval is then calculated as the weighted average of λ_{k-1} and λ_k yielding $\lambda^*_{k-1} = \frac{\Delta_{k-1}\lambda_{k-1} + \Delta_k\lambda_k}{\Delta^*_{k-1}}$ where $\Delta_{k-1} = \tau_k - \tau_{k-1}$, $\Delta_k = \tau_{k+1} - \tau_k$, and $\Delta^*_{k-1} = \Delta_{k-1} + \Delta_k$. Based upon the deletion of interval k, the resulting likelihood ratio is given by

$$\text{lr} = \exp\left\{(\lambda^*_{k-1} - \lambda_{k-1})D_{k-1} + (\lambda^*_{k-1} - \lambda_k)D_k - \exp\{\lambda^*_{k-1}\}S_{k-1}(\boldsymbol{\tau}^*) + [\exp\{\lambda_{k-1}\}S_{k-1}(\boldsymbol{\tau}) + \exp\{\lambda_k\}S_k(\boldsymbol{\tau})]\right\},$$

with prior ratio $\text{pr} = \frac{\text{Poisson}(K|\alpha)}{\text{Poisson}(K+1|\alpha)} \frac{\text{MVN}_{K-1}(\boldsymbol{\lambda}^*|\mu,\sigma^2\boldsymbol{\Sigma}^*)}{\text{MVN}_K(\boldsymbol{\lambda}|\mu,\sigma^2\boldsymbol{\Sigma})}$ and proposal ratio $\text{prop} = \frac{p_B^{(K)}(K+1)}{p_D^{(K+1)}\tau_{K+1}} = \frac{\rho\min\{1,\frac{\alpha}{K+1}\}}{\rho\min\{1,\frac{K+1}{\alpha}\}} \times \frac{K+1}{\tau_{K+1}} = \frac{\alpha}{\tau_{K+1}}$.

Finally, the jacobian of the transformation from $(\lambda_{k-1}, \lambda_k)$ to λ^*_{k-1} is given by

$$\text{J} = \begin{vmatrix} \frac{\text{d}\lambda^*_{k-1}}{\text{d}\lambda_{k-1}} & \frac{\text{d}\lambda^*_{k-1}}{\text{d}\lambda_k} \\ \frac{\text{d}U^*}{\text{d}\lambda_{k-1}} & \frac{\text{d}U^*}{\text{d}\lambda_k} \end{vmatrix} \begin{vmatrix} \frac{\Delta_{k-1}}{\Delta^*_{k-1}} & \frac{\Delta_k}{\Delta^*_{k-1}} \\ -U^*(1-U^*) & U^*(1-U^*) \end{vmatrix} = U^*(1-U^*),$$

where U^* is determined by the relationship $\lambda_k = \lambda_{k-1} + \log[(1-U^*)/U^*]$, implying that a death move is simply the reverse transformation of a birth move. As with a move of type B, we accept the new split time with probability $\min\{1, \text{lr} \times \text{pr} \times \text{prop} \times \text{J}\}$.

Move of type S:

(1) Sampling from the posterior distribution of μ: It can be shown that the conditional posterior distribution of μ is Normal$(b/a, \sigma^2/a)$ with $a = \mathbf{1}^{\text{T}}\boldsymbol{\Sigma}^{-1}\mathbf{1}$ and $b = \mathbf{1}^{\text{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}$, where $\mathbf{1}$ denotes a K-vector of 1's. Thus it is possible to directly obtain samples from the conditional posterior distribution of μ.

(2) Sampling from the posterior distribution of σ^2: Again it is possible to obtain direct samples from the conditional posterior distribution of σ^2. If we assume that the hyperprior of σ^2 is InvGamma(a, b) then the conditional posterior of σ^2 is also InvGamma(a^*, b^*) distribution with $a^* = a + \frac{K}{2}$ and $b^* = b + \frac{Q}{2}$ where $Q = (\boldsymbol{\mu} - \boldsymbol{\lambda})^{\text{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\lambda})$.

(3) Sampling from the posterior distribution of c: The conditional distribution of c is intractable so we employ a Metropolis–Hastings step. Sample $V \sim \text{Unif}(-\delta, \delta)$ and calculate $c^* = c + V$. Then the posterior ratio of c^* vs. c is given by $\text{pr} = \frac{\text{MVN}_K(\boldsymbol{\lambda}|\mu,\sigma^2\boldsymbol{\Sigma}_{c^*})}{\text{MVN}_K(\boldsymbol{\lambda}|\mu,\sigma^2\boldsymbol{\Sigma}_c)}$, where $\sigma^2\boldsymbol{\Sigma}_c$ is the covariance matrix calculated at the value c. By symmetry of the proposal distribution, the proposal ratio is 1 so that c^* is accepted with probability $\min\{1, \text{pr}\}$.

References

Abrams, D., Goldman, A., Launer, C., Korvick, J., Neaton, J., Crane, L., Grodesky, M., Wakefield, S., Muth, K., Kornegay, S., Cohn, D., Harris, A., Luskin-Hawk, R., Markowitz, N., Sampson, J., Thompson, M., Deyton, L., 1994. A comparative trial of didanosine or zalcitabine after treatment with zidovudine in patients with human immunodeficiency virus infection. New Engl. J. Med. 330, 657–662.

Besag, J., Kooperberg, C., 1995. On conditional and intrinsic autoregressions. Biometrika 82, 733–746.

Chow, S.H., Chang, M., 2007. Adaptive Design Methods in Clinical Trials. Chapman & Hall/CRC.

Demets, D.L., Fleming, T.R., Whitley, R.J., Childress, J.F., Ellenberg, S.S., Foulkes, M., Mayer, K.H., O'Fallon, J., Pollard, R.B., Rahal, J.J., Sande, M., Straus, S., Walters, L., Whitley-Williams, P., 1995. The data and safety monitoring board and acquired immune deficiency syndrome (aids) clinical trials. Controlled Clinical Trials 16, 408–421.

Ellenberg, S.S., Fleming, T.R., DeMets, D.L., 2002. Data Monitoring Committees in Clinical Trials: A Practical Perspective. John Wiley.

Fleming, T.R., Harrington, D.P., 1991. Counting Processes and Survival Analysis. Wiley.

Fleming, T.R., Neaton, J.D., Goldman, A., DeMets, D.L., Launer, C., Korvick, J., Abrams, D., 1995. Insights from monitoring the cpcra didanosine/zalcitabine trial. Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology 10, S9–S18.
Follmann, D.A., Albert, P.S., 1999. Bayesian monitoring of event rates with censored data. Biometrics 55 (2), 603–607.
Gillen, D.L., Emerson, S.S., 2005a. Information growth in a family of weighted logrank statistics under repeated analyses. Sequential Analysis 24 (1), 1–22.
Gillen, D.L., Emerson, S.S., 2005b. A note on *P*-values under group sequential testing and nonproportional hazards. Biometrics 61 (2), 546–551.
Gillen, D.L., Emerson, S.S., 2007. Non-transitivity in a class of weighted logrank statistics under non-proportional hazards. Statistics and Probability Letters 77 (2), 123–130.
Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732.
Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109.
Ibrahim, J.G., Chen, M.-H., Sinha, D., 2001. Bayesian Survival Analysis. Springer-Verlag Inc.
Jennison, C., Turnbull, B.W., 2000. Group Sequential Methods With Applications to Clinical Trials. CRC Press.
Mantel, N., 1966. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemo. Rep. 50, 163–170.
McKeague, I.W., Tighiouart, M., 2000. Bayesian estimators for conditional hazard functions. Biometrics 56 (4), 1007–1015.
Müller, H.H., Schäfer, H., 2004. A general statistical principle for changing a design any time during the course of a trial. Statistics in Medicine 23, 2497–2508.
O'Brien, P.C., Fleming, T.R., 1979. A multiple testing procedure for clinical trials. Biometrics 35, 549–556.
Peters, W.P., Rosner, G.L., Vredenburgh, J.J., Shpall, E.J., Crump, M., Richardson, P.G., Schuster, M.W., Marks, L.B., Cirrincione, C., Norton, L., Henderson, I.C., Schilsky, R.L., Hurd, D.D., 2005. A prospective randomized comparison of high-dose chemotherapy with stem-cell support versus intermediate-dose chemotherapy after surgery and adjuvant chemotherapy in women with high- risk primary breast cancer: A report of calgb 9082, swog 9114, and ncic ma-13. Journal of Clinical Oncology 23, 2191–2200.
Pocock, S.J., 1977. Group sequential methods in the design and analysis of clinical trials. Biometrika 64, 191–200.
Prentice, R.L., Pettinger, M., Anderson, G.L., 2005. Statistical issues arising in the women's health initiative. Biometrics 61 (4), 899–911.
Rosner, G.L., 2005. Bayesian monitoring of clinical trials with failure-time endpoints. Biometrics 61 (1), 239–245.
Wakefield, J., Best, N., Waller, L., 2000. Bayesian approaches to disease mapping. In: Elliott, P., Wakefield, J., Best, N., Briggs, D. (Eds.), Spatial Epidemiology: Methods and Applications. Oxford University Press, Oxford, pp. 104–127.