Separate and Joint Modeling of Longitudinal and Event Time Data Using Standard Computer Packages

Xu GUO and Bradley P. CARLIN

Many clinical trials and other medical and reliability studies generate both longitudinal (repeated measurement) and survival (time to event) data. Many well-established methods exist for analyzing such data separately, but these may be inappropriate when the longitudinal variable is correlated with patient health status, hence the survival endpoint (as well as the possibility of study dropout). To remedy this, an earlier article proposed a joint model for longitudinal and survival data, obtaining maximum likelihood estimates via the EM algorithm. The longitudinal and survival responses are assumed independent given a linking latent bivariate Gaussian process and available covariates. We develop a fully Bayesian version of this approach, implemented via Markov chain Monte Carlo (MCMC) methods. We use the approach to jointly model the longitudinal and survival data from an AIDS clinical trial comparing two treatments, didanosine (ddI) and zalcitabine (ddC). Despite the complexity of the model, we find it to be relatively straightforward to implement and understand using the WinBUGS software. We compare our results to those obtained from readily available alternatives in SAS `Procs MIXED`, `NLMIXED`, `PHREG`, and `LIFEREG`, as well as Bayesian analogues of these traditional separate likelihood methods. The joint Bayesian approach appears to offer significantly improved and enhanced estimation of median survival times and other parameters of interest, as well as simpler coding and comparable runtimes.

KEY WORDS: Bayesian approach; Joint model; Longitudinal data; Markov chain Monte Carlo (MCMC); SAS; Survival data; WinBUGS.

1. INTRODUCTION

Many clinical trials generate both longitudinal (repeated measurement) and survival (time to event) data. For example, in AIDS clinical trials, the number of CD4 cells per cubic millimeter of blood is widely used as a biomarker for progression to AIDS when studying the efficacy of drugs to treat HIV-infected patients. It is often measured repeatedly over follow-up periods in large-scale trials. We might be interested in its relationship with time to seroconversion or death, in order to evaluate its appropriateness as a surrogate endpoint (Pawitan and Self 1993).

Many well-established methods exist for analyzing such data separately, including linear mixed effects models for longitudinal data, and Weibull or semiparametric (Cox) proportional hazards models for survival data. But their separate use may be inappropriate when the longitudinal variable is correlated with patient health status, hence the survival endpoint (as well as the possibility of study dropout). Unbiased statistical inferences are more likely to be obtained via a *joint* model (Tsiatis, Degruttola, and Wulfson 1995; Wulfsohn and Tsiatis 1997). For example, Wang and Taylor (2001) induced correlation between the longitudinal and survival processes by including the longitudinal marker as a time-dependent covariate in the proportional hazards survival model. Lin, Turnbull, McCulloch, and Slate (2002) instead employed a latent class model, where each subject's class membership is determined by a multinomial logistic model, after which the longitudinal and survival processes are modeled independently given this membership (marginally, the two processes are dependent, as desired).

This article investigates the approach of Henderson, Diggle, and Dobson (2000), who proposed a very flexible joint model that avoids specifying the class variable, yet allows a very broad range of dependencies between the longitudinal responses and the survival endpoints. These authors model the longitudinal data by including fixed effects, random effects, serial correlation, and pure measurement error, and the survival data by using a semiparametric proportional hazards model with or without frailty terms. The strategy's key idea is to connect the longitudinal and survival processes with a latent bivariate Gaussian process. The longitudinal and survival data are then assumed independent given the linking latent process and any available covariates.

Implementation of a version of the traditional maximum likelihood approach for this joint model is also available using SAS `Proc NLMIXED`; see www.biostat.umn.edu/~brad/software.html and Section 4. The solution is fairly complicated, however, and delivers only point estimates and associated asymptotic standard error estimates for the model parameters. This motivates investigation of a Bayesian alternative that permits full and exact posterior inference for any parameter or predictive quantity of interest. Indeed, this is the usual reason cited for turning to Bayesian methods, but these benefits typically come at a price; namely, the required choice of *prior* distributions for the model parameters, and the higher overhead (both user and machine) of

Xu Guo is Biostatistician for Affymetrix Inc., 3380 Central Expressway, Santa Clara, CA 95051 (E-mail: xu_guo@affymetrix.com). Bradley P. Carlin is Mayo Professor in Public Health at the Division of Biostatistics, School of Public Health, University of Minnesota, MMC 303, Minneapolis, MN 55455 (E-mail: brad@biostat.umn.edu). The work of both authors was supported in part by NIH grant 2–R01–ES07750. The authors are grateful to Sudipto Banerjee and Oliver Schabenberger for helpful discussions and computing assistance that were essential to this project's completion.

  DOI: 10.1198/0003130042854

the associated Markov chain Monte Carlo (MCMC) computational methods.

Although the first of these problems remains vexing in general, the ever-increasing track record of Bayes-MCMC methods in statistical practice now suggests classes of priors that may be thought of as sufficiently "noninformative" in fairly broad settings, where by this we mean a prior that allows the data to dominate the determination of the posterior distribution. As for the second problem, again for a surprisingly broad class of problems, the popular WinBUGS software package (freely available from http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml) offers an intuitive yet general solution. Indeed, comparing the WinBUGS and SAS implementations in our joint longitudinal and survival setting, we will see that the former is actually shorter, more transparent, and takes no more time to run. The Bayesian framework also allows for a more direct comparison among the many possible forms for (and types of association between) the components of the latent bivariate process linking the longitudinal and survival components. In WinBUGS, such comparison is most easily carried out using the DIC criterion (Spiegelhalter, Best, Carlin, and van der Linde 2002), described below. All this suggests the Bayes-MCMC approach not only delivers more here, but for less effort on the user's part.

Section 2 describes the traditional longitudinal and survival models that have been applied to data like ours, and mentions common ways of fitting them. We then go on to develop the fully Bayesian version of the joint modeling approach of Henderson et al. (2000), including its implementation via MCMC methods. Section 3 then describes our dataset in detail, and compares the results of the various approaches mentioned above in this context. Our Bayesian results (for both separate and joint models) are computed entirely in WinBUGS, while our classical results are obtained using the SAS procedures `MIXED`, `NLMIXED`, `PHREG`, and `LIFEREG`. Despite the complexity of the joint model, we find its WinBUGS implementation to be relatively straightforward to implement and understand, and productive of several advantages over the traditional, separate model methods. Finally, Section 4 discusses our findings and suggests avenues for future research in this area.

2. CLASSICAL APPROACHES

Suppose we have a set of m subjects followed over a time interval $[0, \tau)$. The ith subject provides a set of (possibly partly missing) longitudinal quantitative measurements $\{y_{ij},\ j = 1, \ldots, n_i\}$ at times $\{s_{ij},\ j = 1, \ldots, n_i\}$, and a (possibly censored) survival time t_i to a certain endpoint. We now briefly review classical approaches for the separate analysis of longitudinal and survival data, and the joint modeling approach proposed by Henderson et al. (2000).

2.1 Longitudinal Data Models

Linear mixed effects models are widely used to model continuous longitudinal data. The sequence of measurements $y_{i1}, y_{i2}, \ldots, y_{in_i}$ for the ith subject at times $s_{i1}, s_{i2}, \ldots, s_{in_i}$ is modeled as

$$y_{ij} = \mu_i(s_{ij}) + W_{1i}(s_{ij}) + \epsilon_{ij}\ , \qquad (1)$$

where $\mu_i(s) = \mathbf{x}_{1i}^T(s)\boldsymbol{\beta}_1$ is the mean response, $W_{1i}(s) = \mathbf{d}_{1i}^T(s)\mathbf{U}_i$ incorporates subject-specific random effects, and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ is a sequence of mutually independent measurement errors. The $W_{1i}(s)$ can be viewed as the true individual-level CD4 trajectories after they have been adjusted for the overall mean trajectory and other fixed effects. The vectors $\mathbf{x}_{1i}(s)$ and $\boldsymbol{\beta}_1$ represent possibly time-varying explanatory variables and their corresponding regression coefficients, respectively. The $\mathbf{U}_i$ are vectors of random effects corresponding to the explanatory variables $\mathbf{d}_{1i}(s)$ (which may be a subset of $\mathbf{x}_{1i}(s)$) and are typically modeled as iid $N(\mathbf{0}, \boldsymbol{\Sigma})$ random variables. Mixed (i.e., having both fixed and random effects) longitudinal models like this one have a long history in biostatistical theory and practice, dating at least to the seminal work of Laird and Ware (1982).

2.2 Survival Data Models

Both parametric and semiparametric models are available to model survival data. Commonly used parametric models include the exponential and Weibull, which are attractive in their simplicity and the easy interpretability of their components. In practice, however, semiparametric (Cox) proportional hazards models are widely used, since they impose no particular shape on the survival curves (though they do continue to assume the shape is similar across covariate groups).

2.2.1 Weibull Model

In a Weibull model, we assume that the survival time for the ith subject follows a Weibull distribution, $t_i \sim$ Weibull $(r, \mu_i(t))$, where $\log(\mu_i(t)) = \mathbf{x}_{2i}^T(t)\boldsymbol{\beta}_2 + W_{2i}(t)$ and $r > 0$. The vectors $\mathbf{x}_{2i}(t)$ and $\boldsymbol{\beta}_2$ represent (possibly time-dependent) explanatory variables and their corresponding regression coefficients. They may or may not have elements in common with $\mathbf{x}_{1i}$ and $\boldsymbol{\beta}_1$ in the longitudinal model. The form of $W_{2i}(t)$ is similar to $W_{1i}(s)$, including subject-specific covariate effects and an intercept (often called a *frailty*). The event intensity (or *hazard*) at time t is given as

$$\lambda_i(t) = rt^{r-1}\mu_i(t) = rt^{r-1}\exp\left(\mathbf{x}_{2i}^T(t)\boldsymbol{\beta}_2 + W_{2i}(t)\right), \qquad (2)$$

which is monotone in t (decreasing if $r < 1$, increasing if $r > 1$) and reduces to the exponential (constant in t) hazard if $r = 1$. Parameterization (2) is the one used by WinBUGS.

2.2.2 Cox Semiparametric Proportional Hazards Model

In the semiparametric proportional hazards model, we replace the hazard in (2) with

$$\lambda_i(t) = \lambda_0(t)\exp\left(\mathbf{x}_{2i}^T(t)\boldsymbol{\beta}_2 + W_{2i}(t)\right)\ ,$$

where now the baseline hazard function $\lambda_0(t)$ has an arbitrary (instead of Weibull) form. Cox and Oakes (1984, chap. 7) provided full details on traditional methods for fitting such models.

2.3 Joint Model

Association between the longitudinal and survival processes can arise in two ways. One is through common explanatory variables and the other is through stochastic dependence between W_{1i} and W_{2i}. Henderson et al. (2000) proposed to jointly model the two processes via a latent zero-mean bivariate Gaussian process on $(W_{1i}, W_{2i})^T$, which is independent across different subjects. The joint model consists of two linked submodels, which

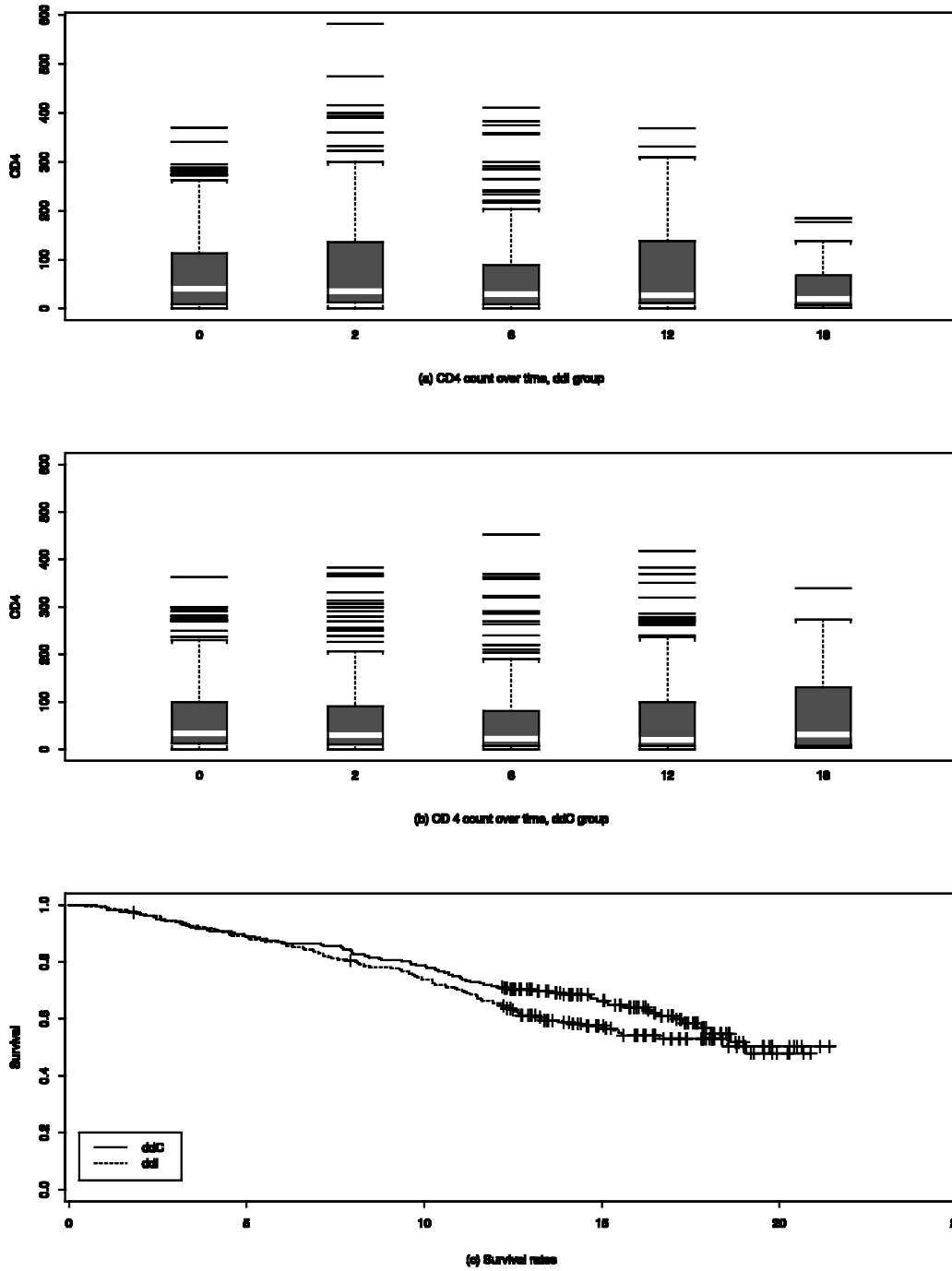

Figure 1. Exploratory plots of longitudinal data and survival data for the ddI/ddc trial.

they refer to as the *measurement model* for the longitudinal process, and the *intensity model* for the survival process. We can apply this joint modeling strategy to connect the classical models for longitudinal data and survival data with each other. When association between the two processes exists, we should obtain less biased and more efficient inferences by using this joint model.

Specifically, our joint model links (1) and (2) by taking

$$W_{1i}(s) \;=\; U_{1i} + U_{2i}\, s\;, \tag{3}$$

and

$$W_{2i}(t) \;=\; \gamma_1 U_{1i} + \gamma_2 U_{2i} + \gamma_3 (U_{1i} + U_{2i}\, t) + U_{3i}\;. \tag{4}$$

The longitudinal model (3) is of the usual Laird–Ware form, with each patient receiving random intercept and slope terms. We emphasize that this form need not be linear in s, and should be well-motivated by the application at hand (in our case we actually experimented with a quadratic form, but found our data could not support it). The parameters γ_1, γ_2, and γ_3 in the survival model (4) measure the association between the two submodels induced by the random intercepts, slopes, and fitted longitudinal value at the event time $W_{1i}(t)$, respectively. As in Subsection 2.1, the pair of latent variables $(U_{1i}, U_{2i})^T$ has a mean-zero bivariate Gaussian distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, while the U_{3i} are independent frailty terms, modeled as iid $N(0, \sigma_3^2)$ variables, independent of the $(U_{1i}, U_{2i})^T$.

Table 1. Classical Separate Analyses for the ddI/ddC Data

Parameter	Point estimate	95% CI	Point estimate	95% CI
	Longitudinal mixed effects model			
Intercept (β_{11})	8.013	(7.325, 8.701)		
Time (β_{12})	−0.167	(−0.207, −0.127)		
Time × Drug (β_{13})	0.030	(−0.037, 0.087)		
Gender (β_{14})	−0.158	(−0.795, 0.479)		
PrevOI (β_{15})	−2.315	(−2.782, −1.848)		
Stratum (β_{16})	−0.131	(−0.592, 0.330)		
	Exponential model		*Proportional hazards (Cox) model*	
Intercept (β_{21})	−3.702	(−4.019, −3.385)	—	—
Drug (β_{22})	0.208	(−0.079, 0.495)	0.217	(−0.070, 0.504)
Gender (β_{23})	−0.169	(−0.410, 0.071)	−0.171	(−0.412, 0.070)
PrevOI (β_{24})	0.620	(0.398, 0.841)	0.646	(0.423, 0.868)
Stratum (β_{25})	0.082	(−0.077, 0.242)	0.079	(–0.082, 0.239)

3. ANALYSIS OF DDI/DDC DATA

3.1 Description of Dataset

Both longitudinal and survival data were collected in a recent clinical trial to compare the efficacy and safety of two antiretroviral drugs in treating patients who had failed or were intolerant of zidovudine (AZT) therapy. In this trial, $m = 467$ HIV-infected patients who met entry conditions (either an AIDS diagnosis or two CD4 counts of 300 or fewer, and fulfilling specific criteria for AZT intolerance or failure) were enrolled and randomly assigned to receive either didanosine (ddI) or zalcitabine (ddC). CD4 counts were recorded at study entry, and again at the 2-, 6-, 12-, and 18-month visits (so that $n_i \leq 5$). The times to death were also recorded. For full details regarding the conduct of the trial the reader is referred to Abrams et al. (1994) and Goldman et al. (1996).

Boxplots of the CD4 counts for the two drug groups over time (Figures 1(a) and (b)) show a high degree of skewness toward high CD4 counts, suggesting a square root transformation. The sample sizes at the five time points are (230, 182, 153, 102, 22) for the ddI group and (236, 186, 157, 123, 14) for the ddC group. There is a sharply increasing degree of missing data over time due to deaths, dropouts, and missed clinic visits. As for the survival rates in the two drug groups, the empirical survival curves (Kaplan–Meier estimates; Figure 1(c)) are very similar during the first six months after randomization. Afterwards, survival in the ddC group is somewhat higher than that in the ddI group through the 18-month visit.

Let y_{ij} denote the square root of the jth CD4 count measurement on the ith patient in the trial, $j = 1, \ldots, n_i$ and $i = 1, \ldots, m$. We include four binary explanatory variables as main effects in our analysis: Drug (ddI = 1, ddC = 0), Gender (male $= 1$, female $= -1$), PrevOI (previous opportunistic infection (AIDS diagnosis) at study entry $= 1$, no AIDS diagnosis $= -1$), and Stratum (AZT failure $= 1$, AZT intolerance $= -1$). Our main goal is to analyze the association among CD4 count, survival time, drug group, gender, AIDS diagnosis at baseline (an indicator of disease progression status), and stratum, accounting for all relevant correlations and subject-specific random effects.

3.2 Results Using Classical Models

We first use the classical models reviewed in Section 2 to model the longitudinal and survival data from the ddI/ddC trial separately. These data analyses are conducted using SAS.

The linear random effects model for square root CD4 count is specified as

$$y_{ij} = \beta_{11} + \beta_{12} s_{ij} + \beta_{13} s_{ij} \times \text{Drug}_i + \beta_{14} \text{Gender}_i + \beta_{15} \text{PrevOI}_i + \beta_{16} \text{Stratum}_i + W_{1i}(s_{ij}) + \epsilon_{ij}, \quad (5)$$

where $W_{1i}(s_{ij}) = U_{1i} + U_{2i} s_{ij}$. Here, $W_{1i}(s_{ij})$ includes the random effects for intercept and slope over time, where the $\mathbf{U}_i = (U_{1i}, U_{2i})^T \overset{\text{iid}}{\sim} N_2(\mathbf{0}, \boldsymbol{\Sigma})$. This specification allows different subjects to have different baseline CD4 counts and different time trends for CD4 counts during the trial.

Results obtained using `Proc MIXED` (using the `type=un` option to obtain the unstructured covariance matrix $\boldsymbol{\Sigma}$) are summarized in the upper half of Table 1. The estimated average regression coefficient of Time for the ddC group is -0.167 with 95% confidence interval of $(-0.207, -0.127)$, suggesting a significant decrease in CD4 count over the study period. The estimated average regression coefficient of Time for ddI group is $-0.167 + 0.030 = -0.137$, which is not significantly different from the corresponding coefficient for the ddC group. PrevOI is also significantly different from 0, having estimated regression coefficient -2.315 and 95% confidence interval $(-2.782, -1.848)$. Hence a patient who was diagnosed with AIDS at study entry has significantly lower CD4 counts than one without an AIDS diagnosis. The other two explanatory variables, Gender and Stratum, are not statistically significant here.

Turning to survival analysis, we began with the full Weibull model but found strong negative association between the Weibull shape parameter r and intercept β_{21}. In the MCMC-Bayes setting, the inability of our data to reliably identify both r and β_{21} if both are included in the sampling order manifests as strong negative correlations between the samples of these two parameters, as well as strong positive autocorrelations in their individual sampled chains. Since convergence failure precludes accurate computation of posterior model summaries, in what follows we have used a Weibull model with $r = 1$ (i.e., an exponential

model) along with a Cox proportional hazards model to analyze the ddI/ddC time-to-event data.

Because none of our covariates are time-varying, the regression equation for the log-relative hazard in the absence of random effects is

$$\log(\mu_i) = \beta_{21} + \beta_{22}\text{Drug}_i + \beta_{23}\text{Gender}_i + \beta_{24}\text{PrevOI}_i + \beta_{25}\text{Stratum}_i \ . \quad (6)$$

We have used the WinBUGS parameterization here; as such, we transform the results from SAS `Proc LIFEREG` (exponential) or `PHREG` (Cox) to be consistent with the parameterization in WinBUGS to permit comparison of separate and joint analyses later. Among the four explanatory variables we have included in the model, only PrevOI is statistically significant at level 0.05 (bottom half of Table 1). Not surprisingly, patients who had a negative AIDS diagnosis at study entry have better average survival rates than those who had a positive diagnosis. The ddC group turns out to have a little better survival than the ddI group (as suggested by Figure 1(c)), but the difference is not significant.

The Cox proportional hazards model uses parameterization (6), except that the intercept is modeled as part of the (nonparametric) baseline hazard, and therefore is not explicitly output by the software (SAS `Proc PHREG`). The results (also bottom half of Table 1) from the proportional hazards model are similar to what we get from the exponential model. PrevOI is again the only statistically significant predictor at level 0.05. The relative hazard is $\exp(.646) = 1.908$ for patients with an AIDS diagnosis, as compared to $\exp(.620) = 1.859$ under the exponential model.

3.3 Results Using Joint Models

Henderson et al. (2000) proposed an expectation–maximization (EM) algorithm to fit the joint model, but appear to have done so using their own computer code, rather than a standard software package. As mentioned earlier, approximate results from this model are available from SAS, but we defer this discussion until Section 4, and focus here on the fully Bayesian version of the joint modeling approach implemented via MCMC methods via the WinBUGS package.

To facilitate fair comparison of our classical and Bayesian analyses (the former of which of course cannot be helped or hindered by prior information), we selected very vague prior distributions in our WinBUGS analysis. That is, we use proper priors, but with hyperparameter values chosen so that the priors will have minimal impact relative to the data. Specifically, in the longitudinal submodel we take multivariate normal and inverse gamma priors for the main effects vector $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{15}, \beta_{16})^T$ and the error variance σ^2_ϵ, respectively, both having very low precision (high variance; readers interested in our specific choices may consult our code at www.biostat.umn.edu/~brad/software.html). We similarly select vague normal and inverse gamma priors for $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \beta_{23}, \beta_{24}, \beta_{25})^T$ and σ^2_3 in the survival submodel. For the parameters common to both models, we take a inverse Wishart (essentially a multivariate gamma; see, e.g., Carlin and Louis 2000, p. 328), which is again vague but does provide at least some shrinkage of the random effects toward $\mathbf{0}$, ensuring good identifiability of the main effects (Carlin and Louis 2000, p. 279). Finally, for the association parameters we choose normal priors for γ_1 and γ_2 that are both quite vague relative to these parameters' likely posterior magnitudes. In summary, our priors are chosen so that our Bayesian analysis mimics a corresponding likelihood analysis, but where likelihoods are restandardized and interpreted as probability distributions on the parameters.

3.3.1 Joint Model Selection

There are many Bayesian approaches to model selection, ranging from formal approaches using posterior model probabilities and Bayes factors (see, e.g., Carlin and Louis 2000, secs. 6.3–4) to informal approaches based on posterior predictive checks (see, e.g., Gelman, Carlin, Stern, and Rubin 1995, sec. 6.3). However, the former are not well-defined using the (often arbitrarily) vague prior distributions we employ, while the latter are not automatic nor easily reduced to a unique, single number summary. As mentioned earlier, we choose the precise nature of the two submodels (i.e., the exact form of $W_1(s)$ and $W_2(t)$) and their association via the DIC (Deviance Information Criterion; Spiegelhalter et al., 2002), a hierarchical modeling generalization of the AIC (Akaike Information Criterion). We prefer this approach for several reasons. First, like AIC it reduces each model to a single number summary, but with a proper Bayesian interpretation even for vague-prior random effects models like ours. These values may be arranged in tables reminiscent of classical ones comparing log-likelihood (deviance) scores via chi-square tests, but where now the models need not be nested to be directly comparable. Last (but certainly not least), DIC is readily available within WinBUGS for all of our models.

Recall that the change in the AIC statistic across two models M_1 and M_2 is given by

$$\Delta\text{AIC} = V - 2(p_2 - p_1) \ ,$$

where p_k is the number of parameters in model M_k, and $V = -2\log\left[\sup_{M_1} f(\mathbf{y}|\boldsymbol{\theta})/\sup_{M_2} f(\mathbf{y}|\boldsymbol{\theta})\right]$, the usual likelihood ratio test statistic. That is, AIC is a penalized likelihood criterion, where the size of the penalty depends on how much "larger" M_2 is than M_1. But a difficulty arises when using this approach to compare models with random effects: how should we "count" parameters? As a simple illustration, consider our U_{3i} frailties: at first glance they appear to contribute m parameters to the total (one for each subject), but if $\sigma^2_3 = 0$, they contribute none (since all would then be identically zero). In fact they contribute the full m only if $\sigma^2_3 = \infty$; for any given dataset, the *effective* number of parameters (which we shall denote by p_D) is somewhere in between these two extremes.

Thinking of $\boldsymbol{\theta}$ and $\mathbf{y}$ as the entire collections of model parameters and data, respectively, Spiegelhalter et al. (2002) showed that p_D turns out to be reasonably defined as

$$p_D = E_{\boldsymbol{\theta}|\mathbf{y}}[D(\boldsymbol{\theta})] - D(E_{\boldsymbol{\theta}|\mathbf{y}}[\boldsymbol{\theta}]) = \overline{D} - D(\bar{\boldsymbol{\theta}}) \ . \quad (7)$$

Here, $D(\boldsymbol{\theta})$ is the *deviance* function, $D(\boldsymbol{\theta}) = -2\log f(\mathbf{y}|\boldsymbol{\theta}) + 2\log h(\mathbf{y})$, where $f(\mathbf{y}|\boldsymbol{\theta})$ is again the likelihood function and $h(\mathbf{y})$ is some standardizing function of the data alone. The sensibility of (7) is easily seen in the case where $\boldsymbol{\theta} \mid \mathbf{y} \sim \text{N}\left(\widehat{\boldsymbol{\theta}}\ , \ -L''_{\widehat{\boldsymbol{\theta}}}\right)$, where $\bar{\boldsymbol{\theta}} \approx \widehat{\boldsymbol{\theta}}$, the maximum likelihood estimates

Table 2. Bayesian Model Selection for the ddI/ddC Data

Model	$W_1(s)$	$W_2(t)$	DIC_1	DIC_2	$\overline{D}$	p_D	DIC_{total}
no random effects							
I	0	0	8201.9	1372.0	9565.5	8.53	9574.0
II	0	U_3	8202.0	1374.7	9561.1	15.60	9576.7
random intercepts							
III	U_1	0	6319.7	1371.0	7261.2	429.5	7690.7
IV	U_1	U_3	6317.2	1374.6	7256.3	435.6	7691.9
V	U_1	$\gamma_1 U_1$	6302.6	1345.8	7222.5	425.8	7648.3
VI	U_1	$\gamma_1 U_1 + U_3$	6305.4	1350.8	7223.1	433.2	7656.2
random intercepts and random slopes							
VII	$U_1 + U_2\, s$	0	6226.3	1372.1	6866.6	731.9	7598.5
VIII	$U_1 + U_2\, s$	$\gamma_1 U_1$	6218.6	1349.8	6839.5	728.9	7568.4
IX	$U_1 + U_2\, s$	$\gamma_2 U_2$	6232.0	1363.6	6859.3	736.4	7595.6
X	$U_1 + U_2\, s$	$\gamma(U_1 + U_2)$	6219.2	1348.3	6838.6	728.9	7567.4
XI	$U_1 + U_2\, s$	$\gamma_1 U_1 + \gamma_2 U_2$	6204.0	1344.3	6803.8	744.5	7548.3
XII	$U_1 + U_2\, s$	$\gamma_1 U_1 + \gamma_2 U_2 + \gamma_3 W_1(t)$	6208.7	1416.5	6959.1	666.1	7625.2

such that $L'_{\hat{\boldsymbol{\theta}}} = 0$. This is the standard normal approximation to the posterior often referred to as the "Bayesian Central Limit Theorem" (Carlin and Louis 2000, sec. 5.1). A second order expansion of $D(\boldsymbol{\theta})$ around $\bar{\boldsymbol{\theta}}$ gives

$$D(\boldsymbol{\theta}) \;\approx\; D(\bar{\boldsymbol{\theta}}) - (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^T L''_{\bar{\boldsymbol{\theta}}}(\theta - \bar{\boldsymbol{\theta}}) \;\approx\; D(\bar{\boldsymbol{\theta}}) + \chi^2_p, \quad (8)$$

since, by the Bayesian Central Limit Theorem, $-(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^T L''_{\bar{\boldsymbol{\theta}}}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})$ has an approximate chi-squared distribution with p degrees of freedom. Rearranging (2) and taking expectations with respect to the posterior distribution of $\boldsymbol{\theta}$ reveals $p_D = \mathrm{E}_{\boldsymbol{\theta}|\mathbf{y}}[D(\boldsymbol{\theta})] - D(\bar{\boldsymbol{\theta}}) \approx p$; that is, that p_D will be approximately the true number of parameters.

Returning to model choice, the DIC approach mimics AIC by setting

$$\mathrm{DIC} = \overline{D} + p_D \ .$$

Here, the fit of a model is summarized in the first term by the posterior expectation of the deviance function, $\overline{D} = E_{\boldsymbol{\theta}|\mathbf{y}}[D(\boldsymbol{\theta})]$, while the complexity of the model is captured in the second term by the effective number of parameters, p_D. Since small values of $\overline{D}$ indicate good fit while small values of p_D indicate a parsimonious model, small values of the sum (DIC) indicate preferred models. As before, in what follows we restrict our attention to exponential models by setting the Weibull shape parameter $r = 1$, so that full Weibull models are not eligible for comparison or selection here.

Table 2 reports the $\overline{D}$, p_D, and DIC score for a variety of joint models for the ddI/ddC data with different forms of the latent processes $W_1(s)$ and $W_2(t)$. In all cases, our results are based on three parallel MCMC sampling chains of 10,000 iterations each, following a 5,000-iteration "burn-in" period (see, e.g., Carlin and Louis 2000, sec. 5.4.6 for a general discussion of MCMC convergence monitoring). By default, WinBUGS assumes the parameter vector $\boldsymbol{\theta}$ contains both the fixed and random effects, and provides the components of DIC for the two submodels (i.e., the terms in the log-likelihood arising from longitudinal and survival model components) to evaluate their relative contributions to the total DIC score; we notate these as DIC_1 and DIC_2. We start with a simple joint model (Model I) with no random effects in the two submodels, which has a large (poor) total DIC score. In Model II, we add a frailty term U_3 in $W_2(t)$, but this does not seem to improve the total DIC at all. A similar relationship exists between Models III and IV and between Models V and VI, all of which differ only in the addition of the frailty term U_3. As such, we do not consider including U_3 in subsequent models.

In Model III, we allow random intercepts in $W_1(s)$, which results in a dramatic improvement in DIC_1 for the longitudinal submodel, and hence the total DIC score for the joint model. Next, we introduce association between $W_1(s)$ and $W_2(t)$ through the common random intercept U_1 in Model V, which leads to a substantial decrease in DIC for both submodels, and hence the total DIC for the joint model. This suggests association between the two submodels, as anticipated for these data. Model VII further allows both random intercepts and slopes in the longitudinal submodel, which substantially decreases DIC_1. The next four models in the table introduce association between the two submodels in different ways, including through random intercepts (Model VIII), random slopes (Model IX), or both random intercepts and slopes with a common association parameter (Model X) or two different association parameters (Model XI). Our final model (Model XII) introduces the full complexity given in Equation (4) by adding the time-varying component $W_1(t)$ to the survival model. This final increase in complexity is apparently not justified for these data, as indicated by the increase in the DIC score and the insignificance of the γ_3 parameter (95% posterior credible interval $(-0.43, 0.26)$). This is also consistent with the aforementioned finding (Carlin and Louis 2000, p. 288) that an exponential model (with its time-constant baseline hazard) is adequate for this very ill population.

Because Model XI emerges with the smallest total DIC, we select it as our final model for the ddI/ddC data. Under this model, it appears that a patient's survival is related to two characteristics driving the patient's longitudinal data pattern, namely the initial CD4 level and the rate of CD4 decrease. This is clinically reasonable, since high CD4 count represents better health status; patients with CD4 counts that are low or in more rapid decline would be expected to have poorer survival.

Table 3. Separate and Joint Bayesian Analysis for the ddI/ddC Data

Parameter	Separate analysis		Joint analysis	
	Posterior mean	95% CI	Posterior mean	95% CI
	Longitudinal submodel		*Longitudinal submodel*	
Intercept (β_{11})	8.04	(7.36, 8.72)	8.06	(7.37, 8.77)
Time (β_{12})	−0.20	(−0.29, −0.10)	−0.27	(−0.36, −0.17)
Time × Drug (β_{13})	0.06	(−0.08, 0.19)	0.03	(−0.11, 0.17)
Gender (β_{14})	−0.14	(−0.78, 0.50)	−0.13	(−0.76, 0.48)
PrevOI (β_{15})	−2.33	(−2.81, −1.85)	−2.34	(−2.81, −1.86)
Stratum (β_{16})	−0.11	(−0.57, 0.36)	−0.12	(−0.59, 0.35)
Σ_{11}	15.63	(13.56, 18.00)	15.54	(13.47, 17.88)
Σ_{22}	0.39	(0.34, 0.46)	0.39	(0.34, 0.45)
$\rho \equiv \Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}}$	−0.11	(−0.22, 0.01)	−0.06	(−0.17, 0.04)
σ^2_ϵ	2.91	(2.60, 3.24)	2.89	(2.60, 3.22)
	Survival submodel		*Survival submodel*	
Intercept (β_{21})	−3.73	(−4.06, −3.41)	−4.09	(−4.52, −3.68)
Drug (β_{22})	0.21	(−0.09, 0.52)	0.27	(−0.10, 0.62)
Gender (β_{23})	−0.16	(−0.39, 0.09)	−0.11	(−0.40, 0.20)
PrevOI (β_{24})	0.62	(0.40, 0.85)	0.76	(0.51, 1.02)
Stratum (β_{25})	0.09	(−0.08, 0.25)	0.08	(−0.12, 0.28)
γ_1	—	—	−0.20	(−0.25, −0.14)
γ_2	—	—	−1.61	(−2.13, −1.08)

3.3.2 *Comparison of Separate and Joint Models*

Having selected a final model, we now compare results obtained under the separate (i.e., ignoring any latent association introduced by W_2) and joint models. Both of these assume the longitudinal model has form (5), while the survival model now takes the form

$$\log(\mu_i) = \beta_{21} + \beta_{22}\text{Drug}_i + \beta_{23}\text{Gender}_i + \beta_{24}\text{PrevOI}_i + \beta_{25}\text{Stratum}_i + \begin{cases} 0 & \text{(separate)} \\ W_{2i}(t) & \text{(joint)} \end{cases},$$

where $W_{2i}(t) = W_{2i} = \gamma_1 U_{1i} + \gamma_2 U_{2i}$. The posterior estimates of the regression coefficients $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ and their 95% confidence intervals are summarized in Table 3. Here the results from the separate and joint analyses are quite similar to each other (and to the classical results in Table 1). In both Table 3 analyses, in the longitudinal submodel only Time and PrevOI are statistically significant at level 0.05 (in the Bayesian sense; 95% credible set excludes 0), while only PrevOI is significant at this level in the survival submodel. However, the posterior estimates of the association parameters in the joint analysis are negative and significantly different from zero, providing strong evidence of association between the two submodels and indicating that both the initial level and slope of CD4 count is negatively associated with the hazard of death.

To investigate this finding, consider a male patient who is AIDS-negative at study entry and intolerant of AZT. We plot the estimated posterior density of the median survival time of this hypothetical patient in Figure 2. In both the separate (panel a) and joint (panel b) analyses, this patient's survival is clearly better if he receives ddC instead of ddI. However, the joint analysis increases the estimated median survival times by roughly 50% in both groups. This is due to this model's correct accounting for the correlation between the longitudinal and survival data; our hypothetical patient has covariate values normally associated with a good prognosis, and this is reflected in the dramatically improved predicted survival times.

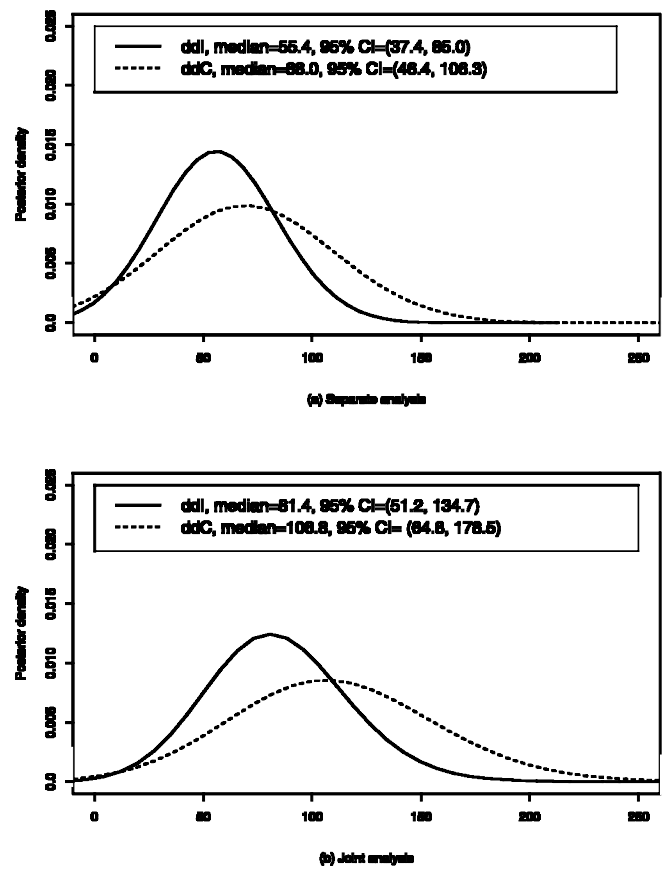

Figure 2. Median survival time for a hypothetical patient (male, negative AIDS diagnosis at study entry, intolerant of AZT): (a) estimated posterior density of median survival time of the patient from separate analysis; (b) estimated posterior density of median survival time of the patient from joint analysis.

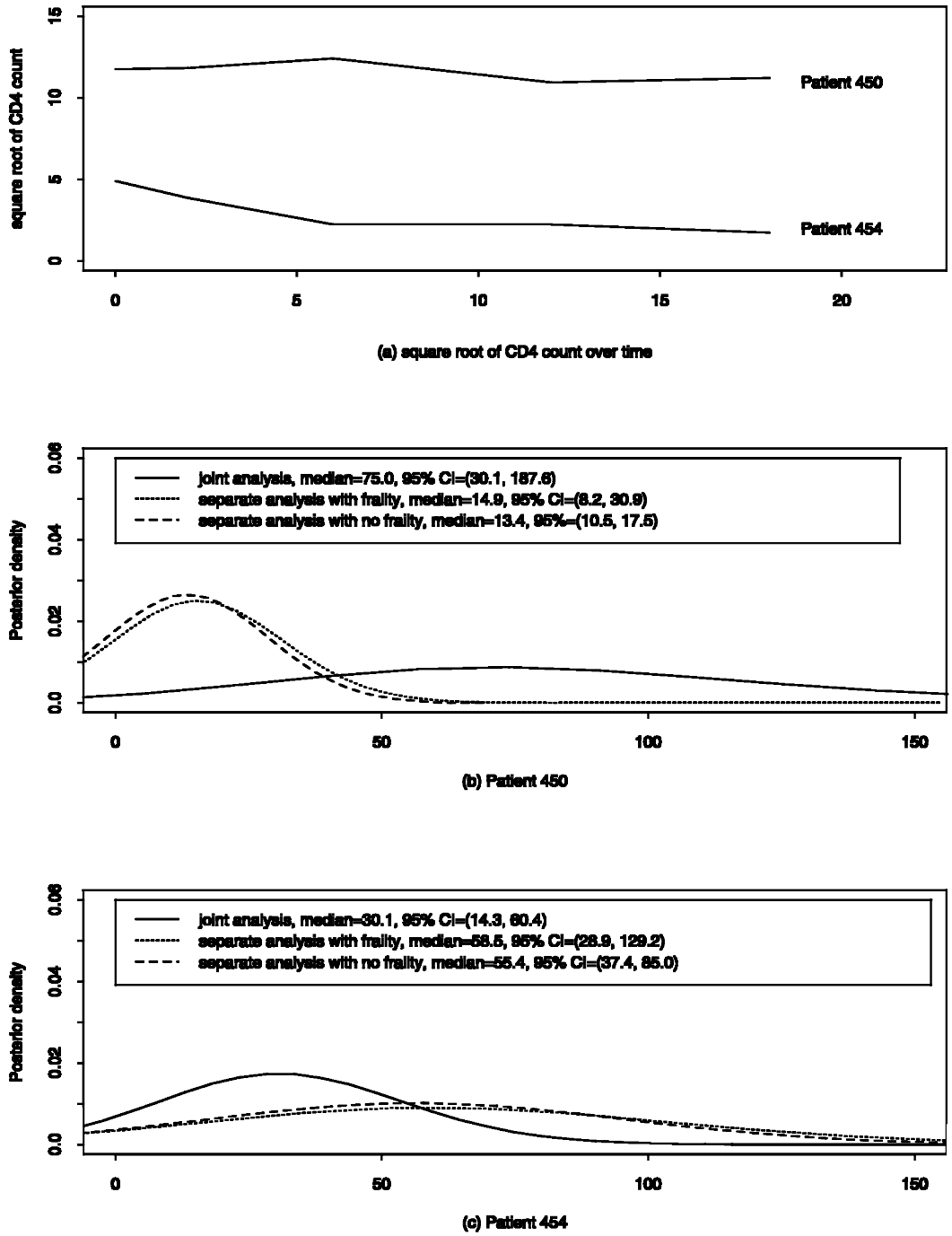

Figure 3. Observed data and estimated posteriors of median survival time for two patients, ddI/ddC study: (a) observed longitudinal data for Patients 450 and 454; (b) estimated posterior densities for Patient 450; (c) estimated posterior densities for Patient 454.

Even more dramatic differences can be observed if we switch from a hypothetical patient to one having a specific observed CD4 trajectory, but an unknown survival time. Figure 3 compares the estimated posterior median survival time distributions for two such patients in our study. The first (Patient 450) is in the ddI group, male, entered the study with an AIDS diagnosis, and had previously failed AZT. The second (Patient 454) is in the ddI group, male, did not have an AIDS diagnosis at study entry, and was intolerant of AZT. Both were still alive at the end of the study, censored at days 571 and 591, respectively. Figure 3(a) shows Patient 450 has a relatively "good" square root CD4 trajectory (starts relatively high and remains that way), while Patient 454 has a "bad" one (starts low and declines). Figures 3(b) and (c) compare the posterior median survival time distributions of the joint model with two separate models (with and without frailty terms U_3) for the two patients. We first note that the two separate analysis curves are virtually identical; as in Table 2, the inclusion of U_3 clearly has little impact. However, the joint results differ much more markedly from the separate results than they did in Figure 2, significantly

increasing the survival time for Patient 450, and decreasing it for Patient 454. Moreover, the joint model actually *reverses* the separate models' findings, in the sense that the patient with the "good" CD4 trajectory is now predicted to survive much longer than the patient with the "bad" trajectory. Such a reversal is quite plausible, yet not possible if we rely only on the fixed effects in the model, since the first patient actually has the "bad" prognostic variable values (AIDS positive, AZT failure), while the situation is reversed for the second patient.

4. DISCUSSION

In this article, we have presented a general Bayesian approach to the joint modeling of longitudinal and survival responses. This general class of models is important in many biostatistical application areas, allowing accurate inference regarding longitudinal responses while adjusting for outcome-dependent study dropout. We can also apply these ideas to problems involving surrogate markers (like CD4 count), where the focus is on using longitudinal measurements to improve prediction of survival prognosis.

The joint models considered in this article can be readily fit using the WinBUGS package, thus avoiding the need for complex EM programming or software and facilitating the models' use in practice. The Web page www.biostat.umn.edu/~brad/software.html contains the WinBUGS code to fit Models XI and XII above; these programs allow the full Weibull model with the shape parameter r following a gamma prior distribution, but as mentioned earlier avoid this complexity for our particular dataset by simply setting $r = 1$. Model comparison and selection is also facilitated via the DIC tool built into the package, enabling direct comparison of nonnested specifications and avoiding the pitfalls of repeated significance testing, and indeed the problems associated with significance testing itself. (We remark here that there was a bug in the WinBUGS 1.3 DIC tool that gave incorrect results for censored data models like ours, but this bug has been repaired in version WinBUGS 1.4.)

Of course, there are some prerequisites to using the WinBUGS package: the user must have at least some background in Bayesian methods, and understand that the basic task is to summarize (correlated) samples from the posterior distributions of interest. Perhaps the most important yet challenging aspect of running the software is learning when it is safe to stop sampling and summarize the output. Indeed, the WinBUGS manual itself begins with a disclaimer that "MCMC sampling is dangerous," in the sense that this is an algorithm that is converging not to a fixed point (like a likelihood maximization algorithm), but to a *random sample* from a particular posterior *distribution.* Fortunately, WinBUGS offers a variety of helpful convergence plots and diagnostics, so with a little practice with some sample models and datasets, the art and science of applied Bayes-MCMC seems well within the reach of many statisticians currently using SAS, S-Plus, and other mainstream packages.

Finally, we hasten to add that a version of the joint analysis we favor is also fittable in SAS. While the `LIFEREG` and `PHREG` procedures do not accommodate random effects, maximum likelihood estimation of the joint model is possible with the `NLMIXED` procedure; indeed, the aforementioned Web page also includes `NLMIXED` code to fit Models XI and XII above. Estimation of the random effects is via empirical Bayes, with associated standard errors obtained by the delta method. Approximate 95% prediction intervals can then be obtained by assuming asymptotic normality. For the median survival times of Patients 450 and 454, under Model XI we obtained point estimates of roughly 72.3 and 22.4, respectively, in rough agreement with Figure 3. However, the asymmetry of the posteriors in this figure (which are similar to the likelihood, due to our vague priors) suggests traditional confidence intervals based on asymptotic normality and approximate standard errors will not be very accurate. Exact results (and corresponding full posterior inference) as available from Figures 3 and 2 still require the fully Bayesian-MCMC (WinBUGS) approach.

[Received March 2003. Revised September 2003.]

REFERENCES

Abrams, D. I., Goldman, A. I., Launer, C., Korvick, J. A., Neaton, J. D., Crane, L. R., Grodesky, M., Wakefield, S., Muth, K., Kornegay, S., Cohn, D. J., Haris, A., Luskin-Hawk, R., Markowitz, N., Sampson, J. H., Thompson, M., Deyton, L., and the Terry Beirn Community Programs for Clinical Research on AIDS (1994), "Comparative Trial of Didanosine and Zalcitabine in Patients with Human Immunodeficiency Virus Infection Who are Intolerant of or Have Failed Zidovudine Therapy," *New England Journal of Medicine*, 330, 657–662.

Carlin, B. P., and Louis, T. A. (2000), *Bayes and Empirical Bayes Methods for Data Analysis* (2nd ed.), Boca Raton, FL: Chapman and Hall/CRC Press.

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

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis*, Boca Raton, FL: Chapman and Hall/CRC Press.

Goldman, A. I., Carlin, B. P., Crane, L. R., Launer, C., Korvick, J. A., Deyton, L., and Abrams, D. I. (1996), "Response of CD4+ and Clinical Consequences to Treatment Using ddI or ddC in Patients with Advanced HIV Infection," *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 11, 161–169.

Henderson, R., Diggle, P. J., and Dobson, A. (2000), "Joint Modelling of Longitudinal Measurements and Event Time Data," *Biostatistics*, 1, 465–480.

Laird, N. M., and Ware, J. H. (1982), "Random-Effects Models for Longitudinal Data," *Biometrics*, 38, 963–974.

Lin, H., Turnbull, B. W., McCulloch, C. E., and Slate, E. H. (2002), "Latent Class Models for Joint Analysis of Longitudinal Biomarker and Event Process Data: Application to Longitudinal Prostate-Specific Antigen Readings and Prostate Cancer," *Journal of the American Statistical Association*, 97, 53–65.

Pawitan, Y., and Self, S. (1993), "Modeling Disease Marker Processes in AIDS," *Journal of the American Statistical Association*, 88, 719–726.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit" (with discussion and rejoinder), *Journal of the Royal Statistical Society*, Ser. B., 64, 583–639.

Tsiatis, A. A., Degruttola, V., and Wulfsohn, M. S. (1995), "Modeling the Relationship of Survival to Longitudinal Data Measured with Error, Applications to Survival and CD4 Counts in Patients with AIDS," *Journal of the American Statistical Association*, 90, 27–37.

Wang, Y., and Taylor, J. M. G. (2001), "Jointly Modeling Longitudinal and Event Time Data with Application to Acquired Immunodeficiency Syndrome," *Journal of the American Statistical Association*, 96, 895–905.

Wulfsohn, M. S., and Tsiatis, A. A. (1997), "A Joint Model for Survival and Longitudinal Data Measured with Error," *Biometrics*, 53, 330–339.