Evaluating Surrogate Endpoints for Clinical Trials: A Bayesian Approach

Mary Kathryn Cowles

Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242, U.S.A.

SUMMARY.

Surrogate endpoints in clinical trials are biological markers or events that may be observed earlier than the clinical endpoints (such as death) that are actually of primary interest. Buyse and Molenberghs (1998) devised two measures for evaluating surrogate endpoints in clinical trials. We propose Bayesian models for extending their methods to settings in which the true endpoint of interest is time to a clinical event and the surrogate endpoint is a continuous marker. The time-to-event component of our models may be a Weibull or log-normal accelerated failure time (AFT) model or a Cox proportional hazards model. Our AFT models can produce posterior predictive distributions for the event times of individuals with censored data.

KEY WORDS: accelerated failure time model; censored data; proportional hazards model; Wishart distribution.

1. Introduction

Surrogate endpoints — biological markers or events that may be observed earlier than the clinical endpoints (such as death) that are actually of primary

email: kate-cowles@uiowa.edu

interest — are widely used to reduce the size and duration of clinical trials. In the last ten years, considerable research has been devoted to the attempt to develop statistical methods for "validating" a surrogate endpoint — i.e. for determining whether a clinical trial based on a particular surrogate endpoint can be expected to reach the same conclusions as would have been reached had the true clinical endpoints been used.

1.1 *The proportion of treatment effect "captured" by a surrogate marker*

To this end, Freedman, Graubard and Schatzkin (1992) (hereafter "FGS") and Lin, Fleming and DeGruttola (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 is appropriate for modeling the clinical endpoint.

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 the point estimate of PTE will lie in (0,1), and 95% confidence intervals for PTE tend to be extremely wide. In addition to these statistical problems, DeGruttola, Fleming, Lin and Coombs (1997) elucidated serious substantive problems that may make the PTE uninterpretable, including 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.

1.2 *The relative effect and the adjusted association*

As an alternative to the PTE, Buyse and Molenberghs (1998) (hereafter "B&M") proposed estimation of two quantities, the relative effect ("RE") of treatment X on the distribution of true endpoints T versus surrogate endpoints S, and "γ_Z," a measure of association between individual patients' true endpoints and surrogate endpoints after controlling for treatment assignment. They presented methods for estimating RE and γ_Z only when either (a) both T and S were binary or (b) both T and S were continuous and could be treated as normally distributed. Buyse et al. (2000) modified the approach for meta-analysis of data from multiple clinical trials or from multiple clinical centers in a single trial. Molenberghs et al. (2001) extended the single-trial and meta-analytic approaches to the case when either T or S was binary or ordinal while the other was continuous. Burzykowski et al. (2001) applied copula models to extend the meta-analytic approach to the situation in which both T and S were failure-time endpoints.

For T and S continuous, B&M proceeded as follows to estimate RE and γ_Z. They first standardized the endpoints and then fit a normal linear seemingly-unrelated-regression (SUR) model in which i indexes patients and x_i is the treatment indicator variable:

$$\begin{aligned} S_i &= \beta_{s,0} + \beta_{s,1} x_i + \epsilon_{S_i} \\ T_i &= \beta_{t,0} + \beta_{t,1} x_i + \epsilon_{T_i} \\ \begin{bmatrix} \epsilon_{S_i} \\ \epsilon_{T_i} \end{bmatrix} &\sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right) \end{aligned} \tag{1}$$

Then the relative effect is $RE = \frac{\beta_{t,1}}{\beta_{s,1}}$ and the adjusted association is $\gamma_Z = \rho$.

If the data values are not standardized, then the covariance matrix in (1)

is $\Sigma = \begin{bmatrix} \sigma_s^2 & \sigma_{st} \\ \sigma_{st} & \sigma_t^2 \end{bmatrix}$ and

$$\begin{aligned} RE &= \frac{\beta_{t,1}/\sigma_t}{\beta_{s,1}/\sigma_s} \\ \gamma_Z &= \frac{\sigma_{st}}{\sigma_s \sigma_t} \end{aligned} \tag{2}$$

B&M (page 1022) pointed out that the desired use for a surrogate endpoint is "to predict the effect of treatment on the true endpoint based on the observed effect of treatment on the surrogate." Thus they regarded a surrogate endpoint for which $RE = 1$ as perfect at the *population* level. In contrast, γ_Z quantifies the individual-level association between S and T after controlling for the treatment effect. B&M considered a surrogate endpoint for which $\gamma_Z = 1$ as perfect at the *individual* level. Even if there were little treatment effect on either the surrogate or the true endpoint, a surrogate could be used to predict individual patients' outcomes if γ_Z were close to 1. (Presumably a marker such as viral load, for which lower marker values are associated with a positive clinical effect, could be considered perfect at the population or individual level if $RE = -1$ or $\gamma_z = -1$.)

B&M pointed out several advantages of RE and γ_Z compared to PTE as measures of surrogacy. For the case (discussed above) of normally-distributed endpoints, they showed that $PTE = \gamma_Z/RE$ — that is, PTE is a composite of the individual-level and population-level aspects of surrogacy and therefore lacks the interpretability of the pair RE and γ_Z. They illustrated by example that, as with the PTE, the confidence interval for RE will be wide unless the treatment effect on the the true endpoint is highly significant; however, γ_Z often may be estimated precisely enough to be useful even using data from smaller trials. B&M also claimed that computation of PTE is ad hoc because

of the necessity to fit two separate models (full and reduced); however Cowles (2002) has shown that the PTE can be calculated from the full model alone so this criticism is less valid. The substantive problems of interpretation of the PTE (see the end of Section 1.1) also apply to RE and γ_Z.

1.3 *Goals of the present paper*

We propose Bayesian models for obtaining the posterior distribution of RE and γ_Z when the surrogate endpoint S is either a single continuous measurement or a single summary measure of the trajectory of longitudinally-evaluated continuous data, and the true endpoint T is time-to-event data with censoring. Some of our models also provide the posterior predictive distribution of exact failure times for patients with censored data.

We propose three types of bivariate normal models for such data. In the simplest models, we apply an appropriate transformation h to the time-to-event data T such that $h(T)$ may be considered normally distributed. Markov chain Monte Carlo (MCMC) methods with data augmentation permit fitting a Bayesian normal linear SUR model when one of the observed response variables is subject to censoring. The data augmentation step in each MCMC sampler iteration involves imputing an exact failure time for each censored observation.

Our second model type is appropriate if a Weibull regression model fits the time-to-event date. It exploits the facts that the exponential density can be expressed as a scale mixture of half normals, and that a Weibull variate is a power of an exponential.

To construct a joint normal-proportional hazards (PH) model, the final approach exploits the Poisson re-expression of the PH model and the rela-

tionship between the Poisson and exponential distribution, and then reapplies the same trick used in the second type.

1.3.1 Scale mixtures of normal and multivariate normal distributions Our proposed methods build on previous theoretical and computational work. Andrews and Mallows (1974) and West (1987) demonstrated that certain univariate and multivariate symmetric probability distributions can be constructed as scale mixtures of normal or multivariate normal distributions. Various authors, including Choy and Smith (1997) and Chen and Shao (1999), used MCMC methods to fit Bayesian models in which these symmetric distributions appeared in the likelihood or as priors. Integration over the mixing distribution was performed implicitly by including in the model a vector of unknown latent parameters drawn from the mixing distribution.

We are not aware of previous use of scale mixtures of *half*-normal distributions, which form the basis of our computational methods.

1.3.2 Outline of the paper Section 2 describes the example dataset to which we apply our methods. Section 3 lays out the likelihood portion of each model, while Section 4 details the priors. Section 5 outlines computing issues. Results of the analysis are presented in Section 6. Finally, Section 7 contains discussion and plans for continuing research.

2. Virology Substudy of AIDS Clinical Trials Group Protocol 175

ACTG 175 (Hammer et al. (1996)) was a randomized, double-blind, placebo-controlled trial comparing monotherapy with either zidovudine (ZDV) or didanosine (ddI) to combination therapy with ZDV plus ddI or ZDV plus

zalcitabine (ddC) in HIV-infected adults with CD4 cell counts on study entry between 200 and 500 cells per cubic millimeter. The primary endpoint was time to death or progression of HIV disease. Multiple measurements of plasma HIV-1 RNA concentration were taken on 391 patients selected for a virology substudy. Two such measurements taken prior to initiation of treatment were averaged to obtain a baseline value of virus load, and additional measurements at weeks 8, 20 and 56 were obtained if the patients were still on the assigned study treatment. Data from the ACTG 175 virology substudy are available for purchase from the National Technical Information Service.

We used data on patients assigned to the ZDV and ZDV+ddC treatment groups, with change from baseline to week 8 on study treatment as the summary measure of longitudinal trajectory of virus load. As was done in the papers on the virology substudy (e.g. Fiscus et al. (1998)), we imputed a value of 200 copies/ml for RNA measurements below the limit of quantitation of the assay used in the study and then applied the $\log_{10}$ transformation to symmetrize and stabilize variance. There were 141 patients for whom valid RNA values were available both at baseline and at week 8 (± 4 weeks) so that change from baseline to week 8 could be evaluated. Of these patients, 27 experienced a clinical endpoint. In analyzing both the marker data and the time-to-event data, we used two binary predictors: $trt_i = 1$ if patient i was assigned to ZDV+ddC and 0 for ZDV, and $strat_i = 1$ if patient i had symptomatic AIDS upon study entry and 0 otherwise.

3. Likelihoods for Models

We consider three different Bayesian joint models for marker data and time-to-event data, which are appropriate when different types of survival models

fit the time-to-event data.

3.1 *Normal or log-normal distribution for failure times*

Simple models for evaluating B&M's RE and γ_Z apply when a transformation h exists such that $h(T)$ may be considered normally distributed. For many AIDS datasets, the identity transformation or a log transformation suffices. We log-transformed the failure times and fit a log-normal accelerated failure time (AFT) model.

For patient i, let s_i denote the change in marker value and $logt_i$ denote the natural log of the time to event, and let trt_i and $strat_i$ be defined as in Section 2. Also denote the log-normal coefficients as $\beta_{l,k}$ $k = 0, 1, 2$. Then the first stage of the Bayesian bivariate normal model required to evaluate RE and γ_Z is as follows:

$$\begin{aligned} p\left(\begin{bmatrix} s_i \\ logt_i \end{bmatrix} \mid \begin{bmatrix} \mu_{s,i} \\ \mu_{t,i} \end{bmatrix}, \Sigma\right) &= N\left(\begin{bmatrix} \mu_{s,i} \\ \mu_{t,i} \end{bmatrix}, \Sigma\right) \\ \mu_{s,i} &= \beta_{s,0} + \beta_{s,1} trt_i + \beta_{s,2} strat_i \\ \mu_{t,i} &= \beta_{l,0} + \beta_{l,1} trt_i + \beta_{l,2} strat_i \end{aligned} \tag{3}$$

Obviously the log failure times could not be standardized prior to the analysis because many were censored. Thus RE and γ_Z were defined as in (2).

3.2 *Weibull AFT model for time to clinical events*

3.2.1 Exponential and Weibull distributions as scale mixtures of the standard half-normal distribution It is well known (Andrews and Mallows (1974), West (1987)) that the double exponential distribution may be expressed as a scale mixture of standard normals. Specifically, if $Z \sim N(0, 1)$ and $Y \sim Exponential(1)$, then $X = \sqrt{2Y} \times Z$ has a double exponential

distribution with parameter 1. It is easily shown that the exponential distribution can be expressed as a scale mixture of *half*-normal distributions (i.e. of zero-mean normals truncated to the positive real line). Specifically, let Λ have the exponential probability density function with parameter 1,

$$f(\lambda) = exp(-\lambda), \quad 0 < \lambda < \infty$$

and let Z have the standard half-normal probability density function

$$f(z) = \frac{2}{\sqrt{2\pi}} exp\left(-\frac{z^2}{2}\right), \quad 0 < z < \infty$$

Consider $T^* = \sqrt{2\Lambda}Z$. Conditional on a given value $\Lambda = \lambda$, the jacobian of the transformation is

$$\frac{dz}{dt^*} = \frac{1}{\sqrt{2\lambda}}$$

and the p.d.f. is

$$f(t^*|\lambda) \;=\; \frac{2}{\sqrt{4\pi\lambda}} exp\left(-\frac{t^{*2}}{4\lambda}\right)$$

Integrating this over the exponential density of Λ yields:

$$f(t^*) \;=\; \int_0^\infty \frac{2}{\sqrt{4\pi\lambda}}\; exp(-\frac{t^{*2}}{4\lambda} - \lambda)\; d\lambda \;=\; exp(-t^*), \quad 0 < t^* < \infty \qquad (4)$$

This can be shown by substituting $u = \sqrt{2\lambda}$ and applying the identity (Andrews and Mallows (1974), equation (2.2)):

$$\int_0^\infty exp\left[-\frac{1}{2}(a^2u^2 + b^2u^{-2})\right] du = \sqrt{\frac{\pi}{2a^2}}\; exp(-|ab|)$$

Thus marginally T^* has an exponential distribution with parameter 1.

Now consider a more complicated transformation, $T = \delta\ (\sqrt{2\Lambda}Z)^{\frac{1}{\alpha}}$, for δ and α positive, real-valued parameters. Conditional on $\Lambda = \lambda$, the jacobian of the transformation is

$$\frac{dz}{dt} = \frac{\alpha}{\sqrt{2\lambda}\ \delta}\left(\frac{t}{\delta}\right)^{\alpha-1}$$

Again integrating over the exponential density of Λ gives:

$$\begin{aligned} f(t) &= \int_0^\infty \frac{2\alpha}{\sqrt{4\pi\lambda}\delta}\left(\frac{t}{\delta}\right)^{\alpha-1} exp\left(-\frac{\left(\frac{t}{\delta}\right)^{2\alpha}}{4\lambda} - \lambda\right) d\lambda \\ &= \frac{\alpha}{\delta}\left(\frac{t}{\delta}\right)^{\alpha-1} exp\left(-\left(\frac{t}{\delta}\right)^{\alpha}\right), \quad 0 < t < \infty \end{aligned} \tag{5}$$

Thus T has a Weibull distribution with scale parameter δ and shape parameter α.

3.2.2 Joint Normal/Weibull models We can proceed as follows to develop a model that captures the relationship between failure times, modeled as Weibulls with scale $exp(\beta_{w,0}+\beta_{w,1}trt_i+\beta_{w,2}strat_i)$, and continuous marker values. We specify a bivariate normal distribution in which the component s_i representing the marker values has unrestricted range, and a latent component z_i^* underlying the failure times marginally has a standard half-normal distribution. That is,

$$\begin{bmatrix} s_i \\ z_i^* \end{bmatrix} = \left(\begin{bmatrix} s_i \\ z_i \end{bmatrix} | z_i > 0\right)$$

where

$$\begin{aligned} \begin{bmatrix} s_i \\ z_i \end{bmatrix} &\sim N\left(\begin{bmatrix} \mu_{s,i} \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_s^2 & \sigma_{st} \\ \sigma_{st} & 1 \end{bmatrix}\right) \\ & i = 1, \ldots, n, \ -\infty < s_i < \infty, \ -\infty < z_i < \infty \end{aligned} \tag{6}$$

and $\mu_{s,i}$ is defined as in (3).

This model is a special case of the multivariate normal models with truncation from below discussed in Horrace (2003). Here the truncation point for s_i goes to the limit of $-\infty$ while for z_i^* the truncation point is 0. Thus results in Section 2 of Horrace (2003) determine that (a) the marginal distribution of

the latent variables z_i^* is standard half normal, and (b) although the resulting *marginal* distribution of the marker values s_i is not normal, their *conditional* distribution given the corresponding z_i^* is indeed normal. That is,

$$z_i^* \sim TN_{(0,\infty)}(0,1)$$

where $TN_{(a,b)}(c,d)$ represents a normal distribution with mean c and variance d, truncated to the interval (a,b); and, letting $\mu_{s_i|z_i^*}$ denote $\mu_{s,i} + \frac{\sigma_{st}}{\sigma_t} z_i^* = \mu_{s,i} + \sigma_{s,t} z_i^*$ and $\sigma^2_{s_i|z_i^*}$ denote $\sigma_s^2 - \frac{\sigma_{st}^2}{\sigma_t} = \sigma_s^2 - \sigma_{st}^2$, the conditional p.d.f. of s_i is

$$p(s_i|z_i^*, \sigma_s^2, \sigma_{st}) = \frac{1}{\sqrt{2\pi}\sigma_{s_i|z_i^*}} exp\left(\frac{-(s_i - \mu_{s_i|z_i^*})^2}{2\sigma^2_{s_i|z_i^*}}\right), \quad -\infty < s_i < \infty$$

Note that the *marginal* p.d.f of s_i is

$$p(s_i) = \frac{\sqrt{2}}{\sqrt{\pi}\sigma_s} exp\left(\frac{-(s_i - \mu_{s_i})^2}{2\sigma_s^2}\right) \times \Phi\left(\frac{\frac{\sigma_{st}}{\sigma_s^2}(s_i - \mu_{s_i})}{\sqrt{1 - \frac{\sigma_{st}^2}{\sigma_s^2}}}\right) \tag{7}$$

which heuristically is proportional to the ordinary normal marginal density of s_i weighted by the probability that the corresponding z_i^* would fall in the correct interval. This point will become important in interpreting the results in Section 6 for $\beta_{s,0}$ in the joint Weibull and PH models.

The remainder of the likelihood specification defines latent exponentially-distributed random variables $\lambda_i,\ i = 1, \ldots n$ and the functional relationship between each observed failure or censoring time $surv_i$ and the corresponding z_i^* and λ_i:

$$\begin{aligned} \lambda_i &\sim Exp(1) \\ surv_i &= (z_i^* \times \sqrt{2\lambda_i})^{\frac{1}{\alpha}} exp(\beta_{w,0} + \beta_{w,0} trt_i + \beta_{w,2} strat_i) \end{aligned} \tag{8}$$

Here $\beta_{w,j},\ j = 0, 1, 2$ denotes Weibull coefficients.

3.3 *Cox Proportional hazards model for time to clinical events*

Exploiting equivalence between Cox's proportional hazards model and a particular Poisson model, Clayton (1994) proposed MCMC methods for Bayesian estimation of the baseline hazard and regression parameters. Such implementation of the PH model in the WinBUGS software package is discussed in the "Leuk" example (Spiegelhalter et al. (1995)). In the counting process notation of Andersen and Gill (1982), a process $N_i(t)$ is observed that counts the number of failures that have been incurred by subject i up to time t. If subject i fails during a small time interval $[t, t + dt)$, then the process increment $dN_i(t) = 1$; otherwise $dN_i(t) = 0$. Under non-informative censoring the likelihood of the data is the same as that resulting if the increments $dN_i(t)$ are considered independent Poisson random variables with means given by an intensity process $I_i(t)$, i.e. $dN_i(t) \sim Poisson(I_i(t)dt)$ where $I_i(t)dt = R_i(t)exp(\mathbf{x}_i(t)^T\boldsymbol{\beta})\ d\Lambda_0(t)$. Here $R_i(t) = 1$ if subject i is in the risk set at time t and 0 otherwise, $d\Lambda_0(t)$ is the increment in the integrated baseline hazard function during the interval $[t, t + dt)$, and $\mathbf{x}_i(t)$ is the vector of subject i's covariates at time t. The likelihood is a product of independent Poisson density evaluations; each subject contributes one term for each distinct failure time for which he or she is in the risk set. Because $d\Lambda_0(t)$ is positive with probability 1, it may be written as $exp(\beta 0_t)$. Thus the model is a Poisson regression with coefficients for indicator variables for the distinct failure times as well as for predictor variables.

Letting n denote the number of subjects in the study, n_i the number of distinct failure times at which subject i was in the risk set, and $fail_{i,j} = 1$ if patient i had an event at time j and 0 otherwise, the likelihood for our

failure-time data is proportional to:

$$\prod_{i=1}^{n}\prod_{j=1}^{n_i} \mu_{i,j}^{fail_{i,j}}\ exp(-\mu_{i,j}) \tag{9}$$

where $\mu_{i,j} = exp(\beta 0_{c,j} + \beta_{c,1} trt_i + \beta_{c,1} strat_i$.

In an exponential model with parameter μ, the contribution to the likelihood of an observation with observed failure time t_i is the exponential density $f(t_i) = \mu\ exp(-\mu t_i)$, while the contribution to the likelihood of an observation with censored failure time t_i is the exponential survivor function $S(t_i) = exp(-\mu t_i)$. Thus (9) is proportional to the product of contributions to the likelihood of independent exponentials

$$\prod_{i=1}^{n}\prod_{j=1}^{n_i} \mu_{i,j}^{fail_{i,j}}\ exp(-\mu_{i,j} t_{i,j}) \tag{10}$$

in which all failure or censoring times $t_{i,j} \equiv 1$. Thus re-expression of an exponential density as a scale mixture of half normals can be used to fit a PH model. In this case, for subject i, a latent vector $\mathbf{z}_i^* = [z_{i,1}^*,\ z_{i,2}^*,\ \ldots, z_{i,n_i}^*]^T$ of standard half-normals underlies the $t_{i,j}$'s in (10). Due to independence, these $z_{i,j}^*$'s are uncorrelated with each other. However, each of them is correlated with s_i, subject i's marker value. Thus

$$\begin{bmatrix} z_{i,1}^* \\ z_{i,2}^* \\ \vdots \\ z_{i,n_i}^* \\ s_i \end{bmatrix} = \left(\begin{bmatrix} z_{i,1} \\ z_{i,2} \\ \vdots \\ z_{i,n_i} \\ s_i \end{bmatrix} \mid z_{i,j} > 0,\ j = 1, \ldots, n_i \right)$$

where the joint distribution of $[\mathbf{z}_i^T, s_i]^T$ is:

$$\begin{bmatrix} z_{i,1} \\ z_{i,2} \\ \vdots \\ z_{i,n_i} \\ s_i \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \mu_{s_i} \end{bmatrix}, \begin{bmatrix} 1 & 0 & \cdots & 0 & \sigma_{s,t} \\ 0 & 1 & 0 & \cdots & \sigma_{s,t} \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & \sigma_{s,t} \\ \sigma_{s,t} & \sigma_{s,t} & \cdots & \sigma_{s,t} & \sigma_s^2 \end{bmatrix} \right) \tag{11}$$

Modeling of the covariances between the marker values s_i and the latent variables $z^*_{i,j}$ in the joint Cox model is difficult. The $z^*_{i,j}$'s are parametric functions of the $t_{i,j}$'s. In the absence of tied failure times, there is only one observed value of $t_{i,j}$ at each distinct failure time j; all the rest are censored and must be imputed. Consequently it is not possible to estimate a separate parameter $\sigma_{s,t,j}$ for the covariance between the s_i's and the $z^*_{i,j}$'s for each j. As a first step, we made the simplifying assumption that there is a common $\sigma_{s,t}$ for all j. In ongoing research, we are considering more realistic models for the covariances.

The joint density in (11) factors into:

$$\begin{aligned} f(z^*_{i,1}, z^*_{i,2}, \ldots, z^*_{i,n_i}, s_i) &= f(z^*_{i,1}) f(z^*_{i,2}|z^*_{i,1}) \ldots f(z^*_{i,n_i}|z^*_{i,1}, z^*_{i,2}, \ldots, z^*_{i,n_i-1}) \\ &\quad \times f(s_i|z^*_{i,1}, z^*_{i,2}, \ldots, z^*_{i,n_i}) \end{aligned}$$

Again applying results from Section 2 of Horrace (2003), we find that

$$\begin{aligned} f(z^*_{i,1}) &= TN_{(0,\infty)}(0,1) \\ f(z^*_{i,j}|z^*_{i,1}, z^*_{i,2}, \ldots, z^*_{i,j-1}) &= TN_{(0,\infty)}(0,1) \\ f(s_i|z^*_{i,1}, z^*_{i,2}, \ldots, z^*_{i,n_i}) &= N(\mu_{s_i} + \sigma_{s,t} \sum_{j=1}^{n_i} z^*_{i,j},\ \sigma^2_s - \sum_{j=1}^{n_i} \sigma^2_{s,t}) \end{aligned}$$

The remainder of the likelihood specification defines latent exponentially distributed random variables $\lambda_{i,j},\ i = 1, \ldots n,\ j = 1, \ldots, n_i$, and the functional relationship between each observed failure or censoring time $t_{i,j}$ and the corresponding $z^*_{i,j}$ and $\lambda_{i,j}$:

$$\begin{aligned} \lambda_{i,j} &\sim Exp(1) \\ t_{i,j} &= z^*_{i,j} \times \sqrt{2\lambda_{i,j}} exp(\beta 0_{c,j} + \beta_{c,1} trt_i + \beta_{c,2} strat_i) \end{aligned} \tag{12}$$

where $\beta 0_{c,j}$ and $\beta_{c,k},\ k = 0, \ldots, 2$, denote Cox PH coefficients.

4. Priors

We placed independent flat priors on the coefficients in the linear model for change in the surrogate marker and on the coefficients in the lognormal, Weibull, and Cox PH time-to-event parts of the respective models:

$$\begin{aligned} p(\beta_{s,k}) &\propto 1, \quad k = 0, 1, 2 \\ p(\beta_{l,k}) &\propto 1, \quad k = 0, 1, 2 \\ p(\beta_{w,k}) &\propto 1, \quad k = 0, 1, 2 \\ p(\beta 0_{c,j}) &\propto 1, \quad j = 1, \ldots, N \\ p(\beta_{c,k}) &\propto 1, \quad k = 1, 2 \end{aligned}$$

where N denotes the number of distinct failure times.

The lognormal model required a prior on Σ, the covariance matrix of the vectors $[s_i,\ logt_i]^T$. WinBUGS parameterizes the multivariate normal distribution in terms of its mean vector and *precision* matrix (inverse of the variance/covariance matrix). Thus our prior on Σ was expressed as a Wishart prior on Σ^{-1} using the WinBUGS parameterization. Our best prior guess was that the variance of the changes in $\log_{10}$ RNA was about 0.25. The lognormal model also involved the natural logs of the failure times, the variance of which we guessed to be close to 1.0. To make the prior vague, we used the smallest integer degrees of freedom that would yield a proper Wishart distribution on a 2×2 matrix, and we set the off-diagonal entries of the prior mean matrix equal to zero to enable the data to drive inference regarding

the sign of the covariance. The resulting prior on the precision matrix was

$$p(\Sigma^{-1}|R,\rho) \;=\; Wishart\left(\left[\begin{array}{cc} 0.25 & 0 \\ 0 & 1.0 \end{array}\right], 4\right) \tag{13}$$

For the Weibull model, a prior was required for the variance-covariance matrix of the vectors $[s_i, z_i^*]^T$, in which the variance of the z_i^*'s is fixed at 1. This is easily accomplished by *partitioning* the implied inverse Wishart prior on Σ according to well-known normal theory laid out for example in Dreze and Richard (1983). If

$$\Sigma = \left[\begin{array}{cc} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{array}\right] \sim IW\left(S = \left[\begin{array}{cc} S_1^2 & S_{12} \\ S_{12} & S_2^2 \end{array}\right], \nu\right),$$

and if $\Sigma_{11.2} \equiv \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2}$ then

$$p(\sigma_2^2, \frac{\sigma_{21}}{\sigma_2^2}, \Sigma_{11.2}) \;=\; p(\sigma_2^2)\; p\left(\frac{\sigma_{21}}{\sigma_2^2}|\Sigma_{11.2}\right)\; p(\Sigma_{11.2})$$

where $p(\sigma_2^2) \;=\; IG(\frac{\nu-1}{2}, \frac{S_2^2}{2})$, $p(\Sigma_{11.2}) \;=\; IG(\frac{\nu}{2}, \frac{S_{11.2}}{2})$, and $p(\frac{\sigma_{21}}{\sigma_2^2}|\Sigma_{11.2}) \;=\; N(\frac{S_{21}}{S_2^2}, \frac{\Sigma_{11.2}}{S_2^2})$. Thus, after fixing σ_2^2 at 1, the Weibull model required a Gamma prior on $\tau_{s|z}^2 \equiv 1/\sigma_{s|z}^2$ and a conditional normal prior on $\delta \equiv \frac{\sigma_{sl}}{\sigma_z^2}|\sigma_{s|z}^2$. Concordant with the entries in (13) these were specified as:

$$\begin{aligned} p(\tau_{s|z}^2) &= G(1, 0.125) \\ p(\delta|\sigma_{s|z}^2) &= N(0, \sigma_{s|z}^2) \end{aligned}$$

Finally, we specified the following diffuse prior on the Weibull shape parameter α:

$$p(\alpha) = Exponential(0.1)$$

5. Model fitting and computing

We used the Bayesian software package WinBUGS (Spiegelhalter et al. (2000)) to fit our models. Complete code, with inline documentation, for all three models may be downloaded from the author's web page,
`www.stat.uiowa.edu/~kcowles`.

For the lognormal and Weibull models, we ran three parallel MCMC chains from overdispersed initial values [the mles for all parameters and the mles plus (minus) four standard errors]. In the log-normal model, the Brooks-Gelman-Rubin (BGR) convergence diagnostic (Brooks and Gelman (1998)) suggested that the three chains began sampling from the same distribution almost immediately. We conservatively discarded 1000 burn-in iterations, ran an additional 20000 iterations, and based inference on the output of iterations 1001-21000 from all three chains. Total runtime for three 21000-iteration chains was under 4 minutes on a 1-Ghz Pentium 3 PC.

In the case of the Weibull model, the first 4000 iterations of each chain were unavailable for use because WinBUGS was adapting the variate-generation algorithm. The BGR convergence diagnostic suggested that the three chains were sampling from the same distribution by the 14000th iteration. We ran an additional 20000 iterations. Total run time was 26 minutes. Our inference is based on the output of iterations 14001-34000 from all three chains.

The situation was quite different for the PH model. As stated in section 3.3, for every distinct failure time j at which subject i is in the risk set and does not fail, a $t_{i,j} > 1$ is simulated in the data augmentation step at the beginning of each MCMC sampler iteration. There are 2973 such combinations of i and j, and only 27 actual events for which $t_{i,j}$ is known. A large amount

of latent data in an MCMC sampler using data augmentation tends to induce high autocorrelation between sampler iterations, and this was clearly the case with the sampler for our PH model. For example, the lag-50 autocorrelation in the output for the parameter $\beta_{c,1}$ was 0.905, and values were similar for other parameters. Consequently, mixing was very slow, and a chain started at poor initial values would take an excruciatingly long time to find the target distribution. To deal with this situation, we chose to run a single very long chain rather than multiple shorter ones. We initialized a chain at the mles and ran 25000 burn-in iterations, followed by an additional 125,000 iterations. Total run time was 46 hours. Application of the Heidelberger and Welch diagnostic (Heidelberger and Welch (1983)) in the Bayesian Output Analysis program (Smith (2002)) to the final 125,000 iterations suggested no systematic trend in the output for any of the parameters of interest. Thus we chose to base inference on those 125,000 iterations.

5.1 *Model comparison and model fit*

The Deviance Information Criterion (DIC) (Spiegelhalter et al. (2002)), which is built into release 1.4 of WinBUGS, could not be used to compare the fit of our joint log-normal, Weibull, and Cox PH models because it is based on log-likelihoods. Although the conditional half-normal likelihoods (given $\lambda_i, \quad i = 1..n$), integrated over the distribution of the λ's, yield the exponential and Weibull likelihoods (as shown in (4) and (5)), the *logs* of the conditional likelihoods do not integrate to any useful expression. Consequently, we used other methods for assessing and comparing the fit of our three models. Since the linear regression model for the marker data was the same for all three versions, we focused on the differences between the three

models for the failure-time data. Evaluating the log likelihoods at the mles indicated that the Weibull model (log likelihood = -194.8) and log-normal model (log likelihood = -194.4) were almost identical in fit.

As an empirical check of the fit of the Weibull distribution to our failure time data, within each of the four strata defined by levels of the two binary predictor variables, we calculated the Kaplan-Meier estimate $\hat{S}(t)$ of the survivor function and plotted $log(-log(\hat{S}(t)))$ versus $log(t)$. Two strata had too few points ($<= 4$) to assess linearity. In the other two strata, the plots were roughly linear, suggesting adequate fit of the Weibull model. The Grambsch-Therneau method of testing the proportional hazards assumption (Grambsch and Therneau (1994)) showed no evidence of lack of fit of the Cox PH model ($p = 0.769$ for the `trt` variable, $p = 0.842$ for the `strat` variable, and $p = 0.972$ globally). We concluded that the Weibull and Cox PH models had satisfactory fit to the failure time data.

6. Results

Tables 1, 3, and 4 summarize the results of fitting our three Bayesian joint models to the data from the ACTG 175 virology study. For comparison, frequentist analyses were also carried out using SAS (SAS Institute (2000)): `proc reg` for linear regression with the RNA-change data, `proc lifereg` for the log-normal and Weibull AFT models, and `proc phreg` for the Cox PH model. Because for the frequentist analysis, the survival analyses were completely separate from the linear regression, frequentist estimates of γ_Z and frequentist confidence intervals for RE could not be computed.

Frequentist linear regression analysis with change in $\log_{10}$ RNA as the response variable indicated that the decline was greater in the ZDV+ddC treat-

ment group but that symptom status at baseline was not a significant predictor. Separate log-normal, Weibull, and Cox PH models for the failure-time data all showed a protective effect of combination therapy with ZDV+ddC compared to monotherapy with ZDV, and increased risk for patients with symptomatic AIDS at study entry compared to those without.

Table 1 compares the results of the Bayesian joint model using the log-normal failure-time component. Posterior means and 95% credible sets for all coefficients in the linear and log-normal regressions are very similar to the corresponding mle's and confidence intervals. The means of the posterior distributions for both RE and γ_z are negative as expected, but the 95% credible sets are quite wide and include positive values.

Table 2 shows the results for the joint model incorporating the Weibull AFT. Note that this model is based on different assumptions regarding the distribution of marker values from those used in the log-normal joint model. Here the assumed marginal distribution of the marker values is as in (7). The substantive interpretation is that marker values that would have been associated with negative values of the latent z^* variable (i.e. with event times occurring *before* the start of the study) are downweighted. The resulting shift to the right of the estimated mean of the underlying normal distribution in the joint analysis compared to the separate frequentist linear regression explains the discrepancy between the estimates of the intercepts $\beta_{s,0}$ in the two versions. Bayesian and frequentist results for the Weibull part of the model are similar, although the asymmetry of the Bayesian credible sets around the posterior means shows skewness in the posterior distributions. The relevant sample size for the survival analysis (27 observed events) is much smaller

than the sample size for estimating the coefficients in the regression for the marker values, n = 141. Thus, for the Weibull coefficients, the symmetric frequentist asymptotic confidence intervals may not be appropriate.

To illustrate how the joint model improves prediction compared to a survival-only model without the marker component, we ran a separate Bayesian Weibull regression model using the treatment and symptom variables as predictors. Results of prediction for a few patients under both models are shown in Table 3. The values of the predictor variables for the first three patients in Table 3 are identical. The first patient was censored at 176 weeks — later than the second and third patients, — but had a much smaller decline in virus load than either of them and thus would be expected to have a poorer prognosis. The joint model predicts successively longer times-to-event for the patients with steeper declines in virus load, whereas the separate model orders the predictions solely according to the censoring times. A similar phenomenon is observable in patients four to six in the table, who share a different set of predictor values. The joint model also provides more precise prediction (narrower 95% prediction intervals) than the separate model.

Finally, Table 4 shows results from the joint model incorporating the Cox PH component. The signs of the coefficients in the Weibull and lognormal models are the reverse of those in the Cox PH model (in which a positive slope means that larger predictor values are associated with larger hazards). Thus inference is substantially the same in the Weibull and Cox PH models.

7. Discussion

We propose bivariate models for uncensored, continuous data and time-to-event data with censoring. These models enable Bayesian estimation of the

measures of marker surrogacy introduced in Buyse and Molenberghs (1998). The joint models with log-normal and Weibull components offer more practical advantages than those with a Cox PH component. The former provide the posterior predictive distribution of exact failure times for patients with censored values. The MCMC sampler for the joint normal/Cox PH model suffers from extremely slow mixing for data with heavy censoring and many distinct failure times. All of our models are accessible for applied work because they can be fit using the software package WinBUGS.

In ongoing research, we are considering more realistic models for the covariances in the joint normal/Cox PH model. An appropriate assumption might be that the covariance is a decreasing parametric function of $(n_i - j)$.

Acknowledgements

The author thanks Shea Watrin and Meeyeon Ahn for computing assistance and Professor George Woodworth for helpful comments.

References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics* **10**, 1100–1120.

Andrews, D. and Mallows, C. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B* **36**, 99–102.

Brooks, S. P. and Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**, 434–455.

Burzykowski, T., Molenberghs, G., Buyse, M., Geys, H. and Renard, D.

(2001). Validation of surrogate endpoints in multiple randomized clinical trials with failure time endpoints. *Journal of Royal Statistical Society* **50**, 405–422.

Buyse, M. and Molenberghs, G. (1998). Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* **54**, 1014–1029.

Buyse, M., Molenberghs, G. and Burzykowski, T. (2000). Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* **56**, 324–325.

Chen, M. H. and Shao, Q. M. (1999). Properties of prior and posterior distributions for multivariate categorical response data models. *Journal of Multivariate Analysis* **71**, 277–296.

Choy, S. T. B. and Smith, A. F. M. (1997). Hierarchical models with scale mixtures of normal distributions. *TEST* **6**, 205–221.

Clayton, D. (1994). Some approaches to the analysis of recurrent event data. *Statistical Methods in Medical Research* **3**, 244–262.

Cowles, M. K. (2002). Bayesian estimation of the proportion of treatment effect captured by a surrogate marker. *Statistics in Medicine* **21**, 811–834.

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

Dreze, J. H. and Richard, J.-F. (1983). *Handbook of Econometrics, Volume I*, chapter Bayesian Analysis of Simultaneous Equation Systems, pages 519–552. North-Holland Publishing Company.

Fiscus, S. A., Hughes, M. D., Lathey, J. L., Pi, T., Jackson, J. B., Rasheed, S., Elbeik, T., Reichman, R., Japour, A., Byington, R., Scott, W., Griffith,

B. P., Katzenstein, D. and Hammer, S. M. (1998). Changes in virologic markers as predictors of CD4 cell decline and progression of disease in human immunodeficiency virus type 1-infected adults treated with nucleosides. *The Journal of Infectious Diseases* **177**, 625–633.

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

Grambsch, P. and Therneau, T. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81**, 515–526.

Hammer, S. M., Katzenstein, D., Hughes, M., Gundacker, H., Schooley, R., Haubrich, R., Henry, W. K., Lederman, M., Phair, J. P., Niu, M., Hirsch, M. S. and Merigan, T. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *The New England Journal of Medicine* **335**, 1081–1090.

Heidelberger, P. and Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research* **31**, 1109–1144.

Horrace, W. C. (2003). Some results on the multivariate truncated normal distribution. *Unpublished Manuscript, Department of Economics, Syracuse University* .

Lin, D.-Y., Fleming, T. R. and DeGruttola, V. (1997). Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine* **16**, 1515–1527.

Molenberghs, G., Geys, H. and Buyse, M. (2001). Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous

outcomes. *Statistics in Medicine* **20**, 3023–3038.

SAS Institute (2000). *SAS/STAT User's Guide, Version 8*. SAS, Inc., Cary, NC.

Smith, B. J. (2002). Bayesian output analysis program (BOA), version 1.0.0 for S-PLUS and R.

Spiegelhalter, D., Thomas, A. and Best, N. (2000). *WinBUGS User Manual, Version 1.3*. MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK.

Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W. (1995). *BUGS Examples, Version 0.5*. MRC Biostatistics Unit, Cambridge.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B* **64**, 583–640.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74**, 646–648.

8. Tables

Table 1: Joint Model with Lognormal AFT for Failure-Time Data

	Bayesian (WinBUGS)			Frequentist (SAS)	
Parameter	Post Mean	MC Error	95% credible set	MLE	95% c.i.
$\beta_{s,0}$	-0.248	0.0003	(-0.385, -0.110)	-0.247	(-0.385, -0.110)
$\beta_{s,1}(trt)$	-0.595	0.0004	(-0.776, -0.414)	-0.585	(-0.776, -0.414)
$\beta_{s,2}(strat)$	-0.157	-0.0005	(-0.419, 0.104)	-0.157	(-0.420, 0.107)
$\beta_{l,0}$	5.719	0.004	(5.299, 6.272)	5.687	(5.228, 6.146)
$\beta_{l,1}(trt)$	0.482	0.002	(-0.023, 1.039)	0.460	(-0.038, 0.957)
$\beta_{l,2}(strat)$	-0.748	0.003	(-1.454, -0.078)	-0.744	(-1.401, -0.087)
RE	-1.499	0.006	(-3.151, 0.071)	NA	
γ_Z	-0.121	0.0008	(-0.343, 0.113)	NA	

Table 2: Joint Model with Weibull AFT for Failure-Time Data

	Bayesian (WinBUGS)			Frequentist (SAS)	
Parameter	Post Mean	MC Error	95% credible set	MLE	95% c.i.
$\beta_{s,0}$	-0.161	0.003	(-0.443, 0.110)	-0.247	(-0.385, -0.110)
$\beta_{s,1}$ (trt)	-0.593	0.0004	(-0.774, -0.411)	-0.585	(-0.776, -0.414)
$\beta_{s,2}$ (strat)	-0.156	-0.0006	(-0.418, 0.107)	-0.157	(-0.420, 0.107)
$\beta_{w,0}$	5.674	0.007	(5.357, 6.058)	5.637	(5.259, 6.015)
$\beta_{w,1}$ (trt)	0.431	0.008	(0.006, 0.868)	0.440	(0.007, 0.873)
$\beta_{w,2}$ (strat)	-0.556	0.008	(-1.066, -0.022)	-0.587	(-0.962, -0.268)
α	1.821	0.008	(1.386, 2.417)	1.859	(1.307, 2.617)
RE	-0.418	0.008	(-0.894, -0.006)	NA	
γ_Z	-0.188	0.005	(-0.637, 0.354)	NA	

Table 3: Prediction using Weibull Joint Model

				From Joint Model		From Univariate model	
trt	strat	Censoring time	RNA chg	median	95% prediction interval	median	95% prediction interval
1	0	175.9	-0.007	385.8	(186.1, 1046.0)	414.8	(188.9, 1332.0)
1	0	142.3	-1.349	410.5	(158.3, 1106.0)	399.1	(157.1, 1321.0)
1	0	139.3	-1.820	428.4	(157.8, 1166.0)	396.8	(154.0, 1326.0)
0	0	142.7	-0.898	291.8	(150.3, 741.1)	282.1	(149.6, 778.9)
0	0	161.9	-0.292	292.3	(167.8, 711.6)	292.6	(168.1, 786.8)
0	0	149.6	-0.555	290.3	(156.3, 723.8)	285.8	(155.7, 782.5)

Table 4: Joint Model with Cox PH Model for Failure-Time Data

	Bayesian (WinBUGS)			Frequentist (SAS)	
Parameter	Post Mean	MC Error	95% credible set	MLE	95% c.i.
$\beta_{s,0}$	-0.132	0.002	(-0.388, 0.121)	-0.247	(-0.385, -0.110)
$\beta_{s,1}$ (trt)	-0.567	0.0005	(-0.754, -0.379)	-0.585	(-0.776, -0.414)
$\beta_{s,2}$ (strat)	-0.190	0.0007	(-0.040, 0.012)	-0.157	(-0.420, 0.107)
$\beta_{t,1}$ (trt)	-0.804	-.023	(-1.569, -0.055)	-0.790	(-1.564, -0.016)
$\beta_{t,2}$ (strat)	1.098	0.022	(0.234, 1.891)	1.088	(0.163, 2.017)
RE	0.788	0.022	(0.052, 1.706)	NA	
γ_Z	-0.014	0.0002	(-0.040, 0.012)	NA	