A Bayesian analysis of doubly censored data using a hierarchical Cox model

Wei Zhang[1,2,*,†], Kathryn Chaloner[1,3], Mary Kathryn Cowles[1,3],
Ying Zhang[1], Jack T. Stapleton[4]

[1] *Department of Biostatistics, University of Iowa, Iowa City, IA*

[2] *Department of Biometrics and Data Management, Boehringer Ingelheim Pharmaceuticals, Ridgefield, CT*

[3] *Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA*

[4] *Department of Internal Medicine, University of Iowa and Iowa City VA Medical Center, Iowa City, IA*

SUMMARY

Two common statistical problems in pooling survival data from several studies are addressed. The first problem is that the data are doubly censored in that the origin is interval censored and the endpoint event may be right censored. Two approaches to incorporating the uncertainty of interval censored origins are developed, and then compared to more usual analyses using imputation of a single fixed

*Correspondence to: Wei Zhang, Department of Biometrics and Data Management, Boehringer Ingelheim Pharmaceuticals, Ridgefield, CT 06877, U.S.A.

†E-mail: wzhang@rdg.boehringer-ingelheim.com

Contract/grant sponsor: NIH/NIAID; contract/grant number: R01 058740

Contract/grant sponsor: National Security Agency; contract/grant number: H98230-04-1-0042

value for each origin. The second problem is that the data are collected from multiple studies and it is likely that heterogeneity exists among the study populations. A random-effects hierarchical Cox proportional hazards model is therefore used. The scientific problem motivating this work is a pooled survival analysis of data sets from three studies to examine the effect of GB virus type C (GBV-C) coinfection on survival of HIV-infected individuals. The time of HIV infection is the origin and for each subject this time is unknown, but is known to lie later than the last time at which the subject was known to be HIV negative, and earlier than the first time the subject was known to be HIV positive. The use of an approximate Bayesian approach using the partial likelihood as the likelihood is recommended because it more appropriately incorporates the uncertainty of interval censored HIV infection times.

KEY WORDS: GBV-C; human immunodeficiency virus; interval censoring; MCMC; Multicenter AIDS Cohort Study; partial likelihood

1. INTRODUCTION

Infection with GB virus type C (GBV-C) in humans is common, but no association between the virus and any known disease state has been demonstrated [1, 2, 3]. Individuals infected with human immunodeficiency virus (HIV) are commonly coinfected with GBV-C, since GBV-C shares the same modes of transmission as HIV. The prevalence of coinfection with GBV-C in HIV-infected individuals ranges from 14% to 43% [4]. Several recent studies of data from early in the epidemic, before the availablility of effective therapy, suggest that coinfection with GBV-C is associated with prolonged survival among HIV-infected people [5, 6, 7, 8]: other studies have concluded that there is no association [9, 10]. A meta-analysis of summary statistics was performed and published in Zhang et al. [11] and this study indicates that persistent GBV-C coinfection is associated with prolonged survival. To further investigate this conclusion, which

remains controversial [12], individual level data from separate studies is modeled here.

Original data sets from three published studies [6, 7, 8] are obtained. These data sets are doubly censored. First, the origin (HIV infection time) Y is interval censored in that it is known to lie in an interval $Y \in [L, U]$. Second, the endpoint (death) time is possibly right censored. Denote survival time $T = E - Y$, where E is the minimum of death time and last followup time. The dependence of T on covariates, and in particular on the indicator of GBV-C infection is of interest. In this paper, approaches are developed for the pooled survival analysis of doubly censored data from multiple studies and applied to the pooled data from the three studies.

A random effects model for the indicator of GBV-C coinfection incorporates the heterogeneity of patient characteristics between the studies. The most popular modeling method in survival analysis, the Cox proportional hazards model [14], avoids making any assumptions about the baseline hazard function $\lambda_0(t)$. Several authors have considered Cox survival models with random effects [15, 16, 17, 18, 19], but these random effects models for survival data either require assumptions regarding the form of the baseline hazard function or restrictions on the classes of models that can be fit. Sargent [20] and Gustafson [21] present a framework through which random effects can be introduced into the Cox model, which uses the Cox partial likelihood [22], and allows very general random-effect structures for the model parameters.

The remainder of this paper is organized as follows. Section 2 gives an introduction to the hierarchical Cox proportional hazards model. Section 3 presents different approaches to incorporating the interval censoring. In Section 4, these approaches are applied to the case study and results are summarized and compared. Section 5 concludes with discussion.

2. HIERARCHICAL COX PROPORTIONAL HAZARDS MODEL

The standard Cox regression represents the relationship between the covariates of interest and the hazard of event at time t through a proportional hazards model:

$$\lambda(t; x_i) \quad = \quad \lambda_0(t) \exp(x_i \beta)$$

where $\lambda(t; x_i)$ is the hazard function, x_i is a $1 \times p$ covariate vector for individual i, and β is a $p \times 1$ vector of coefficients corresponding to fixed effects.

2.1. The hierarchical model

Suppose there are q covariates to be modeled as random effects in addition to p covariates with fixed effects. For each of the q random covariates, there are r_l levels (e.g., trials, centers, studies), for $l = 1, \ldots, q$, with parameter vector $\gamma_l = (\gamma_{l1}, \ldots, \gamma_{lr_l})^T$ corresponding to the l^{th} random covariate. Let w_{il} be a scalar indicating the i^{th} subject's value of the l^{th} random effect covariate. Let $z_{il} = (z_{il1}, \ldots, z_{ilr_l})$, with $z_{ilj} = I_{ilj} w_{il}$, where $I_{ilj} = 1$ if subject i falls in the j^{th} level of the l^{th} random covariate, 0 otherwise. Define $\gamma = (\gamma_1^T, \ldots, \gamma_q^T)^T$ and $z_i = (z_{i1}, \ldots, z_{iq})$. Using this notation, the Cox model with random effects, a reparameterized frailty model [20], can be written as $\lambda(t; x_i, z_i) = \lambda_0(t) \exp(x_i \beta + z_i \gamma)$. Let $\mathbf{D}$ denote a right censored data set, $\mathbf{D} = \{(t_i, \delta_i, x_i, z_i) : i = 1, 2, \ldots, n\}$, where t_i is time from baseline to the minimum of endpoint time and censoring time, δ_i is the indicator for censoring with $\delta_i = 1$ if censored and $\delta_i = 0$ otherwise, and n is the number of observations. Let $\Re_{t_i}$ be the set of subjects at risk at time t_i. If there are no ties, the partial likelihood incorporating random effects is then given by

$$\mathcal{L}(\beta, \gamma | \mathbf{D}) \quad \propto \quad \prod_{i=1}^{n} \left[\frac{\exp(x_i \beta + z_i \gamma)}{\sum_{j \in \Re_{t_i}} \exp(x_j \beta + z_j \gamma)} \right]^{\delta_i}, \tag{1}$$

The partial likelihood in (1) serves as the first stage of the hierarchical model and can

be treated as a likelihood for computing a posterior density. Kalbfleisch [23] demonstrates that treating the partial likelihood as a likelihood leads to a limiting marginal posterior distribution of the regression parameters, assuming an independent increments gamma process prior distribution for the baseline cumulative hazard and independently a uniform distribution on the regression parameters. This result is shown with a different proof in Sinha et al. [24] which extends results to situations with time-dependent covariates, time-varying regression parameters and grouped survival data, and presents a Bayesian justification of a modified partial likelihood for handling ties. See also Chapter 4 of [25]. Chen at al. [26] carry out an in-depth theoretical investigation of Bayesian inference for the Cox regression model and discuss posterior propriety and computation based on Cox's partial likelihood. Sargent [20] and Gustafson [21] present methods for Bayesian analysis of multivariate survival data using (1).

The level-specific parameters γ_{lj} are modeled as draws from a distribution g_l with mean μ_l and variance ν_l. Let g denote the joint density for γ, and assume γ_{lj}'s are independent of each other given μ_l and ν_l, for $l = 1, \ldots, q$ and $j = 1, \ldots, r_l$, then

$$g(\gamma|\mu,\nu) = \prod_{l=1}^{q}\prod_{j=1}^{r_l} g_l(\gamma_{lj}|\mu_l,\nu_l), \tag{2}$$

where $\mu = (\mu_1, \ldots, \mu_q)^T$, and $\nu = (\nu_1, \ldots, \nu_q)^T$.

For the final stage of the hierarchical model, prior distributions need to be specified. A proper prior distribution for the variance component is typically essential for proper posterior distribution and computational stability. Let $f(\mu, \nu|\omega)$ represent this prior distribution, where ω is the vector of hyperparameters and is taken to be known. Also let β have, independent of γ, a uniform prior distribution.

An approximate posterior distribution for the model parameters is then be assumed to be:

$$\pi(\beta, \gamma, \mu, \nu | \mathbf{D}, \omega) \quad \propto \quad \mathcal{L}(\beta, \gamma | \mathbf{D}) \; g(\gamma | \mu, \nu) \; f(\mu, \nu | \omega), \tag{3}$$

where $\mathbf{D}$ is right censored data.

2.2. Estimation of parameters using MCMC methods

The approximate posterior distribution (3) can be estimated with Markov chain Monte Carlo (MCMC) methods. The Metropolis-Hastings algorithm [27, 28] is a general term for a family of MCMC methods that are useful for drawing samples from Bayesian posterior distributions. Let θ denote the set of parameters involved in the hierarchical model. The parameter vector θ is divided into components corresponding to the hierarchy and the single-component Metropolis-Hastings algorithm is used [29].

3. ANALYSIS OF DOUBLY CENSORED DATA

Denote doubly censored data $\mathbf{C} = \{([l_i, u_i], e_i, \delta_i, x_i, z_i) : i = 1, 2, , \ldots, n\}$, where $[l_i, u_i]$ is the interval of dates within which origin y_i falls, and e_i is the minimum of endpoint date and last followup date. Note that dates are defined as length of time from a fixed time point. Further define a function $\mathbf{D}(\cdot, \cdot)$ mapping a doubly censored data set $\mathbf{C}$ and a set of origin dates $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ to a right censored data set; specifically $\mathbf{D}(\mathbf{C}, \mathbf{y}) = \{(t_i, \delta_i, x_i, z_i) : i = 1, 2, \ldots, n\}$, where $t_i = e_i - y_i$ is time from y_i to e_i. A common approach in medical applications [8, 30, 31] is that midpoints of censoring intervals are used to impute interval-censored origins, $\hat{y}_i = (l_i + u_i)/2$, and are then used to compute $\hat{t}_i = e_i - \hat{y}_i$ in analysis as if they are right censored data. In this case, $\mathbf{D}(\mathbf{C}, \hat{\mathbf{y}}) = \{(\hat{t}_i, \delta_i, x_i, z_i) : i = 1, 2, \ldots, n\}$, where $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)^T$. Law and Brookmeyer [32] demonstrate that in HIV studies the Kaplan-

Meier estimate based on this method is notably biased when origin intervals are longer than two years.

Three alternative methods are proposed below which can be implemented using MCMC. If the partial likelihood is used as a likelihood in estimating the posterior distribution using MCMC, the assumption has to be made that this is valid under interval censoring of the origin: an assumption which has not been proved, but which seems reasonable given the results in [23, 24, 26].

3.1. MCMC for imputed data (MCMCid) approach

As an improvement to using the midpoint of censoring interval, MCMCid is proposed here that samples a value of each interval censored origin generated from an estimated distribution of origins. Let G denote the distribution function of origins $\mathbf{y}$. The estimate of G, $\hat{G}$, can be either obtained parametrically using the maximum likelihood estimation based on a known distribution (e.g., Weibull, log Normal), or obtained nonparametrically using Turnbull's self-consistency algorithm [33]. To perform an analysis for doubly censored survival data using MCMCid, for each subject i, $i = 1, \ldots, n$, a value of y_i, denoted $\hat{y}_i$, is randomly sampled from $\hat{G}$, conditional on the interval $[l_i, u_i]$ within which y_i falls. The doubly censored data set $\mathbf{C}$ and imputed origins $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)^T$ are then mapped to a right censored data set $\mathbf{D}(\mathbf{C}, \hat{\mathbf{y}}) = \{(\hat{t}_i, \delta_i, x_i, z_i) : i = 1, 2, \ldots, n\}$, where $\hat{t}_i = e_i - \hat{y}_i$. The hierarchical Cox proportional hazards model can then be fit to the right censored data $\mathbf{D}(\mathbf{C}, \hat{\mathbf{y}})$ using MCMC methods.

The MCMCid approach is straightforward to implement and understand but underestimates the variability of parameter estimates because the uncertainty of imputed origins $\hat{\mathbf{y}}$ is not

incorporated.

3.2. Imputation-embedded MCMC (ieMCMC) approach

The imputation-embedded MCMC (ieMCMC) approach is developed as an alternative to the MCMCid approach. For each MCMC iteration step m, $m = 1, \ldots, M$ (e.g., $M = 20000$), origins $\mathbf{y}^m = (y_1^m, \ldots, y_n^m)^T$ are randomly sampled based on their distribution $\hat{G}$, conditional on the intervals $\{[l_i, u_i] : i = 1, 2, \ldots, n\}$ within which they fall. A right censored data set, $\mathbf{D}(\mathbf{C}, \mathbf{y}^m) = \{(t_i^m, \delta_i, x_i, z_i) : i = 1, 2, \ldots, n\}$, where $t_i^m = e_i - y_i^m$, is generated at each iteration step m. In the ieMCMC approach the origins $\mathbf{y}^m$ vary at each MCMC sampler iteration with a distribution based on $\hat{G}$, but the estimate $\hat{G}$ is still fixed. The uncertainty in estimating G by $\hat{G}$ is not therefore taken into account, although the uncertainty in $\mathbf{y}^m$ conditional on $\hat{G}$ is considered.

3.3. Bayesian partial likelihood (Bayesian-PL) approach

Finally, an approach is proposed with a parametric assumption on origins which is more complete. Denote $h(\mathbf{y}|\xi)$ as the probability density function of the origins, with parameter vector ξ. Let $p(\xi|\varpi)$ represent the prior distribution for ξ, where ϖ is the vector of hyperparameters governing $p(\cdot)$. For a doubly censored data set $\mathbf{C} = \{([l_i, u_i], e_i, \delta_i, x_{i1}, z_i) : i = 1, 2, , \ldots, n\}$, define I_{y_i} to be the indicator function which is equal to 1 if y_i is in $[l_i, u_i]$ and 0 otherwise. The approximate posterior distribution, based on using the partial likelihood as the likelihood, for all the model parameters, which include the interval censored unknown

origins $\mathbf{y}$, can be expressed as

$$\pi(\beta,\gamma,\mu,\nu,\xi,\mathbf{y}|\mathbf{C},\omega,\varpi) \quad \propto \quad \mathcal{L}(\beta,\gamma|\mathbf{D}(\mathbf{C},\mathbf{y}))\ g(\gamma|\mu,\nu) \\ \{\prod_i \frac{h(y_i|\xi)\ I_{y_i}}{[H(u_i|\xi)-H(l_i|\xi)]}\}\ f(\mu,\nu|\omega)\ p(\xi|\varpi), \tag{4}$$

where $H(\cdot|\xi)$ is cumulative density function of $\mathbf{y}$. At each iteration of the MCMC sampler, the origins $\mathbf{y}$ are drawn from their full conditional distribution, given all other model quantities. The resulting analysis provides an updated estimate of the distribution of origins, as well as correctly capturing the effect of the uncertainty in origins on estimation of the parameters of primary interest. Because (4) is an approximation of the posterior distribution, in what follows we refer to this as the Bayesian-PL method.

4. CASE STUDY

The pooled dataset consists of doubly censored data sets from three studies [6, 7, 8] where GBV-C is measured late in HIV disease. Each study corresponds to a different population. The "late" data set of the Williams study [8] is used, from the Multicenter AIDS Cohort Study (MACS) which has a documented HIV seroconversion window of approximately 6 months on average. In the MACS study, the intervals are based on the retrospective testing of stored blood samples obtained on a regular basis, including those before and after HIV seroconversion. In the other two studies, testing is also on stored blood samples, and the date of subjects first known positive HIV test is used as the right limit of the interval, and January 1st 1978 (or date of birth for subjects born after January 1st 1978) is treated as the left limit of the interval. January 1st 1978 was chosen because an analysis of stored blood samples from a study in San Francisco indicates an extremely low prevalence of HIV infection before this date [34]. The

sample sizes of [6, 7, 8] are 362, 197 and 138, respectively. A summary of these three studies can be found in [11]. All studies follow subjects through the time period before the advent of highly effective therapy for HIV in 1996.

Fitting the regular Cox model to three imputed data sets separately, the estimated log hazard ratio of GBV-C coinfection, controlling for baseline log(CD4+ count in cells/mL) and age at HIV infection, is -1.23, -1.62 and -0.97 for [6, 7, 8], respectively. For both Xiang and Tillmann studies [6, 7], HIV infection time is heavily interval censored, with a mean interval width of about 10 years. In contrast, HIV infection time in the Williams study [8] has much narrower intervals. The subjects in the Tillmann study are from Germany, and the subjects in the other two studies are from the U.S.A. The differences among the studies could be due to several reasons, including the fact that GBV-C testing is not standardized and each study used a different primer for a qualitative test. A recent study [35] indicates that the sensitivity and specificity of each test varies and the sensitivity of one particular test depends on GBV-C RNA levels. This motivates the need for an analysis that has the ability to account for the possibly differing effect of GBV-C infection within each population. The primary endpoint for this pooled analysis is overall survival. All 695 eligible patients are included in the analysis.

Let x_{i1}, x_{i2}, and w_i denote log(CD4+ count in cells/mL), age at HIV infection and GBV-C coinfection status for subject i, respectively. Let $I_{ij} = 1$ if subject i is from study j, 0 otherwise, with $j = 1, 2, 3$ corresponding to the three studies [6, 7, 8]. Define $z_i = (z_{i1}, z_{i2}, z_{i3})$, where $z_{ij} = I_{ij} w_i$. Let $\beta = (\beta_1, \beta_2)^T$ denote the fixed effects of covariates x_{i1} and x_{i2}. Let $\gamma = (\gamma_1, \gamma_2, \gamma_3)^T$ denote the random effects of covariates $z_i = (z_{i1}, z_{i2}, z_{i3})$, with constraint $\gamma_j \sim N(\mu, \sigma^2)$, where μ is the population effect of GBV-C infection and σ^2 population variance.

Specifically, the hazard function for individual i at time t_k is given by

$$\lambda(t_k; x_{i1}, x_{i2}, z_i) \quad = \quad \lambda_0(t_k) \exp(x_{i1}\beta_1 + x_{i2}\beta_2 + z_i\gamma). \tag{5}$$

Note that age at HIV infection x_{i2} is a known deterministic function of infection time y_i. Fixed effects, rather than random effects, are used for the age and CD4+ cell count as the effects from each data set of the three studies are very similar. This is unlike the effect of GBV-C in each study which varies more.

In what follows approximate posterior density, mean and corresponding approximate highest posterior density region (HDR) are calculated using the approximation based on the partial likelihood. These will be referred to as posterior density, mean and HDR for simplicity, without explicitly qualifying that they are approximations.

4.1. Prior distributions

Other than the three studies for which we have data, there are four additional studies [5, 9, 10, 38] providing only summary statistics (hazard ratio and corresponding 95% confidence interval). A meta-analysis of summary statistics for these four studies was done, similarly to the meta-analysis of all summary statistics in [11]. The estimated combined effect of GBV-C in these four studies is -0.41 with estimated standard error 0.42. This result helps postulate the prior distribution for μ. To be conservative, the standard error is multiplied by 2, so that μ is normally distributed as $N[-0.41, (0.42 \times 2)^2]$.

A proper prior distribution for σ^2 is used for the sake of computational stability of the MCMC methods and to generate a proper approximate posterior distribution. A gamma distribution on $\tau = \sigma^{-2}$ is used with τ distributed as $\Gamma(0.25, 0.005)$. This distribution is specified by considering what values for the random effects are reasonable. For example, a

belief representing moderate heterogeneity between the studies would be that γ_j $(j = 1, 2, 3)$ vary around μ by ± 0.10. Using 0.10 as an estimate of the standard error σ leads to a prior estimate of $\tau = 100$. A prior belief that presents substantial heterogeneity might be that the γ_j vary around μ by ± 1.0. Using 1.0 as the prior standard error of γ_j leads to a prior estimate of $\tau = 1$. The prior distribution $\Gamma(0.25, 0.005)$, with mean 50 and standard error of 100, gives reasonable weight to these extremes. A flat prior distribution is used for β.

4.2. Joint posterior distributions

A parametric model for the distribution of $\mathbf{y}$ is implemented in the example instead of the non-parametric method. Reasons for this choice include that the non-parametric maximum likelihood estimator $\hat{G}$ given by Turnbull in [33] is only unique up to an equivalence class and also has discrete components. The parametric model gives a smoother distribution for the times of infection, which is thought to more realistically model the reality of the spread of HIV infection. In addition, for the pooled data set, Turnbull's estimate of the survival function is not very different from the maximum likelihood estimate based on assuming a Weibull distribution (Figure 1).

4.2.1. MCMCid and ieMCMC approaches The distribution function G of HIV infection times $\mathbf{y}$ is assumed to be a $Weibull(\alpha, \gamma)$. Conditional on the maximum likelihood estimates ($\hat{\alpha}$, $\hat{\lambda}$) and intervals for infection times, infection times $\hat{\mathbf{y}}$ are randomly sampled. Let $\mathbf{C}$ denote the doubly censored data $\mathbf{C} = \{([l_i, u_i], e_i, \delta_i, x_{i1}, z_i) : i = 1, 2, , \ldots, n\}$. Following the prior distributions in Section 4.1, the joint approximate posterior density of all parameters for the

MCMCid approach is given by

$$\pi(\beta,\gamma,\mu,\tau|\mathbf{D}(\mathbf{C},\hat{\mathbf{y}})) \propto \mathcal{L}(\beta,\gamma|\mathbf{D}(\mathbf{C},\hat{\mathbf{y}}))\prod_{j=1}^{3}\tau^{\frac{1}{2}}\exp(-\frac{(\gamma_j-\mu)^2\tau}{2}) \exp(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2})\,\tau^{a_0-1}\exp(-\frac{\tau}{b_0}), \quad (6)$$

where $\mathbf{D}(\mathbf{C},\hat{\mathbf{y}}) = \{(\hat{t}_i,\delta_i,x_{i1},\hat{x}_{i2},z_i) : i = 1,2,\ldots,n\}$, $\hat{t}_i = e_i - \hat{y}_i$, $\hat{x}_{i2}$ is a function of $\hat{y}_i$, and $(a_0,b_0,\mu_0,\sigma_0) = (0.25, 0.005, -0.41, 0.42\times 2)$.

The joint posterior density of all parameters for the ieMCMC approach is the same as (6), except that $\mathbf{D}(\mathbf{C},\hat{\mathbf{y}})$ is replaced by $\mathbf{D}(\mathbf{C},\mathbf{y}^m)$, where $\mathbf{D}(\mathbf{C},\mathbf{y}^m) = \{(t_i^m,\delta_i,x_{i1},x_{i2}^m,z_i) : i = 1,2,\ldots,n\}$, $t_i^m = e_i - y_i^m$, and x_{i2}^m is a function of y_i^m. In the ieMCMC approach, $\mathbf{D}(\mathbf{C},\mathbf{y}^m)$ changes at each MCMC iteration m, while $\mathbf{D}(\mathbf{C},\hat{\mathbf{y}})$ is fixed during the process of the MCMCid approach.

For both MCMCid and ieMCMC approaches, given data and other parameters in the model, the full conditional posterior distribution for τ has a gamma distribution $\pi(\tau|\cdot) \propto \Gamma[a_0+\frac{3}{2}, b_0+\sum_{j=1}^{3}\frac{(\gamma_j-\mu)^2}{2})]$, and the full conditional posterior distribution for μ has a normal distribution $\pi(\mu|\cdot) \propto N[\lambda\mu_0+(1-\lambda)\bar{\gamma},(1-\lambda)(3\tau)^{-1}]$, where $\lambda = \frac{(3\tau)^{-1}}{(3\tau)^{-1}+\sigma_0^2}$ and $\bar{\gamma} = \sum_{j=1}^{3}\gamma_j/3$.

It is straightforward to perform the Gibbs sampling on μ and τ, but there is no direct way to draw from parameters β and γ, and the single-component Metropolis-Hastings algorithm [29] for the sampling of these two parameter components is used.

4.2.2. Bayesian-PL approach The distribution of $\mathbf{y}$ is also assumed to be a $Weibull(\alpha,\lambda)$ for the purpose of a fair comparison to the MCMCid and ieMCMC approaches, and the prior distributions for α and λ are specified independently as log normal distributions: $LogNorm(\mu_\alpha,\sigma_\alpha^2)$ and $LogNorm(\mu_\lambda,\sigma_\lambda^2)$, respectively. For $y_i \in [l_i,u_i]$, $i = 1,\ldots,n$, the joint approximate posterior density of all parameters for the Bayesian-PL approach is then given

by

$$\begin{aligned}\pi(\beta,\gamma,\mu,\tau,\alpha,\lambda,\mathbf{y}|\mathbf{C}) \quad \propto \quad & \mathcal{L}(\beta,\gamma|\mathbf{D}(\mathbf{C},\mathbf{y}))\prod_{j=1}^{3}\tau^{\frac{1}{2}}\exp[-\frac{(\gamma_j-\mu)^2\tau}{2}] \\ & \exp[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}]\ \tau^{a_0-1}\exp(-\frac{\tau}{b_0}) \\ & \prod_{i=1}^{n}\{\alpha\lambda^{-\alpha}y_i^{\alpha-1}\exp[-(\frac{y_i}{\lambda})^{\alpha}]\}\{\exp[-(\frac{l_i}{\lambda})^{\alpha}]-\exp[-(\frac{u_i}{\lambda})^{\alpha}]\}^{-1} \\ & \alpha^{-1}\exp[-\frac{(\log(\alpha)-\mu_\alpha)^2}{2\sigma_\alpha^2}]\ \ \lambda^{-1}\exp[-\frac{(\log(\lambda)-\mu_\lambda)^2}{2\sigma_\lambda^2}], \qquad (7)\end{aligned}$$

where $(\mu_\alpha,\sigma_\alpha,\mu_\lambda,\sigma_\lambda)=(1.31,0.4,3.51,0.5)$.

4.3. Results of primary analysis

All methods are implemented in R [39], using the Metropolis-within-Gibbs algorithm. The posterior full conditional distributions of μ and τ are normal and gamma respectively, so these parameters were drawn using Gibbs sampling. The approximate posterior full conditional distributions for parameters using the single-component Metropolis-Hastings algorithm [29] in the Bayesian-PL approach are given in the Appendix. Code is available from the first author. The WinBUGS software package [40] could not be used easily because of the doubly censored data complicated by an interval censored covariate. Three independent chains are generated for each of the 3 approaches (MCMCid, ieMCMC and Bayesian-PL approach). Each chain consists of 14,000 iterations after a series of 6,000 burn-in iterations. The Brooks and Gelman convergence diagnostic [41] indicates that there is no evidence against the convergence of sampler for each parameter in all approaches.

Table I summarizes results from the MCMCid, ieMCMC and Bayesian-PL approaches based on the hierarchical Cox proportional hazards model. The point estimates from the three approaches are similar for each parameter except for σ. For each parameter estimate,

the standard error from the Bayesian-PL approach is, appropriately, the largest among three approaches, while the standard error from the MCMCid approach is the smallest one. Consequently, the 95% HDR from the Bayesian-PL approach is generally wider than the one from the MCMCid or ieMCMC approach; the 95% HDR from the MCMCid approach tends to be the narrowest. The differences are substantial, especially for the parameters of most interest: μ and σ.

Results from all three approaches indicate that GBV-C infection is associated with prolonged survival. From the Bayesian-PL approach, the estimated hazard ratio for GBV-C viremia is $e^{-\hat{\mu}} = e^{-0.891} = 0.41$ with 95% probability falling into the interval $(e^{-1.423}, e^{-0.335}) = (0.24, 0.72)$ after adjusting for baseline log(CD4+ count in cells/mL) and age at HIV infection. Baseline log(CD4+ count in cells/mL) is also associated with prolonged survival: estimated hazard ratio $e^{\hat{\beta}_1} = e^{-0.645} = 0.52$, with 95% probability falling into the interval $(e^{-0.842}, e^{-0.444}) = (0.43, 0.64)$.

4.4. Results of sensitivity analysis

To examine the behavior of the estimate of μ from the Bayesian-PL approach when the distribution for infection times $\mathbf{y}$ is modified, the hierarchical Cox proportional hazards model was fit using different distributions for $\mathbf{y}$. Specifically, we used a flat distribution for $\mathbf{y}$, a single $Weibull(\alpha, \lambda)$ for $\mathbf{y}$, and a set of three Weibull distributions, one for each of the three studies, $Weibull(\alpha_j, \lambda_j)$ for $\mathbf{y}_j$, $j = 1, 2, 3$. The choice of $\Gamma(0.25, 0.005)$ as the prior for τ is also examined by using $\Gamma(0.001, 0.001)$. Overall, there are $3 \times 2 = 6$ scenarios in the sensitivity analyses, with the first one corresponding to the primary analysis (see Table II and Figure 2). The value of $\hat{\mu}$, the estimate for the logarithm of hazard ratio of GBV-C infection, changes

only slightly, as does the standard error and 95% HDR. This analysis suggests that the results are insensitive to the choice of prior distribution.

5. DISCUSSION

The imputation-embedded MCMC (ieMCMC) and the Bayesian-PL approaches are developed to deal with doubly censored survival data and compared to an MCMC analysis of a right censored data set constructed by imputing a single value for each interval censored origin (MCMCid). The MCMCid approach considerably underestimates the variability of the estimates. The ieMCMC allows for some uncertainty of imputed origins to play a role but again results in underestimation of the variability of the parameter estimates. In comparison the Bayesian-PL approach treats unobservable origins $\mathbf{y}$ as unknown quantities with a parametric distribution G. Prior distributions are then assigned for the hyperparameters of G. Interval censoring is treated by data augmentation [42] with $\hat{\mathbf{y}}$ drawn from their posterior predictive distribution. The results from the Bayesian-PL approach more appropriately reflect uncertainty than the MCMCid and ieMCMC approaches. This paper demonstrates the ability of the Bayesian-PL approach to incorporate the uncertainty of imputed origins in doubly censored survival data. Our sensitivity study shows the results from the Bayesian-PL approach are reasonably insensitive to the specification of the parametric form of G (Figure 2).

Härkänen *et al.* [43] presented a non-parametric Bayesian intensity model for doubly censored data in the fully Bayesian framework, which treats unobservable origins $\mathbf{y}$ as unknown quantities with piece-wise constant hazard functions. Hazard functions are assigned gamma prior distributions. Komárek *et al.* [44] applied a modified version of this approach to doubly censored dental data to examine the effect of fluoride-intake on the time to caries development

in children. However, these approaches are complex to implement and computationally demanding.

It should be noted that a formal justification of using the Bayesian-PL approach has not been provided for interval censored origins and random effects. Given the results in [23, 24, 26] this assumption is not unreasonable, but additional work needs to be done.

A parametric model could have been used for the case study and an estimate of the survival curve obtained. The primary interest in this data analysis is however the question of whether or not co-infection with GBV-C is associated with prolonged survival of individuals infected with HIV disease. All three studies, and the other studies in the meta-analysis [11] use data before the advent of highly effective therapy for HIV infection, and so the survival curve itself is not of current interest.

The Bayesian hierarchical Cox model has accommodated random study-specific effects and therefore incorporated between-study heterogeneity. Through the specification of prior distributions, the prior information relevant to the parameters of interest has been taken into account.

The methods for doubly censored survival data developed in this paper have enabled the analysis of these data sets and lead to the conclusion that the hazard ratio with GBV-C infection is approximately 40% of the hazard without GBV-C infection, and the hypothesis of no difference in hazard can be ruled out with high probability. The pooled analysis of the individual subject data therefore augments and supports the meta-analysis result of the summary statistics previously reported in [11]. Biological plausibility for a beneficial mechanism and *in vitro* evidence in inhibiting HIV replication are provided in [6, 45, 46, 47, 48, 49]. However, as in all observational data, the results of our analyses do

not provide evidence that GBV-C is causally related to improved survival, and it is possible that GBV-C infection is not the reason that HIV-positive individuals coinfected with GBV-C live longer but, rather, that it serves as a biological marker of a different factor related to HIV disease progression. This warrants further investigation and is a subject of debate in the scientific literature [9, 12, 13, 49, 50].

APPENDIX

The posterior full conditional distributions of most of the parameters in the Bayesian-PL model were of nonstandard forms and were sampled using Metropolis or Metropolis-Hastings updates. This appendix lists these full conditional distributions and any unusual features of the sampling algorithms used.

A.1 Full conditional distribution of $\mathbf{y}$

The unnormalized full conditional distribution of $\mathbf{y}$ is

$$\pi(\mathbf{y}|\beta,\gamma,\mu,\tau,\alpha,\lambda,\mathbf{C}) \quad \propto \quad \mathcal{L}(\beta,\gamma|\mathbf{D}(\mathbf{C},\mathbf{y}))(\prod_{i=1}^{n} y_i)^{\alpha-1}\exp[-\sum_{i=1}^{n}(\frac{y_i}{\lambda})^{\alpha}],$$

where $\mathbf{C} = \{([l_i,u_i],e_i,\delta_i,x_i,z_i) : i = 1,2,,\ldots,n\}$. At each iteration m, candidates y_i^{new}, $i = 1,\ldots,n$, are generated independently from truncated normal densities:

$$y_i^{new}|y_i^{m-1} \quad \sim \quad TN(y_i^{m-1},\sigma_y^2|l_i,u_i),$$

where $[l_i,u_i]$ is the interval within which subject i's infection time y_i is known to lie. Then the entire vector $\mathbf{y}^{new} = (y_1^{new},\ldots,y_n^{new})$ is accepted based on the acceptance probability

$$R = min\{1, \frac{\mathcal{L}(\beta^{m-1},\gamma^{m-1}|\mathbf{D}(\mathbf{C},\mathbf{y}^{new}))}{\mathcal{L}(\beta^{m-1},\gamma^{m-1}|\mathbf{D}(\mathbf{C},\mathbf{y}^{m-1}))} \times (\frac{\prod_{i=1}^{n} y_i^{new}}{\prod_{i=1}^{n} y_i^{m-1}})^{(\alpha^{m-1}-1)}$$
$$\times \exp[\sum_{i=1}^{n}(\frac{y_i^{m-1}}{\lambda^{m-1}})^{\alpha^{m-1}} - \sum_{i=1}^{n}(\frac{y_i^{new}}{\lambda^{m-1}})^{\alpha^{m-1}}] \times \prod_{i=1}^{n} \frac{\Phi(\frac{u_i-y_i^{new}}{\sigma_y}) - \Phi(\frac{l_i-y_i^{new}}{\sigma_y})}{\Phi(\frac{u_i-y_i^{m-1}}{\sigma_y}) - \Phi(\frac{l_i-y_i^{m-1}}{\sigma_y})}\},$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Note that a new set of infection times $\mathbf{y}$ results in a new set of calculated failure and censoring times. This in turn produces new values of distinct survival times $\{t_k : k = 1,\ldots,K\}$, so that different individuals appear in the risk set $\Re_k$ involved in the partial likelihood calculations in (2).

A.2 Full conditional distributions of β, γ, α, and λ

The unnormalized full conditional distributions for β, γ, α, and λ are given below. Each of these parameters is sampled using a random-walk Metropolis update in which a candidate value is drawn from a normal or multivariate normal density centered at the value from the previous iteration.

- $\pi(\beta|\gamma,\mu,\tau,\alpha,\lambda,\mathbf{C},\mathbf{y}) \propto \mathcal{L}(\beta,\gamma|\mathbf{D}(\mathbf{C},\mathbf{y}))$
- $\pi(\gamma|\beta,\mu,\tau,\alpha,\lambda,\mathbf{C},\mathbf{y}) \propto \mathcal{L}(\beta,\gamma|\mathbf{D}(\mathbf{C},\mathbf{y})) \prod_{j=1}^{3} \exp[-\frac{(\gamma_j-\mu)^2\tau}{2}]$
- $\pi(\alpha|\beta,\gamma,\mu,\tau,\lambda,\mathbf{C},\mathbf{y}) \propto \alpha^{n-1}\lambda^{-n\alpha}(\prod_{i=1}^{n} y_i)^{\alpha-1} \exp[-\sum_{i=1}^{n}(\frac{y_i}{\lambda})^{\alpha}] \exp[-\frac{(\log(\alpha)-\mu_\alpha)^2}{2\sigma_\alpha^2}]$
- $\pi(\lambda|\beta,\gamma,\mu,\tau,\alpha,\mathbf{C},\mathbf{y}) \propto \lambda^{-n\alpha-1} \exp[-\sum_{i=1}^{n}(\frac{y_i}{\lambda})^{\alpha}] \exp[-\frac{(\log(\alpha)-\mu_\alpha)^2}{2\sigma_\alpha^2}]$.

ACKNOWLEDGEMENTS

The authors also wish to thank Dr Hans Tillmann from the University of Leipzig, Germany, and the Multicenter AIDS Cohort Study (MACS) for providing data. The MACS has centers located at: The Johns Hopkins Bloomberg School of Public Health (Joseph Margolick); Howard Brown Health Center and Northwestern University Medical School (John Phair); University of California, Los Angeles (Roger Detels); University of Pittsburgh (Charles Rinaldo); and Data Analysis Center (Lisa Jacobson). This research was supported by NIH/NIAID (R01 058740) and National Security Agency (H98230-04-1-0042).

Table I. Comparison of results from different approaches for pooled analysis using the hierarchical Cox proportional hazards model.

	MCMCid*		ieMCMC*		Bayesian-PL**	
Parameter	Mean(SE)	95% HDR	Mean(SE)	95% HDR	Mean(SE)	95% HDR
μ	-0.797(0.174)	(-1.136, -0.475)	-0.846(0.208)	(-1.248, -0.436)	-0.891(0.277)	(-1.423, -0.335)
σ	0.327(0.413)	(0.027, 1.016)	0.425(0.519)	(0.029, 1.292)	0.727(0.783)	(0.028, 1.947)
γ_1	-0.779(0.145)	(-1.067, -0.494)	-0.845(0.158)	(-1.162, -0.544)	-0.998(0.185)	(-1.360, -0.623)
γ_2	-1.073(0.359)	(-1.812, -0.486)	-1.227(0.416)	(-2.123, -0.596)	-1.585(0.476)	(-2.497, -0.775)
γ_3	-0.636(0.274)	(-1.111, -0.039)	-0.623(0.336)	(-1.190, 0.109)	-0.463(0.412)	(-1.189, 0.339)
β_1	-0.538(0.086)	(-0.707, -0.371)	-0.569(0.098)	(-0.756, -0.377)	-0.645(0.101)	(-0.842, -0.444)
β_2	0.013(0.006)	(0.001, 0.027)	0.017(0.007)	(0.002, 0.030)	0.008(0.007)	(-0.006, 0.021)

* Random imputation for infection times $\mathbf{y}$ based on the estimated $Weibull(\hat{\alpha}, \hat{\lambda})$.

** Prior distribution for infection times $\mathbf{y}$: $Weibull(\alpha, \lambda)$, where $\alpha \sim LogNorm(1.31, 0.4)$ and $\lambda \sim LogNorm(3.51, 0.5)$.

Table II. Sensitivity analysis for the Bayesian-PL approach using different prior distributions for τ and $\mathbf{y}$.

Analysis	τ	Infection times $\mathbf{y}$	$\hat{\mu}$ (SE)	95% HDR
Tau1Y1	$\Gamma(0.25, 0.005)$	$Weibull(\alpha, \lambda)$	-0.891(0.277)	(-1.423, -0.335)
Tau1Y2	$\Gamma(0.25, 0.005)$	$Weibull(\alpha_j, \lambda_j), j = 1, 2, 3$	-0.866(0.288)	(-1.398, -0.263)
Tau1Y3	$\Gamma(0.25, 0.005)$	flat prior	-0.924(0.272)	(-1.417, -0.350)
Tau2Y1	$\Gamma(0.001, 0.001)$	$Weibull(\alpha, \lambda)$	-0.876(0.271)	(-1.385, -0.326)
Tau2Y2	$\Gamma(0.001, 0.001)$	$Weibull(\alpha_j, \lambda_j), j = 1, 2, 3$	-0.878(0.283)	(-1.392, -0.279)
Tau2Y3	$\Gamma(0.001, 0.001)$	flat prior	-0.911(0.250)	(-1.360, -0.412)

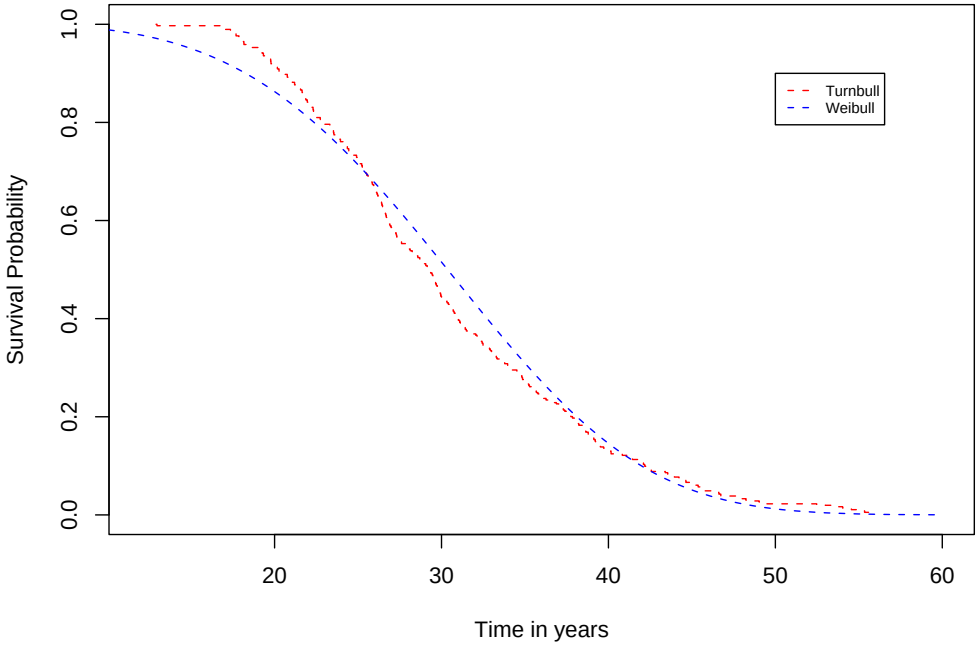

Figure 1. Estimated survival curves for HIV infection time of the pooled data set based on different distribution assumptions.

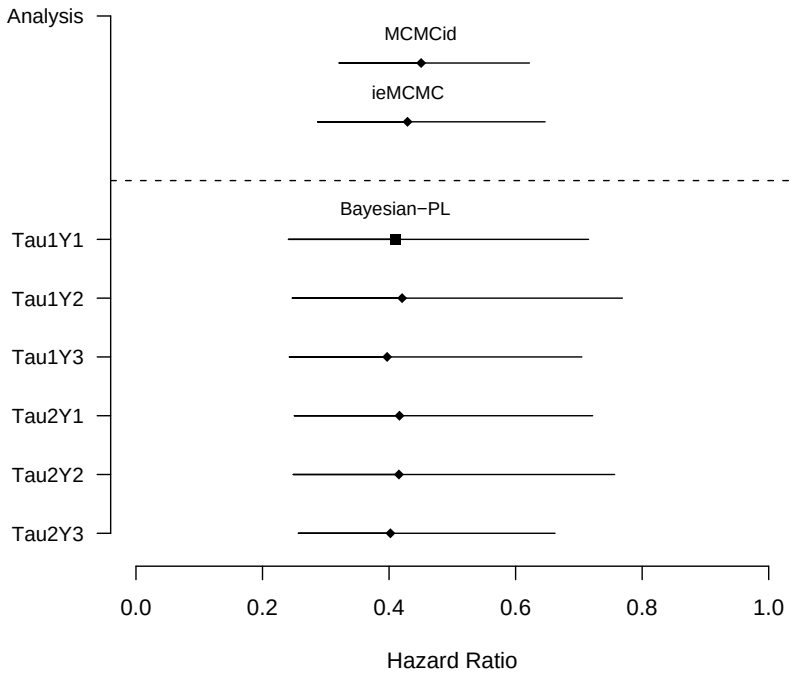

Figure 2. Estimated hazard ratio and 95% HDR from the MCMCid and ieMCMC approaches (above the dashed line), and from sensitivity analysis with different prior distributions for τ and HIV infection times $\mathbf{y}$ using the Bayesian-PL approach (below the dashed line). The primary analysis is Tau1Y1.

REFERENCES

1. Alter HJ. The cloning and clinical implications of HGV and HGBV-C. *New England Journal of Medicine* 1996; **334**:1536-1537.
2. Rambusch EG, Wedemeyer H, Tillmann HL, Heringlake S, Manns MP. Significance of coinfection with hepatitis G virus for chronic hepatitis C–a review of the literature. *Z Gastroenterol* 1998; **36**:41-53.
3. Tillmann HL, Heringlake S, Trautwein C, *et al.* Antibodies against the GB virus C envelope 2 protein before liver transplantation protect against GB virus C de novo infection. *Hepatology* 1998; **28**:379-384.
4. Stapleton JT. GB virus type C/hepatitis G virus. *Semin Liver Disease* 2003; **23**:137-148.
5. Lefrère JJ, Roudot-Thraval F, Morand-Joubert L, *et al.* Carriage of GB virus C/Hepatitis G virus RNA is associated with a slower immunologic, virologic, and clinical progression of human immunodeficiency virus disease in coinfected persons. *Journal of Infectious Diseases* 1999; **179**:783-789.
6. Xiang J, Wünschmann S, Diekema DJ, *et al.* Effect of coinfection with GB virus C on survival among patients with HIV infection. *New England Journal of Medicine* 2001; **345**:707-714.
7. Tillmann HL, Heiken H, Knapir-Botor A, *et al.* Infection with GB virus C and reduced mortality among HIV-infected patients. *New England Journal of Medicine* 2001; **345**:715-724.
8. Williams CF, Klinzman D, Yamashita TE, *et al.* Persistent GB virus C infection and survival in HIV-infected men. *New England Journal of Medicine* 2004; **350**:981-990.
9. Björkman P, Flamholc L, Nauclér A, *et al.* GB virus C during the natural course of HIV-1 infection: viremia at diagnosis does not predict mortality. *AIDS* 2004; **18**:877-886.
10. Birk M, Lindback S, Lidman C. No influence of GB virus C replication on the prognosis in a cohort of HIV-1-infected patients. *AIDS* 2002; **16**:2482-2485.
11. Zhang W, Chaloner K, Tillmann HL, Williams CF, Stapleton JT. Effect of early and late GBV-C viremia on survival of HIV infected individuals: a meta-analysis. *HIV Medicine* 2006; **7**:173-180.
12. Stapleton JT, Chaloner K, Williams CF. GB virus C infection and survival in the Amsterdam Cohort Study. *Journal of Infectious Diseases* 2005; **191**:2157-2158.
13. Van der Bij AK, Kloosterboer N, Prins M, *et al.* Reply to George and Stapleton et al. *Journal of Infectious Diseases* 2005; **191**:2158-2160.
14. Cox DR. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* 1972; **34**:187-200.
15. Clayton DG. A model for association in biovariate life tables and its applications in epidemiological studies

of familial tendency in chronic disease indigence. *Biometrika* 1978; **65**:141-151.

16. Clayton DG, Cuzick J. Multivariate associations of the proportional hazards model. *Journal of the Royal Statistical Society, Series A* 1985; **148**:82-108.
17. Gustafson P. A Bayesian analysis of bivariate survival data from a multicenter cancer clinical trial. *Statistics in Medicine* 1995; **14**:2523-2535.
18. Stangl D. Prediction and decision making using Bayesian hierarchical models. *Statistics in Medicine* 1995; **14**:2173-2190.
19. Stangl D, Greenhouse J. Assessing placebo response using Bayesian hierarchical survival models. *Lifetime Data Analysis* 1998; **4**:5-28.
20. Sargent DJ. A general framework for random effects survival analysis in the Cox proportional hazards setting. *Biometrics* 1998; **54**:1486-1497.
21. Gustafson P. Large hierarchical Bayesian analysis of multivariate survival data. *Biometrics* 1997; **53**:230-242.
22. Cox DR. Partial likelihood. *Biometrika* 1975; **62**:269-275.
23. Kalbfleisch JD. Nonparametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B* 1978; **40**:214-221.
24. Sinha D, Ibrahim JG, Chen M. A Bayesian justification of Cox's partial likelihood. *Biometrics* 2003; **90**:629-641.
25. Ibrahim JG, Chen M-H, Sinha D. *Bayesian Survival Analysis*. Springer-Verlag Inc: New York, 2001.
26. Chen M-H, Ibrahim JG, Shao Q-M. Posterior propriety and computation for the Cox regression model with applications to missing covariates. *Biometrika* 2006; **93**:791-807.
27. Metropolis N, Rosenbluth AW, Rosenbluth MN, *et al.* Equations of state calculations by fast computing machines. *Journal of Chemical Physics* 1953; **21**:1087-1092.
28. Hastings WK. Monte Carlo sampling methos using Markov chains and their applications. *Biometrika* 1970; **57**:97-109.
29. Gilks WR, Richardson S, Spiegelhalter DJ. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 1996.
30. Liu KJ, Darrow WW, Rutherford GW. A model-based estimate of the mean incubation period for AIDS in homosexual men. *Science* 1988; **240**:1333-1335.
31. Mariotto AB, Mariotti S, Pezzotti P, *et al.* Estimation of the acquired immunodeficiency syndrome incubation period in intravenous drug users: a comparison with male homosexuals. *American Journal*

of Epidemiology 1992; **135**:428-437.

32. Law CG, Brookmeyer R. Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in Medicine* 1992; **11**:1569-1578.
33. Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B* 1976; **38**:290-295.
34. Jaffe HW, Darrow WW, Echenberg DF, *et al.* The Aquired Immunodeficiency Syndrome in a cohort of homosexual men: a six year follow-up study. *Annals of Internal Medicine* 1985; **103**:210-214.
35. Souza IE, Allen JB, Xiang J, *et al.* Effect of primer selection on estimates of GB virus C (GBV-C) prevalence and response to antiretroviral therapy for optimal testing for GBV-C viremia. *Journal of Clinical Microbiology* 2006; **44**:3105-3113.
36. Zhang W. Analysis of doubly censored survival data with applications to GBV-C and HIV studies. Ph.D. dissertation, 2005, University of Iowa, Iowa City, Iowa.
37. Gauvreau K, DeGruttola V, Pagano M, *et al.* The effect of covariates on the induction time of AIDS using improved imputation of exact seroconverson times. *Statistics in Medicine* 1994; **13**:2021-2030.
38. Toyoda H, Fukuda Y, Hayakawa T, Takamatsu, Saito H. Effect of GB virus C/hepatitis G virus coinfection on the course of HIV infection in hemophilia patients in Japan. *Journal of Acquired Immune Deficiency Syndrome and Human Retrovirology* 1998; **17**:209-213.
39. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. IBSN 3-900051-07-0, URL http://www.r-project.org, 2005.
40. Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS User Manual, 2003. URL http://www.mrc-bsu.cam.ac.uk/bugs.
41. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 1998; **7**:434-455.
42. Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. *Journal of American Statistical Association* 1987; **82**:528-540.
43. Härkänen T, Virtanen JI, Arjas E. Caries on permanent teeth: a non-parametric Bayesian analysis. *Scandinavian Journal of Statistics* 2000; **27**:577-588.
44. Komárek A, Lesaffre E, Härkänen T, Declerck D, Virtanen JI. A Bayesian analysis of multivariate doubly-interval-censored dental data. *Biostatistics* 2005; **6**:145-155
45. Nattermann J, Nischalke HD, Kupfer B, *et al.* Regulation of CC chemokine receptor 5 in hepatitis G virus infection. *AIDS* 2003; **17**:1457-1462.

46. Xiang J, George SL, Wünschmann S, *et al.* Inhibition of HIV-1 replication by GB virus C infection through increases in RANTES, MIP-1alpha, MIP-1beta, and SDF-1. *Lancet* 2004; **363**:2040-2046.

47. Jung S, Knauer O, Donhauser N, *et al.* Inhibition of HIV strains by GB virus C in cell culture can be mediated by CD4 and CD8 T-lymphocyte derived soluble factors. *AIDS* 2005; **19**:1267-1272.

48. Xiang J, Sathar MA, McLinden JH, Klinzman D, Chang Q, Stapleton JT. South African GB virus C isolates: interactions between genotypes 1 and 5 isolates and HIV. *Journal of Infectious Diseases* 2005; **192**:2147-2151.

49. George SL, Varmaz D, Stapleton JT. GB virus C replicates in primary T and B lymphocytes. *Journal of Infectious Diseases* 2006; **193**:451-454.

50. Stiehm ER. Disease versus disease: how one disease may ameliorate another. *Pediatrics* 2006; **117**:184-191.