Evaluating Surrogate Markers for Use as Clinical-Trial Endpoints: A Bayesian Approach

Mary Kathryn Cowles
Department of Statistics and Actuarial Science
University of Iowa

Iowa City, Iowa 52242

Phone: 319-335-0727

Fax: 319-335-3017

kcowles@stat.uiowa.edu

This research was supported by NIAIDS grant R01 AI46962.

Short title: Bayesian evaluation of surrogacy

Abstract

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 (1) 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 either a Weibull or log-normal accelerated failure time model. Our models also can produce posterior predictive distributions for individual patients' event times given values of markers and other covariates.

KEYWORDS: accelerated failure time model, censored data, 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 interest – are widely used to reduce the size and duration of clinical trials. The FDA Modernization Act of 1997 specifically permits the FDA to approve a marketing application for a new drug "upon a determination that the product has an effect on a clinical endpoint or on a surrogate endpoint that is reasonably likely to predict clinical benefit."

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 (2) (hereafter "FGS") and Lin, Fleming, and DeGruttola (3) (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 (4) generalized the above methods for estimating PTE to any setting in which a generalized linear model (GLM) 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 $\widehat{PTE}$ itself will lie in (0,1), and 95% confidence intervals for PTE tend to be extremely wide. In addition to these statistical problems, the PTE presents serious substantive problems that may make it 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. (5)

1.2 The relative effect and the adjusted association

As an alternative to the PTE in assessing surrogacy, Buyse and Molenberghs (1) (correction (6)) (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. The approach has been modified for the situation when data from multiple clinical trials (or possibly from multiple clinical centers in a single trial) can be combined in a meta-analysis (7). Both the single-trial and the meta- analytic approaches have been extended to the case when either T or S was binary or ordinal while the other was continuous (8). Finally, copula models were applied to extend the meta-analytic approach to the situation in which both T and S were failure-time endpoints (9).

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) point 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 regard 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, and B&M consider 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 point out several advantages of RE and γ_Z compared to PTE as a measure of surrogacy. For the case (discussed above) of normally-distributed endpoints, they show that

$$PTE = \frac{\gamma_Z}{RE}$$

— 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 illustrate by example that, as with the PTE, the confidence interval for RE is 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 claim that construction of confidence intervals for PTE is ad hoc because of the necessity to fit two separate models (full and reduced) in order to compute PTE; however (4) 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.

We propose two types of bivariate normal models for such data. In the simpler models, we apply an appropriate transformation h to the time-to-event data T such that $h(T)$ may be considered normally distributed. For many AIDS datasets, the identity transformation or a log transformation suffices for this purpose. 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 at the beginning of each MCMC sampler iteration involves imputing exact failure times for each of the censored observations.

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 truncated normals, and that a Weibull variate is a power of an exponential.

Section 2 describes the example dataset to which we apply our methods. Section 3 lays out the models. Section A explains how to specify priors and handle censored data in order to fit the normal or Weibull models using WinBUGS (10). Results of the analysis are presented in Section 6.

Finally, Section 7 contains discussion and plans for continuing research.

2 AIDS Clinical Trials Group Protocol 152

ACTG 152 (11) was a randomized, double-blind, placebo-controlled trial comparing a three drug regimens (zidovudine alone, didanosine alone, and combination therapy with zidovudine plus didanosine) in symptomatic HIV-infected children three months to 18 years of age. The 831 randomized patients were stratified according to age (< 30 months or ≥ 30) months at study entry. The primary endpoint was length of time to death or to progression of HIV disease.

In addition, blood specimens were collected at baseline and at weeks 4, 8, 12, 24, 36, 48, 72, and 96 during follow-up for analysis of CD4 counts, p24 antigen, and RNA viral load.

When an interim analysis showed a significantly higher risk of disease progression or death in the zidovudine group than in either of the other treatment groups, the study arm with zidovudine alone was stopped and unblinded. The other two treatment arms were continued.

The ACTG 152 dataset available for purchase from the National Technical Information Service as of June, 2003, includes clinical endpoints and data on CD4 counts and p24 antigen for all patients but does not include any RNA data.

We used the ACTG 152 data from the NTIS to evaluate CD4 count and p24 antigen separately as surrogate markers. We used the change from baseline to week 12 on study treatment as the summary measure of longitudinal trajectory of each marker. Consistent with the analysis in the primary results paper (11), we applied the $\log_{10}$ transformation to both markers to symmetrize and stabilize variance. We restricted our analysis to data available at the time of the interim analysis and to patients in the ZDV-only and the combination-therapy treatment groups. We further restricted our p24 analysis to patients who had detectable p24 antigen at study entry. There were 477 patients in our CD4 analysis and 161 patients in our p24 analysis for whom valid marker values were available both at baseline and within a 4-week window around week 12.

3 Models

We consider two different joint models, 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

3.1.1 Likelihood

Simple models for evaluating Buyse and Molenberghs' RE and γ_Z apply when a transformation h exists such that $h(T)$ may be considered normally distributed. We fit two such models — one for change in p24 and one for change in CD4. In both cases, the failure times were log-transformed, and modeled by a log-normal accelerated failure time (AFT) model.

For patient i, let s_i denote the change in marker value, t_i denote the natural log of the time to event, trt_i be a (0,1) indicator of treatment group, and $strat_i$ be a (0,1) indicator of CD4 count stratification level. Then the first stage of the Bayesian bivariate normal model required to evaluate RE and γ_Z is as follows:

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

Obviously the log failure times could not be standardized prior to the analysis because many were censored. Thus the quantities RE and γ_Z for which we sought posterior distributions were defined as in 2.

3.1.2 Priors

In the second stage, we placed vague independent normal priors on the βs:

$$\begin{aligned} p(\beta_{s,k}) &\sim N(0, 10^6), \quad k = 0, 1, 2 \\ p(\beta_{t,k}) &\sim N(0, 10^6), \quad k = 0, 1, 2 \end{aligned} \tag{4}$$

WinBUGS parameterizes the multivariate normal distributions in terms of its mean vectors and *precision* matrix (inverse of the variance/covariance matrix). Thus our prior on Σ was expressed as a Wishart prior on Σ^{-1}. To express our prior beliefs about Σ using the WinBUGS parameterization of the Wishart distribution we proceeded as follows. Our best prior guesses of the components of the variance covariance matrix of the vectors $[s_i,\ logt_i]^T$ in the log-normal models for p24 change were that the variance of the s_is was .25 and the variance of the $logt_i$s was 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

$$p(\Sigma^{-1}|R, \rho) = Wishart(\begin{bmatrix} 0.25 & 0 \\ 0 & 1.0 \end{bmatrix}, 4)$$

Similar reasoning produced the following prior on Σ^{-1} for the analysis of CD4 data:

$$p(\Sigma^{-1}|R, \rho) = Wishart(\begin{bmatrix} 0.01 & 0 \\ 0 & 1.0 \end{bmatrix}, 4)$$

See Section A for a discussion of Wishart and inverse Wishart priors, and the modeling of censored data in WinBUGS.

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 (12; 13) 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 half normal probability density function

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

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

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

Integrating this over the exponential density of λ yields:

$$f(t^*) \;=\; \int_0^\infty \frac{2}{\sqrt{4\pi\lambda}} exp(-\frac{t^{*2}}{4\lambda} - \lambda) d\lambda \;=\; exp(-t^*), \quad 0 < t^* < \infty \tag{5}$$

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:

$$f(t) \;=\; \int_0^\infty \frac{2\alpha}{\sqrt{4\pi\lambda}\delta}\left(\frac{t}{\delta}\right)^{\alpha-1} exp(-\frac{\left(\frac{t}{\delta}\right)^{2\alpha}}{4\lambda} - \lambda)d\lambda \;=\; \frac{\alpha}{\delta}\left(\frac{t}{\delta}\right)^{\alpha-1} exp\left(-\left(\frac{t}{\delta}\right)^{\alpha}\right), \quad 0 < x < \infty \quad (6)$$

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

3.2.2 Joint Normal/Weibull models

Suppose we wished to fit to our failure-time data a Weibull model with shape $exp[-(\beta_{0,t} + \beta_{1,t}trt_i + \beta_{2,t}strat_i)]$. Then we can proceed as follows to develop a model that captures the relationship between the continuous marker values and the failure times. 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 is truncated to the positive real line.

$$\begin{aligned}\begin{bmatrix} s_i \\ z_i^* \end{bmatrix} \;&\sim\; N\left(\begin{bmatrix} \mu_{s,i} \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_s^2 & \sigma_{st} \\ \sigma_{st} & 1 \end{bmatrix}\right) \\ &i = 1, \ldots, n, \; -\infty < s_i < \infty, 0 < z_i^* < \infty \end{aligned} \quad (7)$$

where $\mu_{s,i}$ is defined as in 3.

This model is a special case of the multivariate normal models with truncation from below discussed in (14). Here the truncation point for s_i is $-\infty$ and for z_i^* is 0. Thus results in Section 2 of (14) 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.

$$z_i^* | s_i, \sigma_s^2, \sigma_{st} \;\sim\; TN_{(0,\infty)}(\frac{\sigma_{st}}{\sigma_s^2}(s_i - \mu_{s,i}),\; 1 - \frac{\sigma_{st}^2}{\sigma_s^2})$$

where $TN_{(a,b)}(c,d)$ represents a normal distribution with mean c and variance d, truncated to the interval (a,b).

Letting $\mu_{s_i|z_i^*}$ denote $\frac{\sigma_{st}}{1} z_i^*$ and $\sigma^2_{s_i|z_i^*}$ denote $sigma_s^2 - \frac{\sigma_{st}^2}{\sigma_s^2}$, the conditional p.d.f. of s_i is proportional to

$$p(s_i|z_i^*, \sigma_s^2, \sigma_{st}) \propto \frac{1}{\sigma_{z,i|s_i}} \frac{1}{\left(1 - \Phi\left(\frac{-\mu_{z,i|s_i}}{\sigma_{z|s_i}}\right)\right)} exp\left(\frac{-(z_i^* - \mu_{z,i|s_i})^2}{2\sigma^2_{z|s_i}}\right), \quad 0 < z_i^* < \infty$$

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:

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

Then letting $mult_i$ denote $\sqrt{2\lambda_i} exp[-(\beta_{0,t} + \beta_{1,t} trt_i + \beta_{2,t} strat_i)$, and applying standard transformation-of-variables methods, the conditional density of $surv_i$ given $s_i,\ \lambda_i$ is proportional to

$$p(surv_i|s_i, \sigma_s^2, \sigma_{st}, \lambda_i) \propto \frac{1}{mult_i \sigma_{z,i|s_i}} \frac{1}{\left(1 - \Phi\left(\frac{-\mu_{z,i|s_i}}{\sigma_{z|s_i}}\right)\right)} exp\left(\frac{-(\frac{surv_i}{mult_i} - \mu_{z,i|s_i})^2}{2\sigma^2_{z|s_i}}\right) \quad 0 < surv_i < \infty \tag{9}$$

4 Priors

In the analyses of both the p24 and the CD4 data, 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 and Weibull time-to-event parts of the respective models:

$$\begin{aligned} p(\beta_s, k) &\propto 1, \quad k = 0, 1, 2 \\ p(\beta_l, k) &\propto 1, \quad k = 0, 1, 2 \end{aligned}$$

$$p(\beta_w, k) \quad \propto \quad 1, \quad k = 0, 1, 2$$

Recall that

Section A.3 discusses how to partition a Wishart prior to accommodate one component being fixed.

5 Model fitting and computing

The Appendix explains how to use the Bayesian software package WinBUGS (10) to fit our models. Complete code for both models may be downloaded from the author's web page

`www.stat.uiowa.edu/~kcowles`

We initialized our samplers at the maximum likelihood estimates of the parameters, obtained from the SAS analyses described in 6. Due to our vague priors, the mles are likely to be draws from a high-posterior-density region of the parameter space; consequently little or no sampler burn-in should be required. However, WinBUGS automatically discards the first 4000 iterations of the samplers for the Weibull models, because the sampling algorithm for some parameters is being tuned during those iterations and the resulting output is not truly from a Markov chain.

In the version of the code used for sampling from posterior predictive distributions of failure times, the MCMC sampling algorithm involves a data augmentation step in which an exact value of each censored failure time is simulated at each iteration of the sampler, conditionally on the values of $\boldsymbol{\beta}_t$ and other parameters from the previous iteration. New values of $\boldsymbol{\beta}_t$ are then simulated conditionally on the failure times (observed and imputed). When, as with our data, the proportion of censored failure times is large (66% in the the p24 dataset and 80% in the CD4 dataset), this sampling algorithm produces a slowly-mixing sampler, and very large numbers of iterations are required to obtain reasonably precise estimation of posterior means and quantiles. For example,

lag 1 and lag 25 autocorrelations in the output for FIND THESE NUMBERS $\beta_{t,1}$ in the exponential AFT model for CD4 data were 0.997 and 0.923 respectively. Consequently, we chose to run each of our samplers for 304,000 iterations and to base our inference on the last 300,000 iterates for each parameter. This brought the autocorrelation-adjusted standard errors of the posterior mean estimates (the quantity that WinBUGS calls "MC error") below ± 0.033 for all parameters in all models.

The sample size relevant for survival analysis is the number of observed failure times, rather than the total sample size. Thus, the sample sizes for survival analysis are moderate (96 events in the CD4 analysis and 55 in the p24 analysis). The relevant sample sizes for estimating the coefficients in the regressions for the marker values are larger in both cases — 477 for CD4 and 161 for p24. Tables 1-2 show results of the Bayesian joint SUR models for CD4 and failure times. For comparison, frequentist analyses were also carried out using SAS (15). A linear regression model was fit to the CD4 change data using `proc reg`. `Proc lifereg` was used for the log-normal and Weibull AFT models. In contrast to the Bayesian joint model, for the frequentist analysis, the survival analyses were completely separate from the linear regression. Thus frequentist estimates of γ_Z were not available, and frequentist confidence intervals could not be estimated for RE.

5.1 Model comparison

The fact that the Deviance Information Criterion (DIC) (Spiegelhalter et al., 1998), which is built into release 1.4 of WinBUGS, is based on log-likelihoods rendered it unusable to compare the fit of our log-normal, Weibull, and Cox PH models. This is because, although the conditional half-normal likelihoods (given $\lambda_i, \quad i = 1..n$), integrated over the distribution of the *lambda*s, yield the exponential and Weibull likelihoods (as shown in 5 and 6), the *logs* of the conditional likelihoods do not integrate to any useful expression.

6 Results

For the Weibull model, we ran three parallel MCMC chains from overdispersed initial values (the mles for all parameters and the plus (minus) four standard errors. The Brooks-Gelman-Rubin convergence diagnostic (ref) as implemented in WinBUGS suggested that the three chains were sampling from the same distribution by the 14000th iteration. We ran an additional 15000 iterations (total run time for three chains for 29000 was 22 minutes 17 seconds). Our inference is based on the output of iterations 14001-29000 from all three chains.

7 Discussion

8 Tables and Figures

Table 1: CD4 change: Log-Normal AFT for Failure-Time Data

	Bayesian (WinBUGS)			Frequentist (R)	
Parameter	Post Mean	MC Error	95% credible set	MLE	95% c.i.
$\beta_{s,0}$	-0.0104	0.0003	(-0.0443, 0.0251)	-0.0106	(-0.0452, 0.0241)
$\beta_{s,1}$	0.0579	0.0004	(0.0155, 0.0982))	0.0578	(0.0172, 0.0984)
$\beta_{s,2}$	0.0414	0.0003	(0.0006, 0.0822)	0.0418	(0.0012, 0.0825)
$\beta_{t,0}$	3.737	0.0029	(3.486, 4.024)	3.7029	(3.4515, 3.9543)
$\beta_{t,1}$	0.3198	0.024	(0.0415, 0.6232)	0.3082	(0.0344, 0.5820)
$\beta_{t,2}$	0.4852	0.0023	(0.1981, 0.7891)	0.4666	(0.1864, 0.7467)
RE	1.436	0.0751	(0.1185, 4.901)	NA	
γ_Z	-0.0249	0.0007	(-0.1557, 0.1047)	NA	

Table 2: CD4 change: Weibull AFT for Failure-Time Data

	Bayesian (WinBUGS)			Frequentist (R)	
Parameter	Post Mean	MC Error	95% credible set	MLE	95% c.i.
$\beta_{s,0}$	-0.0103	0.0004	(-0.0453, 0.0247)	-0.0106	(-0.0452, 0.0241)
$\beta_{s,1}$	0.0578	0.0004	(0.0170, 0.0996))	0.0578	(0.0172, 0.0984)
$\beta_{s,2}$	0.0415	0.0003	(0.0011, 0.0813)	0.0418	(0.0012, 0.0825)
$\beta_{t,0}$	3.794	0.0028	(3.588, 4.041)	3.7573	(3.5441, 3.9705)
$\beta_{t,1}$	0.3421	0.0025	(0.0803, 0.625)	0.3287	(0.06898, 0.5883)
$\beta_{t,2}$	0.4739	0.0023	(0.207, 0.7774)	0.4525	(0.1811, 0.7239)
α	1.576	0.0033	(1.302, 1.882)	1.6177	(1.3508, 1.9374)
RE	1.989	0.227	(0.2621, 4.98)	NA	
γ_Z	-0.0641	0.0015	(-0.2917, 0.1689)	NA	

References

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

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

[3] D.-Y. Lin, T. R. Fleming, V. DeGruttola, Estimating the proportion of treatment effect explained by a surrogate marker, Statistics in Medicine 16 (3) (1997) 1515–1527.

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

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

Table 3: p24 change: Log-Normal AFT for Failure-Time Data

	Bayesian (WinBUGS)			Frequentist (SAS)	
Parameter	Post Mean	MC Error	95% credible set	MLE	95% c.i.
$\beta_{s,0}$	-0.4086	0.0012	(-0.5444, -0.2778)	-0.4088	(-0.5386, -0.2790)
$\beta_{s,1}$	-0.3499	0.0015	(-0.5248, -0.1702)	-0.3503	(-0.5248, -0.1759)
$\beta_{s,2}$	0.0933	0.0012	(-0.0995, 0.2895)	0.0936	(-0.0982, 0.2855)
$\beta_{t,0}$	3.274	0.0027	(2.989, 3.600)	3.2309	(2.9511, 3.5107)
$\beta_{t,1}$	0.4069	0.0035	(0.0056, 0.8379)	0.3802	(-0.00083, 0.7613)
$\beta_{t,2}$	0.4202	0.0030	(-0.0471, 0.8939)	0.4111	(-0.0209, 0.8431)
RE	-0.7091	0.0080	(-1.795, -0.0091)	NA	
γ_Z	-0.0575	0.0011	(-0.2687, 0.1448)	NA	

Table 4: ACTG 175 RNA: Weibull AFT for Failure-Time Data

	Bayesian (WinBUGS)			Frequentist (SAS)	
Parameter	Post Mean	MC Error	95% credible set	MLE	95% c.i.
$\beta_{s,0}$			(-0.5443, -0.2819)	-0.247	(-0.385, -0.110)
$\beta_{s,1}$			(-0.5247, -0.1726)	-0.585	(-0.776, -0.414)
$\beta_{s,2}$			(-0.1029, 0.2880)	-0.157	(-0.420, 0.107)
$\beta_{t,0}$			(3.251, 3.791)	5.637	(5.259, 6.015)
$\beta_{t,1}$			(0.0849, 0.937)	0.440	(0.007, 0.873)
$\beta_{t,2}$			(-0.1044, 0.8287)	-0.587	(-0.962, -0.268)
α				1.859	(1.307, 2.617)
RE			(-0.085, 1.809)	NA	
γ_Z			(-0.331, 0.419)	NA	

Table 5: ACTG 175 RNA: Weibull AFT for Failure-Time Data

			Bayesian (WinBUGS)			Frequentist	
trt strat[] obs.t[] s[]	joint mean	std	.025	.5	.975	univ mean	std	.025	.5	.975
1	0	175.9	-0.007			444.2				2
1	0	142.3	-1.349			465.9				2
1	0	139.3	-1.820			486.5				2
0	0	142.7	-0.898			330.9				1
0	0	161.9	-0.292			329.7				1
0	0	149.6	-0.555			328.5				1

Table 6: ACTG 175 RNA: Cox PH Model for Failure-Time Data

	Bayesian (WinBUGS)			Frequentist (SAS)	
	Post	MC	95%		95%
Parameter	Mean	Error	credible set	MLE	c.i.
$\beta_{s,0}$			()	-0.247	(-0.385, -0.110)
$\beta_{s,1}$			()	-0.585	(-0.776, -0.414)
$\beta_{s,2}$			()	-0.157	(-0.420, 0.107)
$\beta_{t,1}$			()	-0.790	(-1.564, -0.016)
$\beta_{t,2}$			()	1.088	(0.163, 2.017)
RE			()	NA	
γ_Z			()	NA	

Table 7: p24: Weibull AFT for Failure-Time Data

	Bayesian (WinBUGS)			Frequentist (SAS)	
	Post	MC	95%		95%
Parameter	Mean	Error	credible set	MLE	c.i.
$\beta_{s,0}$	-0.4102	0.0012	(-0.5443, -0.2819)	-0.4088	(-0.5386, -0.2790)
$\beta_{s,1}$	-0.3497	0.0015	(-0.5247, -0.1726)	-0.3503	(-0.5248, -0.1759)
$\beta_{s,2}$	0.0945	0.0014	(-0.1029, 0.2880)	0.0936	(-0.0982, 0.2855)
$\beta_{t,0}$	3.495	0.0026	(3.251, 3.791)	3.4337	(3.1948, 3.6727)
$\beta_{t,1}$	0.4757	0.0034	(0.0849, 0.937)	0.4480	(0.0715, 0.8245)
$\beta_{t,2}$	0.3327	0.0033	(-0.1044, 0.8287)	0.3030	(-0.1193, 0.7252)
α	1.476	0.0031	(1.138, 1.843)	1.5434	(1.2223, 1.9489)
RE	0.784	0.015	(-0.085, 1.809)	NA	
γ_Z	0.048	0.00087	(-0.331, 0.419)	NA	

[6] M. Buyse, G. Molenberghs, T. Burzykowski, D. Renard, H. Geys, The validation of surrogate endpoints in meta-analyses of randomized experiments, Biostatistics 1 (2000) 49–68.

[7] M. Buyse, G. Molenberghs, T. Burzykowski, Criteria for the validation of surrogate endpoints in randomized experiments, Biometrics 56 (1) (2000) 324–325.

[8] G. Molenberghs, H. Geys, M. Buyse, Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcomes, Statistics in Medicine 20 (20) (2001) 3023–3038.

[9] T. Burzykowski, G. Molenberghs, M. Buyse, H. Geys, D. Renard, Validation of surrogate endpoints in multiple randomized clinical trials with failure time endpoints, Journal of Royal Statistics Society 50 (4) (2001) 405–422.

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

[11] J. Englund, C. Baker, C. R. R. McKinney, B. Petrie, M. Fowler, D. Pearson, A. Gershon, G. M. E. A. J. Schlizoberg, J. Sullivan, Zidovudine, didanosine, or both as the initial treatment for symptomatic hiv-infected children, The New England Journal of Medicine 336(24) (1997) 1704–1712.

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

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

[14] W. C. Horrace, Some results on the multivariate truncated normal distribution, Unpublished Manuscript, Department of Economics, Syracuse University .

[15] SAS, Inc., Cary, NC, SAS/STAT User's Guide, Version 8 (2000).

[16] G. E. P. Box, G. C. Tiao, Bayesian Inference in Statistical Analysis, Addison-Wesley, Reading, MA, 1973.

[17] J. H. Dreze, J.-F. Richard, Handbook of Econometrics, Volume I, North-Holland Publishing Company, 1983, Ch. Bayesian Analysis of Simultaneous Equation Systems, pp. 519–552.

A Using WinBUGS to fit SUR models with censored data

A.1 Priors on the precision matrix

All of our models include an unknown variance/covariance matrix of a bivariate normal distribution. WinBUGS requires parameterizing such models in terms of the *precision matrix* (inverse of the variance/covariance matrix). The Wishart distribution is the conjugate prior for the precision matrix of a multivariate normal distribution with known mean. It is the standard choice of prior for precision matrices in realistic multivariate-normal-based models with means (and possibly many other parameters) unknown because it leads to a Wishart full conditional distribution for the precision matrix that simplifies MCMC-based model fitting. The two parameters of the Wishart distribtution are a mean matrix and a scalar parameter called the degrees of freedom.

If X denotes a $p \times p$ symmetric, positive definite random matrix, R is a fixed $p \times p$ symmetric, positive definite matrix, ν is a strictly positive scalar, and the p.d.f. of X is

$$p(X|R,\nu) \;\propto\; |R|^{\frac{\nu}{2}}|X|^{\frac{\nu-p-1}{2}} \exp\left[-\frac{1}{2}tr(RX)\right] \tag{10}$$

then, with the parameterization used in WinBUGS, X has a Wishart distribution with parameters R and ν: $X \;\sim\; dwish(R,\nu)$

In what follows, we use the WinBUGS parameterization. The Wishart distribution is proper if

$\nu \geq p$. If $X \sim dwish(R, \nu)$, then the moments are as follows:

$$\begin{aligned} E(X_{ij}) &= \nu(R^{-1})_{ij} \\ Var(X_{ij}) &= \nu\left[(R^{-1})_{ij}^2 + (R^{-1})_{ii}(R^{-1})_{jj}\right] \\ Cov(X_{ij}, X_{kl}) &= \nu\left[(R^{-1})_{ik}(R^{-1})_{jl} + (R^{-1})_{il}(R^{-1})_{jk}\right] \end{aligned}$$

Note that the gamma distribution is a special (one-dimensional) case of the Wishart. If X and R are scalars, and the p.d.f of X is proportional to $x^{\frac{\nu}{2}-1}exp\left(-\frac{Rx}{2}\right)$ then

$$W(R, \nu) = G(\frac{\nu}{2}, \frac{R}{2})$$

WinBUGS does not allow the use of its Wishart distribution with one-dimensional matrices, however.

If $X \sim dwish(R, \nu)$, then X^{-1} has an *inverse Wishart* distribution with the same parameters, where

$$E\left((X_{ij}^{-1}\right) = \frac{R_{ij}}{\nu - p - 1}$$

The inverse Wishart distribution is always proper; however, it has a degenerate form if $\nu < p$, and obviously the first moment is negative or infinite unless $\nu > p + 1$.

We use the following steps to specify a prior on a covariance matrix, say Σ, in WinBUGS:

1. Let R equal the prior guess for the mean of the $p \times p$ *variance/covariance* matrix Σ.

2. Choose a degrees-of-freedom parameter ν $(> p + 1)$ that roughly represents an "equivalent prior sample size" – your belief in R as the value of Σ is as strong as if you had seen ν previous vectors with sample covariance matrix R.

3. Define a matrix $S = (\nu - p - 1)R$.

4. In WinBUGS, put a Wishart prior with parameters S and ν on the corresponding precision matrix Σ^{-1}.

5. then

   - $E(\Sigma_{i,j}) = R_{i,j}$
   - the variance of the prior will be decreasing in ν
   - $E((\Sigma)^{-1}_{i,j}) = \frac{\nu}{\nu-p-1}(R^{-1})_{i,j}$

For example, our best prior guesses of the components of the variance covariance matrix of the vectors $[s_i,\ logt_i]^T$ in the log-normal model for RNA change were that the variance of the s_is was .25, the variance of the $logt_i$s was 1.0, and the covariance between them was 0.125). 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. The resulting prior on the precision matrix was

$$Wishart(\begin{bmatrix} 0.25 & 0.125 \\ 0.125 & 1.0 \end{bmatrix}, 4)$$

A.2 Accommodating censored data

Because the vast majority of patients in ACTG 320 had not experienced clinical events by the end of the study, their failure times are censored. This results in missing values of $logt_i$ in the log-normal AFT and of $surv_i$ in the exponential AFT. There are even more censored values in our version of the PH model, because there is a censored exponential for each subject for each distinct failure time preceding the time at which he or she had an event. Because WinBUGS release 1.3 can handle missing or censored values in scalar quantities with a univariate normal distribution but not in vectors following a *multivariate* normal distribution, the joint distributions $p([s_i,\ logt_i]^T)$ must be expressed as $p(s_i)p(logt_i|s_i)$, and similarly for the joint distribution $p([s_i,\ surv_i]^T)$.

At each iteration, the WinBUGSS sampler will simulate an exact value of each censored $logt_i$ or $surv_i$ from its full conditional distribution. Censoring in WinBUGS is denoted with the notation $I(lower, upper)$, with $I(lower,)$ indicating an observation known to lie above `lower`. Thus, two

vectors — one representing failure times and one representing censoring times — must be included in the data list. Entries for the exponential and log-normal models are shown in the table below.

	Exponential model		Log normal model	
Observation type	Entry in `surv`	Entry in `cent`	Entry in `logt`	Entry in `cent`
Failed	Observed failure time	0	Log observed failure time	-1000
Censored	NA	Observed censoring time	NA	Log observed censoring time

The WinBUGS code fragment below defines the likelihood in the log-normal model. `Mus[i]` and `precs` denote the marginal mean and marginal precision of the marker value s_i, while `mut[i]` and `prect` denote the conditional mean and precision of $logt_i$ given s_i. The quantity `regcoef` is (using the notation in 7) $\frac{\sigma_{st}}{\sigma_s^2}$.

```
for (i in 1:N) {
    mus[i] <- betas[1] + betas[2] * trt[i] + betas[3] * strat[i]
    s[i] ~ dnorm( mus[i], precs )
    mut[i] <- betat[1] + betat[2] * trt[i] + betat[3] * strat[i] +
           regcoef * (s[i] - mus[i])
    logt[i] ~ dnorm( mut[i], prect)I(cent[i], )
}
```

A.3 Partitioning an inverse Wishart prior

When the likelihood is expressed in this format, priors are required for `precs`, `prect`, and `regcoef`, rather than for Σ^{-1} per se. This is easily accomplished by *partitioning* the implied inverse Wishart prior on Σ according to well-known normal theory laid out for example in (16; 17). If

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \sim IW(S = \begin{bmatrix} S_1^2 & S_{12} \\ S_{12} & S_2^2 \end{bmatrix}, \nu)$$

, and if $\Sigma_{22.1} \equiv \sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^2}$ then

$$p(\sigma_1^2, \frac{\sigma_{12}}{\sigma_1^2}, \Sigma_{22.1}) = p(\sigma_1^2)p(\frac{\sigma_{12}}{\sigma_1^2}|\Sigma_{22.1}), p(\Sigma_{22.1})$$

where

$$\begin{aligned} p(\sigma_1^2) &= IG(\frac{\nu-1}{2}, \frac{S_1^2}{2}) \\ p(\Sigma_{22.1}) &= IG(\nu 2, \frac{S_{22.1}}{2}) \\ p(\frac{\sigma_{12}}{\sigma_1^2}|\Sigma_{22.1}) &= N(\frac{S_{12}}{S_1^2}, \frac{\Sigma_{22.1}}{S_1^2}) \end{aligned}$$

The following code fragment from the log-normal model converts a Wishart prior on the precision matrix Σ^{-1} into equivalent priors on `precs`, `prect`, and `regcoef`.

```
# The next lines are equivalent to
#     Siginv[1:2,1:2] ~ dwish( R[1:2,1:2], nu)

as <- (nu-1) /2
bs <- R[1,1] / 2
precs ~ dgamma(as,bs)

at <- nu / 2
bt1 <- S[2,2] - pow(S[1,2], 2) / S[1,1]
bt <- bt1 / 2
prect ~ dgamma(at, bt)

priorprec <- S[1,1] * prect
priormean <- S[1,2] / S[1,1]
```

```
regcoef ~ dnorm(priormean, priorprec)

# transformations to quantities of interest
Sigma[1,1] <- 1 / precs
Sigma[1,2] <- regcoef * Sigma[1,1]
Sigma[2,2] <- 1/prect + pow(Sigma[1,2],2) / Sigma[1,1]
RE3 <-( betat[2] / sqrt(Sigma[2,2] ))/ (betas[2] * sqrt(precs))
gammaz <- Sigma[1,2] / sqrt(Sigma[1,1] * Sigma[2,2])
```

In the exponential model, `Sigma` represents the variance/covariance matrix of $[s_i, z_i^*]^T$, and consequently `Sigma[2,2]` is fixed at 1. Then `surv[i]` is z_i^* with its variance multiplied by (or, as WinBUGS requires, its precision divided by) $2\lambda_i exp[-2(\beta_{0,t} + \beta_{1,t} trt_i + \beta_{2,t} strat_i)]$. The corresponding sections of code for the exponential model are as follows:

```
for (i in 1:N) {
        lambda[i] ~ dexp(1)
        mus[i] <- betas[1] + betas[2] * trt[i] + betas[3] * strat[i]
        s[i] ~ dnorm( mus[i], precs )

        mut[i] <-  regcoef  * (s[i] - mus[i])

        transprect[i] <- prect * exp(  2 * (betat[1] +
        betat[2] * trt[i ] + betat[3] * strat[i]) ) / (2 * lambda[i])

        surv[i] ~ dnorm(mut[i], transprect[i])I(cent[i], )
}
```

```
as <- (nu-1) /2
bs <- R[1,1] / 2
precs ~ dgamma(as,bs)

# define prect to force Sigma[2,2] = 1
prect <- 1 / (1 - pow(regcoef,2) * Sigma[1,1])

priormean <- R[1,2] / R[1,1]
priorprec <- R[1,1] * prect
regcoef ~ dnorm(priormean, priorprec)

# transformations to quantities of interest
Sigma[1,1] <- 1 / precs
RE3 <-betat[2] / (betas[2] * sqrt(precs)  )
gammaz <- regcoef * sqrt( Sigma[1,1])
```

Complete WinBUGS code for all three models is available on the author's webpage:
www.stat.uiowa.edu/ kcowles