Hierarchical Proportional Hazards Regression Models for Highly Stratified Data

Bradley P. Carlin and James S. Hodges

Division of Biostatistics, School of Public Health, University of Minnesota, Box 303 Mayo Building, Minneapolis, Minnesota 55455, U.S.A.

August 2, 1999

SUMMARY

In clinical trials conducted over several data collection centers, the most common statistically defensible analytic method, a stratified Cox model analysis, suffers from two important defects. First, identification of units which are outlying with respect to the baseline hazard is awkward, since this hazard is implicit (rather than explicit) in the Cox partial likelihood. Second (and more seriously), identification of modest treatment effects is often difficult, since the model fails to acknowledge any similarity across the strata. In this paper we consider a number of hierarchical modeling approaches which preserve the integrity of the stratified design while offering a “middle ground” between traditional stratified and unstratified analyses. We investigate both fully parametric (Weibull) and semiparametric models, the latter based not on the Cox model but on an extension of an idea by Gelfand and Mallick (1995, Biometrics 51, 843–852) which models the integrated baseline hazard as a mixture of monotone functions. We illustrate the methods using data from a recent multicenter AIDS clinical trial, comparing their ease of use, interpretation, and degree of robustness with respect to estimates of both the unit-specific baseline hazards and the treatment effect.

Key words: Baseline hazard function; Bayesian methods; Cox model; Markov chain Monte Carlo; Partial likelihood.
1 Introduction

In February 1996, a data safety and monitoring board considered a clinical trial comparing two treatments for *Mycobacterium avium* complex (MAC), a disease common in late stage HIV-infected people. (The trial was Community Programs for Clinical Research on AIDS (CPCRA) trial 027, reported by Cohn et al., 1999.) Table 1 summarizes the primary-endpoint data available to the board. In this trial, eleven clinical centers (“units”) have enrolled between 1 and 13 patients in the trial, for a total of 69. Eighteen patients have died, five in treatment group (“tx group”) 1 and 13 in treatment group 2; the times of death or censoring (in half-days) are given in Table 1. We suppose the trial’s randomization to be stratified by unit, and thus begin our analysis with a Cox proportional hazards regression stratified by unit, using the partial likelihood associated approximations (Kalbfleisch and Prentice, 1980, Sections 4.4, 4.8).

If the largest time in a stratum is an event time, not a censoring time (or more generally, if one risk set is depleted), the partial likelihood derives no information from that event (Cox and Oakes, 1984, Section 7.2). To see this, suppose \( d_i \) failures (deaths) are observed among the \( n_i \) subjects in stratum \( i \). Denoting the \( k^{th} \) risk set (subjects not failed or censored at the instant just preceding the \( k^{th} \) failure) by \( R_{ik} \) for \( k = 1, \ldots, d_i \), the contribution to the partial likelihood from stratum \( i \) is

\[
L(\beta) = \prod_{k=1}^{d_i} \left( \frac{e^{\beta^T x_{ik}}}{\sum_{j \in R_{ik}} e^{\beta^T x_{ij}}} \right),
\]

where \( x_{ij} \) is the covariate vector associated with the \( j^{th} \) individual, and \( \beta \) is the corresponding parameter vector. If the largest time in this stratum corresponds to a death, the ratio inside the parentheses in equation (1) equals 1.

In our trial, four of the 18 events have the largest time in their strata. All four are in
treatment group 2; two of the four are the only event in their strata. The stratified analysis using 14 events estimates a relative risk (hazard ratio) of 1.9 for group 2 compared to group 1, with a 95% confidence interval of (0.6, 5.9) and a $P$-value of 0.24. By comparison, an unstratified analysis using all 18 events estimates a relative risk of 3.1, a 95% confidence interval of (1.1, 8.7), and a $P$-value of 0.02. (This de-attenuation in the regression effect is the opposite of the usual effect of moving from stratified to unstratified analysis.) In the actual study, the $P$-value of 0.02 was small enough to persuade the data and safety monitoring board to stop the trial, especially because two previous trials had found evidence that treatment 2 had a higher death rate than treatment 1. However, a $P$-value of 0.24 would have left the board little choice but to continue the trial.

The board seems to face a dichotomous choice: stratified or unstratified analysis? We argue that this is a false dichotomy, and that the stratified and unstratified analyses are better understood as limiting cases. (Whether a stratified randomization means that the strata should be modeled with different baseline hazards is a distinct issue, outside this paper’s scope.) The stratified analysis pays too high a price by implicitly assuming the strata are completely unrelated; the unstratified analysis is risky when the strata may indeed have different baseline hazards. As an alternative, we present analytic methods in which the stratum-specific baseline hazards are treated as draws from a population of hazard functions, and the data provide information about variability of the hazard function between units. If the functions are highly variable, the analysis moves toward the stratified analysis; if the functions are similar to each other, it moves toward the unstratified analysis; and if the situation is middling or unclear, it gives an answer between the two, much like shrinkage estimators in simpler problems (e.g., Efron and Morris, 1975). Our main interest is in situations with many strata and few events, because the efficiency cost of stratification diminishes as the total number of events increases in a fixed number of strata.
A simple compromise between stratified and unstratified analyses is to include a collection of unit-specific dummy variables, one for each center. However, this fixed-effects approach is generally inferior to a *frailty model*, which replaces the hazard function for subject $j$ in stratum $i$, $h(t|\beta, x_{ij}) = h_0(t) \exp(\beta^T x_{ij})$, by $h(t|\beta, x_{ij}, \gamma_i) = h_0(t) \gamma_i \exp(\beta^T x_{ij})$, where the frailties $\gamma_i$ are typically assumed to be i.i.d. variates from a specified distribution (e.g., gamma). Our work generalizes this idea by treating the *entire* baseline hazard as a random effect. We investigate both parametric and semiparametric forms for $h_{0i}(t)$, and illustrate advantages over the standard frailty model in differentiation and interpretation of the strata.

Previous research on the problem of many small strata has two streams. The first generalizes rank tests to improve their asymptotic approximations (Woolson and Lachenbruch, 1980; Schoenfeld and Tsiatis, 1987). The second stream is more parametric. Lee et al. (1993)'s method allows regressors besides the treatment indicator and gives confidence intervals. However, it assumes lognormal failure times and uses large-sample estimating equation methods like those of Liang and Zeger (1986). Thall and Lachin (1986) use the semiparametric partial-likelihood approach, but do not examine their methods’ small-sample performance. By contrast, our methods use exact (up to Monte Carlo error) Bayesian calculations, allow covariates, produce intervals, and have the desired shrinkage feature. Moreover, our semiparametric method retains flexible specifications for the baseline hazard functions, the most attractive feature of Cox regression.

Section 2 describes parametric (Weibull) and semiparametric versions of our approach, and details the associated computing strategies. Section 3 illustrates both approaches using the data in Table 1 and discusses how we tuned the computing algorithms. Finally, Section 4 discusses our findings and suggests directions for future research.
Hierarchical modeling of stratified survival data

We begin by considering a fully parametric model with proportional hazards and a Weibull baseline hazard. The Weibull parameters are easily interpretable and permit reasonable flexibility. We then consider a semiparametric model that retains proportional hazards but uses a more flexible baseline hazard function, with some accompanying disadvantages.

2.1 Parametric approaches

Let \( t_{ij} \) be the time to death or censoring for subject \( j \) in stratum \( i, j = 1, \ldots, n_i, i = 1, \ldots, k \). Let \( \gamma_{ij} \) be a death indicator (0 if alive, 1 if dead), and \( x_{ij} \) a treatment indicator (we use -1 for one treatment, 1 for the other, for reasons to be clarified later). With proportional hazards and a Weibull baseline hazard in each stratum, stratum \( i \)'s hazard is

\[
h(t; x_{ij}) = h_0(t) \exp(\beta_1 x_{ij}) = \rho_i t^{\rho_i - 1} \exp(\rho_i \beta_0 + \beta_1 x_{ij}),
\]

where \( \rho_i > 0 \) and \( \beta = (\beta_0, \beta_1)' \in \mathbb{R}^2 \). This parametrization has integrated baseline hazard \( H(t) = \{t \exp(\beta_0)\}^\alpha \), so \( \beta_0 \) is the log of a time scale accelerator and the \( \rho_i \) are shape parameters. This parameterization is computationally convenient because it partially orthogonalizes the likelihood in \( \beta_0 \) and the \( \rho_i \) while making them functionally independent (Louis, 1987). The strata may be similar, so we model the shape parameters as exchangeable, i.e., \( \rho_i \overset{iid}{\sim} G(\alpha, 1/\alpha) \), where \( G(a, b) \) denotes the gamma distribution with mean \( ab \) and variance \( ab^2 \). Thus, we model the \( \rho_i \)'s with mean 1, corresponding to a constant baseline hazard, and variance \( 1/\alpha \). Placing a flat prior on \( \beta \) and a low-information but proper \( G(c, d) \) prior on \( \alpha \) completes our hierarchical specification.

Denoting vectors with boldface, the joint posterior distribution for \( \beta, \rho, \) and \( \alpha \) is

\[
p(\beta, \rho, \alpha \mid t, x, \gamma) \propto L(\beta, \rho ; t, x, \gamma) \left\{ \prod_{i=1}^{k} p(\rho_i \mid \alpha) \right\} p(\alpha)
\]
Markov chain Monte Carlo (MCMC) methods allow sampling from this posterior (Della-portas and Smith, 1993; Carlin and Louis, 1996, Sec. 8.1.4). The likelihood’s complexity precludes closed-form full conditional posteriors, so each parameter is most easily sampled using univariate Metropolis-Hastings steps (Tierney, 1994; Carlin and Louis, 1996, Sec. 5.4.4). A simple algorithm uses Metropolis random walk steps with Gaussian proposals for $\beta_0$ and $\beta_1$, and Hastings independence steps with gamma proposals for $\alpha$ and the $\rho_i$’s. For instance, at iteration $g$ we generate a candidate $\beta_0^g \sim N(\beta_0^{(g-1)}, \sigma_{\beta_0}^2)$ and accept this candidate with probability $p(\beta_0^g, \beta_1^{(g-1)}, \rho^{(g-1)}, \alpha^{(g-1)} \mid t, x, \gamma) / p(\beta_0^{(g-1)}, \beta_1^{(g-1)}, \rho^{(g-1)}, \alpha^{(g-1)} \mid t, x, \gamma)$. Here, $\sigma_{\beta_0}^2$ is a tuning constant chosen to allow acceptance of roughly 50% of the candidates. With the treatment indicator $x_{ij}$ centered around 0, the two components of $\beta$ are roughly uncorrelated a posteriori, so generating them separately should not slow the algorithm’s convergence. The parametric models in this section can also be fit using the BUGS programming language; a Windows-based version of BUGS is freely available at the website http://www.mrc-bsu.cam.ac.uk/bugs/Welcome.html.

As a generalization of model (2), consider

$$h(t; x_{ij}) = \rho_i t^{\rho_i - 1} \exp(\rho_i \beta_{0i} + \beta_1 x_{ij}) ,$$

which has unit-specific random time accelerators $\beta_{0i}$ in addition to unit-specific random shape effects $\rho_i$. We model the $\beta_{0i}$’s as exchangeable draws from $N(\mu_0, \sigma_0^2)$ with a flat prior on $\mu_0$ and an $IG(\epsilon_0, f_0)$ prior on $\sigma_0^2$, where $IG$ denotes the inverse (reciprocal) gamma
distribution. The resulting changes to the unnormalized posterior (3) are easily handled by expanding the MCMC algorithm. (MCMC computations could accommodate non-normal distributions for the $\beta_i$, including Student’s $t$ and bimodal densities.) In particular, the $\beta_i$ are readily sampled using univariate Metropolis steps with Gaussian proposals, while $\mu_0$ and $\sigma_0^2$ can be drawn using Gibbs steps from their normal and inverse gamma full conditional posteriors, respectively.

A second generalization of this parametric model is

$$h(t; x_{ij}) = \rho_i t^{\rho_i-1} \exp(\rho_i \beta_{0i} + \beta_{1i} x_{ij}),$$

which adds a third set of unit-specific random effects, the treatment effects $\beta_{1i}$. This allows the size of the treatment effect to vary across strata. One might expect these effects to be fairly similar, suggesting an exchangeable $N(\mu_1, \sigma_1^2)$ model for the $\beta_{1i}$. Adding flat and $IG(\epsilon_1, f_1)$ priors for $\mu_1$ and $\sigma_1^2$ respectively, the unnormalized posterior is again straightforward, analogous to (3). MCMC implementation again involves Hastings steps with gamma proposals for $\alpha$ and the $\rho_i$, Metropolis steps with Gaussian proposals for the $\beta_0$ and $\beta_{1i}$, and Gibbs steps for $\mu_0$, $\mu_1$, $\sigma_0^2$ and $\sigma_1^2$.

Finally, if covariate information $z$ is available, we can include it either as a single fixed effect, $h(t; x_{ij}, z) = \rho_i t^{\rho_i-1} \exp(\rho_i \beta_{0i} + \beta_{1i} x_{ij} + \delta' z)$, or as a collection of stratum-specific random effects by replacing $\delta$ with $\delta_i$. The latter may be modeled as exchangeable but need not be. Fitting such models is again straightforward using MCMC methods.

### 2.2 Semiparametric approaches

The foregoing parametric models offer substantial flexibility, especially when they include random effects; to get the additional flexibility of semiparametric models, we move toward the Cox proportional hazards model (Cox and Oakes, 1984, Section 7.2). Within
the Bayesian framework, several authors (e.g., Carlin et al., 1993) have proposed using the partial likelihood as a likelihood to obtain a posterior distribution for the treatment effect. Kalbfleisch (1978) justified this asymptotically by placing a gamma process prior on the baseline hazard function (independent of the prior for the regression parameters), and showing that the marginal posterior for the regression parameters approaches a form proportional to the partial likelihood as the gamma process prior becomes arbitrarily diffuse. Gustafson (1997) and Sargent (1998) extended this approach to allow (among other things) stratum-specific treatment effects $\beta_{1i}$.

However, the Cox model does not allow fully hierarchical modeling of stratum-specific baseline hazards, because the baseline hazard is implicit in the partial likelihood computation. As an alternative, we use the approach developed by Gelfand and Mallick (1995), which flexibly models the integrated (cumulative) baseline hazard as a mixture of monotone functions. To develop the likelihood, let $h_0_i(t)$ be the baseline hazard in stratum $i$ at time $t$, and let $H_0(t)$ be the corresponding integrated baseline hazard. Define $J_0(t) = a_0 H_0_i(t) / \{a_0 H_0_i(t) + b_0\}$, which conveniently takes values in $[0, 1]$. Given $J_0(t)$, a known function of $t$ which transforms the time scale to the interval $[0, 1]$, we model $J_0(t)$ as a mixture of Beta cumulative distribution functions (cdf’s) “centered” around $J_0(t)$, namely $J_0(t) = \sum_{i=1}^{m} v_i I(B(J_0(t) ; r_i, s_i))$, where $\sum_{i=1}^{m} v_i = 1$ for all $i$ and $I(B(\cdot ; a, b)$ denotes the incomplete Beta function (i.e., the cdf of a $Beta(a,b)$ distribution). The idea is that any distribution function on $[0, 1]$ can be approximated arbitrarily well by a finite mixture of Beta cdf’s (Diaconis and Ylvisaker, 1985), so the same is true for $J_0$, an increasing function that maps $[0, 1]$ onto itself. Fix $m$, the number of Beta cdf’s to be mixed, at some integer and select $\{(r_i, s_i)\}$ so the resulting Beta densities cover the interval $[0, 1]$ (e.g., using the “evenly spaced” choices $r_i = \psi l$ and $s_i = \psi (m - l + 1)$ for some $\psi > 0$). The likelihood is thus a function of the fixed effects $\beta$ and the stratum-specific mixing weight vectors.
\( \mathbf{v}_i = (v_{i1}, \ldots, v_{im})', i = 1, \ldots, k. \)

To evaluate the likelihood \( L(\mathbf{\beta}, \mathbf{v}_1, \ldots, \mathbf{v}_k) \), we need expressions for \( H_0(t) \) and \( h_0(t) \) in terms of \( J_0(t) \) (i.e., in terms of the \( \mathbf{v}_i \)'s). The former is

\[
H_0(t) = \frac{b_0 J_0(t)}{a_0 (1 - J_0(t))} = \frac{b_0 \sum_{i=1}^m v_{ii} IB(J_0(t); r_i, s_i)}{a_0 \{1 - \sum_{i=1}^m v_{ii} IB(J_0(t); r_i, s_i)\}}, \tag{6}
\]

while the latter is obtained by differentiating (6) using the quotient and chain rules:

\[
h_0(t) = \frac{b_0 \frac{\partial}{\partial t} J_0(t)}{a_0 (1 - J_0(t))^2} = \frac{b_0 \frac{\partial}{\partial t} J_0(t) \sum_{i=1}^m v_{ii} Be(J_0(t); r_i, s_i)}{a_0 \{1 - \sum_{i=1}^m v_{ii} IB(J_0(t); r_i, s_i)\}^2}, \tag{7}
\]

where \( Be \) denotes the Beta probability density.

The leverage in this approach comes from modeling the \( \mathbf{v}_i \), which determine the stratum-specific baseline hazards. It is natural to model the \( \mathbf{v}_i \) as draws from a \( Dirichlet(\phi_1, \ldots, \phi_k) \), where we take \( \phi_1 = \cdots = \phi_k = \phi \). As shown by Mallick (1995), this choice combined with the evenly spaced \( r_i \) and \( s_i \) mentioned above implies that the model for the \( H_0(t) \) is centered by \( \bar{J}_0(t) \), in the sense that \( E[J_0(t)] = \bar{J}_0(t) \). It is intuitive, then, to specify \( \bar{J}_0(t) \) to represent a plausible central function around which the \( J_0(t) \) are distributed. We use the flat hazard function of the exponential distribution, which, recalling the presence of the intercept \( \beta_0 \) in \( \mathbf{\beta} \), is simply \( \bar{h}_0(t) = 1 \). Thus \( \bar{H}_0(t) = t \), and we can specify \( \bar{J}_0(t) = a_0 t / (a_0 t + b_0) \). Because the \( t_{ij} \) in Table 1 range from 4 to 468, we use scaling constants \( a_0 = 1 \) and \( b_0 = 100 \), leading to \( \bar{J}_0(t) \) values that largely cover the interval (0,1). The model is then completed by specifying a flat prior for \( \mathbf{\beta} \) and a prior \( h(\phi) \) on \( \phi \).

Writing \( \mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_k) \), the joint posterior is

\[
p(\mathbf{\beta}, \mathbf{v}, \phi \mid \mathbf{t}, \mathbf{x}, \gamma) \propto L(\mathbf{\beta}, \mathbf{v}; \mathbf{t}, \mathbf{x}, \gamma) \prod_{i=1}^k \tilde{p}(v_i \mid \phi) h(\phi)
\]
\[
\propto \prod_{i=1}^{k} \prod_{j=1}^{n_i} \{h_{0i}(t_{ij}) \exp(\beta_0 + \beta_1 x_{ij})\}^{n_{ij}} \exp\{-H_{0i}(t_{ij}) \exp(\beta_0 + \beta_1 x_{ij})\} \\
\times \left[ \prod_{i=1}^{k} \left\{ \frac{\Gamma(m_0)}{(\Gamma(\phi))^m} \prod_{l=1}^{m} v_{il}^{\phi-1} \right\} \right] h(\phi),
\]

where \( \sum_i v_{ii} = 1 \) for each \( i \) and \( H_{0i}(t) \) and \( h_{0i}(t) \) are in (6) and (7), respectively. Again we may sample from this posterior using univariate MCMC steps: Metropolis random walk steps with Gaussian proposals for the two components of \( \beta \), and Hastings independence steps with gamma and Dirichlet proposals for \( \phi \) and the \( v_i \), respectively. (The Dirichlet proposals are easily generated using \( m \) gamma random variates; see Devroye, 1986, p. 594.) The most time-consuming aspect of evaluating (8) is computing the incomplete Beta functions in \( h_{0i}(t_{ij}) \) and \( H_{0i}(t_{ij}) \), because they are themselves numerical integrals. Fortunately, the arguments of these functions (\( \tilde{J}_0(t_{ij}), r_i, \) and \( s_i \)) are not random variables, so the incomplete Betas can be computed and stored at the outset of the algorithm; only the mixing weights \( v_i \) must be updated from iteration to iteration.

3 Numerical illustrations

3.1 Parametric models

We now fit the foregoing models to the dataset in Section 1, beginning with the parametric models in Section 2.1. Consider first the model with unit-specific hazards \( \rho_i \) and time accelerators \( \beta_{0i} \), given in (4). To specify the \( \text{Gamma}(c, d) \) prior for \( \alpha \), we do not refer to past datasets but simply set \( c = 3 \) and \( d = 10 \). That is, our prior guess for the standard deviation of the \( \rho_i \)'s is \( 1/\sqrt{30} \approx 0.18 \), allowing a fairly broad region of values centered around 1. We also take \( \epsilon_0 = 3 \) and \( f_0 = 0.5 \), a vague but proper prior for \( \sigma_0^2 \) having unit mean and standard deviation. We use independent univariate Gaussian proposals for \( \beta_1 \) and the \( \beta_{0i} \) having standard deviations 0.25 and 0.5, respectively, and \( \text{Gamma}(4, 9) \) and
Gamma(20, 0.05) proposals for $\alpha$ and the $\rho_i$, respectively, producing Metropolis-Hastings acceptance ratios near 50%, as suggested by Gelman et al. (1996) and by MCMC “folklore.”

Running five initially overdispersed parallel chains of our algorithm for 10,000 iterations each, we found fairly rapid convergence to the stationary distribution, as measured by sample autocorrelations within the chains, cross-correlations between parameters, and by plots of the sampled values. Thus we discarded only the first 50 samples from each chain, leaving 49,750 samples for posterior inference. These samples give a posterior median for $\beta_1$ of 0.57, and a 95% equal-tail credible interval of (0.07, 1.15), suggesting a statistically significant elevation in log-relative hazard in treatment group 2. (A Bayesian decision on whether to stop the trial at this point would depend on the posterior probability mass lying inside an indifference zone for $\beta_1$, within which we would be indifferent between treatments 1 or 2; see Spiegelhalter et al., 1994.) Our point estimate of the relative risk is thus $\exp\{\hat{\beta}_1 - (-\hat{\beta}_1)\} = \exp(2\hat{\beta}_1) = 3.1$, identical (to one decimal place) to our Section 1 unstratified analysis estimate.

Figure 1 shows the stratum-specific baseline hazards. The first three panels plot the posterior mean, posterior standard deviation, and approximate pointwise upper 95% confidence bound (the pointwise mean plus two pointwise standard deviations) for the baseline hazard functions, for each of the 11 strata. Clearly the strata differ modestly; most have nearly flat hazards, corresponding to the exponential model, around which we centered the $\rho_i$.

The lower right panel of Figure 1 plots $E(\beta_{0i}^*|t, x, \gamma)$ versus $E(\rho_i|t, x, \gamma)$, where $\beta_{0i}^* = \rho_i\beta_0$. We use this transformation so that the baseline hazard in (4) becomes $h(t; x_{ij}) = \rho_i t^{\rho_i - 1} \exp(\beta_{0i}^* + \beta_1 x_{ij})$, i.e., $\beta_{0i}^*$ is a true “intercept” in the model, augmenting the hazard by a constant amount over time regardless of $\rho_i$. (By contrast, the effect of $\beta_{0i}$ on the hazard depends on whether $\rho_i$ is greater or less than 1, making it hard to interpret. Our transformation thus illustrates a key advantage of Monte Carlo methods, namely the ease with which samples from a computationally convenient scale may be converted to a new scale
The vertical reference line is at -6.82, the posterior median for \( \mu_0 \), while \( \rho_i = 1 \) is marked by a reference line. The \( E(\rho_i|t, x, \gamma) \) are all in the interval (0.90, 1.06), substantially narrower than the 95% prior interval of (0.69, 1.37); similar shrinkage occurs for the \( \beta_{0i} \) posterior means. A generally decreasing trend is also evident in the lower right panel, implying that units with higher overall risk (as measured by \( \beta_{0i}^* \)) tend to have decreasing baseline hazards over time (\( \rho_i < 1 \)). The exemplars of this trend are units F and I, which had several early deaths but also several long-term survivors.

Next, we consider the model with unit-specific \( \rho_i, \beta_{0i}, \) and \( \beta_{1i} \), given in (5). We specify the prior for \( \sigma_i^2 \), the variance of \( \beta_{1i} \), by setting \( \epsilon_1 = 3 \) and \( f_1 = 1 \), again a vague specification with mean and standard deviation equal to \( \sqrt{1/2} \approx 0.7 \), slightly smaller than for \( \sigma_0^2 \) because we expect less variation in the treatment effects than in the intercepts.

Using the same MCMC control parameters, convergence to the posterior distribution was unequivocal, although the larger model had higher posterior correlations among parameters and hence slower convergence. The point and interval estimates for \( \mu_1 \), the center of the distribution of the \( \beta_{1i} \), are 0.65 and (-0.05, 1.40). While the point estimate is comparable to that for \( \beta_1 \) in the previous model, the slightly wider interval estimate appears to be the result of \textit{bona fide} variation in the treatment effects across strata. The fitted hazard functions by stratum are given in the top row of Figure 2, while the bottom row gives scatter plots summarizing the posterior means of the \( \rho_i, \beta_{0i}^*, \) and \( \beta_{1i} \). For the latter, reference lines are at \( \rho_i = 1 \) and the posterior medians of \( \mu_0 \) and \( \mu_1 \). Unit F continues to be anomalous, with high overall risk, decreasing baseline hazard, and relatively low treatment effect. By contrast, unit D emerges from the pack with a high treatment effect; comparison to Figure 1 indicates that the previous model allocated some of unit D’s treatment effect to overall risk (\( \beta_{0D}^* \)).

Ignoring the outlying units I and F (which, not coincidentally, have the most events), the \( \beta_{1i} \) have more posterior variability (i.e., less shrinkage) than the \( \beta_{0i}^* \). This suggests our
dataset contains more information about the treatment effect than about either aspect of the baseline hazard. Nonetheless, all 11 of the $\beta_i$ have positive posterior means, implying that, after accounting for the differences between the strata, treatment 1 is superior in every unit. Finally, the posterior median of $\sigma_0$ was 0.70, with a central 95% interval of (0.47, 1.13), scarcely different from the previous model, while $\sigma_1$ had posterior median and 95% interval of 0.57 and (0.36, 1.00), respectively.

A rough idea of the relative fit of these two models is provided by the expected posterior log-likelihood score, $E[\log L(\beta, \rho) | t, x, \gamma]$. Monte Carlo estimates of these scores are reported in the legends of Figures 1 and 2. While the usual asymptotics justifying chi-square differences for such scores (or their penalized counterparts, AIC and BIC) are not appropriate in our random effects model setting, they do still provide an overall measure of fit that is consistent across models. Here, the scores differ by only about 1 point, suggesting similar fit. We prefer model (5) for its greater insight into the workings of the various units, as highlighted by the second row of plots in Figure 2.

### 3.2 Semiparametric model

We now turn to the semiparametric stratified model having posterior proportional to (8). After experimenting with $m = 3$ Beta components, we settled on $m = 5$ and set $r = (1, 2, 3, 4, 5)$ and $s = (5, 4, 3, 2, 1)$ for an evenly spaced collection of Beta cdf’s. We began with a flat prior for the Dirichlet parameter $\phi$, $h(\phi) = 1$, and the following Metropolis-Hastings control parameters: Gaussian proposals with standard deviation 1 and 0.2 for $\beta_0$ and $\beta_1$, respectively, $\text{Dirichlet}(\phi^*1)$ proposals for the $\mathbf{v}$; with $\phi^* = 1$, and $\text{Gamma}(\epsilon^*, f^*)$ proposals for $\phi$ with $\epsilon^* = f^* = 1$ (the standard exponential distribution). Five initially overdispersed parallel sampling chains, each of 1000 iterations, showed poor convergence for $\phi$ and the $\mathbf{v}$’s, suggesting a proper prior might be required for $\phi$. Switching $\phi$’s prior $h(\phi)$ to a $\text{Gamma}(\epsilon^*, f^*)$ (identical to the $\phi$ candidate distribution) helped somewhat, but
the posterior range of $\phi$ values sampled was nearly as wide as that of the prior. We tried replacing $r$ and $s$ by $\psi r$ and $\psi s$ for various $\psi > 0$, thus changing the peakedness of the Beta mixands, but again, convergence for $\phi$ was disappointing.

To check whether our data truly had insufficient information to estimate $\phi$, we fit model (8) with $\phi$ assumed known. We investigated the sensitivity of the expected posterior log-likelihood score for this model over a range of choices for $\phi$ and $\psi$. For a given $\psi$, $E[\log L(\beta, \nu) | t, x, \gamma]$ varied by only a few points for $\phi \in [0.1, 10]$; thus, our data contain little information about $\phi$. They are slightly more informative about $\psi$, though even here, $\psi \in [3, 5]$ produce roughly equivalent scores. (Estimating $\psi$ as part of the MCMC algorithm is a real computational inconvenience because its location inside the incomplete Beta functions requires reevaluation of these functions at every iteration.)

Figure 3 plots results from one of the best-fitting (and quickest-converging) models we considered, namely $\phi = 0.5$ and $\psi = 3$. The first row of this figure gives the IB and Be mixands and the fitted $J_0_i(t)$ as functions of $J_0(t)$. The second row gives the fitted mean, standard deviation, and approximate 95% upper confidence limits for the hazard functions by unit, similar to the figures in Section 3.1. While the fitted IB mixture $J_0_i(t)$ is monotone, the corresponding baseline hazard need not be – a feature not shared by the parametric models. However, the semiparametric model also produces an odd dip in the hazard, down to 0 at $t = 0$. This is because at $t = 0$, equation (7) reduces to $h_{0_i}(t) = \sum_{i=1}^{m} v_{il} Be(0; \psi r_i, \psi s_i)$. For $r$ and $s$ as chosen above, this mixture is small when $\psi > 1$, because all of the mixands equal 0. Conversely, it is quite large if $\psi < 1$, because the first mixand is infinite. This behavior near the origin seems to be a drawback of the method.

On the positive side, this model’s results for $\beta_0$ and $\beta_1$ are very similar to those under the parametric models. For example, for the ($\phi = 0.5$, $\psi = 3$) semiparametric model we obtained point and 95% interval estimates of $-7.07$ and $(-8.01, -6.16)$ for $\beta_0$, while for $\beta_1$ the
estimates were 0.59 and (0.09, 1.17). These are quite similar to the corresponding results from the previous subsection, especially for the drug effect, which is of primary interest. Moreover, these results varied little across the models considered in our sensitivity analysis, suggesting robustness in our estimate of the treatment effect with respect to the baseline hazard. Finally, all our methods lean strongly toward ignoring the stratification; our overall conclusion about the difference between the treatment groups is, as one would hope, very close to what the unstratified analysis said in Section 1.

4 Discussion

We have demonstrated methods for censored survival data having many strata and few patients per stratum. The methods allow flexible models and small-sample inferential summaries computed with arbitrary accuracy. The price is more intensive computing, but the computing methods are not exotic and run quickly when coded in compiled languages.

The extra flexibility of the semiparametric model had little effect on our estimate of the treatment effect and was rather more opaque than the parametric method. Our dataset appears to have little information about the baseline hazard functions, which may be an inherent feature of situations with many strata and few cases per stratum. Also, the anomalous behavior of the hazard at $t = 0$ is awkward. However, radical changes to $r$ and $s$ in the baseline hazard (results not shown) had almost no effect on the posterior of the treatment effect $\beta_1$. This approach also does not sacrifice 4 of the 18 events, as does the stratified partial likelihood analysis. Thus, when the baseline hazard is of little inherent interest, but one would like to allow different baseline hazards across strata and to leave their shape largely unspecified, the semiparametric model may be useful.

Both of our approaches change the interpretation of proportional hazards regressions somewhat. In a Cox regression with a treatment indicator as the only explanatory variable,
the test of the treatment effect is formally identical to the log-rank test (Kalbfleisch and Prentice, 1980, Section 4.2) and can thus be interpreted as an approximate randomization test. Our approach breaks this connection and has no randomization interpretation.

Computation by MCMC is flexible enough to permit the analysis of more complex datasets and other models for the baseline hazard. As an illustration of the former, in clinical trial networks it is common to have the same units engaged in several protocols simultaneously. By adding one more level (i.e., protocol) to our model hierarchy, we could identify “high-flying” and “low-flying” units (i.e., ones which consistently find or fail to find treatment effects when they exist) while still taking full advantage of the stratified design.

Regarding the second situation, some trials stratify patients by clinical unit and by a baseline characteristic, such as viral load above or below 10,000 equivalents/ml. It is common to analyze such data with two strata per clinical unit (e.g., one per baseline viral load stratum), so the number of strata can quickly grow large. An alternative to this analysis or to one of our models would be a model that “crosses” the stratification, or assumes a product formulation such as $h_{0i}(t) = h_{0i}(t) \times h_{0i}(t)$, where the two subscripts represent units and viral-load groups. We hope to report on these and other issues in a subsequent paper.

Acknowledgements

The research of the first author was supported by National Institute of Allergy and Infectious Diseases (NIAID) Grant R01-AI41966, while that of the second author by NIAID Contract NO1-AI05073. The authors thank Prof. Tom Louis for helpful discussions, and the Community Programs for Clinical Research on AIDS, also supported by NIAID Contract NO1-AI05073, for permission to analyze the MAC data set.


<table>
<thead>
<tr>
<th>unit</th>
<th>tx group</th>
<th>time</th>
<th>unit</th>
<th>tx group</th>
<th>time</th>
<th>unit</th>
<th>tx group</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>74+</td>
<td>E</td>
<td>1</td>
<td>214</td>
<td>H</td>
<td>1</td>
<td>74+</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>248</td>
<td>E</td>
<td>2</td>
<td>228+</td>
<td>H</td>
<td>1</td>
<td>88+</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>272+</td>
<td>E</td>
<td>2</td>
<td>262</td>
<td>*</td>
<td>H</td>
<td>148+</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>344</td>
<td>*</td>
<td>F</td>
<td>1</td>
<td>6</td>
<td>H</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>4+</td>
<td>F</td>
<td>2</td>
<td>16+</td>
<td>I</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>156+</td>
<td>F</td>
<td>1</td>
<td>76</td>
<td>I</td>
<td>2</td>
<td>16+</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>100+</td>
<td>F</td>
<td>2</td>
<td>202</td>
<td>I</td>
<td>1</td>
<td>120+</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>20+</td>
<td>F</td>
<td>1</td>
<td>268+</td>
<td>I</td>
<td>2</td>
<td>174+</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>64</td>
<td>F</td>
<td>2</td>
<td>368+</td>
<td>I</td>
<td>1</td>
<td>268+</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>88</td>
<td>F</td>
<td>1</td>
<td>380+</td>
<td>I</td>
<td>2</td>
<td>396+</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>148+</td>
<td>F</td>
<td>1</td>
<td>424+</td>
<td>I</td>
<td>1</td>
<td>286+</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>162+</td>
<td>F</td>
<td>2</td>
<td>428+</td>
<td>I</td>
<td>1</td>
<td>366</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>184+</td>
<td>F</td>
<td>2</td>
<td>436+</td>
<td>I</td>
<td>2</td>
<td>396+</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>188+</td>
<td>G</td>
<td>2</td>
<td>32+</td>
<td>I</td>
<td>1</td>
<td>468+</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>198+</td>
<td>G</td>
<td>1</td>
<td>64+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>382+</td>
<td>G</td>
<td>1</td>
<td>102</td>
<td>J</td>
<td>1</td>
<td>18+</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>436+</td>
<td>G</td>
<td>2</td>
<td>162+</td>
<td>J</td>
<td>1</td>
<td>36+</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>50+</td>
<td>G</td>
<td>2</td>
<td>182+</td>
<td>J</td>
<td>2</td>
<td>160+</td>
</tr>
<tr>
<td>E</td>
<td>2</td>
<td>64+</td>
<td>G</td>
<td>1</td>
<td>364+</td>
<td>J</td>
<td>2</td>
<td>254</td>
</tr>
<tr>
<td>E</td>
<td>2</td>
<td>82</td>
<td></td>
<td></td>
<td></td>
<td>K</td>
<td>1</td>
<td>28+</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>186+</td>
<td>H</td>
<td>2</td>
<td>22+</td>
<td>K</td>
<td>1</td>
<td>70+</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>214+</td>
<td>H</td>
<td>1</td>
<td>22+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>2</td>
<td>106+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Survival data from CPCRA 027; time to death in half-days. Here, “+” indicates a censored observation, and “*” indicates that the largest observed time in the given stratum is a death.
Figure 1: Posterior summaries, parametric (Weibull) model with two sets of random effects. Mean log-likelihood score = -135.23; upper bounds for posterior standard deviations: $\beta_0^*$, 1.67; $\rho_i$, 0.23; upper bounds for Monte Carlo standard error estimates: $\beta_0^*$, 0.033; $\rho_i$, 0.0016.
Figure 2: Posterior summaries, parametric (Weibull) model with three sets of random effects. Mean log-likelihood score = -131.16); upper bounds for posterior standard deviations: $\beta_{0i}$, 1.64; $\beta_{1i}$, 0.70; $\rho_{0i}$, 0.21; upper bounds for Monte Carlo standard error estimates: $\beta_{0i}$, 0.040; $\beta_{1i}$, 0.016; $\rho_{i}$, 0.0015.
Figure 3: Posterior summaries, nonparametric model (mean log-likelihood score = -132.39).