Bayesian Estimation of the Proportion of Treatment Effect Captured by Longitudinal Surrogate Markers

Short title: Longitudinal Surrogate Marker Evaluation

Mary Kathryn Cowles
Department of Statistics and Actuarial Science
241 SH
University of Iowa
Iowa City, Iowa 52242
Phone: 319-335-0727
Fax: 319-335-3017
kcowles@stat.uiowa.edu

Bayesian Estimation of the Proportion of Treatment Effect Captured by Multiple Longitudinal Surrogate Markers

SUMMARY

"Surrogate markers" or "surrogate endpoints" in clinical trials are biological measurements or events observable earlier than the clinical endpoints (such as death) that are actually of primary interest. In many clinical trials, such markers are measured repeatedly on each patient. The "proportion of treatment effect captured" by a surrogate endpoint (PTE) is a measure intended to address the question of whether trials based on a surrogate endpoint reach the same conclusions as would have been reached using the true endpoint.

We extend the method of Cowles (2002) for Bayesian estimation of the PTE to the case in which the true endpoint is time to a clinical event and one or more continuous-valued markers are measured on a fixed schedule. Either an accelerated-failure-time model or a proportional-hazards model may be used for the time-to-event data, and the metric of interest for one or more of the markers may be area under the curve (cumulative exposure). We use the software package WinBUGS to fit our models to viral-load, CD4-count, and disease-progression data from an AIDS clinical trial.

KEYWORDS: accelerated failure time model, area under the curve, generalized linear model, numerical integration, proportional hazards model, time-varying covariates

1 Introduction

To reduce the size and duration of clinical trials, laboratory values that can be measured early and often are commonly used as primary endpoints instead of clinical events that more directly measure treatment effect on patients' health. Laboratory data used in this way are called "surrogate markers" or "surrogate endpoints." In particular, the Federal Drug Administration grants initial approval to new drug regimens in AIDS based on clinical trials with viral load (possibly in combination with CD4 count) as their primary endpoints. CD4 cells are the immune-system cells that are infected by and destroyed by HIV. The CD4 count is the number of CD4 cells per cubic millimeter of blood, and the CD4 percent is the percentage of white blood cells that are of the CD4 type. Viral load is the number of copies of HIV genetic material per milliliter of plasma. Both CD4 count data and viral load data are subject to measurement error, composed of short-term biological fluctuation as well as inevitable inaccuracy in laboratory procedures. Furthermore, longitudinal marker data can be measured only when clinical-trial participants attend their scheduled visits and thus are subject to nonrandom missingness that may introduce bias into estimation of patient-specific marker trajectories and treatment-group averages.

When both marker data and time-to-event data are collected in a clinical trial, several research questions may be of scientific interest, including (1) the treatment effect on time to clinical events, (2) the treatment effect on the marker trajectories, (3) the relationship between marker trajectories and time to clinical events, and (4) the relationship between treatment effect on marker trajectories and treatment effect on time to clinical events.

1.1 Measurement-error models for longitudinal surrogate markers

A number of papers have presented models that address the issue of measurement error in marker data. Lange et al. (1992) pioneered the use of MCMC methods to fit Bayesian normal hierarchical models to transformed longitudinal CD4 count data. They used simple linear random effects models with a changepoint.

Taylor et al. (1994) proposed a model for the true latent process underlying observed CD4-count measurements. In addition to fixed and random effects, their model included an additive two-parameter integrated Ornstein-Uhlenbeck (OU) process. Sy et al. (1997) extended this model to the case of multivariate longitudinal data by incorporating a multivariate integrated OU process. Applying their model to CD4-count and beta-2-microglobulin data from the Multicenter AIDS Cohort Study, they found that these markers appeared to follow a special case of the integrated OU process, namely

bivariate Brownian motion.

Although these papers aim to model an underlying "measurement-error-free" trajectory of longitudinal marker data, their estimates still may be biased due to nonrandom dropout and other nonignorable missingness. Nonrandom dropout occurs when a patient dies or becomes too ill to attend clinic visits. Intermittent missingness may also be nonignorable (Laird, 1988) if sicker patients are more likely than healthier patients to miss visits. Thus patients with fewer observations of marker data may tend to have poorer marker values (e.g. higher viral load or lower CD4) than patients with more observed data.

1.2 Jointly modeling longitudinal marker data and time-to-event data

Substantial recent statistical research has involved joint models for longitudinal marker data and event-time data. Faucett and Thomas (1996, p. 1664) describe the advantages of such joint models (referring to the marker as the "covariate"):

> In a survival analysis setting, where the covariate of interest is time-dependent, either the entire history of the covariate for every subject, or, minimally, measurements of the covariate at each time of disease occurrence for all subjects in the corresponding risk set, are necessary. This extensive measurement of covariates is rarely, if ever, executed and the values obtained are typically subject to measurement error. By modelling the covariates over time, we can enhance the survival analysis since we can interpolate covariate values between the observed measurements to the specific times of disease occurrence, with use of the entire covariate history of the subjects. Modelling the covariate also allows adjustment for covariate measurement error, which is known to result in biased estimates of relative risk parameters (Prentice, 1982). By accounting for measurement error, the standard error of the relative risk estimate will reflect correctly the uncertainty in the measurements of the covariate. Conversely, utilizing the survival data in the covariate tracking model will yield improved covarate tracking parameter estimates by allowing adjustment for informative right censoring of the repeated measurements by the disease process.

1.2.1 Frequentist joint models

Pawitan and Self (1993) used a Weibull accelerated failure time (AFT) model for time from HIV infection to diagnosis of AIDS in combination with a random effects model for trajectories of CD4 counts. Because they factored their joint model into the marginal Weibull AFT model times the conditional model for

CD4 count given time to event, they did not have to deal with time-varying covariates in their AFT model.

DeGruttola and Tu (1994) jointly modeled CD4 counts and survival times using normally-distributed random effects; survival times, possibly after suitable transformation, were assumed to follow a normal distribution.

Wulfsohn and Tsiatis (1997) employed the EM (expectation-maximization) algorithm to estimate simultaneously the parameters of a normal random effects model for longitudinal CD4 counts and a Cox proportional hazards (PH) model for survival time. In the latter, the log of risk of death was assumed to depend linearly on the true, unobserved value of CD4 count.

Henderson et al. (2000) proposed a class of joint models in which a latent bivariate stationary Gaussian stochastic process underlies the longitudinal marker data and the time to event data. They extended Wulfsohn and Tsiatis's EM algorithm to fit this class of models.

1.2.2 Bayesian joint models

Berzuini (1995) fit discrete-time Bayesian PH models in which longitudinal marker data and other covariates were used to predict time to failure.

Faucett and Thomas (1996) used Gibbs sampling to fit a Bayesian version of Wulfsohn and Tsiatis's joint random effects/PH model. They cite several advantages of the Bayesian approach (p. 1665):

> ...we obtain estimates of variability, and in fact the entire marginal or joint posterior distributions, of all model parameters in the Gibbs sampling framework without complex derivations or simplifying assumptions. Thus, the variance estimates for the disease risk parameters correctly reflect the uncertainty inherent in the covariate tracking model parameters, and conversely, the variance estimates of the covariate tracking model parameters reflect the uncertainty in the parameters of the disease risk model. Also, we can incorporate informative priors if desired for fully Bayesian analyses.

Wang and Taylor (2001) extended the work of Faucett and Thomas (1996) by including a stochastic process (IOU) as well as random effects in a flexible model for longitudinal marker data, in combination with a PH model for time to events.

Xu and Zeger (001a) employed a similar Bayesian joint model combining a Gaussian stochastic process model (assumed to be stationary and therefore excluding IOU) for longitudinal marker data

with a PH model for survival time data. In extending their model to multiple markers, Xu and Zeger (001b) dropped the stochastic process component and combined a multivariate random effects model with a PH model.

Except for Berzuini (1995), all of the Bayesian models described in this section required special-purpose MCMC samplers and could not be fit using standard software such as WinBUGS.

1.3 Estimating the proportion of treatment effect "captured" by a surrogate marker

The crucial question when drugs are approved based on surrogate-endpoint trials is whether the same conclusion would have been reached had actual clinical endpoints been used. Aiming to address this question, Freedman et al. (1992) (hereafter "FGS") and Lin et al. (1997) (hereafter "LFD") developed statistical methods for estimating the "proportion of treatment effect captured" (PTE) by a surrogate endpoint. FGS dealt with logistic regression and LFD with proportional hazards models.

Cowles (2002) generalized the above methods for estimating PTE to any setting in which a generalized linear model (GLM) is appropriate for modeling the clinical endpoint. Because not only linear, logistic, and Poisson regression but also survival analysis may be cast in the GLM framework, GLMs can be used for virtually all clinical trial endpoints. Under a GLM, conditional on model parameters θ_i, the response variables Y_i defined by the clinical outcomes of patients i, $i = 1, \ldots, n$, are assumed to be independent draws from a distribution in a natural exponential family, with probability density function for each observed value y_i:

$$f(y_i; \theta_i, \phi) = exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right] \tag{1.1}$$

Covariates are incorporated into the model through a monotonic, differentiable link function g, which relates a linear predictor $\mathbf{z}_i^T\boldsymbol{\beta}$to a transformation of the expectation of the response variable:

$$E(Y_i) \;=\; g^{-1}(\mathbf{z}_i^T\boldsymbol{\beta}),$$

where $\mathbf{z}_i$ is the vector of covariates for subject i and $\boldsymbol{\beta}$ is a vector of coefficients.

If g is the "canonical link," then the relationship with (1.1) is $\theta_i = \mathbf{z}_i^T\boldsymbol{\beta}$. Under the simplest possible *full model* for computing PTE, which includes the marker covariate,

$$E(Y_i \mid x_i, s_i) = g^{-1}(\beta_0 + \beta_1 x_i + \beta_2 s_i) \tag{1.2}$$

where x_i is an indicator variable for treatment group and s_i is the marker value.

Standard computation of PTE also employs the *reduced model*, with the same exponential-family density and link function as the full model but omitting the marker covariate from the linear predictor:

$$E(Y_i \mid x_i) = g^{-1}(\beta_{R,0} + \beta_{R,1} x_i) \tag{1.3}$$

Then PTE is estimated as

$$\widehat{PTE} \;=\; 1 \;-\; \frac{\hat{\beta}_1}{\hat{\beta}_{R,1}}$$

FGS suggested that a lower 95% confidence limit for PTE greater than a pre-chosen proportion, perhaps 0.75, validates the usefulness of the surrogate endpoint. Unfortunately, there is no guarantee that $\widehat{PTE}$ itself will lie in (0,1), and 95% confidence intervals for PTE, tend to be wide. Reasonably precise estimates of PTE are possible by this method only if the estimated unadjusted treatment effect $\hat{\beta}_{R,1}$ is at least 4 times its standard error. Because an interaction between the marker effect and the treatment effect would make PTE meaningless, testing for such an interaction is necessary.

1.4 Markov chain Monte Carlo and Bayesian estimation of PTE

Markov chain Monte Carlo (MCMC) methods enable generating samples from the joint and marginal posterior distributions of unknown quantities in Bayesian models and of functions of these unknowns. Cowles (2002) showed how to use MCMC methods to produce draws from the posterior distribution of PTE, $p(PTE \mid \mathbf{Y})$, by generating samples from the joint posterior distribution of β_1 and $\beta_{R,1}$ and computing the value of PTE corresponding to each pair. Let $\boldsymbol{\beta}$ denote the coefficients in the full model, which includes one or more markers, a treatment group indicator, and possibly other prognostic covariates. Let $\boldsymbol{\beta_R}$ denote the coefficients in the reduced model, which contains all predictor variables *except* the marker(s). The identity

$$p(\boldsymbol{\beta}, \boldsymbol{\beta_R}|\mathbf{Y}) = p(\boldsymbol{\beta}|\mathbf{Y})\; p(\boldsymbol{\beta_R}|\boldsymbol{\beta}, \mathbf{Y})$$

suggests this strategy: fit the full model and at each iteration, say k, after drawing values $\boldsymbol{\beta}^k$ from the joint posterior distribution $p(\boldsymbol{\beta} \mid \mathbf{Y})$, draw $\boldsymbol{\beta_R}^k$ from the distribution $p(\boldsymbol{\beta_R} \mid \boldsymbol{\beta}, \mathbf{Y})$ conditional on those values. Then the paired values of β_1^k and $\beta_{R,1}^k$ generated by the sampler at successive iterations will constitute draws from the joint posterior distribution of these two parameters, and draws from the posterior distribution of PTE may be computed as

$$PTE^k = 1 - \frac{\beta_1^k}{\beta_{R,1}^k} \tag{1.1}$$

For GLMs with canonical links and $a(\phi)$ in (1.1) equal for all observations, $p(\boldsymbol{\beta_R} \mid \boldsymbol{\beta}, \mathbf{Y})$ is a degenerate distribution; i.e., $\boldsymbol{\beta_R}^k$ may be computed deterministically given $\boldsymbol{\beta}^k$. The reduced-model likelihood equations enable solving for the reduced-model coefficients by equating expectations of sufficient statistics under the full and reduced models. Suppose the full model includes m (≥ 1) marker variables and $p-1$ other covariates, including treatment group indicator. Let the $(p+m)$-vector $\boldsymbol{\beta}^k$ denote the draw from the posterior distribution of the coefficients in the full model at the kth iteration of an MCMC sampler, and let $\mu^k_{i,\ full} \equiv g^{-1}(\sum_{l=0}^{p+m-1} x_{il}\ \beta^k_l)$. Then substituting $\mu^k_{i,\ full}$ for $y_i,\ i = 1, \ldots, n$, into the likelihood equations for the reduced model produces a system of p nonlinear equations

$$\sum_i x_{ij}\ \mu^k_{i,\ full} = \sum_i x_{ij}\ g^{-1}(\sum_{l=0}^{p-1} x_{il}\beta^k_{R,l}), \quad j = 0, \ldots, p-1 \tag{1.2}$$

that may be solved uniquely for $\boldsymbol{\beta}^k_R$, the corresponding p-vector of coefficients under the reduced model. Thus $\boldsymbol{\beta}^k_R$ is a nonlinear transformation of the full-model coefficients and the design matrix; $\boldsymbol{\beta}^k_R = T(\boldsymbol{\beta}^k)$.

Because an iterative algorithm such as iteratively reweighted least squares (IRLS) is required to solve the above system of nonlinear equations, the computations cannot be carried out within a WinBUGS program. However, an excellent approximation may be calculated noniteratively.

Appendix 2 of Cowles (2002) shows exactly for the normal linear model with σ^2 known and asymptotically for the log-linear Poisson regression model, that the posterior distribution $p(T(\boldsymbol{\beta}) \mid \mathbf{Y})$ obtained by transforming the full-model coefficients is equal to $p(\boldsymbol{\beta_R} \mid \mathbf{Y})$ under either a normal prior or a locally uniform prior on $\boldsymbol{\beta}$. The same appendix describes the transformation required when an informative prior is placed on the full-model coefficients.

1.5 Goals of the present paper

The present paper extends Cowles's Bayesian approach to enable assessing the proportion of treatment effect on a time-to-event endpoint that is captured by the the longitudinal trajectories of one or more markers measured with error. Computation of the PTE is embedded in a joint model that combines a discrete-time version of the bivariate longitudinal model of Sy et al. (1997) with either a PH model or an exponential accelerated failure time (AFT) model for the event-time data. The time-varying value of interest for one or more of the markers may be cumulative exposure (area under the curve) rather than instantaneous value. To make our models accessible to applied statisticians, we provide WinBUGS code for fitting them.

The study data to which our models will be fit — from AIDS Clinical Trials Group (ACTG) Protocol

320 — is described in Section 2. The models, and the methodology for computing the PTE within them, are presented in Section 3. In Section 4, results from four Bayesian models for the analysis of ACTG 320 data are compared. The Bayesian analyses were carried out using the statistical software package WinBUGS (Spiegelhalter et al., 1995). The code is available for download from the author's webpage, `www.stat.uiowa.edu/~kcowles`.

2 AIDS Clinical Trials Group Protocol 320

ACTG 320 (Hammer et al., 1997) was a randomized, double-blind, placebo-controlled trial comparing a three-drug regimen (indinavir, lamivudine, and either zidovudine or stavudine) with a two-drug regimen (zidovudine and lamivudine) in HIV-infected adults with CD4 counts <= 200 and at least 3 months of prior zidovudine therapy. The 1156 randomized patients were stratified according to their CD4 count ($\leq$ 50 cells/mm^3 or 50-200 cells/mm^3) at study entry. The primary endpoint was occurrence of an AIDS-defining event (according to the CDC definition) or death. In addition, blood specimens were collected at baseline and at weeks 4, 8, 24, and 40 during follow-up for analysis of CD4 counts and viral load. The ACTG 320 dataset available for purchase from the National Technical Information Service includes clinical endpoints and CD4 data for all patients but viral load data on only 198 patients who were randomly selected for a virology substudy.

Although frequentist analysis of the full ACTG 320 dataset indicates that the three-drug regimen was significantly better than the two-drug regimen in forestalling AIDS-defining events (relative risk = 0.416, 95% c.i 0.249 - 0.696, p-value = 0.0008) in a proportional hazards model including the stratification factor and treatment assigment as the only covariates), in the subgroup of patients included in the virology study, the treatment effect was not significant at the .05 level (relative risk = 0.538, 95% c.i. 0.180 - 1.606, p-value = 0.267). This has serious ramifications for our intended analysis since, as mentioned in section 1.3, lack of a strong treatment effect in the reduced model leads to wide intervals for the PTE.

Because blood could be drawn for RNA and CD4 evaluation after patients had experienced clinical progression (other than death, of course), the dataset included some marker values that postdated failure times. These marker values were deleted in all of our analyses. Also, one patient who had no RNA measurements prior to clinical failure was removed from the dataset, leaving 197 patients.

3 Models

3.1 Bivariate longitudinal model for the RNA and CD4 data

Our model for the bivariate RNA and CD4 data is similar to the bivariate stochastic-process model involving Brownian motion of Sy et al. (1997) (hereafter "STC.") Because marker measurements in our clinical trial data were taken on a fixed schedule, rather than irregularly as in the cohort study data used by STC, we used a bivariate random walk (discrete-time Brownian motion) as the stochastic-process element in our model. The $\log_{10}$ transformation is commonly applied to RNA data to symmetrize and to stabilize variance. We followed STC in using the 4th-root transformation for CD4-count data.

In the following, $i = 1, \ldots, N = 197$ indexes patients and $j = 1, \ldots, T = 5$ indexes the fixed marker-measurement times m_j in weeks since start of treatment ($m_1 = 0$, $m_2 = 4$, $m_3 = 8$, $m_4 = 24$, and $m_5 = 40$).

In the first stage of the longitudinal part of the model, the symbols $y_{i,j,1}$ and $y_{i,j,2}$ respectively represent $\log_{10}$ of RNA and fourth root of CD4 measured on patient i at week j. The binary variable trt_i equals 1 if patient i is in the two-drug treatment group and 2 if the patient is in the three-drug group. The measurement errors $\epsilon_{i,j,1}$ and $\epsilon_{i,j,2}$ are assumed independent across marker type, patients, and repeated measurements on the same patient. Patient-specific intercepts of $\log_{10}$ RNA and fourth root CD4 are $\alpha_{i,1}$ and $\alpha_{i,2}$. As in STC, for parameter identifiability, patient-specific slopes on time are *not* used in combination with the stochastic process increments $w_{i,j,1}$ and $w_{i,j,2}$. Instead, all patients in each treatment group g ($g = 1$ or 2) share the same slope, $\mu_{g,3}$ for $\log_{10}$ RNA and $\mu_{g,4}$ for fourth root CD4, for group indicator $g = 1$ or 2. The first stage of the longitudinal part of the model is:

$$
\begin{aligned}
y_{i,j,1} &= \alpha_{i,1} + \mu_{trt_i,3}\, m_j + w_{i,j,1} + \epsilon_{i,j,1} \\
y_{i,j,2} &= \alpha_{i,2} + \mu_{trt_i,4}\, m_j + w_{i,j,2} + \epsilon_{i,j,2}, \quad i = 1, \ldots, N;\ j = 1, \ldots, T \\
\epsilon_{i,j,1} | \sigma^2_{y,1} &\sim N\left(0, \sigma^2_{y,1}\right) \\
\epsilon_{i,j,2} | \sigma^2_{y,2} &\sim N\left(0, \sigma^2_{y,2}\right)
\end{aligned}
$$

The second stage of the longitudinal model defines distributional assumptions on the vectors $\boldsymbol{\alpha}$ and $\mathbf{w}$. The patient-specific intercepts $\boldsymbol{\alpha_i}$ are assumed normally distributed around treatment-group specific mean intercepts $[\mu_{g,1}, \mu_{g,2}]^T$. The random walk increments at time 0 are fixed at 0; the $\boldsymbol{\alpha_i}$ account for between-patient differences in RNA and CD4 values at study entry.

$$
\begin{aligned}
\begin{bmatrix} w_{i,1,1} \\ w_{i,1,2} \end{bmatrix} &\equiv \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\
\begin{bmatrix} w_{i,j,1} \\ w_{i,j,2} \end{bmatrix} \mid \begin{bmatrix} w_{i,j-1,1} \\ w_{i,j-1,2} \end{bmatrix} &\sim N\left(\begin{bmatrix} w_{i,j-1,1} \\ w_{i,j-1,2} \end{bmatrix}, (m_j - m_{j-1})\Sigma_w \right), \quad i = 1, \ldots, N;\ j = 2, \ldots, T \\
\begin{bmatrix} \alpha_{i,1} \\ \alpha_{i,2} \end{bmatrix} \mid \begin{bmatrix} \mu_{trt_i,1} \\ \mu_{trt_i,2} \end{bmatrix} &\sim N\left(\begin{bmatrix} \mu_{trt_i,1} \\ \mu_{trt_i,2} \end{bmatrix}, \Sigma_\alpha \right)
\end{aligned}
$$

The third stage of the longitudinal model specifies priors on the remaining parameters. Independent vague normal priors are placed on the treatment-group-specific mean intercepts and slopes of $\log_{10}$ RNA and fourth root CD4, the μs. Weak semi-conjugate inverse Wishart priors, parameterized such that the distribution $IW(\mathbf{R}, \rho)$ for random matrices of dimension d has mean $(\rho - d - 1)\mathbf{R}$, are used for the covariance matrices of the random effects and the random walk increments, and weak inverse gamma priors, parameterized such that the distribution $IG(a, b)$ has mean $\frac{b}{a-1}$ are placed on the measurement-error variances.

$$
\begin{aligned}
\mu_{g,k} &\sim N(0, 1000), \quad g = 1, 2;\ k = 1, \ldots, 4 \\
\Sigma_\alpha &\sim IW\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, 3 \right) \\
\Sigma_w &\sim IW\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, 3 \right) \\
\sigma_{y,k} &\sim IG(0.1\ , 0.1), \quad k = 1, 2
\end{aligned}
$$

3.2 Models for the time-to-event data

The joint model is completed by specification of the distribution of the time-to-event data, conditional on the parameters of the longitudinal model. The quantities $y^*_{i,j,1} = \alpha_{i,1} + \mu_{trt_i,3}\ m_j + w_{i,j,1}$ and $y^*_{i,j,2} = \alpha_{i,2} + \mu_{trt_i,4}\ m_j + w_{i,j,2}$ are interpretable as the unobservable, "measurement-error-free" values of $y_{1,i,j}$ and $y_{2,i,j}$ respectively. The times at which markers must be evaluated as time-varying covariates in the failure-time models do not coincide with the actual times at which blood was drawn for marker measurement. Consequently, the measurement-error-free values of the markers must be interpolated.

The instantaneous value of CD4 count (suitably transformed) is considered appropriate to use as a predictor of time to clinical events. However, for viral load, some clinicians believe that cumulative exposure (area under the curve or AUC) is a better predictor than instantaneous value, since the former relates directly to the amount of damage that has been done to the immune system, the nervous system, and other bodily functions and organs. The integral required for computing area under the curve of $\log_{10}$ RNA from study entry (week 0) to any arbitrary time t, $\int_0^t [\alpha_{i,1} + \mu_{trt_i,3}\ s + w_{i,1}(s)]\, ds$ can be

evaluated analytically given values of $\alpha_{i,1}$, $\mu_{trt_i,3}$, and $w_{i,j,1}$, $j = 1, \ldots, h$, where h is the largest value of j such that $m_{j-1} \leq t$. Code for this purpose is identified in comments in the WinBUGS program.

We fit four versions of our joint model. The longitudinal part was the same in all of them, and the four versions of the time-to-event portion were:

Model	Survival model	Time-varying covariates	
		$\log_{10}$ RNA	4th root CD4
1	Proportional hazards (PH)	instantaneous	instantaneous
2	PH	AUC	instantaneous
3	Accelerated failure time (AFT)	instantaneous	instantaneous
4	AFT	AUC	instantaneous

3.2.1 PH model for time-to-clinical event data

The intuition behind the most-commonly-used semi-parametric PH models is that the effect of covariates is multiplicative on the hazard. In the ACTG 320 virology dataset, 14 patients experienced clinical endpoints, but due to ties there were only 12 distinct failure times. These were the times at which measurement-error-free values of the RNA and CD4 covariates had to be imputed.

The Poisson formulation of the PH model, described in Whitehead (1980) was used. In this formulation, for each patient (here indexed by i) and each distinct failure time (ordered from earliest to latest and here indexed by l), a binary indicator variable $fail_{i,l}$ equals 1 if patient i failed at failure time l and 0 otherwise. The $fail_{i,l}$s are treated as independent Poisson random variables, so the first stage of this part of the model is:

$$fail_{i,l} \sim Poisson(R_{i,l} exp(\beta_{0,l} + \beta_1 trt0_i + \beta_2 strat_i + \beta_3 y^*_{i,l,1} + \beta_4 y^*_{i,l,2})), \quad i = 1, \ldots, N, \; l = 1, \ldots, T$$

Here $R_{i,l} = 1$ if patient i is in the risk set at time t_l and 0 otherwise; $trt0_i = 1$ if patient i is in the three-drug treatment group and 0 otherwise; $strat_i = 1$ if patient i's CD4 count at study entry was above 50 and 0 otherwise; $y^*_{i,l,1}$ is the modeled value of $\log_{10}$ RNA (or AUC of $\log_{10}$ RNA) at time t_l; and $y^*_{i,l,2}$ is the modeled value of 4th root CD4 at time t_l.

We specified independent vague normal priors on the coefficients:

$$\begin{aligned} \beta_{0,l} &\sim N(0, 100), \quad l = 0, \ldots, T \\ \beta_k &\sim N(0, 10000), \quad k = 1, \ldots, 4 \end{aligned}$$

Bayesian computation of PTE requires draws from the the joint posterior distribution of β_1 from the above Poisson model and $\beta_{R,1}$ from the following reduced model:

$$fail_{i,l} \sim Poisson(R_{i,l} \, exp(\beta_{R,0,l} + \beta_{R,1} trt0_i + \beta_{R,2} strat_i))$$

For the case in which treatment group indicator is the only covariate in the reduced model, Cowles (2002) derived a close approximation to the value of $\beta_{R,1}|\ \boldsymbol{\beta}$ that may be carried out within a WinBUGS sampler. We extend this to the case in which the reduced model includes an additional binary covariate as well as treatment group indicator. The link function g is the log function with inverse the exponential function. Let $\Sigma_{p,q}$ denote summation over all i such that $trt0_i = p$ and $strat_i = q$, where p and q can take on values 0 or 1. Thus, at each distinct failure time $l = 1, \ldots, T$, we have for the full model at the k^{th} iteration of an MCMC sampler:

$$E(\sum_{p,q} fail_{il}) = \sum_{p,q} R_{il}\ exp(\beta_{0l}^k + \beta_1^k trt0_i + \beta_2^k strat_i + \beta_3^k y_{i,l,1}^* + \beta_4^k y_{i,l,2}^*) \equiv \sum_{p,q} R_{il}\mu_{il,full}^k$$

Letting n_{pql} denote the number of patients in treatment group p and stratification group q who are in the risk set at time l, equating expectations under the full and reduced models, we have:

$$\sum_{0,0} R_{il}\mu_{il,full}^k = \sum_{0,0} R_{il} exp(\beta_{R,0l}^k + \beta_{R,1}^k trt0_i + \beta_{R,2}^k strat_i) = n_{00l} \times exp(\beta_{R,0l}^k + \beta_{R,1} \times 0\ +\ \beta_{R,2} \times 0)$$

and similarly

$$\begin{aligned}
\sum_{1,0} R_{il}\mu_{il,full}^k &= n_{10t} \times exp(\beta_{R,0t}^k + \beta_{R,1} \times 1\ +\ \beta_{R,2} \times 0) \\
\sum_{0,1} R_{il}\mu_{il,full}^k &= n_{01l} \times exp(\beta_{R,0}^k + \beta_{R,1} \times 0\ +\ \beta_{R,2} \times 1) \\
\sum_{1,1} R_{il}\mu_{il,full}^k &= n_{11l} \times exp(\beta_{R,0l}^k + \beta_{R,1} \times 1\ +\ \beta_{R,2} \times 1)
\end{aligned}$$

Thus

$$\beta_{R,1}^k|\beta_{01}^k, \ldots, \beta_{0T}^k, \beta_1^k, \beta_2^k, \beta_3^k, \beta_4^k, Y \simeq \sum_{q=0}^{1} \left\{ \sum_{l=1}^{T_q^*} \left[log\left(\frac{\sum_{1,q}\ R_{il}\mu_{il,full}^k}{n_{1ql}} \right) - log\left(\frac{\sum_{0,q} R_{il}\mu_{il,full}^k}{n_{0ql}} \right) \right] /T_q^* \right\} /2 \tag{3.3}$$

Here T_q^* is the number of distinct failure times at which members of both treatment groups in stratification level q are at risk. If either n_{0ql} or n_{1ql} is 0, then failure time l contributes no information toward the estimation of β_1 and $\beta_{R,1}$ in stratification level q; the corresponding summand in (3.3) is undefined and must be omitted from the summation.

In our dataset, patients from all combinations of treatment group and stratification level were represented in all risk sets, so $T_0^* = T_1^* = T = 12$.

3.2.2 Exponential AFT model for time-to-clinical event data

In contrast to PH models, the intuition behind AFT models is that the effect of covariates is multiplicative on the time-to-event itself rather than on the hazard. AFT models (Cox and Oakes, 1984, Section 5.2) posit that each patient i has a "baseline failure time" $t_i^{(0)}$ — what his or her failure time would have been had all covariate values been equal to 0. In an AFT with constant (i.e., non-time-varying) covariates only, the relationship between the baseline failure (or censoring) time $t_i^{(0)}$ and the observed failure or censoring time $obs.t_i$ is usually given as:

$$t_{i,NTV}^{(0)} \;=\; exp(\mathbf{z}_i^T\boldsymbol{\beta})obs.t_i \tag{3.4}$$

where $\mathbf{z}_i$ is patient i's vector of covariate values and $\boldsymbol{\beta}$ is the corresponding vector of coefficients. The AFT model likelihood is completed by specifying a probability distibution on the $t_i^{(0)}$s. If the exponential distribution, which implies a constant hazard, is chosen, then this could be written as $T_i^{(0)} \;\sim\; Exponential(exp(\beta_0))$, where β_0 is the log of the hazard.

With time-varying covariates, the relationship becomes:

$$t_{i,TV}^{(0)} \;=\; \int_0^{obs.t_i} exp(\mathbf{z}(s)_i^T\boldsymbol{\beta})ds \tag{3.5}$$

where $\mathbf{z}(t)_i$ denotes the covariate values of subject i at time t.

Let $fail_i$ be an indicator variable such that $fail_i = 1$ if patient i failed at time $obs.t_i$ and $fail_i = 0$ if patient i was censored at time $obs.t_i$. Then under an exponential AFT, the contribution to the log likelihood made by patient i is

$$\left(\mathbf{z}(obs.t_i)^T \;\; \boldsymbol{\beta}\;\right) \times fail_i - \int_0^{obs.t_i} exp(\mathbf{z}(s)_i^T\boldsymbol{\beta})ds \tag{3.6}$$

Numerical evaluation of the integral in (3.6) is required in order to fit either a Bayesian or a frequentist model. We chose to use composite Simpson's algorithm (Burden and Faires, 1989, Section 4.4, for example) for this purpose. To approximate $F = \int_a^b f(t)dt$, composite Simpson's algorithm requires partitioning the interval [a,b] into an even number n of subintervals of equal length with endpoints $t_0 = a,\; t_1, \ldots, t_{n-1},\; t_n = b$. Then the approximation to the integral is

$$F \simeq \frac{\left(f(t_0) + f(t_n) + 4\sum_{j=0}^{n/2-1} f(t_{2j+1}) + 2\sum_{j=1}^{n/2-1} f(t_{2j})\right)(b-a)}{3n}$$

Thus numerical approximation of the integral requires imputation of the time-varying covariates at each endpoint, $t_0, \ldots, t_n$. Preliminary simulation studies suggested that using more than $n = 32$

subintervals did not improve the accuracy of the approximation to the integrals required. Therefore, in our WinBUGS programs for AFT models, we partitioned each patient's $obs.t_i$ into 32 subintervals and, at each iteration of the sampler, we used the longitudinal model to impute values of $\log_{10}$ RNA (or AUC) and fourth-root CD4 at all the subinterval endpoints. The integral was approximated in the WinBUGS program by taking the inner product of the vector of composite-Simpson's-algorithm multipliers $(1, 4, 2, \ldots, 2, 4, 1)$ with the vector of function evaluations. This process is identified in comments in the WinBUGS code.

In fitting our exponential AFT models, we used the same independent vague normal priors on the coefficients that we had used with the PH models:

$$\beta_k \sim N(0, 10000), \quad k = 0, \ldots, 4$$

Note that Cowles's (2002) method for computing PTE requires only the *reduced* model to be a GLM with canonical link and $a(\phi)$ in (1.1) the same for all observations. Although the *full* model presented here involves time-varying covariates, the *reduced* model is an exponential AFT model with exclusively constant covariates (treatment group and stratification factor) and as such may be cast as a Poisson model. Aitkin and Clayton (1980) point out that, with the indicator variable $fail_i$ defined as above, the contribution to the likelihood from individual i may be written as

$$t_{i,NTV}^{(0)\ fail_i}\ exp(-t_{i,NTV}^{(0)})\ obs.t_i^{-fail_i} \tag{3.7}$$

where $t_{i,NTV}^{(0)}$ is defined as in (3.4). Since the last term in (3.7) does not involve $\boldsymbol{\beta}$, the likelihood is proportional to a Poisson likelihood with $fail_i$ as the random variable and mean $t_{i,NTV}^{(0)}$. Therefore, we may redefine our reduced model for the ACTG 320 data as:

$$fail_i \sim Poisson(\ exp(\beta_{R,0} + \beta_{R,1} trt0_i + \beta_{R,2} strat_i + log(obs.t_i)))$$

Thus, Cowles's (2002) method of computing the values of reduced-model coefficients given values of the full-model coefficients is applicable here. Again, this is done exactly by solving a variant of the reduced-model likelihood equations in which the expectations of the sufficient statistics under the reduced model are set equal to the expectations of the same sufficient statistics given the current draw, say k, of values of the full-model coefficients. Letting

$$t_{i,full}^{(0),k} = exp(\beta_0^k + \beta_1^k trt0_i + \beta_2^k strat_i) \times \int_0^{obs.t_i} \left[exp(\beta_3^k y_{i1}^*(s) + \beta_4^k y_{i2}^*(s)\right] ds, \tag{3.8}$$

$z_{i0} = 1, z_{i1} = trt0_i$, and $z_{i2} = strat_i$, the resulting system of nonlinear equations may be written:

$$\sum_i z_{ij} t_{i,full}^{(0)k} \;=\; \sum_i z_{ij} exp(\beta_{R,0}^k + \beta_{R,1}^k trt0_i + \beta_{R,2}^k strat_i + log(obs.t_i)), \quad j = 0, 1, 2$$

An excellent approximation to $\beta_{R,1}|\boldsymbol{\beta}$ may be computed noniteratively in WinBUGS. As in the previous section, let $\Sigma_{p,q}$ denote summation over all i such that $trt0_i = p$ and $strat_i = q$, and compute within each group defined by values of p and q the sums of expectations under the full model of the $fail_i$:

$$E(\sum_{p,q} fail_i) = \sum_{p,q} t_{i,full}^{(0)k}$$

where $t_{i,full}^{(0)}$ is defined as in (3.8).

Now equate these sums of expectations with the same sums under the reduced model:

$$E(\sum_{0,0} fail_i) = \sum_{0,0} exp(\beta_{R,0}^k + \beta_{R,1}^k \times 0 + \beta_{R,2}^k \times 0) obs.t_i \;=\; exp(\beta_{R,0}^k) \sum_{0,0} obs.t_i \;=\; \sum_{0,0} t_{i,full}^{(0)k}$$

and similarly

$$\begin{aligned} E(\sum_{1,0} fail_i) &= exp(\beta_{R,0}^k + \beta_{R,1}^k) \sum_{1,0} obs.t_i \;=\; \sum_{1,0} t_{i,full}^{(0)k} \\ E(\sum_{0,1} \;=\; fail_i) \;&=\; exp(\beta_{R,0}^k + \beta_{R,2}^k) \sum_{0,1} obs.t_i \;=\; \sum_{0,1} t_{i,full}^{(0)k} \\ E(\sum_{1,1} fail_i) \;&=\; exp(\beta_{R,0}^k + \beta_{R,1}^k + \beta_{R,2}^k) \sum_{1,1} obs.t_i \;=\; \sum_{1,1} t_{i,full}^{(0)k} \end{aligned}$$

This leads to the approximation:

$$\beta_{R,1}^k | \beta_0^k, \beta_1^k, \beta_2^k, \beta_3^k, \beta_4^k, Y \simeq \sum_{q=0}^{1} \left[log\left(\frac{\sum_{1,q} t_{i,full}^{(0)k}}{\sum_{1,q} obs.t_i} \right) - log\left(\frac{\sum_{0,q} t_{i,full}^{(0)k}}{\sum_{0,q} obs.t_i} \right) \right] /2$$

4 Results

To fit each of our four models, using WinBUGS we ran three parallel chains started from overdispersed initial values. Because WinBUGS automatically adapted the Metropolis-Hastings candidate-generating densities during the first 4000 sampler iterations, the output from those iterations could not be used for inference. We ran the Brooks, Gelman, and Rubin convergence diagnostic (built into WinBUGS) beginning at iteration 4001 to determine how many additional burn-in iterations had to be discarded. We ran enough post-burn-in iterations that the sampling error in estimating the posterior mean (called

"MC error" in WinBUGS output) was less than a tenth of the estimated posterior standard deviation for all parameters. Tables 1 and 2 summarize posterior inference under each of the four models.

Table 1
Survival-model parameters

Predictor	Model 1 PH Instantaneous RNA	Model 2 PH AUC RNA	Model 3 AFT Instantaneous RNA	Model 4 AFT AUC RNA
	Coefficient: posterior mean (95% credible set)			
Treat	0.093 (-1.294, 1.375)	-0.0782 (-1.332, 1.115)	0.292 (-1.061, 1.470)	-0.238 (-1.486, 0.913)
Strat	0.427 (-1.264, 2.328)	0.2847 (-1.331, 1.940)	0.539 (-1.167, 2.293)	0.153 (-1.501, 1.827)
Inst RNA	0.380 (-0.386, 1.146)		0.741 (-0.048, 1.513)	
AUC RNA		0.0174 (-0.024, 0.060)		-0.0067 (-0.185, 0.0040)
CD4	-1.010 (-1.863, -0.263)	-0.946 (-1.644, -0.198)	-0.905 (-1.759, -0.006)	-0.974 (-1.708, -0.270
$\beta_{R,1}$	-0.711 (-1.886, 0.380)	-0.716 (-1.910, 0.397)	-0.734 (-1.979, 0.411)	-0.669 (-1.927, 0.417)

Table 2
PTE

	Model 1 PH Instantaneous RNA	Model 2 PH AUC RNA	Model 3 AFT Instantaneous RNA	Model 4 AFT AUC RNA
Posterior median	0.998	0.716	1.197	0.379
95% credible set	(-9.408, 11.60)	(-7.447, 8.802)	(-10.73, 13.60)	(-6.066, 5.475)
50% credible set	(0.475, 1.705)	(0.385, 1.299)	(0.695, 2.161)	(0.246, 0.963)

In all four models, 95% credible sets for the coefficients of treatment group and stratification factor are wide and centered near 0, indicating that neither covariate is a useful predictor of time to clinical events after controlling for the marker variables in the model.

In contrast, higher instantaneous CD4 values are clearly protective, as shown by the fact that the 95% credible sets for log relative risks for this covariate lie entirely to the left of zero in all models. Examples of CD4 cell counts that differ by one fourth-root unit are 110 cells versus 25 cells and 300 cells versus 100 cells. Point estimates for the relative risk of disease progression for these differences are between $e^{-1.01} = 0.364$ and $e^{-0.905} = 0.405$ in all four models.

Somewhat unexpectedly, AUC of $\log_{10}$ RNA (models 2 and 4) appears to be a poorer predictor than instantaneous value of $\log_{10}$ RNA (models 1 and 3) for time to clinical events in this group of patients. This may be because AUC could be evaluated only over the duration of patients' participation in this study, rather than over the entire period since their HIV infection.

For point estimates of PTE, posterior medians are presented in preference to posterior means. Since sampled values of $\beta_{R,1}$ were often close to zero and $\beta_{R,1}$ appears in the denominator of the computation for PTE, some values of PTE were very extreme; therefore, means were not central values. As expected given the insignificance of treatment effect in the reduced model (see Section 2), 95% credible sets for PTE are very wide. However, the lower endpoint of the 50% credible sets for PTE in model 3 implies that $Pr(PTE > 0.695|\mathbf{y}) = .75$; i.e., that there is 75% probability that the combination of instantaneous values of $\log_{10}$ RNA and 4th-root CD4 captures at least 69.5% of the treatment effect on time to clinical events when failure time is modeled with an AFT. This is consistent with accumulating evidence of the usefulness of this combination marker.

5 Discussion

This paper has developed Bayesian models that can be easily implemented using the WinBUGS software package for evaluating the treatment effect on survival time that is captured by multiple longitudinal surrogate markers. Different marker metrics (including AUC), different types of failure-time models (PH and AFT), and the possibility of additional time-varying covariates, are accommodated. In so doing, this paper lays the foundation for Bayesian models incorporating longitudinal measures of adverse side effects, changes in treatment regimen, and patient compliance into joint models for markers and clinical outcomes. These models will begin to address critical problems affecting the interpretation of PTE as a criterion of surrogate marker validity (DeGruttola et al., 1997). These include the facts that net treatment effect on clinical endpoints includes unintended side effects and that patients may change treatment assignment or compliance with treatment between the assessment time for marker values and that for clinical outcomes.

ACKNOWLEDGMENTS

This research was supported by NIAIDS grant R01 AI46962 and in part by NSF/EPA grant R 826887-01-0. The author thanks Brian Smith and Meeyeon Ahn for programming assistance.

References

Aitkin, M. and D. Clayton: 1980, 'The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM'. *Applied Statistics* **29**, 156–163.

Berzuini, C.: 1995, *Markov Chain Monte Carlo in Practice*, Chapt. Medical Monitoring, pp. 321–337. New York: Chapman and Hall.

Burden, R. L. and J. D. Faires: 1989, *Numerical Analysis*. Boston: PWS-Kent Publishing Co., fourth edition.

Cowles, M. K.: 2002, 'Bayesian Estimation of the Proportion of Treatment Effect Captured by a Surrogate Marker'. *Statistics in Medicine* p. to appear.

Cox, D. R. and D. Oakes: 1984, *Analysis of survival data*. New York: Chapman and Hall.

DeGruttola, V., R. Fleming, D. Y. Lin, and R. Coombs: 1997, 'Perspective: Validating surrogate markers – are we being naive?'. *Journal of Infectious Diseases* **175**(2), 237–246.

DeGruttola, V. and X. M. Tu: 1994, 'Modelling progression of CD-4 lymphocyte count and its relationship to survival time'. *Biometrics* **50**, 1003–1014.

Faucett, C. O. and D. C. Thomas: 1996, 'Simultaneously Modelling Censored Survival Data and Repeatdly Measured Covariates: A Gibbs Sampling Approach'. *Statistics in Medicine* **15**, 1663–1685.

Freedman, L. S., B. I. Graubard, and A. Schatzkin: 1992, 'Statistical validation of intermediate endpoints for chronic diseases'. *Statistics in Medicine* **11**(2), 167–178.

Hammer, S. M., K. E. Squires, M. D. Hughes, J. M. Grimes, L. M. Demeter, J. S. Currier, J. J. Eron, J. E. Feinberg, H. H. Balfour, L. R. Deyton, J. A. Chodakewitz, and M. A. Fischl: 1997, 'A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200per cubic millimeter or less'. *The New England Journal of Medicine* **337**(11), 725–733.

Henderson, R., P. Diggle, and A. Dobson: 2000, 'Joint Modelling of Longitudinal Measurements and Event Time Data'. *Biostatistics* **1**(4), 465–480.

Laird, N. M.: 1988, 'Missing Data in Longitudinal Studies'. *Statistics in Medicine* **7**, 305–315.

Lange, N., B. P. Carlin, and A. E. Gelfand: 1992, 'Hierarchical Bayes Models for the Progression of HIV Infection Using Longitudinal CD4 T-Cell Numbers (with discussion)'. *Journal of the American Statistical Association* **87**, 615–632.

Lin, D.-Y., T. R. Fleming, and V. DeGruttola: 1997, 'Estimating the proportion of treatment effect explained by a surrogate marker'. *Statistics in Medicine* **16**(3), 1515–1527.

Pawitan, Y. and S. Self: 1993, 'Modeling disease marker processes in AIDS'. *Journal of the American Statistical Association* **88**, 719–726.

Prentice, R. L.: 1982, 'Covariate Measurement Errors and Parameter Estimation in a Failure Time Regression Model'. *Biometrika* **69**, 331–342.

Spiegelhalter, D., A. Thomas, N. Best, and W. Gilks: 1995, 'BUGS:Bayesian inference using Gibbs sampling, Version 0.5'. MRC Biostatistics Unit, Cambridge.

Sy, J. P., J. M. G. Taylor, and W. G. Cumberland: 1997, 'A Stochastic Model for the Analysis of Bivariate Longitudinal AIDS Data'. *Biometrics* **53**, 542–555.

Taylor, J. M. G., W. G. Cumberland, and J. P. Sy: 1994, 'A Stochastic Model for Analysis of Logitudinal AIDS Data'. *Journal of the American Statistical Association* **89**(427), 727–736.

Wang, Y. and J. M. G. Taylor: 2001, 'Jointly Modeling Longitudinal and Event Time Data with Application to Acquired Immunodeficiency Syndrome'. *Journal of the American Statistical Association* **96**(455), 895–904.

Whitehead, J.: 1980, 'Fitting Cox's Regression Model to Survival Data using GLIM'. *Applied Statistics* **29**(3), 268–275.

Wulfsohn, M. S. and A. A. Tsiatis: 1997, 'A joint model for survival and longitudinal data measured with error'. *Biometrics* **53**, 330–339.

Xu, J. and S. L. Zeger: 2001a, 'Joint Analysis of Longitudinal Data Comprising Repeated Measures and Times to Event'. *Applied Statistics* **50**(part 3), 375–387.

Xu, J. and S. L. Zeger: 2001b, 'The Evaluation of Multiple Surrogate Endpoints'. *Biometrics* **57**, 81–87.