Biometrics 59, 221–228
June 2003

A Bayesian Semiparametric Joint Hierarchical Model for Longitudinal and Survival Data

Elizabeth R. Brown[*]

Department of Biostatistics, University of Washington, Seattle,
Washington 98195, U.S.A.
[*]*email:* elizab@u.washington.edu

and

Joseph G. Ibrahim

Department of Biostatistics University of North Carolina, Chapel Hill,
North Carolina 27599, U.S.A.

Summary. This article proposes a new semiparametric Bayesian hierarchical model for the joint modeling of longitudinal and survival data. We relax the distributional assumptions for the longitudinal model using Dirichlet process priors on the parameters defining the longitudinal model. The resulting posterior distribution of the longitudinal parameters is free of parametric constraints, resulting in more robust estimates. This type of approach is becoming increasingly essential in many applications, such as HIV and cancer vaccine trials, where patients' responses are highly diverse and may not be easily modeled with known distributions. An example will be presented from a clinical trial of a cancer vaccine where the survival outcome is time to recurrence of a tumor. Immunologic measures believed to be predictive of tumor recurrence were taken repeatedly during follow-up. We will present an analysis of this data using our new semiparametric Bayesian hierarchical joint modeling methodology to determine the association of these longitudinal immunologic measures with time to tumor recurrence.

Key words: Dirichlet process; Joint longitudinal and survival model; Semiparametric Bayes.

1. Introduction

Often in clinical trials where the primary endpoint is time to an event, patients are also monitored longitudinally with respect to one or more biologic endpoints throughout the follow-up period. These endpoints may be immunologic measures in a vaccine trial or may be immunologic and virologic measures in a study of infectious diseases. These longitudinal measures may be associated with survival, but may be difficult to include in a model because they are often incomplete and prone to measurement error. They may also be very diverse and, therefore, difficult to model with known probability distributions. A model that links the hazard to these longitudinal measures that can also incorporate information about missingness and measurement error and is robust to incorrect distributional assumptions on the longitudinal measures is becoming increasingly essential in many applications. To motivate these models, we examine three different settings where such methodology may be advantageous. These are HIV/AIDS trials, cancer vaccine trials, and quality of life studies.

In clinical trials of therapies for diseases associated with human immunodeficiency virus (HIV), immunologic and virologic markers are measured repeatedly over time on each patient. These markers are prone to measurement error and are high within patient variability due to biological fluctuations. Modeling these covariates over time is preferable to using the raw data (Tsiatis et al., 1992; DeGruttola et al., 1993; DeGruttola and Tu, 1994; Tsiatis, DeGruttola, and Wulfsohn, 1995; LaValley and DeGruttola, 1996). Many HIV clinical trials focus on the opportunistic infections (OI) associated with HIV disease where the study endpoint is the time to development of the OI. In these trials, immunologic and virologic markers might be utilized as time-varying covariates. DeGruttola and Tu (1994), Tsiatis et al. (1995), Faucett and Thomas (1996), Wulfsohn and Tsiatis (1997) and Wang and Taylor (2001) have taken approaches to jointly modeling HIV-related outcomes and immunologic measures.

In health related quality of life (HRQOL) studies, subjects are often followed until the occurrence of some event. One method of assessing HRQOL is to administer questionnaires intermittently during follow-up. Often the patients do not complete the questionnaire, leaving all or part of it incomplete. The result is an incomplete measure of quality of life. HRQOL may also be assessed by measuring adverse effects of therapy and how long these effects last; therefore, treatment may affect HRQOL, with some treatments having a greater impact than others. HRQOL may also predict time to event. As a patient's HRQOL measure declines, they may be moving closer in time to the event. Since both disease status and treatment may affect a patient's HRQOL, and HRQOL may be associated with survival, we might want to include this information in a survival model. However, since HRQOL may be measured with error or have missing information, a method that can jointly model this information may give greater insight into the relationship between HRQOL and survival.

Although the methodology we present here can be used in HIV/AIDS or HRQOL studies, throughout we will focus on cancer studies for a clearer focus and exposition. Cancer vaccine trials are becoming increasingly popular due to recent advances in biological research. As well as being less toxic than traditional chemotherapies, these vaccines may induce an antitumor response that may protect the patient against tumor relapse. In cancer vaccine studies, vaccinations are given to patients to raise the patients' antibody levels against the tumor cells. A successful vaccine increases the patient's immune system's antibody production to help eradicate and prevent future tumors. Examining the association between the antibody measures and survival may shed light on the the biological pathways of the disease. However, the antibody measures are prone to measurement error; therefore, the raw data should not be used in a survival analysis. Hence, methods which can model the association between the antibody measures and time to tumor recurrence, while accounting for the error in the antibody measures, is essential.

There can be a great deal of diversity in immune responses between patients, and so we may need more flexibility than a parametric model would allow. For example, in the data set from the melanoma cancer vaccine clinical trial, which we discuss in more detail in Section 3, there is a lot of diversity in the immune response as represented by the IgM titre levels. IgM is an antibody produced by plasma cells and circulates in the blood. The IgM titre level is a measure of the amount of IgM circulating in the blood. Figures 1 gives a clear illustration of the diversity of responses in this population. While many of the patients' IgM levels increase in response to treatment, 45 of the 224 patients from the melanoma study have no increase in IgM titre levels.

Histograms of the IgM measures at different points in time suggest nonnormality in their distributions.

After determining the distributional form of the longitudinal measures, we need an appropriate model to link the event times to these longitudinal covariates. The proportional hazards model relates the hazard to time-dependent covariates (Cox, 1972) by taking

$$\lambda\{t \mid \mathcal{G}(t)\} = \lambda_0(t) f\{\mathcal{G}(t), \theta\},$$

where $f\{\mathcal{G}(t), \beta\}$ is a function of the covariate history specified up to an unknown parameter, or vector of parameters, θ. Usually in biomedical applications, the true covariate history, $\mathcal{G}(t)$, is not available; however, we may have observations, $Y(t)$, representing some function of the true covariate, $g(t)$, to which we refer here as the *trajectory function,* which is measured with error.

Statistical packages are widely available to perform survival analyses with time-dependent covariates. However, if the covariates are measured with error, the analysis becomes more complex. Simply including the raw measurements in the survival analysis leads to bias (Prentice, 1982).

DeGruttola and Tu (1994) proposed an approach to extend the general random effects model to the analysis of longitudinal data using informative censoring. A similar model for informatively censored longitudinal data was proposed by Schluchter (1992). DeGruttola and Tu (1994) extend this method to jointly model survival times and disease progression using normally distributed random effects. Assuming that these two outcomes are independent given the random effects, the joint likelihood is easily specified. The maximum likelihood estimates of the unknown parameters are obtained using the EM algorithm. Tsiatis et al. (1995) present a computationally straightforward and easy to implement approach which reduces the bias in a model with time-varying covariates measured with error. They use asymptotic approximations to show consistency of estimates obtained by modeling longitudinal data separately, then plug the estimates into a Cox proportional hazards model. Estimation and inference for the survival model is carried out using the partial likelihood theory developed by Cox (1972, 1975). Wulfsohn and Tsiatis (1997) and Faucett and Thomas (1996) both assume a proportional hazards model for survival conditional on the longitudinal measure. They then obtain a joint model by multiplying this conditional survival distribution times the distribution of the longitudinal measures. Wulfsohn and Tsiatis (1997) take a frequentist approach to fitting this model, while Faucett and Thomas (1996) take a Bayesian approach using Gibbs sampling. Although this is an improvement over the two-stage model presented by Tsiatis et al. (1995), it still may not allow enough flexibility in the model for the longitudinal measures. Additional Bayesian joint modeling approaches were developed by Ibrahim, Chen, and Sinha (2001), who extended the longitudinal model to the multivariate case, and Wang and Taylor (2001), who presented a more flexible approach to modeling the longitudinal measures. Approaching the joint modeling problem from the Bayesian perspective is more natural and straightforward. It avoids the many complicated approximations required by the frequentist approach. Yet, with noninformative priors, we can still get maximum likelihood–type estimates.

However, none of these approaches address many of the issues that are often present when modeling data from a highly heterogeneous population. They all assume a parametric trajectory function, except for Wang and Taylor (2001) who include an integrated Ornstein-Uhlenbeck (IOU) stochastic process in the trajectory model. They build a linear trajectory using random intercepts and a fixed slope. The IOU process allows the trajectory to vary around the parametric line. This formulation allows for deviation from a parametric fit; however, it may not allow for enough between-patient variability. We will take a different approach to increasing the robustness of the longitudinal model. Instead of allowing variation around the parametric trajectory, we will allow the parameters of the trajectory to come from unspecified distributions. This will allow more flexibility and robustness in the model. Another problematic issue arises when observations from individual trajectories do not have the same distribution. This can occur as patients reach the endpoint and leave the study. The longitudinal observations of the remaining patients should not be assumed to have the same distribution as those patients who have left the study. Also, the usual assumption that observations are normally distributed may, in fact, not be true. They may come from a mixture of normal distributions, a distribution with heavier tails, or some other distribution which cannot be easily specified. All these possibilities may be difficult to model parametrically. Often, we are not even interested in the parameter estimates for the longitudinal covariate and therefore a strong distributional assumption is typically not necessary. In this article, we will present a flexible and robust semiparametric model to address these concerns.

2. A New Semiparametric Model

In this section, we review the background and setup for our new model. First, we specify the likelihood for the joint model. After a brief review of the Dirichlet process (DP) prior, we will show how specifying a DP prior on the parameters of the longitudinal trajectory results in a novel semiparametric joint survival and longitudinal model.

2.1 *Likelihood*

Since we are interested in determining the impact of the longitudinal measures on the survival outcome, we construct the joint likelihood as the product of the time-to-event likelihood conditional on the longitudinal measure multiplied by the likelihood of the longitudinal measure. We build the likelihood for the longitudinal measure by constructing individual trajectories for each patient. Each subject has m_i longitudinal measures, denoted as $y_{ij}, j = 1, \ldots, m_i$, taken at times t_{ij}. The longitudinal model is given by

$$y_{ij} = \psi_\beta(t_{ij}) + \epsilon_{ij},$$

$$\epsilon_{ij} \sim N(0, \sigma^2),$$

where $\psi_\beta(t_{ij})$ is the trajectory function. An individual's contribution to the likelihood is then given by

$$f(Y_i \mid \psi_\beta, \sigma) \propto \frac{1}{\sigma^{m_i}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{m_i}\{y_{ij} - \psi_\beta(t_{ij})\}^2\right\},$$

where $Y_i = (y_{i1}, \ldots, y_{im_i})$ is an individual's vector of longitudinal responses.

The trajectory function can take on many forms. In this article, we will use a quadratic form given by

$$\psi_\beta(t_{ij}) = \beta_{0i} + \beta_{1i}t_{ij} + \beta_{2i}t_{ij}^2, \qquad (1)$$

which reflects an initial increase in antibody levels in response to cancer vaccine therapy, followed by a decline as the treatment begins to wear off. This is a common response of immunologic measures to therapy; therefore, we may expect this model to be appropriate in many other settings, not just cancer vaccine trials. Tsiatis et al. (1995) also used a quadratic trajectory when modeling CD4 counts. The hazard function is expressed as

$$h(t \mid Y) = \lambda(t)\exp\{\gamma\psi_\beta(t) + x'\alpha\}, \qquad (2)$$

where γ is a scalar parameter linking the trajectory to the hazard function, $\lambda(t)$ is the baseline hazard, and α is a parameter vector linking a vector x of baseline covariates to the failure time. These covariates may be categorical, such as treatment, or continuous, such as age. Building the hazard in this manner sets the stage for modeling survival and longitudinal data simultaneously in a proportional hazards setting. This is different than previous approaches considered by Tsiatis et al. (1995), who model the longitudinal measures first and then plug them into the proportional hazards model, and different than DeGruttola and Tu (1994), who model both outcomes jointly, but do not use a proportional hazards structure and place strong distributional assumptions on the survival outcome.

The specification of the hazard in (2) leads to the following distribution for the survival component given the trajectory function:

$$f(s_i, \nu_i \mid Y_i) = \lambda(s_i)^{\nu_i}\exp[\nu_i\{\gamma\psi_\beta(s_i) + x_i'\alpha\}] \times \exp\left\{-\int_0^{s_i}\lambda(u)e^{\gamma\psi_\beta(u) + x_i'\alpha}\,du\right\},$$

where s_i is the survival time for the ith subject and ν_i is the censoring indicator for subject i.

If we assume the baseline hazard function is piecewise constant so that

$$\lambda(u) = \lambda_j, \quad u_{j-1} \le u < u_j, \quad j = 1, \ldots, J,$$

where u_j, $j = 0, \ldots, J$, define the intervals for $\lambda(u)$ and are selected based on the quantiles of the observed event times, then the cumulative hazard,

$$\int_0^{s_i}\lambda(u)e^{\gamma\psi_\beta(u) + x_i'\alpha}\,du,$$

can be rewritten as

$$e^{x_i'\alpha}\sum_{j=1}^{J}H_{ij}(\beta, \gamma, \lambda), \qquad (3)$$

where

$$H_{ij}(\beta, \gamma, \lambda) = I\{s_i > u_{j-1}\}\lambda_j\int_{u_{j-1}}^{\min(u_j, s_i)}e^{\gamma\psi_\beta(u)}\,du \qquad (4)$$

and $I\{s_i > u_j\}$ is an indicator function which equals 1 if the event time occurs in or later than the jth interval, and 0 otherwise. The integral in (4) does not have an analytical solution when the trajectory is quadratic. Instead, we use GNU Scientific Library (GSL) (Galassi, Gough, and Jungman, 2001) to perform nonadaptive Gauss-Kronrod integration to numerically calculate the integral.

We can now write subject i's contribution to the joint likelihood function as

$$f(Y_i, s_i, \nu_i) = \lambda(s_i)^{\nu_i}\exp\left\{\nu_i\{\gamma\psi_\beta(s_i) + x_i'\alpha\} - e^{x_i'\alpha}\sum_{j=1}^{J}H_{ij}(\beta, \gamma, \lambda)\right\} \times \frac{1}{(2\pi\sigma^2)^{m_i/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{m_i}\{y_{ij} - \psi_\beta(t_{ij})\}^2\right\}. \qquad (5)$$

2.2 *Mixture of Dirichlet Process Model*

The parametric distributions of the elements of $\beta_i = (\beta_{0i}, \beta_{1i}, \beta_{2i})'$ in the trajectory function (1) are difficult to check and justify. A nonparametric specification of the distribution of the β_i's allows for a more flexible modeling scheme and, therefore, may be more desirable and realistic. As shown in Section 1, we cannot be sure that the β_i's all come from the same distribution or that our distributional assumption is correct. Also, the distribution of the β_i's may not remain constant over time and, in many settings, there is evidence of nonnormality in the data. To overcome these concerns, we will consider a Dirichlet process prior for the β_i's, in order to relax the distributional assumption on these parameters and therefore relax the assumptions about the trajectory function.

2.2.1 Dirichlet process prior. The Dirichlet process (DP) prior is a natural approach to building a semiparametric model. Because we can easily obtain posterior estimates using standard MCMC approaches such as Gibbs sampling (Gelfand and Smith, 1990), the model is also easy to fit. For a review of Dirichlet processes and mixture of Dirichlet processes (MDP) models, see Escobar (1994) and Escobar and West (1998).

We can use the MDP model to build a semiparametric random effects model (Bush and MacEachern, 1996; Ibrahim and Kleinman, 1998; Kleinman and Ibrahim, 1998) that can easily be incorporated into the joint model developed in Section 2.1. This new joint model allows a more flexible and robust approach to examining the relationship between the longitudinal measures and survival time.

We specify noninformative proper priors on the parameters in the likelihood (5). Specifically, we use the conjugate prior for the underlying hazard, $\lambda_k \sim \Gamma(a_k, b_k), k = 1, \ldots, J$, and for the error variance, $\sigma^2 \sim IG(a, b)$, where $\Gamma(a, b)$ denotes the gamma distribution with shape parameter a and scale parameter b, and $IG(a, b)$ denotes the inverse gamma distribution with shape parameter a and scale parameter b. We specify a normal prior $N(\mu_\gamma, \sigma_\gamma^2)$ for γ, the parameter linking the trajectory to the hazard. A normal prior is also appropriate for the baseline covariate parameter vector, $\alpha \sim N(\mu_\alpha, \Sigma_\alpha)$.

We can relax the distributional assumption on the β_i's in the model of the trajectory function in the joint model (5) by assuming they come from some unspecified distribution G and place a Dirichlet process prior on this distribution given by

$$\beta_i \sim G, \quad G \sim DP(MG_0) \quad \text{and} \quad G_0 = N_3(b_0, V_0), \qquad (6)$$

where $N_3(a, b)$ is the 3-dimensional multivariate normal distribution with mean vector a and variance-covariance matrix b. The full conditional of β_i is given by

$$p(\beta_i \mid \beta_{-i}, Y_i, s_i, \nu_i, rest) \propto \sum_{j \neq i} q_j \delta(\beta_j) + q_0 g_0(\beta_i \mid Y_i, s_i, \nu_i), \qquad (7)$$

where

$$g_0(\beta_i \mid Y_i, s_i, \nu_i) \propto g_0(\beta_i) f(Y_i, s_i, \nu_i \mid \beta_i),$$

$$g_0(\beta_i) = N_3(b_0, V_0),$$

rest denotes the rest of the model parameters, and $q_j \propto f(Y_i, s_i, \nu_i \mid \beta_j)$ is the likelihood evaluated with $\beta_i = \beta_j$. We use the notation $f(Y_i, s_i, \nu_i \mid \beta_i)$ and $f(Y_i, s_i, \nu_i \mid \beta_j)$ to distinguish between subject i's contribution to the likelihood evaluated with its trajectory parameter vector, β_i, and subject i's contribution to the likelihood evaluated with subject j's trajectory parameter vector, $\beta_i = \beta_j$. Also,

$$q_0 \propto M \int \cdots \int f(Y_i, s_i, \nu_i \mid \beta_i)\, dg_0(\beta_i). \qquad (8)$$

We complete our Bayesian hierarchical model by specifying prior distributions for b_0 and V_0. The priors should be selected with care in order to ensure that the model is identifiable. In this setting, we use $b_0 \sim N(m, v)$ and $V_0^{-1} \sim Wishart(S_0, \nu_v)$, where $Wishart(a, b)$ denotes the Wishart distribution with scale matrix a and b degrees of freedom.

The parametric model can be obtained as a special case from the MDP model by letting $M \to \infty$. In this case, the posterior distribution for the parameters of the trajectory function are then given by

$$p(\beta_i \mid \beta_{-i}, Y_i, s_i, \nu_i, rest) \propto g_0(\beta_i) f(Y_i, s_i, \nu_i \mid \beta_i).$$

We use the Gibbs sampler (Gelfand and Smith, 1990) to obtain estimates from the posterior. Equation (7) lends us a straightforward approach to sample from the posterior of the MDP model when q_0 has a closed-form solution. This is very often the case in models where the base measure is conjugate to the likelihood. However, in more complex models, such as the model we present, it is not possible to formulate conjugate base measures to the likelihood. No matter what base measure we choose, q_0 will have no closed-form solution. This presents a dilemma which can only be solved computationally. We must use either a numeric method to approximate q_0 (8) or find an alternative method for sampling from the posterior. MacEachern and Müller (1998) and Neal (2000) present similar algorithms for sampling from the conditional posterior of an MDP model when q_0 does not have a closed-form solution. For this implementation, we use Neal's Algorithm 8. This is summarized in the Gibbs sampling scheme presented in the Appendix.

The value of M takes on different meanings in different models or data sets. In one model, a specified value of M may result in a given number of clusters, but the same value of M for another model may result in a different number of clusters. Therefore, the value of M itself should not be interpreted as a measure of how "nonparametric" the model is. To get an estimate of the amount of mixing in the model, it is better to examine the posterior distribution of the number of clusters. It is also important to note that very small values of M may lead to a model indistinguishable from a fixed effects model (e.g., $y_{ij} = b_{00} + b_{01}t_{ij} + b_{02}t_{ij}^2 + \epsilon_{ij}$). Small values of M can reduce the sampling to one cluster with very low probability of an observation moving out of this cluster. We may also specify a prior distribution on M and sample M from the posterior distribution using the Gibbs sampler. Escobar and West (1995) derive a sampling scheme for M with a gamma prior, $M \sim \Gamma(M_a, M_b)$, which we will use in this model. The posteriors and sampling scheme for the joint model with Dirichlet process priors are presented in the Appendix. The program to fit the model was written in C and is available from the authors.

3. Application

To illustrate the joint semiparametric model, we examine data from an intergroup trial of the Eastern Cooperative Oncology Group, the Southwest Oncology Group, and Cancer and Leukemia Group B. Study E1694 (Kirkwood et al., 2001) was designed determine if GM2, a ganglioside which is serologically well-defined to be a melanoma antigen, coupled to keyhole limpet hemocyanin (KLH) and administered with the adjuvant QS-21 (altogether known as the GMK vaccine) is superior to interferon-α2b with respect to relapse-free survival (time to tumor recurrence) and overall survival, with a secondary goal to determine the association of preexisting and vaccine-induced IgM and IgG antibodies with relapse-free and overall survival. IgG, as with IgM, is also an antibody produced by plasma cells and circulates in the blood. IgM and IgG measures were taken at baseline, days 29 and 85, and at 6, 9, 12, 18, and 24 months. Some patients had an additional measure taken at the time of relapse.

For this analysis, we used IgM as the longitudinal measure because it is believed to respond more quickly to therapy than IgG. We conducted an analysis using patients from the GMK vaccine treatment arm only and included only those patients with three or more longitudinal observations, using relapse-free survival as the primary time-to-event endpoint. Of the 224 patients included in this analysis, 76 had observed event times. 139 patients had 3 longitudinal measures, 59 had 4, 21 had 5, and 5 had 6 longitudinal measures. Figure 1 shows the observed longitudinal measures plotted against time for the 224 patients included in the analysis. The heterogeneity of the patients' immunological responses is apparent in this plot. The distributions of the mean, median, and maximum of the log(IgM) measures of the patients do not appear to conform to any known distributions and may be best modeled without parametric distributional assumptions.

We fit this data to the new model presented in Section 2 with the underlying hazard function, $\lambda(t)$, estimated using $J = 8$ intervals. The time points defining the intervals were taken to be quantiles of the observed survival times. The priors for the parameters in (5) and (6) were taken as $b_0 \sim N_3\{(0, 4, 3)', \text{diag}(1, 10, 10)\}$, $V_0^{-1} \sim Wishart(I, 5)$, $\gamma \sim N(0, 20)$, $\lambda_k = \Gamma(0.1, 0.1), k = 1, \ldots, J$, $\sigma^2 \sim IG(0.01, 0.01)$, and $M \sim \Gamma(1, 1)$.

The posterior median of the precision parameter, M, is 0.55 (0.07, 2.13), corresponding to a median of $k = 3$ clusters. We also performed the analysis with $M \sim \Gamma(10, 10)$ and $M \sim \Gamma(0.1, 0.1)$ and obtained very similar results to those we present here. The estimated association of the longitudinal measures and relapse-free survival (γ) is greater in the parametric model than in the MDP model. This suggests the parametric model may overestimate the association between the IgM immune response and relapse-free survival. A more intuitive approach to understanding the relationship of log(IgM) and relapse-free survival is through the hazard ratio. A hazard ratio of 1 would indicate no effect of the IgM level on the hazard over a one-unit change in log(IgM). A hazard ratio <1 ($\gamma < 0$) indicates increased protection against relapse as the value of log(IgM) increases. For a one-unit change in log(IgM), the estimated hazard ratio for the MDP model is 0.99 with a 95% credible interval equal to (0.88, 1.11), which suggests that there is no impact of IgM on the time to tumor recurrence. The estimated hazard ratio with 95% credible interval for the parametric model is 0.91 (0.77, 1.03). Figure 2 gives more insight into the posterior distribution of γ. The MDP model has decreased variability in the in the posterior distribution of γ as well as smaller estimates in magnitude. While the parametric model suggests an association between IgM and time to tumor recurrence, the MDP model does not.

The posterior density estimates of the hyperparameters of the trajectory function (b_{00}, b_{01}, b_{02}) are shown in Figure 3. There is greater variability in the estimates from the MDP model. The 95% credible intervals for b_{02} do not contain 0 for either model suggesting that the quadratic assumption is correct. Since we included only one arm of the trial, we cannot make statements about the efficacy of the GMK vaccine or the relationship of the IgM levels to treatment and their combined effect on relapse-free survival. However, we can say that for the patients being administered the GMK vaccine, their IgM levels do not appear to have a strong association with relapse-free survival. It is important to keep in mind that only patients with 3 or more observations were included in this analysis. These results cannot therefore be extrapolated to all of the

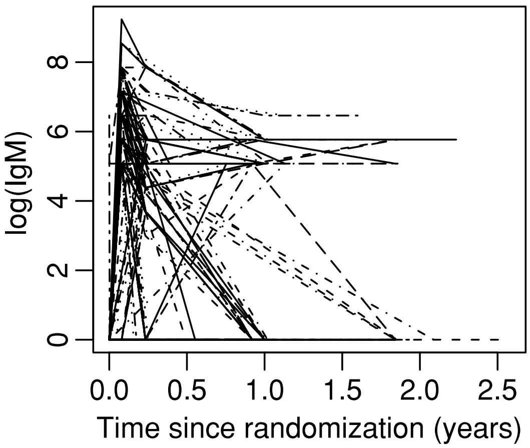

Figure 1. Observed trajectories of IgM for all 224 patients.

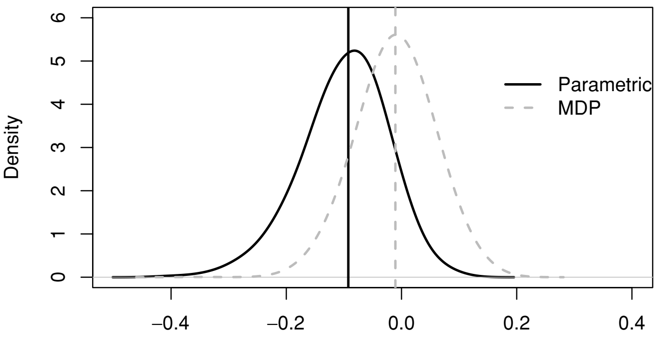

Figure 2. Posterior densities of γ for study E1694 for the MDP model (dashed line) and the parametric model (solid line). The vertical lines represent the posterior medians.

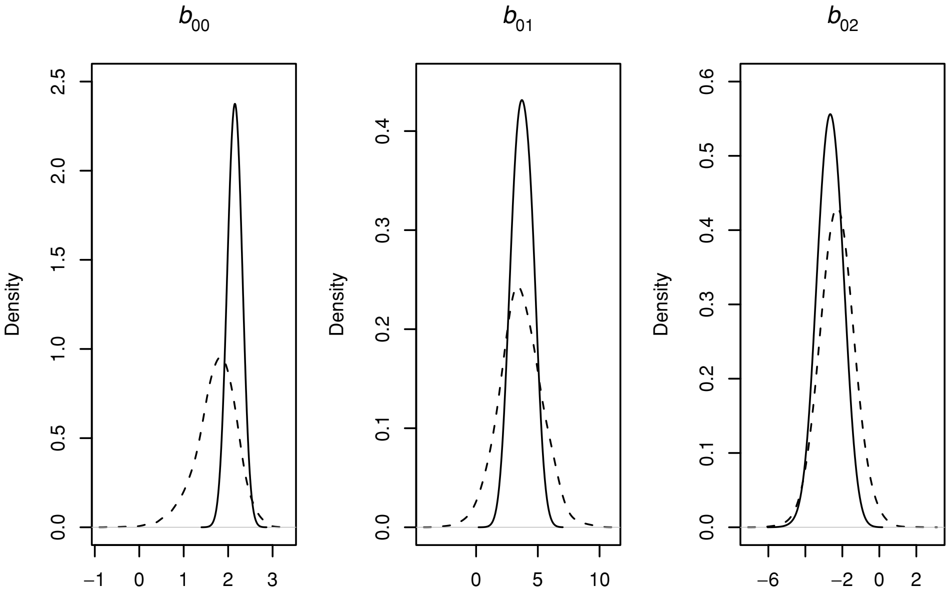

Figure 3. Posterior density estimates of $b_0 = (b_{00}, b_{01}, b_{02})'$ from the MDP model (dashed line) and the parametric model (solid line).

patients receiving GMK in this trial. Also, many patients do not have longitudinal measures at later time points, making it difficult to fit a longitudinal model that can be guaranteed to accurately estimate the trajectory at later event times. This may result in an underestimate of the relationship between IgM and time to tumor recurrence. The plots in Figure 4 of fitted trajectories for 6 patients clearly show a quadratic trend in the IgM measures over time. We also see that in the

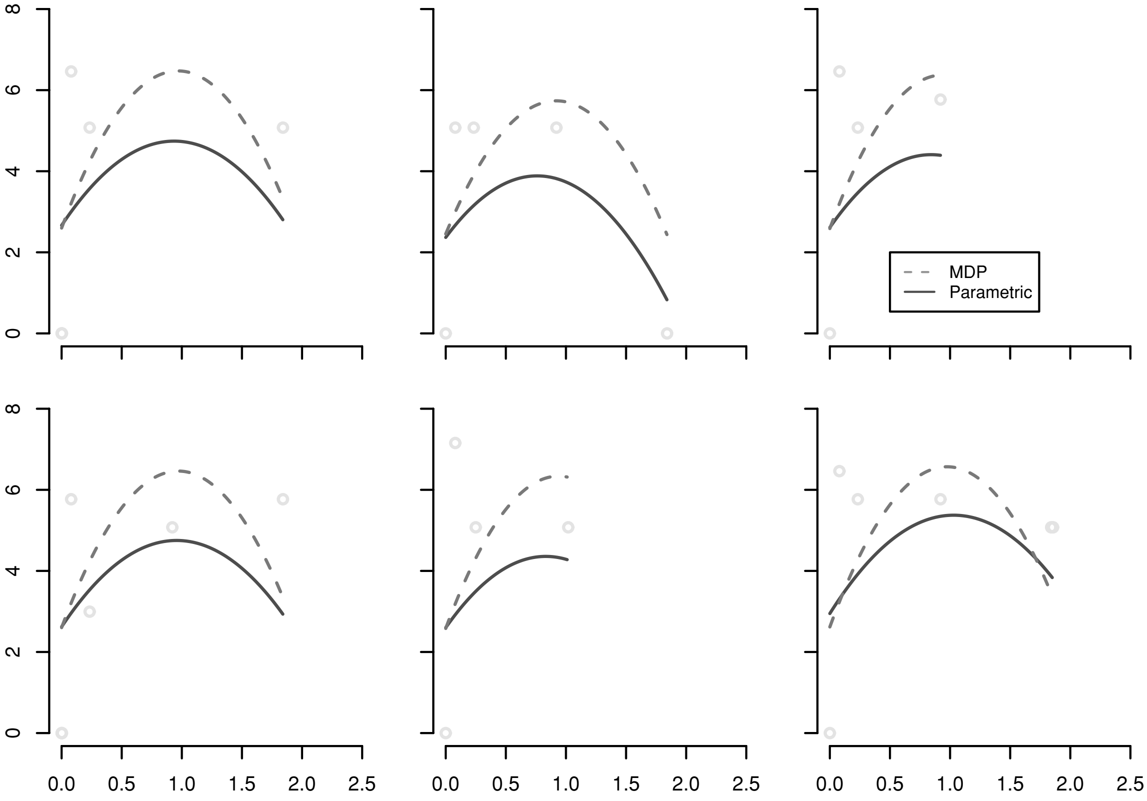

Figure 4. Sample trajectories and their fits with $J = 8$ for 6 patients. The circles represent the observed data.

MDP model, the fitted trajectory is closer to the observed values.

4. Discussion

With more studies being conducted that repeatedly take measures over time in an effort to evaluate a patient's health status or risk for some event, a joint modeling approach is essential. With the ready availability of powerful desktop computers, the obstacles to fitting these complicated models are easily overcome and more complicated joint models are computationally feasible. The MDP joint model which may have been computationally infeasible until recently is now attainable. With this model, we have a flexible and robust approach to fitting the longitudinal measures in the joint model. Our model is also useful when there is uncertainty about the distributional assumptions. The new MDP joint model also works well in a real data setting. This model could also be used to validate a parametric model. If, in fact, the MDP model shows no improvement over the parametric model, we can be more confident that our parametric assumptions are valid. The trajectory function easily could be extended to incorporate other covariates that may help to model the trajectory more clearly. For example, if the trajectory is expected to be different across treatments, it may be appropriate to include treatment as a covariate in the trajectory model.

We have presented an appealing and novel approach to extending the joint longitudinal and survival model to the semiparametric case. The MDP joint model relaxes distributional assumptions on the parameters of the trajectory function and can improve estimates in cases where parametric distributional assumptions are inappropriate. It is easily implemented using Gibbs sampling. The MDP model can also be used as a validation tool for the parametric model.

Acknowledgements

The authors wish thank the editor, the associate editor, and the two referees for their helpful comments and suggestions which have led to an improvement in this article. This work was partially supported by NIH grants CA 70101 and CA 74015.

Résumé

Dans cet article, nous proposons un nouveau modèle semi paramétrique bayésien hiérarchique pour la modélisation simultanée de données longitudinales et de données de survie. Nous nous affranchissons de toute hypothèse sur les distributions pour le modèle longitudinal en utilisant un processus de Dirichlet, comme distribution a priori, pour les paramètres qui le définissent. La distribution a posteriori qui en découle pour ces paramètres est libre de toute contrainte paramétrique, ce qui permet d'obtenir des estimations plus robustes. Les approches de ce type deviennent de plus en plus essentielles pour beaucoup d'applications, comme par exemple dans les essais de vaccination contre le HIV et le cancer, où les réponses de patients au traitement sont extrêmement variables et ne peuvent être aisément modélisées à partir de distributions connues. Nous présentons un exemple à partir d'un essai de vaccination dans le cancer, pour lequel la durée étudiée est le délai jusqu'à la récidive de la tumeur. Des mesures immunologiques, reconnues pour être prédictives des récidives tumorales, ont été régulièrement recueillies pendant la période de suivi. L'étude avait pour but d'étudier l'association des mesures immunologiques répétées avec le délai de récidive. Nous présentons les résultats de cette analyse des données faite à partir de ce nouveau modèle semi paramétrique bayésien hiérarchique.

References

Bush, C. A. and MacEachern, S. N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika* **83,** 275–285.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62,** 269–276.

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

DeGruttola, V., Wulfsohn, M., Fischl, M., and Tsiatis, A. (1993). Modeling the relationship between survival and CD4 lymphocytes in patients with AIDS and ARC. *Journal of Acquired Immune Deficiency Syndromes* **6,** 359–365.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet Process prior. *Journal of the American Statistical Association* **89,** 268–277.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90,** 577–588.

Escobar, M. D. and West, M. (1998). Computing nonparametric hierarchical models. In *Practical Nonparametric and Semiparametric Bayesian Statistics.* Lecture Notes in Statistics, Volume 133. New York: Springer-Verlag.

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

Galassi, M., Gough, B., and Jungman, G. (2001). *GNU Scientific Library Reference Manual.* Bristol, U.K.: Network Theory.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85,** 398–409.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41,** 337–348.

Ibrahim, J. G. and Kleinman, K. P. (1998). Semiparametric Bayesian methods for random effects models. In *Practical Nonparametric and Semiparametric Bayesian Statistics.* Lecture Notes in Statistics, Volume 133. New York: Springer-Verlag.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian methods for joint modelling of longitudinal and survival data with applications to cancer vaccine studies.* Technical report, Harvard University, Cambridge, Massachusetts.

Kirkwood, J. M., Ibrahim, J., Sosman, J. A., Sondak, V. K., Agarwala, S. S., Ernstoff, M. S., and Rao, U. (2001). High-dose interferon alfa-2b significantly prolonged relapse–free and overall survival compared with the GM2-KLH/QS-21 vaccine in patients with resected stage IIB–III melanoma: Results of intergroup trial E1694/S9512/C509801. *Journal of Clinical Oncology* **19,** 2370–2380.

Kleinman, K. P. and Ibrahim, J. G. (1998). A semiparametric Bayesian approach to the random effects model. *Biometrics* **54,** 921–938.

LaValley, M. and DeGruttola, V. (1996). Models for empirical Bayes estimators of longitudinal CD4 counts. *Statistics in Medicine* **15,** 2289–2305.

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet Process models. *Journal of Computational and Graphical Statistics* **7,** 223–238.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9,** 249–265.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69,** 331–342.

Schluchter, M. D. (1992). Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine* **11,** 1861–1870.

Tsiatis, A. A., Dafni, U., DeGruttola, V., Propert, K., Strawderman, R. L., and Wulfsohn, M. (1992). The relationship of CD4 counts over time to survival in patients with AIDS: Is CD4 a good surrogate marker? In *AIDS epidemiology: Methodological issues,* N. Jewell, K. Kietz, and V. Farewell (eds), 256–274. Boston: Birkhäuser-Boston.

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 modelling longitudinal and event time data with application to AIDS studies. *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.

Received July 2001. *Revised October* 2002.
Accepted October 2002.

Appendix

We use Gibbs sampling to sample from the joint posterior distribution of the parameters: β, γ, λ, b_0, σ^2, and V_0. The joint posterior does not have a closed form; however, given that the conditional posteriors either have a closed form or are log-concave and can be sampled using adaptive rejection sampling (ARS) (Gilks and Wild, 1992), implementation of the Gibbs sampler is straightforward. Let D denote the data and *rest* denote the remaining parameters. Then at each iteration of the Gibbs sampler, we proceed as follows:

1. Sample $[\phi_j \mid rest, D]$ from $\log\{p(\phi_j \mid rest, D)\} \propto \sum_{i=1}^{n} \log\{p(\beta_i | rest, D\} I\{z_i = j\}, j = 1, \ldots, k$, using ARS, where $p(\beta_i \mid rest, D)$ is the conditional posterior of β_i from the parametric model.
2. Sample $[b_0 \mid rest, D] \sim$ $N_3\{(kV_0^{-1} + C_1^{-1})^{-1}(V_0^{-1} \sum_{i=1}^{k} + C_1^{-1} C_0, (kV_0^{-1} + C_1^{-1})^{-1}\}$.
3. Sample $[V_0^{-1} \mid rest, D] \sim$ $Wishart[\{S_0^{-1} + \sum_{i=1}^{k} (\beta_i - b_0)(\beta_i - b_0)'\}^{-1}, \nu_v + k]$.
4. Sample $[z \mid rest, D]$ and set $\beta_i = \phi_{z_i}$ using Neal's algorithm.
5. Sample $[(1/\sigma^2) \mid rest, D] \sim$ $\Gamma(a_\epsilon + \sum_{i=1}^{n} m_i/2, b_\epsilon + \sum_{i=1}^{n} \sum_{j=1}^{m_i} \{Y_{ij} - \psi_\beta(t_{ij})\}^2/2)$.
6. Sample $[\gamma \mid rest, D]$ from $\log\{p(\gamma \mid \cdot)\} \propto \gamma \sum_{i=1}^{n} \nu_i \psi_\beta(s_i) - \sum_{i=1}^{n} e^{x'_i\alpha} \sum_{j=1}^{J} H_{ij}(\beta, \gamma, \lambda) - (\gamma^2 - 2\mu_\gamma \gamma)/(2\sigma_\gamma^2)$ using ARS.
7. Sample $[\lambda_j \mid rest, D] \sim$ $\Gamma\{n_j + a_j, b_j + \sum_{i=1}^{n} e^{x'_i\alpha} H_{ij}(\beta, \gamma, \lambda)\}, j = 0, \ldots, J$.
8. Sample $[\alpha \mid rest, D]$ from $\log\{p(\alpha \mid \cdot)\} \propto \sum_{i=1}^{n} \nu_i x'_i \alpha - \sum_{i=1}^{n} e^{x'_i\alpha} \sum_{j=1}^{J} H_{ij}(\beta, \gamma, \lambda) - (1/2)(\alpha - \mu_\alpha)' \Sigma_\alpha^{-1} (\alpha - \mu_\alpha)$.
9. Sample M in two steps.
   (a) Sample the latent variable η from $[\eta \mid k, M] \sim \text{Beta}(M + 1, N)$, where $\text{Beta}(a, b)$ is the beta distribution.
   (b) Sample M from $[M \mid \eta, k] \sim$ $\pi_\eta \Gamma\{M_a + k, M_b = \log(\eta)\} + (1 - \pi_\eta)\Gamma\{M_a + k - 1, M_b - \log(\eta)\}$ where $(\pi_\eta)/(1 - \pi_\eta) = (M_a + k - 1)/[N\{M_b - \log(\eta)\}]$.