# A Multiresolution Hazard Model for Multicenter Survival Studies: Application to Tamoxifen Treatment in Early Stage Breast Cancer

Peter BOUMAN, Xiao-Li MENG, James DIGNAM, and Vanja DUKIĆ

In multicenter studies, one often needs to make inference about a population survival curve based on multiple, possibly heterogeneous survival data from individual centers. We investigate a flexible Bayesian method for estimating a population survival curve based on a semiparametric multiresolution hazard model that can incorporate covariates and account for center heterogeneity. The method yields a smooth estimate of the survival curve for "multiple resolutions" or time scales of interest. The Bayesian model used has the capability to accommodate general forms of censoring and a priori smoothness assumptions. We develop a model checking and diagnostic technique based on the posterior predictive distribution and use it to identify departures from the model assumptions. The hazard estimator is used to analyze data from 110 centers that participated in a multicenter randomized clinical trial to evaluate tamoxifen in the treatment of early stage breast cancer. Of particular interest are the estimates of center heterogeneity in the baseline hazard curves and in the treatment effects, after adjustment for a few key clinical covariates. Our analysis suggests that the treatment effect estimates are rather robust, even for a collection of small trial centers, despite variations in center characteristics.

KEY WORDS: Bayesian survival analysis; Breast cancer; Clinical trials; Hazard estimation; Kaplan–Meier estimator; Meta-analysis; Multicenter study; Multiresolution models; Posterior predictive check; Tamoxifen.

## 1. INTRODUCTION

Analysts of multicenter clinical trials data are faced with a series of statistical challenges in ascertaining treatment effects while accounting for the possibly confounding influence of both measured and unmeasured patient- and clinic-specific characteristics. The effects of treatment and other patient covariates may vary significantly from center to center, and the unrecorded patient and/or center characteristics may influence trial endpoints. In addition, the baseline hazard function may display nonunimodal shape, in contrast to the common assumption of unimodality common to many parametric models. Furthermore, in the arena of Bayesian models that are designed to address some of these challenges, methods for diagnosing departures from model assumptions are in need of further development.

In this article we present an analysis of a large multicenter randomized placebo-controlled clinical trial to evaluate the effect of tamoxifen in treatment of women with early stage breast cancer (Fisher et al. 1989). The effects of tamoxifen and a few other clinically relevant patient covariates will be of interest, along with variations in treatment effect and baseline hazard from center to center. Whereas the number of patients enrolled per center varies considerably, a hierarchical Bayes approach is used to accommodate borrowing of information among large and small trial centers. For this analysis, we extend a semiparametric hazard estimator that was explored in Bouman, Dukić, and Meng (2005), which allows us to flexibly model the hazard function and incorporate a priori hazard shape and smoothness assumptions. For the purposes of model criticism and refinement, we also examine a method for detecting departures from the proportional-hazards assumption, based on the posterior predictive distribution (e.g., Gelman, Meng, and Stern 1996).

The plan of this article is as follows: Section 2 details some of the key statistical issues in multicenter survival modeling and introduces the breast cancer problem and data. Section 3 develops the multiresolution survival model, while Section 4 gives technical details of the Markov chain Monte Carlo (MCMC) model implementation and model criticism criteria based on the posterior predictive distribution. Section 5 presents the analysis of the multicenter breast cancer clinical trial, including our strategy for model checking and comparison. Finally, Section 6 concludes the article with a discussion of modeling issues raised in the analysis.

## 2. STATISTICAL ISSUES IN MULTICENTER TRIALS

Clinical trials for diseases such as early stage cancer usually involve multiple enrollment sites so as to obtain sufficient numbers of patients to detect modest but potentially clinically meaningful treatment benefits over a reasonable time period. Although the trial protocol document generally defines specific patient entry criteria and treatment procedures, there will invariably be some heterogeneity in patient mixture and treatment delivery among centers. These may be due to differences in patient demographic and clinical characteristics, as well as deviations from the protocol by the treating centers, variations in treatment compliance by patients, differences in follow-up and event monitoring, and other unknown or unmeasured factors. This evaluation of the intended treatment under a variety of conditions may be viewed as a strength of multicenter clinical trials, because summaries of treatment effect that incorporate this "natural" heterogeneity may more realistically represent the impact of the treatment in practice.

There are several reasons why one should attempt to describe these variations and consider the impact of enrollment sites

Peter Bouman is Assistant Professor, Marketing Department, Kellogg School of Management, Northwestern University, Evanston, IL 60208 (E-mail: *p-bouman@kellogg.northwestern.edu*). Xiao-Li Meng is Professor, Department of Statistics, Harvard University, Cambridge, MA 02138 (E-mail: *meng@stat.harvard.edu*). James Dignam (E-mail: *jdignam@health.bsd.uchicago.edu*) and Vanja Dukić (Email: *vdukic@health.bsd.uchicago.edu*) are Assistant Professors, Department of Health Studies, University of Chicago, Chicago, IL 60637. A major portion of this work was a part of Bouman's doctoral dissertation at Department of Statistics, The University of Chicago. All authors thank NSF and NIH for partial support. The authors especially thank Dr. Gray for sharing his software.


Journal of the American Statistical Association
December 2007, Vol. 102, No. 480, Applications and Case Studies
DOI 10.1198/016214506000000951

on estimated treatment effects. First, presence of heterogeneity can challenge the validity of the overall trial findings that are based on data that are aggregated without consideration of the center-specific effects (Localio, Berlin, Have, and Kimmel 2001). Second, strong heterogeneity may suggest that different subpopulations of treatment effects are present, which may be due to differences in patient populations or may indicate problems with protocol implementation and adherence within some centers. Third, treatment effects are quite often modest, and so from the perspective of any single center or investigator, treatment comparisons may be equivocal or reversed. In trials where treatment assignment is not blinded, this may cause inappropriate alterations in trial conduct or diminished enthusiasm for continuing enrollment and follow-up. Finally, if heterogeneity is suggested, one might then examine the extent to which the variations are attributable to patient characteristics, adherence to treatment, follow-up reporting delinquency, or other factors. These insights can help to determine how the summary treatment effect should be estimated and reported. Our approach, by pooling information from all centers, provides a flexible way to obtain more reliable estimates of center-specific effects, as well as an appropriately constructed overall treatment effect estimate in the presence of heterogeneity.

### 2.1 Multicenter Randomized Clinical Trial for Early Stage Breast Cancer

In 1982, the National Surgical Adjuvant Breast and Bowel Project (NSABP), a National Cancer Institutes sponsored multicenter cancer cooperative group, initiated Protocol B-14, a clinical trial to evaluate the efficacy of the drug tamoxifen after surgery for breast cancer. A total of 2,892 women with estrogen-receptor-positive breast tumors and axillary lymph nodes histologically negative for tumor cells were randomized after surgery to receive either a placebo or tamoxifen (1,453 and 1,439 women, respectively) between January 1982 and January 1988. Primary findings were first obtained in 1989, showing a significant reduction in breast cancer recurrence risk for patients who received tamoxifen (Fisher et al. 1989). Longer follow-up of these patients eventually revealed a survival advantage for those who received tamoxifen (Fisher et al. 1996).

The trial was conducted across 168 centers in the United States and Canada; of these, 110 trial centers contributed at least two patients to each of the placebo and tamoxifen groups. In this analysis we included only patients enrolled in these 110 centers (constituting 96% of the cohort of 2,817 protocol-eligible patients) to facilitate comparative analysis of our results with other potential analysis of the trial by conventional stratified methods. Across these centers, the number of patients per center ranged from 4 to 241, with a median of 15. Primary endpoints for the trial were overall survival, defined as time from surgery to death from any cause, and disease-free survival (DFS), defined as time to first breast cancer recurrence at any local, regional, or distant anatomic site, occurrence of a tumor in the opposite breast, occurrence of other second primary cancers, or death prior to these events (i.e., time to first event of any kind).

In this article we first perform an analysis using the DFS endpoint. However, over extended follow-up, DFS naturally exhibits nonproportionality with respect to the tamoxifen treatment effect, because patients who experience reduction in breast cancer recurrence hazard due to tamoxifen consequently remain at risk to fail later from the other events that comprise DFS (second primary cancers or deaths from noncancer causes). Thus, we also present results for an additional endpoint, breast cancer-free survival (BCFS), defined as time to breast cancer recurrence or occurrence of a new tumor in the opposite breast, treating the other event types as censored observations. Modeling the cause-specific hazard for breast cancer events only may have more clinical relevance, and with the exception of endometrial cancer, which occurs in less than 1.5% of patients but is more frequent among women taking tamoxifen, rates for nonbreast cancer events are essentially equal between the two treatment groups (Fisher et al. 1996). For BCFS and similar breast cancer specific endpoints, the tamoxifen treatment effect appears to follow the proportional-hazards assumption quite well through at least 15 years of follow-up (Fisher et al. 2004). In this analysis follow-up was administratively censored at 10 years, so that proportionality holds reasonably for both endpoints.

Among the 2,705 patients in the analysis, 733 experienced breast cancer recurrence or a tumor in the opposite breast, 254 experienced other failure events (second primary cancers, deaths prior to any other event), 89 were lost to follow-up prior to 10 years (and treated as censored at their respective loss times), and 1,629 were event-free at 10 years.

In a previous study (Bryant, Fisher, Gündüz, Costantino, and Emir 1998), tumor size, tumor progesterone receptor level, and age at enrollment were found to be prognostic for DFS. In Table 1, which summarizes key patient characteristics, established clinical categories were used for the first two covariates (tumor size and progesterone receptor level), and linear and quadratic terms were used to model effects of age at enrollment, which was standardized with sample mean $\bar{x} = 54.7$ years and sample standard deviation $s = 10.0$ years. The main objective of this work, as detailed in Section 5, will be to flexibly estimate for both DFS and BCFS the baseline hazard and the effects of treatment, tumor size, progesterone receptor level, and age, while accounting for center heterogeneity in both the baseline hazard and the treatment effect. The model variables for both endpoints are coded so that the baseline group will consist of patients enrolled in the placebo trial arm, age 54.7 years at enrollment, with tumor size $\leq$2.0 cm, and progesterone receptor level less than 10 fmol/mg.

Table 1. Characteristics of 2,705 patients from NSABP B-14

| Covariate | Placebo | Tamoxifen | Total |
|---|---|---|---|
| Tumor size | | | |
| $\leq$2 cm | 790 | 782 | 1,572 |
| 2.1–4 cm | 489 | 511 | 1,000 |
| $\geq$4.1 cm | 71 | 62 | 133 |
| Progesterone receptor level | | | |
| $<$10 fmol/mg | 313 | 295 | 608 |
| $\geq$10 fmol/mg | 1,037 | 1,060 | 2,097 |
| Age | | | |
| 25–34 | 44 | 30 | 74 |
| 35–44 | 200 | 202 | 402 |
| 45–54 | 358 | 372 | 730 |
| 55–64 | 496 | 500 | 996 |
| 65+ | 252 | 251 | 503 |

### 2.2 Multicenter Survival Analysis

An important goal of survival analysis is estimation of the survival curve $S(t)$ and its transform, the cumulative hazard $H(t) = -\log(S(t))$. Standard survival analysis (Cox and Oakes 1984) accounts for differences among patients in this hazard function through the proportional-hazards model (Cox 1972), under which patient covariate effects multiply the baseline hazard $H_{\text{base}}(t)$. In multicenter survival studies, where survival data from different centers must be combined into one inferential framework, methods that do not account for center heterogeneity via center-specific covariates and/or center-specific effects may misestimate overall uncertainty of the hazard parameters. Lagakos and Schoenfeld (1984), among others, discussed the lack of collapsibility of the hazard ratio when important covariates are omitted from the model, illustrating that not acknowledging between-center heterogeneity may be highly misleading.

Existing literature provides a number of approaches to combining survival data from multiple centers, although most have been formulated for a meta-analysis setting, which is geared toward combining one-dimensional summaries of individual studies, rather than for a multicenter setting where models for combining data or curves from centers are required. Parmar, Torri, and Stewart (1998) outlined how to estimate treatment–control log-hazard ratios and associated variances from available summary statistics or published Kaplan–Meier curves from multiple clinical trials. Their method, intended for meta-analytic applications, however, depends on assumptions of uniform right-censoring over discrete time intervals, and does not accommodate patient-level or study-level covariates. A method for estimating the baseline hazard function is given, but only for data stratified into treatment and control groups. Hunink and Wong (1994) also described estimation of baseline hazard from multiple studies, but they only account for covariate effects through stratification, without allowing for patient characteristics observed on a continuous scale. In addition, none of these methods accounts for potential heterogeneity from study to study, beyond that attributed to the observed (discretized) study covariates. See also Earle, Pham, and Wells (2000) for a comparison of the performance of five survival meta-analysis methods against an analysis of individual patient data. In the area of multicenter analysis, Glidden and Vittinghoff (2004) investigated the use of a gamma frailty model to account for center-specific effects.

To the best of our knowledge, though, the closest model to the one presented in this article was given by Gray (1994), who developed a Bayesian analysis of variation in patient survival by study center in multicenter trials, allowing for heterogeneity in both control and treatment groups by placing a bivariate normal prior on center-specific effects. Our study extends that approach by employing a resolution-invariant method that accounts for different forms of missing data beyond simple right-censoring and can incorporate virtually arbitrary prior assumptions about the hazard. In addition, we focus on developing measures for diagnostic testing of key model assumptions. Details of the multiresolution approach are summarized in the following section, while a full treatment and discussion of its theoretical properties, including a comparison of its performance to some of the common nonparametric hazard estimators, can be found in Bouman et al. (2005). The application in that article involves an AIDS dataset with a more complex censoring and truncation mechanism than the dataset in this current article.

## 3. MULTIRESOLUTION MODEL FORMULATION

In this section we describe a strategy for estimating baseline population survival, given possibly censored failure times and covariate data for patients from $K$ separate sources or studies (e.g., centers), while accounting for study-specific heterogeneity beyond that attributable to the study-specific covariates. To do so, we adopt a Bayesian proportional-hazards model that allows for general censoring of the survival times and reflects multiple sources of uncertainty in the posterior estimate of the common population survival curve.

For this analysis we will choose a fixed and ordered set of time horizons $t_j$ (the "time resolution") and seek estimates of the underlying *baseline* survival probability at the chosen $t_j$, $S_{\text{base}}(t_j)$. Following standard practice, we focus on the cumulative hazard $H_{\text{base}}(t)$ and the discrete hazard increments $d_j \equiv H_{\text{base}}(t_j) - H_{\text{base}}(t_{j-1})$. Posterior estimates of $d_j$ can then easily be transformed into survival function estimates via the identity $d_j = \int_{t_{j-1}}^{t_j} h_{\text{base}}(s)\, ds$, where the function $h_{\text{base}}(t)$ is the hazard rate at time $t$.

### 3.1 Multiresolution Prior for Baseline Hazard Increments

A standard Cox proportional-hazards model estimates covariate effects for survival, but treats the baseline hazard function as a nuisance parameter that can have a large or infinite number of dimensions. The model we describe here estimates the baseline cumulative hazard $H_{\text{base}}(t)$ at times $t_j$, along with covariate effects, where the resolution, consisting of the time points $0 < t_1 < t_2 < \cdots < t_J$, has been chosen in advance according to clinical interest. For the purposes of this model, assume that $J = 2^M$ for $M > 0$, with the number of bins $J$ chosen in proportion to $N$, the total sample size available, so that there are multiple observations per bin. The spacing of the $t_j$ is usually chosen according to the time scale that is most reflective of the analysis needs and the relevant assumptions about the underlying hazard function; in particular, the time points are not required to be evenly spaced. In the case that the resolution cannot be specified with meaningful prior input based on clinical needs and knowledge, the optimal level of resolution may be chosen using model selection criteria such as the DIC of Spiegelhalter, Best, Carlin, and van der Linde (2002); see Bouman et al. (2005) for an example of choosing an appropriate number of bins for estimating the delay in AIDS case reports to the Centers for Disease Control.

In each interval $j$, $j = 1, \ldots, J$, for all times $t$ such that $t_{j-1} < t \le t_j$, a constant hazard rate $d_j$ is assumed and thus the baseline cumulative hazard is linearly interpolated at time $t$. For times $t > t_J$, $S(t)$ is not defined under our model and the failure times after $t_J$ are right-censored at $t_J$. Hence, the number of bins $J = 2^M$ and bin widths $t_j - t_{j-1}$ should be chosen so that the resulting piecewise-constant hazards assumption is mathematically reasonable and meets the substantive goals of the analysis, while $t_J$ should be chosen so that a minimal amount of information is lost to "closure" censoring. We note that the assumption of piecewise-constant hazard has been adopted in a

number of articles in the literature such as, for example, Walker and Mallick (1997).

Once the $J$ time intervals have been chosen, the hazard increments $d_j$ should be specified in a way that makes it easy to formulate our prior beliefs about the shape and smoothness of the underlying hazard curve. Let $H \equiv H_{\text{base}}(t_J)$ be the cumulative hazard at the final time horizon $t_J$ and let $d_j, j = 1, \ldots, J$, be its $J$ increments. For the multiresolution parameterization, write $H_{0,0} \equiv H(t_{2^M}) = H(t_J)$, and for $m = 1, \ldots, M$, $p = 0, \ldots, 2^{m-1} - 1$ (where $m$ is the level of resolution and $p$ is the position in that level) decompose $H_{m-1,p}$ into the dyadic summands $H_{m,2p} + H_{m,2p+1}$. At the highest level of resolution, note that $H_{M,0} \equiv d_1$ and $H_{M,1} \equiv d_2, \ldots, H_{M,2^M-1} \equiv d_J$, the hazard increments we want to estimate. Now let $R_{m,p} \equiv H_{m,2p}/H_{m-1,p}$ and parameterize the hazard increments $d_1, \ldots, d_j$ by $H_{0,0}$ (hereafter $H$) and the "splits" $R_{1,0}, \ldots, R_{M,2^{M-1}-1}$ (hereafter denoted $R_{m,p}$). For example, when $M = 3$ and $J = 2^3 = 8$,

$$\begin{aligned} d_1 &= HR_{1,0}R_{2,0}R_{3,0}, \\ d_2 &= HR_{1,0}R_{2,0}(1 - R_{3,0}), \\ &\vdots \\ d_8 &= H(1 - R_{1,0})(1 - R_{2,1})(1 - R_{3,3}). \end{aligned}$$

A simple diagram of the two-level multiresolution prior is given in Figure 1. Split parameters higher in the hierarchy govern coarser scale details of the cumulative hazard function, while lower level parameters control finer scale differences. Motivated by the development in Nowak and Kolaczyk (2000), we place Beta priors on the $R_{m,p}$'s and a Gamma prior on $H$. The shape parameters of the Beta prior for each $R_{m,p}$ are chosen to center the multiresolution prior on a given discrete hazard $d_j^*, j = 1, \ldots, J$. To control for the amount of smoothing in the multiresolution prior, we multiply the shape parameter of the Beta priors at each additional level of the hierarchy by a hyperparameter $k$. For example, the priors for $H$ and $R_{m,p}$ given $M = 3$ and $J = 8$ would be

$$H \sim \mathcal{G}a(a, \lambda), \tag{1}$$

$$R_{1,0} \sim \mathcal{B}e\big(2\gamma_{1,0}ka, 2(1 - \gamma_{1,0})ka\big), \tag{2}$$

$$R_{2,p} \sim \mathcal{B}e\big(2\gamma_{2,p}k^2a, 2(1 - \gamma_{2,p})k^2a\big), \qquad p = 0, 1, \tag{3}$$

$$R_{3,p} \sim \mathcal{B}e\big(2\gamma_{3,p}k^3a, 2(1 - \gamma_{3,p})k^3a\big), \qquad p = 0, 1, 2, 3. \tag{4}$$

The use of arbitrary prior means, $E(R_{m,p}) = \gamma_{m,p}$, allows us to a priori "center" the baseline hazard at any desired value $d_j^*$. This is done by noting that because the $H$ and $R_{m,p}$'s are a priori independent, we can make $E(d_j) = d_j^*$ by specifying the prior means of $H$ and $R_{m,p}$ separately. Specifically, we first set recursively $D_{m-1,p} = D_{m,2p} + D_{m,2p+1}$, with $D_{M,0} = d_1^*$ and $D_{M,1} = d_2^*, \ldots, D_{M,2^M-1} = d_J^*$. We then let

$$\gamma_{m,p} = \frac{D_{m,2p}}{D_{m-1,p}} \qquad \text{and} \qquad a\lambda = \sum_{j=1}^{J} d_j^*. \tag{5}$$

This particular formulation of successive Beta priors also has two desirable properties: (1) prior resolution invariance; and (2) prior correlation of the $d_j$ dependent on the choice of $k$ and $a$; for proofs of these properties see the Appendix. Prior resolution invariance implies that the multiresolution prior on a particular $R_{m,p}$ does not depend on the number of levels $M$. In other words, by integrating out higher resolution parameters, one would obtain exactly the same prior as if those parameters had not been considered in the first place. When our model assumptions hold approximately, this invariance is reasonably preserved in the posterior inference as well, as explained in detail in Bouman et al. (2005).

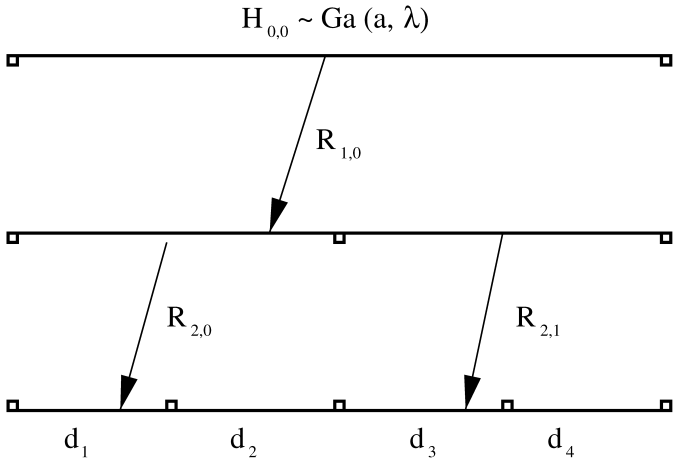


Figure 1. Annotated diagram of the two-level multiresolution prior.

As proved for a simpler version of this prior in Bouman et al. (2005), when $k = .5$, the baseline hazard increments $d_j$ are a priori uncorrelated and, in fact, independently Gamma distributed. Choosing $k$ less or greater than .5 yields, respectively, negative or positive prior correlation among the $d_j$'s. Positive prior correlation induces smoothing of the baseline hazard function, with hazard increments borrowing strength from the neighbors, which could be desirable in the presence of heavy censoring.

### 3.2 Hyperpriors for Hyperparameters $a$, $k$, and $\lambda$

Although a fixed $k$ can be specified, it is often possible to estimate $k$ from the data by putting a hyperprior on it. An exponential hyperprior with mean $\mu_k$ leads to the full conditional distribution for $k$ (conditioning on all other model parameter $k^-$),

$$\begin{aligned} \pi(k|k^-) \propto \exp\left(-\frac{k}{\mu_k}\right) & \\ \times \prod_{m=1}^{M} \prod_{p=0}^{2^{m-1}-1} & \left[\frac{\Gamma(2ak^m)}{\Gamma(2\gamma_{m,p}ak^m)\Gamma(2(1-\gamma_{m,p})ak^m)}\right. \\ & \left. \times (R_{m,p})^{2ak^m\gamma_{m,p}}(1 - R_{m,p})^{2ak^m(1-\gamma_{m,p})}\right], \end{aligned} \tag{6}$$

implying that the information in the data for $k$ will come from the joint posterior of $a$ and all the "splits" $R_{m,p}$.

Multiresolution priors are based on a treelike structure, which can induce a blocky correlation pattern. Namely, in a simple multiresolution model, where $a$ is fixed, it is possible to have a situation in which two neighboring hazard increments are less correlated than those further apart. To compensate for this undesirable property, Bouman et al. (2005) proposed mixing over the shape parameter $a$ to balance out the correlations among increments.

For this reason, we place a zero-truncated Poisson (ZTP) hyperprior on $a$. The ZTP prior was chosen mostly for computational convenience, because integer shape parameters will suffice for most practical purposes. The density for ZTP with parameter $\mu_a$ is $e^{-\mu_a}{\mu_a}^a/[a!(1-e^{-\mu_a})]$, resulting in the full conditional distribution for $a$ proportional to

$$\frac{\mu^a}{\Gamma(a+1)}\frac{H^a}{\Gamma(a)} \times \prod_{m=1}^{M}\prod_{p=0}^{2^{m-1}-1}\left[\frac{\Gamma(2ak^m)}{\Gamma(2\gamma_{m,p}ak^m)\Gamma(2(1-\gamma_{m,p})ak^m)} \times (R_{m,p})^{2ak^m\gamma_{m,p}}(1-R_{m,p})^{2ak^m(1-\gamma_{m,p})}\right]. \quad (7)$$

To model the prior uncertainty about the scale parameter $\lambda$ that governs the mean of the cumulative hazard $H$, we use an exponential prior on $\lambda$ with mean $\mu_\lambda$, yielding the full conditional distribution

$$\pi(\lambda|\lambda^-) \propto \exp\left(-\frac{\lambda}{\mu_\lambda}\right)\frac{\exp(-H(t_J)/\lambda)}{\lambda^a}. \quad (8)$$

We note that with the foregoing specification, to set the prior mean of $d_j$ at $d_j^*$ given the hyperparameters, we need to choose $\mu_a$ and $\mu_\lambda$ such that [based on (5)]

$$\frac{\mu_a\mu_\lambda}{1-e^{-\mu_a}} = \sum_{i=1}^{J} d_j^*.$$

### 3.3 Log-Linear Proportional-Hazards Likelihood With Center-Specific Effects

Our model handles general forms of failure time censoring via Bayesian imputation. Because our model only describes the baseline hazard on the interval $[0, t_J]$, we make inference on the model parameters conditional on imputing censored failure times before $t_J$, while assuming that all subsequent patient events are administratively right-censored and integrated out of the likelihood. In this section we derive the continuous-time complete-data likelihood, which is needed for posterior inference. We employ the proportional-hazards assumption (Cox 1972) that $h(t|X,\psi) = \exp(X'\psi)h_{\text{base}}(t)$. (The use of a discrete time proportional-hazards likelihood with the multiresolution prior is explored in Bouman et al. 2005.)

When the $i$th failure time $T_i$ is observed without censoring, the conditional likelihood function for $T_i \in [0, t_J]$, with $X = X_i$, is

$$L(\psi|T_i, X_i) = f(T_i|X_i,\psi) = h(T_i|X_i,\psi)S(T_i|X_i,\psi) = \exp(X_i'\psi)h_{\text{base}}(T_i)S_{\text{base}}(T_i)^{\exp(X_i'\psi)}, \quad (9)$$

using $f(T) = h(T)S(T)$. When a right-censored time $T_i > t_{\text{cens}}$ is observed, the conditional likelihood becomes

$$L(\psi|T_i, X_i) = S(t_{\text{cens}}|X_i,\psi) = S_{\text{base}}(t_{\text{cens}})^{\exp(X_i'\psi)}. \quad (10)$$

The covariates $X_i$ in the preceding likelihoods represent patient $i$'s information, such as age and tumor size, for which the effects on survival are not expected to vary from study center to study center. Whereas this is a multicenter study, we also allow a center-specific hazard multiplier $\exp(\eta_{0,c})$ to model the intrinsic survival heterogeneity that is due to the effects of other unobserved covariates. Following standard practice, we model the center effects $\boldsymbol{\eta_0} = (\eta_{0,1}, \ldots, \eta_{0,C})^\top$ as iid $\text{N}(0, \tau_0^{-2})$ and model $\tau_0^2$ (the precision) as an exponential variable with mean $\mu_0$.

Similarly, for clinical trials in which patients are administered one of two treatments (usually a standard and an experimental therapy), we also account for heterogeneity by center in the treatment/standard log-hazard ratio $\beta_{\text{treat}}$. Center-specific departures from $\beta_{\text{treat}}$ are denoted by $\boldsymbol{\eta_1} = (\eta_{1,1}, \ldots, \eta_{1,C})^\top$ so that the log-hazard ratio for center $c$ becomes $\beta_{\text{treat}} + \eta_{1,c}$. In this model, therefore, subjects in the standard arm of center $c$ of the trial will have the conditional hazard rate $\exp(\eta_{0,c} + X'\boldsymbol{\beta})h_{\text{base}}(t)$, while those in the experimental arm will have the conditional hazard rate $\exp(\eta_{0,c} + \eta_{1,c} + \beta_{\text{treat}} + X'\boldsymbol{\beta})h_{\text{base}}(t)$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_L)^\top$ denotes the vector of covariate effects other than that of treatment. Again, assuming that trial centers are drawn at random from a larger population, the $\eta_{1,c}$ are modeled as $\text{N}(0, \tau_1^{-2})$, with $\tau_1^2$ an exponential variable similar to the one used for $\tau_0^2$. Although we model the two center-specific effects as a priori independent mostly because the center heterogeneity appears rather low, correlated center-specific effects can be employed by using a multivariate normal prior. The parameter $\boldsymbol{\beta}$ is given a vague multivariate normal prior.

## 4. MODEL FITTING AND MODEL CHECKING

Given the multiresolution prior and survival time likelihood described in the previous section, we now detail the Markov chain Monte Carlo estimation procedures for estimating the parameter posteriors for $H$, $k$, and the $R_{m,p}$. In addition, we outline a model checking procedure based on the posterior predictive distribution.

### 4.1 Markov Chain Monte Carlo Bayesian Model Estimation

For patient $i$, we observe the failure or censoring time $T_i$, a censoring indicator $\delta_i$ (0 for censoring, 1 for observed failure), center number $c$ ($c = 1, \ldots, C$), and covariates $X_i$. The likelihood contribution from the $i$th patient is

$$L(T_i|\delta_i, \boldsymbol{\beta}, \beta_{\text{treat}}, \boldsymbol{\eta_0}, \boldsymbol{\eta_1}, H, R_{m,p}, X_i, c, A_i) = [P_i h_{\text{base}}(T_i)]^{\delta_i}\exp\big(-P_i H_{\text{base}}(\min(T_i, t_J))\big),$$

where $P_i = \exp(X_i'\boldsymbol{\beta} + \eta_{0,c} + A_i(\eta_{1,c} + \beta_{\text{treat}}))$ and $A_i$ is a dummy variable that denotes participation in the experimental treatment trial arm for the $i$th patient. Note that $P_i$ is the hazard proportion for the $i$th patient, depending on the effects for overall treatment effect, center-specific effects, and effects for other covariates. Patients with failure time $T_i > t_J$ are considered administratively censored ($\delta_i = 0$) at time $t_J$ and contribute $P_i\exp(-H_{\text{base}}(t_J))$ to the total likelihood. For $N$ observed patients, the total log-likelihood expression then becomes

$$\boldsymbol{\delta}'[\mathbf{X}\boldsymbol{\beta} + \Gamma_{\mathbf{0}}\boldsymbol{\eta_0} + \Gamma_{\mathbf{1}}\boldsymbol{\eta_1} + \beta_{\text{treat}}\mathbf{A} + \mathbf{F}\Pi\mathbf{R}] - \sum_{i=1}^{N}\exp(X_i'\boldsymbol{\beta} + \Gamma_{0,i}\boldsymbol{\eta_0} + \Gamma_{1,i}\boldsymbol{\eta_1} + \beta_{\text{treat}}A_i) \times H_{\text{base}}(\min(T_i, t_J)), \quad (11)$$

where $\mathbf{X} = (X_1, X_2, \ldots, X_N)'$ is the $N \times L$ matrix of covariates other than treatment assignment; $\Gamma_\mathbf{0}$ and $\Gamma_\mathbf{1}$ are $N \times C$ indicator matrices of membership in the $c$th study and the treatment arm of the $c$th study, respectively: $\mathbf{A}$ is the vector of $N$ treatment assignments; $\Pi$ is the $2^M \times (2^{M+1} - 1)$ multiresolution matrix for which the $(i, j)$th element is 1 when $j = 1$ or $i \in [1 + 2^{M-m}(j \bmod 2^m), \ldots, 2^{M-m} + 2^{M-m}(j \bmod 2^m)]$, $m = \lfloor \log_2(j) \rfloor$, and 0 otherwise; $\mathbf{R}$ is the multiresolution parameter vector $(\log(H), \log(R_{1,0}), \log(1 - R_{1,0}), \ldots, \log(R_{M,0}), \log(1 - R_{M,0}), \ldots, \log(R_{M,2^M-1}), \log(1 - R_{M,2^M-1}))$; $\mathbf{F}$ is an $N \times 2^M$ matrix for which the $(i, j)$th element is 1 if patient $i$ has an event (observed failure or right-censoring) at a time $T_i \in (t_{j-1}, t_j]$ and 0 otherwise. (Patients with failure or right-censoring times $T_i > t_J$ have $F_{i,j} = 0, j = 1, \ldots, J$.) Note that we use the piecewise-constant hazard rate assumption to compute the cumulative hazard $H(T_i)$ at all times $T_i < t_J$.

The Gibbs sampler steps (Geman and Geman 1984) for the parameters $H$, $R_{m,p}$, and $k$ are as follows:

1. Draw $H$ from the full conditional $\pi(H|\lambda, R_{m,p}) = \mathcal{G}a((a + \sum_{i=1}^N \delta_i), 1/[(1/\lambda) + \sum_{i=1}^N F(T_i)])$, with mean $\mu = (a + \sum_{i=1}^N \delta_i)/[(1/\lambda) + \sum_{i=1}^N F(T_i)]$, where $F(T_i) = H(\min(T_i, t_J))/H(t_J)$, a function of $T_i$ and $R_{m,p}$.
2. Draw the $R_{m,p}$ (in any order) from $\pi(R_{m,p}|k, H)$ with log full conditional (for given $m$ and $p$)

$$\left(ak^m - 1 + \sum_{i=1}^N \delta_i \Pi_{i,r}\right) \log(R_{m,p}) + \left(ak^m - 1 + \sum_{i=1}^N \delta_i \Pi_{i,r'}\right) \log(1 - R_{m,p}) - \sum_{i=1}^N H(\min(T_i, t_J)), \quad (12)$$

   where $r$ and $r'$ are the columns of $\mathbf{\Pi}$ that correspond to $\log(R_{m,p})$ and $\log(1 - R_{m,p})$, respectively. Observe that this distribution is *not* Beta, because the terms $H(\min(T_i, t_J))$ depend on the $R_{m,p}$ as well as $H$ when $T_i < t_J$.
3. Draw $k$ from $\pi(k|R_{m,p})$, $\lambda$ from $\pi(\lambda|H)$, and $a$ from $\pi(a|H, R_{m,p})$ as described in Section 3.2.

The conditional posterior distributions for $H$, $\tau_0$, and $\tau_1$ are conjugate Gamma. The full conditionals for $R_{m,p}$, $\eta_{0,c}$, $\eta_{1,c}$, and each $\beta_l$ and $\beta_{\text{treat}}$ are log concave and therefore can be sampled by the adaptive rejection sampling (ARS) algorithm of Gilks and Wild (1992). To sample from the full conditional distributions for the hyperparameters $\lambda$ and $k$, which are in general *not* log concave, we use an extension of ARS known as adaptive rejection Metropolis sampling, which was described by Gilks, Best, and Tan (1995).

### 4.2 Posterior Predictive Model Checking

The use of the posterior predictive distribution (PPD) for model checking was investigated in Gelman et al. (1996), where distributions of realized discrepancy statistics were used to diagnose directions of inadequate model fit to data. The PPD for a future, replicate observation $y^{\text{rep}}$ given the vector of observed censorings, failure times $y$, and model $M$ is formally defined as

$$P(y^{\text{rep}}|y, M) = \int P(y^{\text{rep}}|\boldsymbol{\theta}, M) P(\boldsymbol{\theta}|y)\, d\boldsymbol{\theta}, \quad (13)$$

where $\boldsymbol{\theta}$ is the vector of parameters for model $M$ and $P(\boldsymbol{\theta}|y)$ is its posterior given the data. We usually condition the PPD on the collection of covariates $\mathbf{X}$, treatment assignments $\mathbf{A}$, and center memberships $c(i)$ for each patient, and write

$$P\big(y^{\text{rep}}|y, M, \mathbf{X}, \mathbf{A}, c(i)\big) = \int P\big(y^{\text{rep}}|\boldsymbol{\theta}, M, \mathbf{X}, \mathbf{A}, c(i)\big) P\big(\boldsymbol{\theta}|y, \mathbf{X}, \mathbf{A}, c(i)\big)\, d\boldsymbol{\theta}. \quad (14)$$

Given a set of Monte Carlo draws $\boldsymbol{\theta}^g, g = 1, \ldots, G$, from the parameter posterior, for each $g$ we draw one $y^{\text{rep},g}$ from $P(y^{\text{rep}}|\boldsymbol{\theta}^g, M, \mathbf{X}, \mathbf{A}, c(i))$. For our model, this amounts to drawing a set of $N$ survival times $T_i^{\text{rep},g}$ from $P(T_i^{\text{rep},g}|\beta_{\text{treat}}^g, \boldsymbol{\beta}^g, \boldsymbol{\eta_0}^g, \boldsymbol{\eta_1}^g, \mathbf{d}^g, M, \mathbf{X}, \mathbf{A}, c(i))$, where each $T_i^{\text{rep},g}$ is in one of $(t_{j-1}, t_j]$, $j = 1, \ldots, J$, or $(t_J, \infty]$, because we only work with a discrete approximation to the baseline hazard function.

We can develop univariate statistics [functions of the original data $y$ or replicate $y^{\text{rep},g}$ from the PPD (13)] that measure departures from model assumptions in directions meaningful to applied practitioners. One class of functions is of the form $T(y^{\text{rep},g}|t_j) = f(\widehat{S}_{\text{treat}}^{\text{rep},g}(t_j)) - f(\widehat{S}_{\text{control}}^{\text{rep},g}(t_j))$, where $f(\cdot)$ is a chosen function, $t_j$ is a given time horizon, and $\widehat{S}_{\text{treat}}^{\text{rep},g}$ and $\widehat{S}_{\text{control}}^{\text{rep},g}$ are estimated survival functions in treatment and control arms given the replicate data. When $f(s) = s$, our statistic is a difference in survival probabilities; when $f(s) = \log(s)$, it is the difference in cumulative hazards at time $t_j$; when it is the complementary log–log function $f(s) = \log(-\log(s))$, we obtain the log ratio of cumulative hazards between treatment and control arms. Due to the popularity (and interpretability as an estimate of $\beta_{\text{treat}}$) of the log-hazard ratio, we use it here as an additional check of the validity of our proportional-hazards assumptions.

## 5. ANALYSIS OF BREAST CANCER MULTICENTER TRIAL

For all analyses performed in this article we chose a 32-bin model, with the resolution corresponding to a constant 3.75-month hazard rate. We chose this resolution because it is the closest to the 3- to 6-month monitoring intervals of the trial protocol. The model parameters were estimated using output from five Gibbs sampler chains with 1,000,000 iterations each, of which the first 500,000 draws were discarded as burn-in. Of the remaining 500,000 draws, every fifth iteration was retained to reduce correlation, resulting in 100,000 draws for our analysis (although the thinning is unnecessary for most parts of our analysis). Standard Gelman–Rubin diagnostics performed separately for each parameter were used to check convergence. We discuss the results for DFS in Section 5.1 and then for the BCFS endpoint in Section 5.2.

### 5.1 Results for the DFS Analysis

For DFS analysis, we applied our model to all centers together, as well as to the three center strata separately. The strata were constructed by sorting the 110 trial centers according to their size and then dividing them into three groups in such a

Table 2. Posterior credible intervals for predictor effects: All centers combined; 10-year DFS

| | Trt. | Tumor: med. | Tumor: lg. | PGR high | Age lin. | Age quad. |
|---|---|---|---|---|---|---|
| Pooled analysis | | | | | | |
| 2.5% | −.58 | .17 | .30 | −.44 | −.01 | .05 |
| 50% | −.44 | .29 | .55 | −.29 | .07 | .11 |
| 97.5% | −.30 | .42 | .80 | −.15 | .13 | .16 |
| Large centers | | | | | | |
| 2.5% | −.69 | .05 | −.19 | −.43 | −.08 | −.01 |
| 50% | −.46 | .27 | .31 | −.19 | .04 | .08 |
| 97.5% | −.21 | .49 | .80 | .09 | .16 | .17 |
| Medium centers | | | | | | |
| 2.5% | −.70 | .13 | .23 | −.59 | −.02 | .07 |
| 50% | −.47 | .36 | .66 | −.35 | .09 | .17 |
| 97.5% | −.23 | .57 | 1.09 | −.10 | .19 | .27 |
| Small centers | | | | | | |
| 2.5% | −.64 | −.03 | .11 | −.61 | −.07 | −.02 |
| 50% | −.39 | .23 | .57 | −.35 | .07 | .08 |
| 97.5% | −.14 | .47 | .99 | −.09 | .20 | .17 |

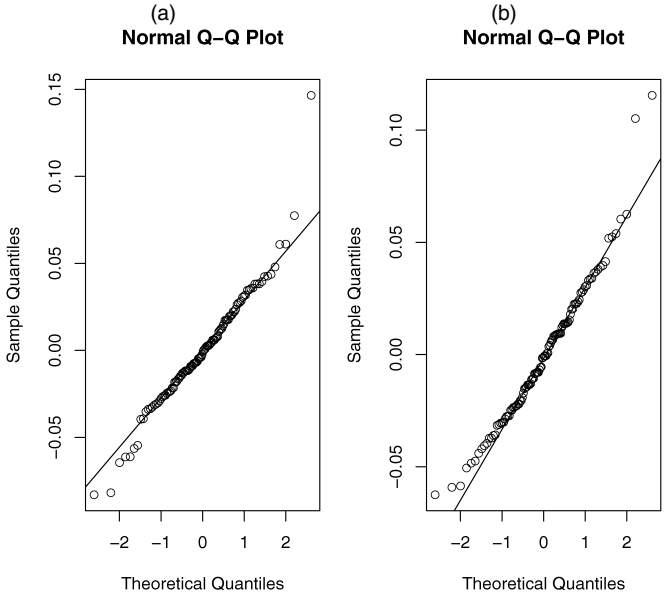


Figure 2. The Q–Q plots of center-specific baseline heterogeneity: 10-year DFS (a) and BCFS (b) endpoints.

way that roughly the same number of patients are in each group. This grouping resulted in the following strata: a "large size" stratum with 9 centers (size varies from 51 to 241) and 956 patients, a "medium size" stratum with 29 centers (size 19–50) and 987 patients, and a "small size" stratum with 72 centers (size 4–18) and 762 patients. The results are shown in Table 2, which presents marginal 95% posterior credible intervals for $\beta_{\text{treat}}$ and the overall (pooled) tamoxifen/placebo log-hazard ratio estimated across trial centers, as well as each of the other covariate effects $\beta_l$.

The 95% posterior credible interval estimate of (.56, .74) for the pooled treatment–control hazard ratio $\exp(\beta_{\text{treat}})$ indicates a clear protective effect for tamoxifen. Tumor size is also shown to be prognostic of 10-year disease-free survival, with the posterior median of the parameter increasing with increasing tumor size at time of enrollment. A higher concentration of progesterone receptors is shown to be protective, and there is posterior evidence of a quadratic effect in age above and below the mean age of 54.7 years. These results agree with those of the analysis in Bryant et al. (1998), which suggested that the quadratic effect was due to an increased frequency of recurrence for younger women, who tend to have aggressive tumors, coupled with increased failure risk for older women due to other causes, such as a second primary cancer or death from noncancer causes. Due to little center heterogeneity in the data, our covariate effect estimates agree very closely with the estimates produced by a simple Cox proportional-hazards marginal model fitted without any center-specific effects. In addition, the Q–Q plots of those center-specific median effects in both DFS and BCFS analyses, shown in Figure 2, reveal no apparent contradiction to the normality of center effects, an assumption convenient for flagging extreme center effects based on the classic "two sigma" rule.

The stratified analysis results largely agree with the overall analysis, especially regarding the treatment estimates. As a partial check of our results, we also examined (Fig. 3) Kaplan–Meier estimates of survival curves for all centers pooled, as well as for each stratum. These curves closely resemble each other, lending support to our findings in Table 2.

Figure 4 displays pooled and stratified median posterior estimates for the 32 baseline hazard increments themselves (corresponding to constant 3.75-month hazard rates) for DFS endpoint. Pointwise posterior credible intervals are also shown, where the boxplots give posterior medians, quartiles, and 2.5th and 97.5th estimated posterior percentiles, yielding what we like to refer to as *caterpillar plots*. Note that this 32-dimensional vector is a discrete approximation to the baseline hazard rate $h_{\text{base}}(t)$, for which we estimate the hazard increment $d_j = \int_{t_{j-1}}^{t_j} h_{\text{base}}(s)\,ds$. Note that we show the 32 pointwise estimates of the hazard increments, although aggregation of these increments or further postanalysis smoothing could be performed as desired. The estimate in Figure 4 is consistent with the noted pattern of a rising hazard to the end of the second year of treatment (24 months), followed by a gradual decline in the following years; in a number of other large-scale cohorts of early stage breast cancer patients who receive surgical treatment, the time-varying rate of recurrence seems to peak around 2 years and then decrease (Hess, Pusztai, Buzdar, and Hortobagyi 2003). The apparent nonmonotonic pattern may be due to an increase in length of the screening interval beginning after 4 years on study.

## 5.2 Results for the BCFS Analysis

Results for the BCFS endpoint are somewhat different. First, notice from looking at the pooled analysis results in Table 3 that the restriction of the outcomes from DFS to the BCFS results in clinical predictors having slightly stronger effects and, in particular, that the median estimate of protective treatment effect in the BCFS analysis is larger than that for the DFS endpoint. We hypothesize that because the DFS endpoint includes events other than breast cancer recurrence, the protective effect of tamoxifen should, on average, be correspondingly smaller

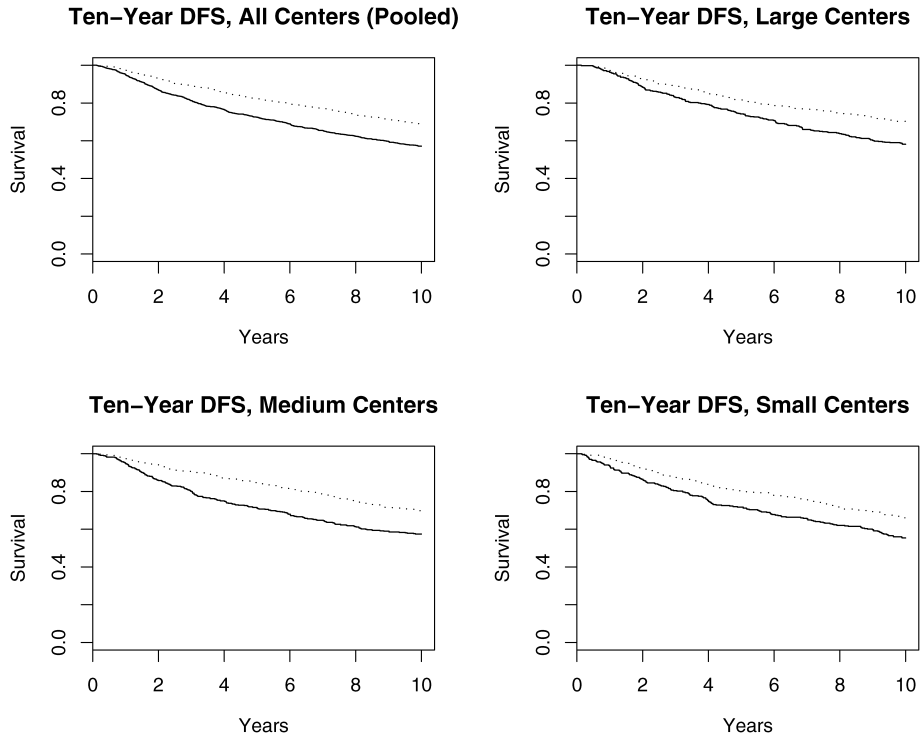


Figure 3. Kaplan–Meier plots for 10-year disease-free survival: All centers combined and stratified by size (—— control; ····· treatment).

than that in the BCFS analysis, with a similar reasoning for the effects of high progesterone receptor level.

Second, from Table 3 one can see that the BCFS endpoint also displays a higher degree of heterogeneity in parameter estimates across different center strata. Although the large and small centers generally resemble the pooled survival estimates in the two arms, medium centers seem to have a slightly higher treatment effect. This contrast is also demonstrated in Figure 5, where Kaplan–Meier BCFS estimates across treatment arms for the pooled and the three strata are shown. Although not sta-

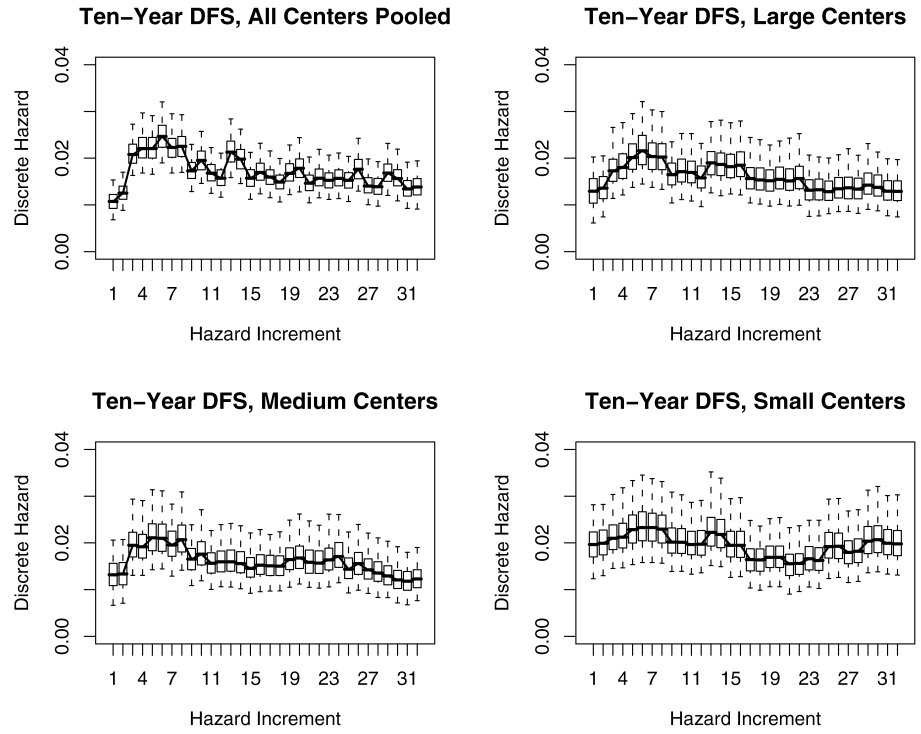


Figure 4. Caterpillar plots of posterior pointwise estimates of baseline hazard increments for DFS data (with 2.5%/25%/50%/75%/97.5% credible boxplots).

Table 3. Posterior credible intervals for predictor effects: All centers combined, 10-year BCFS

| | Trt. | Tumor: med. | Tumor: lg. | PGR high | Age lin. | Age quad. |
|---|---|---|---|---|---|---|
| Pooled analysis | | | | | | |
| 2.5% | −.73 | .27 | .38 | −.52 | −.14 | .02 |
| 50% | −.58 | .44 | .71 | −.36 | −.06 | .08 |
| 97.5% | −.41 | .59 | .99 | −.19 | .03 | .15 |
| Large centers | | | | | | |
| 2.5% | −.83 | .25 | −.14 | −.55 | −.22 | −.06 |
| 50% | −.48 | .50 | .52 | −.25 | −.08 | .06 |
| 97.5% | −.19 | .77 | 1.13 | .05 | .07 | .17 |
| Medium centers | | | | | | |
| 2.5% | −1.00 | .24 | .35 | −.72 | −.14 | .05 |
| 50% | −.72 | .51 | .85 | −.46 | −.01 | .17 |
| 97.5% | −.45 | .80 | 1.30 | −.15 | .14 | .28 |
| Small centers | | | | | | |
| 2.5% | −.79 | −.04 | −.04 | −.69 | −.24 | −.09 |
| 50% | −.52 | .26 | .54 | −.38 | −.09 | .04 |
| 97.5% | −.22 | .55 | 1.04 | −.05 | .08 | .15 |

tistically significant, this potential difference across the three strata might warrant further investigation. However, perhaps the most interesting finding of all is the robustness of the tamoxifen treatment effect. It is indeed rather remarkable, at least in this trial, that even in a collection of small center studies, despite expected individual center variations, one can obtain similar results as in larger, presumably well-established center, studies.

The caterpillar plots in Figure 6, show pooled and stratified posterior median and pointwise 95% credible interval estimates for the 32 baseline hazard increments themselves (corresponding to constant 3.75-month hazard rates) for the BCFS endpoint. The hazard estimates in Figure 6 show a similar pattern to the DFS hazards, with the baseline BCFS hazard increments peaking around 22 months and declining thereafter.

### 5.3 Posterior Predictive Model Checking

Figure 7 displays the posterior predictive distribution of the treatment–control log-hazard ratio $T(y^{\text{rep}}|t_j)$ at eight time horizons $t_j$ ($t_j \in \{7.5, 15, 22.5, 30, 37.5, 45, 52.5, 60\}$ months). Each histogram shows 500 draws from the posterior predictive distribution of

$$T(y^{\text{rep}}|t_j) = \log\big(-\log(S^{\text{rep}}_{\text{treat}}(t_j))\big) - \log\big(-\log(S^{\text{rep}}_{\text{control}}(t_j))\big),$$

where each $y^{\text{rep}}$ is a replicate dataset of size $N = 2{,}705$. (In repeated trials, a sample size of 500 draws was found to give a reasonably stable Monte Carlo estimate of the PPD.) The vertical line in the histogram indicates the observed value $T(y|t_j)$ for each time horizon, and its position can be used to judge the agreement between the observed data and the PPD reference distribution under the proportional-hazards assumption. If proportional hazards gave a poor fit to the data, we would expect that the PPD would indicate a roughly constant log-hazard ratio over time, where the data would show stronger time variation. We see that PPD reveals no strong contradictions from the 15-month horizon onward, but at 7.5 months there seems to be some evidence of nonproportionality, with the realized value of the log cumulative hazard ratio appearing in the far left tail of the PPD reference distribution. At this early follow-up time, however, very few failure events have occurred, making the corresponding hazard estimates more unstable. As a non-Bayesian check of proportional hazards, we also produced plots

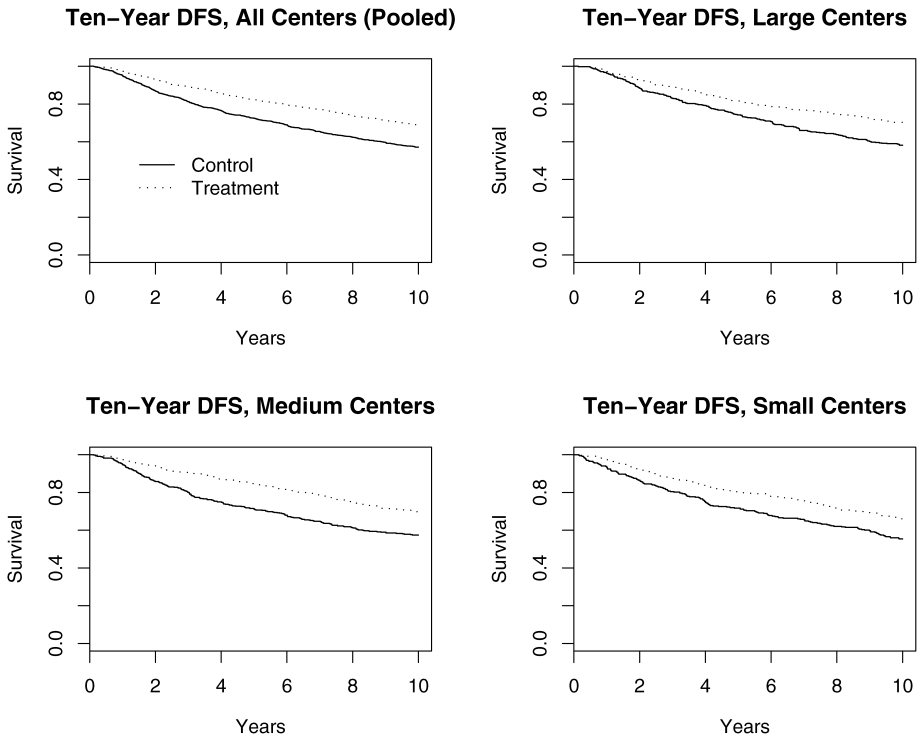


Figure 5. Kaplan–Meier plots for 10-year breast-cancer-free survival: All centers combined and stratified by size (—— control; ····· treatment).

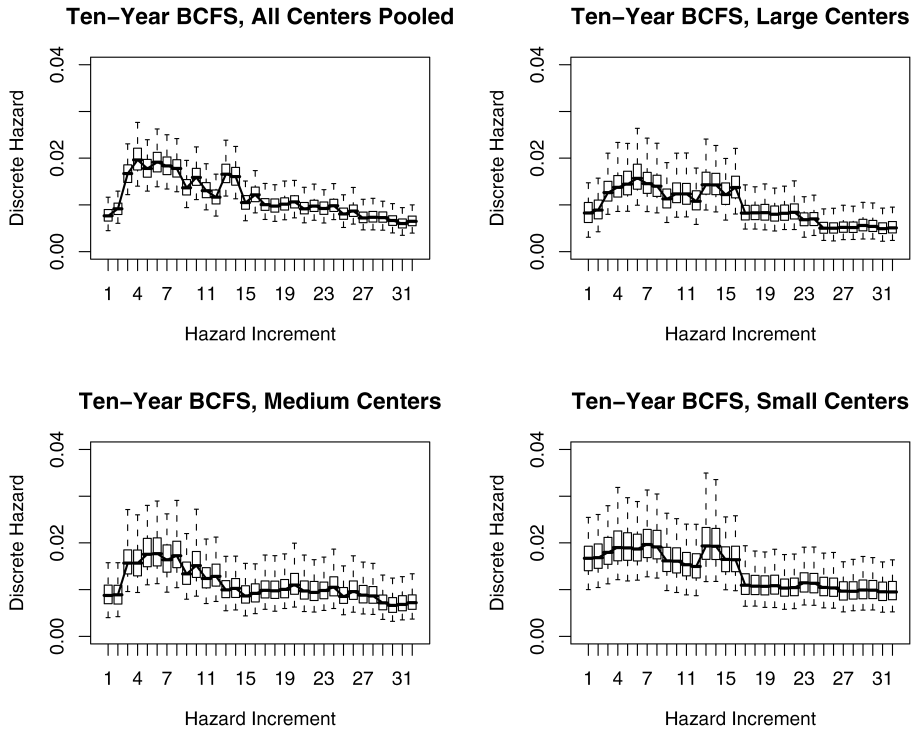


Figure 6. Caterpillar plots of posterior pointwise estimates of baseline hazard increments for BCFS data (with 2.5%/25%/50%/75%/97.5% credible boxplots).

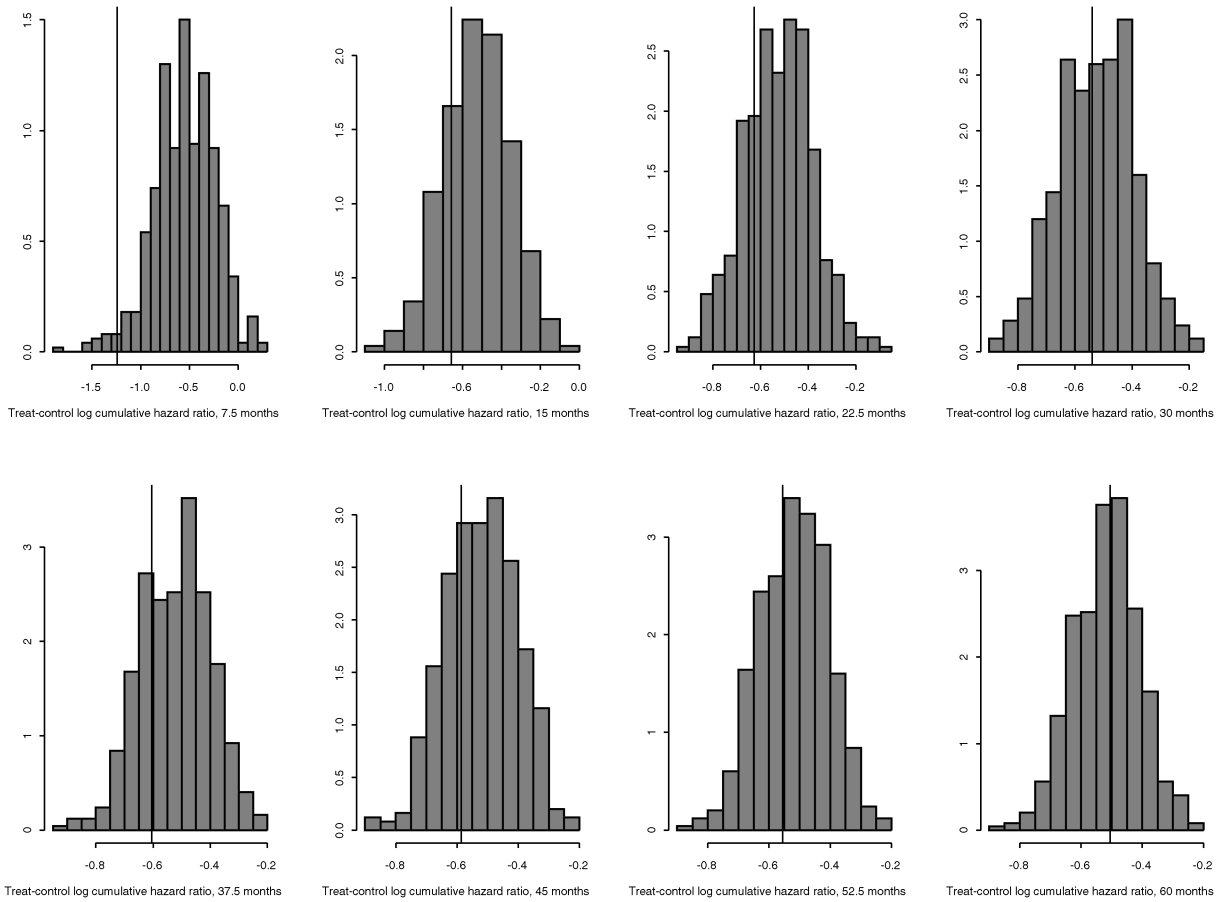


Figure 7. Posterior predictive distributions for treatment–control log cumulative hazard ratios: Eight time points.

of Schoenfeld residuals for the treatment effect (not shown) for each stratum and endpoint, but there was no significant evidence of nonproportionality in the data.

### 5.4 Sensitivity to Prior Mean

To illustrate the nature of different priors we have considered, in Figure 8 we present 10 draws from an exponential-centered prior and from a Weibull-centered (with shape parameter equal to 3) prior. The two priors seem to differ most notably in the later time period, with the Weibull-centered prior having a tendency to rise toward the end of the 10-year period. The effect of the two priors on the final baseline DFS hazard estimates is shown in Figure 9. Although the amount of smoothness in the two estimates seems different (the Weibull-based hazard estimate seems to be less smooth than the exponential-based one, as expected), there seems to be very little practical difference among the two sets of covariate estimates (Table 4). We thus do not repeat the sensitivity analysis for the BCFS data.

### 5.5 Comparison to Gray's Model

Our model of multicenter clinical trial outcomes can be compared to the Bayesian hierarchical model explored in Gray (1994). Following Gray's notation, we model survival times in $N$ trial centers, each with $n_i$ patients with covariates $x_{ijk}$ for the $k$th covariate of the $j$th patient in the $i$th center. We assume that there are $p-1$ (equal to $L$ in our notation) other covariates in addition to the treatment assignment, for a total of $p$ variables per patient. To model the baseline hazard, time is divided into $m$ discrete intervals with boundaries $0 < t_1 < \cdots < t_m$ and corresponding discrete hazard increments $e^{\alpha_i}$. (In the analysis in Gray 1994, $m$ was set to 30, which was found to be large enough and not to have a significant impact on effect estimation.) The vector of proportional effects for the $p$ covariates is written as $\beta = (\beta_1, \ldots, \beta_p)'$, and the center-specific baseline and treatment effects are, respectively, denoted $\theta_{i0}$ and $\theta_{i1}$ for $i = 1, \ldots, N$. Note that the total treatment effect for center $i$ is then $\beta_p + \theta_{i1}$. The hazard for subject $ij$ is then written as

$$\lambda(t|x_{ij}, \alpha, \beta, \theta_i) = \exp\left\{\sum_{q=1}^{m} \alpha_q I_q(t) + \theta_{i0} + \sum_{k=1}^{p} \beta_k x_{ijk} + \theta_{i1} x_{ijp}\right\}, \quad (15)$$

where $I_q(t) = 1$ if $t_{q-1} < t \le t_q$ and is 0 otherwise.

For this parameterization, Gray (1994) placed independent normal priors on $\beta_k$ and a zero-centered bivariate normal prior on the center effects $\theta_i$, whose covariance is given an inverse-Wishart hyperprior. Successive log-hazard increments $\alpha_q - \alpha_{q-1}$ are modeled as independent normal, with an inverse-Gamma hyperprior on the variance. Writing $\gamma_{ijq} = \alpha_q + \theta_{i0} + \sum_{k=1}^{p} \beta_k x_{ijk} + \theta_{i1} x_{ijp}$, $\delta_{ijq}$ as the failure indicator for patient $ij$ in interval $q$, and $u_{ijq}$ as the total follow-up time in interval $q$, the likelihood term for center $i$ becomes

$$L_i(\alpha, \beta, \theta_i) = \exp\left\{\sum_{j=1}^{n_i} \sum_{q=1}^{m} [\delta_{ijq}\gamma_{ijq} - u_{ijq}\exp(\gamma_{ijq})]\right\}. \quad (16)$$

Gray (1994) estimated the posterior distribution for $(\alpha, \beta, \theta_i)$ via Gibbs sampling.

We fitted Gray's model of center-specific heterogeneity for the 10-year DFS endpoint, for all centers together, and confirmed our findings that there was little heterogeneity in these data. Table 5 gives posterior estimates for the treatment effect and other covariates for Gray's model. As can be seen, they agree closely with our estimates in Table 2.

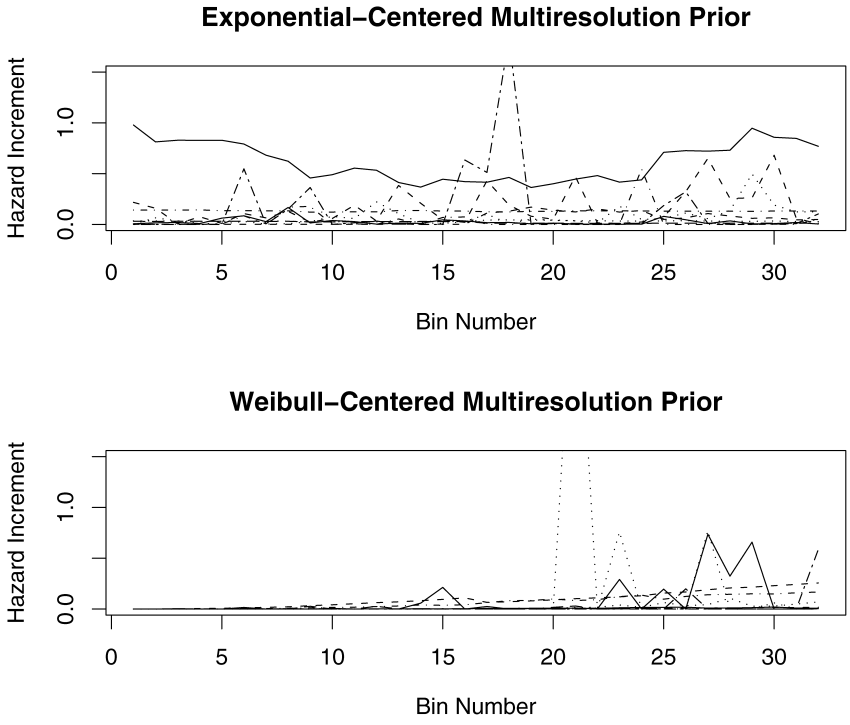


Figure 8. Ten random draws from multiresolution priors centered on $H(t) \propto t$ and $H(t) \propto t^3$.

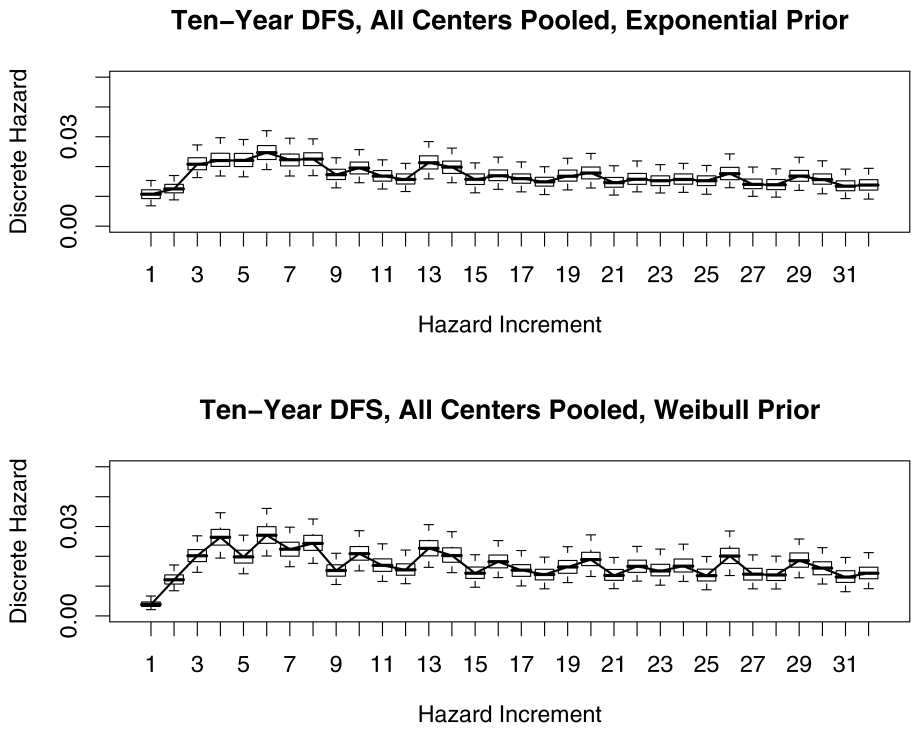


Figure 9. Caterpillar plots of the baseline hazard estimate for the DFS data under the exponential and Weibull priors.

## 6. DISCUSSION

Because there is usually a strict study protocol in a randomized trial, the problem addressed in this article is somewhat different from the well-known phenomenon of intrahospital variation in treatment outcomes, which may be attributable to a wide variety of factors, including attraction of difficult cases to certain hospitals or differential skills and resources across hospitals. However, like observational studies, variation in outcomes in the randomized trial setting may be due to differences in patient baseline characteristics that are not controlled by trial entry criteria. Furthermore, deviation from the protocol prescribed treatment and follow-up regimen could account for apparent variations in treatment efficacy.

Our stratified analysis is an attempt to include a measure of case availability and possibly cancer treatment expertise. Known prognostic factors, which are balanced by treatment group overall through stratified randomization, may differ by center and were included in modeling. Although DFS stratified analyses (and, in particular, the baseline hazard rate estimates) were very similar to the pooled analysis, for the BCFS endpoint we observe indications of larger variation among clinical effects in the three strata based on size of institution. For the DFS endpoint, which may tend to reflect patient heterogeneity in risks for other diseases, this heterogeneity was much smaller, and both the effect of treatment and some clinical covariates were more modest. Further analyses of this finding will entail consideration of other possible explanatory factors.

The B-14 trial included several large centers, with the 12 largest accruing centers accounting for nearly 40% of the over 2,800 patients enrolled. However, in many cases, recruitment of participants to multicenter trials is highly diffuse and has many small individual contributions. This is, in fact, a necessity in trials for less common diseases, where a large catchment area (with many individual contributing centers) and lengthy accrual period are necessary to obtain a sufficient sample size. In trials with more complex treatment regimens and numerous, mostly small centers, heterogeneity of treatment effects might appear substantial. Methods such as those proposed here could be used to evaluate the significance of intracenter variations in treatment efficacy.

While investigating this multicenter dataset, we have also adapted and explored some of the properties of a multiresolution estimator for a discrete time proportional-hazards model.

Table 4. Posterior credible intervals for predictor effects: 10-year DFS all centers; exponential-versus Weibull-centered hazard prior

| Prior | Trt. | Tumor: med. | Tumor: lg. | PGR high | Age lin. | Age quad. |
|---|---|---|---|---|---|---|
| Exponential-centered | | | | | | |
| 2.5% | −.58 | .17 | .30 | −.44 | −.01 | .05 |
| 50% | −.44 | .29 | .55 | −.29 | .07 | .11 |
| 97.5% | −.30 | .42 | .80 | −.15 | .13 | .16 |
| Weibull-centered | | | | | | |
| 2.5% | −.57 | .17 | .29 | −.44 | −.01 | .06 |
| 50% | −.44 | .29 | .56 | −.29 | .06 | .11 |
| 97.5% | −.31 | .42 | .81 | −.15 | .13 | .16 |

Table 5. Posterior credible intervals for predictor effects under Gray's model: All centers combined; 10-year DFS

| | Trt. | Tumor: med. | Tumor: lg. | PGR high | Age lin. | Age quad. |
|---|---|---|---|---|---|---|
| 2.5% | −.56 | .17 | .31 | −.43 | .00 | .06 |
| 50% | −.43 | .30 | .57 | −.29 | .07 | .11 |
| 97.5% | −.30 | .43 | .81 | −.15 | .13 | .16 |

Our method furthers the existing multicenter approaches by allowing for more generally censored (e.g., interval-censored or truncated) data. For survival estimation in large public-health databases, we often work with discretized covariates and survival times. In some situations, using "binned" failure times and making the underlying population survival curve a finite-dimensional parameter may be a desirable goal by itself. In other situations, continuously observed outcomes are simply not available. In any case, the analysis must take into account the effects of discretization on inference and the resulting loss in precision.

We have also demonstrated the use of the posterior predictive distribution in model criticism. We believe, in particular, that such techniques can be used to detect and justify adjustment for heterogeneity in treatment and other covariate effects across multiple studies, or multiple centers in a single study, as we have applied them. Of course, the center-specific effects estimated in our model are approximations to the actual heterogeneity in survival times, but including such effects can still give some indication of how individual centers depart from the overall effect. Future work will include investigation of heterogeneity in covariate effects other than the treatment effect we have looked at in this study.

## APPENDIX: PROPERTIES OF THE MULTIRESOLUTION PRIOR

Here we give proofs of two properties of the generalized multiresolution prior given in (1)–(4) in Section 3.1 when $0 < \gamma_{m,p} < 1$. (These arguments are generalizations of those given in Bouman et al. 2005 for the simpler case in which $\gamma_{m,p} \equiv .5$.)

First, when $k = .5$, we observe that in the prior, $H \sim \mathcal{G}a(a, \lambda)$ and $R_{1,0} \sim \mathcal{B}e(\gamma_{1,0}a, (1 - \gamma_{1,0})a)$, with the "next level" hazard increments $H_{1,0} = HR_{1,0}$ and $H_{1,1} = H(1 - R_{1,0})$. A change-of-variables calculation, using the fact that $H$ is independent of $R_{1,0}$, shows that $H_{1,0}$ and $H_{1,1}$ are independently distributed as $\mathcal{G}a(\gamma_{1,0}a, \lambda)$ and $\mathcal{G}a((1 - \gamma_{1,0})a, \lambda)$, respectively. By following this argument recursively to level $M$ in the prior, we see that the $d_j$ are independently Gamma-distributed, with shape parameters that depend on the values of the $\gamma_{m,p}$.

Second, our prior for any $H_{m,p}$ does not depend on our choice of $M$, so that the prior is resolution-invariant under aggregation of the $d_j$'s. For $m = 0$, this statement is trivially true, because the Gamma prior for $H_{0,0} \equiv H$ does not depend on $M$. For a fixed level $0 < m < M$, the prior for $H_{m,p}$ will be $\mathcal{G}a(a\prod_{i=1}^{m} b_{i,p_i}, \lambda)$, where $b_{i,p_i} = \gamma_{i,p_i}$ or $1 - \gamma_{i,p_i}$, depending on whether the splitting $R_{i,p_i}$ variable in forming $H_{m,p}$ is in the left branch or the right branch (see Fig. 1). Clearly, this distribution does not depend on $M$ or any parameters for the "lower levels" of the prior; that is, for fixed $m$, the joint prior of the $H_{m,p}$ is invariant to the choice of $M > m$.

*[Received December 2003. Revised December 2005.]*

## REFERENCES


Bouman, P., Dukić, V., and Meng, X.-L. (2005), "A Bayesian Multiresolution Hazard Model With Application to an AIDS Reporting Delay Study," *Statistica Sinica*, 15, 325–357.

Bryant, J., Fisher, B., Gündüz, N., Costantino, J., and Emir, B. (1998), "S-Phase Fraction Combined With Other Patient and Tumor Characteristics for the Prognosis of Node-Negative, Estrogen-Receptor Positive Breast Cancer," *Breast Cancer Research and Treatment*, 51, 47–61.

Cox, D. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society*, Ser. B, 34, 187–220

Cox, D., and Oakes, D. (1984), *Analysis of Survival Data*, New York: Chapman and Hall.

Earle, C., Pham, B., and Wells, G. (2000), "An Assessment of Methods to Combine Published Survival Curves," *Medical Decision Making*, 20, 104–111.

Fisher, B., Costantino, J., Redmond, C., Poisson, R., Bowman, D., Couture, J., Dimitrov, N., Wolmark, N., Wickerham, D., Fisher, E., Margolese, R., Robidoux, A., Shibata, H., Terz, J., Paterson, A., Feldman, M., Farrar, W., Evans, J., Lickley, H., and Ketner, M. (1989), "A Randomized Clinical Trial Evaluating Tamoxifen in the Treatment of Patients With Node-Negative Breast Cancer Who Have Estrogen-Receptor-Positive Tumors," *New England Journal of Medicine*, 320, 479–484.

Fisher, B., Dignam, J., Bryant, J., DeCillis, A., Wickerham, D., Wolmark, N., Costantino, J., Redmond, C., Fisher, E., Bowman, D., Deschenes, L., Dimitrov, N., Margolese, R., Robidoux, A., Shibata, H., Terz, J., Paterson, A., Feldman, M., Farrar, W., Evans, J., and Lickley, H. (1996), "Five versus More Than Five Years of Tamoxifen Therapy for Breast Cancer Patients With Negative Lymph Nodes and Estrogen Receptor-Positive Tumors," *Journal of the National Cancer Institute*, 88, 1529–1542.

Fisher, B., Jeong, J., Bryant, J., Anderson, S., Dignam, J., Fisher, E., and Wolmark, N. (2004), "Treatment of Lymph Node-Negative, Oestrogen-Receptor-Positive Breast Cancer: Long-Term Findings From National Surgical Adjuvant Breast and Bowel Project Randomised Clinical Trials," *The Lancet*, 364, 858–868.

Gelman, A., Meng, X., and Stern, H. (1996), "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies" (with discussion), *Statistica Sinica*, 6, 733–807.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Gilks, W., Best, N., and Tan, K. (1995), "Adaptive Rejection Metropolis Sampling," *Applied Statistics*, 44, 455–472.

Gilks, W., and Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Applied Statistics*, 41, 337–348.

Glidden, D., and Vittinghoff, E. (2004), "Modelling Clustered Survival Data From Multicentre Clinical Trials," *Statistics in Medicine*, 23, 369–388.

Gray, R. (1994), "A Bayesian Analysis of Institutional Effects in a Multicenter Cancer Clinical Trial," *Biometrics*, 50, 244–253.

Hess, K., Pusztai, L., Buzdar, A., and Hortobagyi, G. (2003), "Estrogen Receptors and Distinct Patterns of Breast Cancer Relapse," *Breast Cancer Research and Treatment*, 78, 105–118.

Hunink, M., and Wong, J. (1994), "Meta-Analysis of Failure-Time Data With Adjustment for Covariates," *Medical Decision Making*, 14, 59–70.

Lagakos, S., and Schoenfeld, D. (1994), "Properties of Proportional-Hazards Score Tests Under Misspecified Regression Models," *Biometrics*, 40, 1037–1048.

Localio, A., Berlin, J., Have, T. T., and Kimmel, S. E. (2001), "Adjustments for Center in Multicenter Studies: An Overview," *Annals of Internal Medicine*, 135, 112–123.

Nowak, R., and Kolaczyk, E. (2000), "A Statistical Multiscale Framework for Poisson Inverse Problems," *IEEE Transactions on Information Theory*, 46, 1811–1825.

Parmar, M., Torri, V., and Stewart, L. (1998), "Extracting Summary Statistics to Perform Meta-Analyses of the Published Literature for Survival Endpoints," *Statistics in Medicine*, 17, 2815–2834.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 64, 583–616.

Walker, S., and Mallick, B. (1997), "Hierarchical Generalized Linear Models and Frailty Models With Bayesian Nonparameteric Mixing," *Journal of the Royal Statistical Society*, Ser. B, 59, 845–860.