Paper 252-2010

Survival Analysis: Overview of Parametric, Nonparametric and Semiparametric approaches and New Developments

Joseph C. Gardiner, Division of Biostatistics, Department of Epidemiology, Michigan State University, East Lansing, MI 48824

ABSTRACT

Time to event data arise in several fields including biostatistics, demography, economics, engineering and sociology. The terms *duration analysis*, *event-history analysis*, *failure-time analysis*, *reliability analysis*, and *transition analysis* refer essentially to the same group of techniques although the emphases in certain modeling aspects could differ across disciplines. SAS® procedures LIFETEST, LIFEREG, PHREG, RELIABILITY, and QLIM have different capabilities for analyzing duration data. Methods include Kaplan-Meier estimation, accelerated life-testing models, and the ubiquitous Cox model. Recent developments in SAS extend their reach to include analyses of multiple failure times, recurrent events, frailty models, Markov models and use of Bayesian methods. We present an overview of these methods with examples illustrating their application in the appropriate context.

INTRODUCTION

Survival Analysis is a collection of methods for the analysis of data that involve the time to occurrence of some event, and more generally, to multiple durations between occurrences of different events or a repeatable (recurrent) event. From their extensive use over decades in studies of *survival times* in clinical and health related studies and failures times in industrial engineering (e.g., reliability studies), these methods have evolved to special applications in several other fields, including demography (e.g., analyses of time intervals between successive child births), sociology (e.g., studies of recidivism, duration of marriages), and labor economics (e.g., analysis of spells of unemployment, duration of strikes). Books and monographs continue to be published in this area that attest to its rich methodology and versatility. See references for a partial list.

The typical context in biostatistics is a data gathering process that records an event time T measured from a specified time origin in a sample of patients. However, when follow up ends the event may not have occurred in some patients resulting in *right censored* event times. What we know is that *T exceeds U*, where U is the follow up time. The survival times of these patients are *censored*, and U is called the *censoring time*. Censoring will also occur if say a patient dies from causes unrelated to the endpoint under study, or withdraws from study for reasons not related to the endpoint. Such patients are *lost to follow up*. When there is a competing risk for the endpoint of death, it is important to ascertain whether death is due to the cause under study. Other forms of censoring are possible depending on the type of study. For example, if the true event time T is not observed but is known to be less than or equal to V, we have a case of *left censoring*. If all that is known about T is that it is somewhere between two times U and V ($U<V$), we say it is *interval censored.*

Generally, one records a number of covariates **z** (e.g., age, gender, comorbidity, treatment assignment etc.) whose influence on the distribution of T is of interest. Due to the longitudinal feature of the data gathering process some covariates are time-invariant while others could be time-varying. The latter may arise from intermediate events that influence the distribution of T. *Multi-state models* provide a means of analyzing data with multiple event times.

Despite our best intention in recording all covariates relevant to a specific analysis, we might encounter heterogeneity in patient samples that cannot be explained by the observed covariates alone. *Unobserved heterogeneity* is likely in observational studies. *Frailty models* and *finite-mixture models* can be very informative in this regard.

For time-fixed covariates **z,** the survival distribution $S(t \mid \mathbf{z}) = P[T > t \mid \mathbf{z}] = \exp(-H(t \mid \mathbf{z}))$ is expressed in terms of the cumulative hazard $H(t \mid \mathbf{z}) = \int_0^t h(u \mid \mathbf{z})du$ where $h(t \mid \mathbf{z})$ denotes the hazard function. (The relationship between S and H is more subtle when the distribution T is not continuous). We may interpret $h(t \mid \mathbf{z})\Delta t$ as a conditional probability because $h(t \mid \mathbf{z})\Delta t \approx P[T < t + \Delta t \mid T \geq t, \mathbf{z}]$. For this reason $h(t \mid \mathbf{z})$ is often referred to as the *instantaneous risk* of the event happening at time *t*. Other useful summary quantities in survival analysis are (suppressing dependence on **z**):

Mean survival time, $\mu = E(T) = \int_0^\infty S(t)dt$

Mean survival restricted to time L, $\mu_L = E(\min(T, L)) = \int_0^L S(t)dt$

Percentiles of survival distribution, $t_p = \inf\{t > 0 : S(t) \leq 1 - p\}$, 0<*p*<1

Mean residual life at time t, $r(t) = E(T - t \mid T > t) = \{S(t)\}^{-1} \int_t^\infty S(u)du$

Just as H determines S, the relationship between r and S is $S(t) = r(0)\{r(t)\}^{-1} \exp\left(-\int_0^t \{r(u)\}^{-1} du\right)$. Because survival data are often quite skewed with long right tails, the restricted mean survival or the median survival time are generally preferred as summary statistics.

The objectives in a survival analysis may include estimation of one or more of these statistics at specified covariate profiles and quantifying the influence of **z** (e.g., treatments, demographics) on survival. These goals can be achieved through modeling how **z** impacts T directly or indirectly through for example, the hazard $h(t \mid \mathbf{z})$. However, an initial analysis would typically employ nonparametric methods to estimate the survival function and summary statistics, and a comparison across several groups or sub-populations.

I. NONPARAMETRIC ANALYSIS

Procedure LIFETEST is the mainstay of nonparametric survival analysis. For right censored data it computes the Kaplan-Meier (product limit) estimator of the survival distribution S, its quartiles and the restricted mean μ_L . It provides tests of comparison of the survival distribution across two or more populations including adjustment of the p-value for multiple comparisons if warranted, and tests of trend for ordered alternatives. Using ODS graphics with the PLOTS= option can produce exquisite graphs for estimates for S, its derivatives, and pointwise confidence intervals or a confidence band for S.

ILLUSTRATIVE EXAMPLE 1

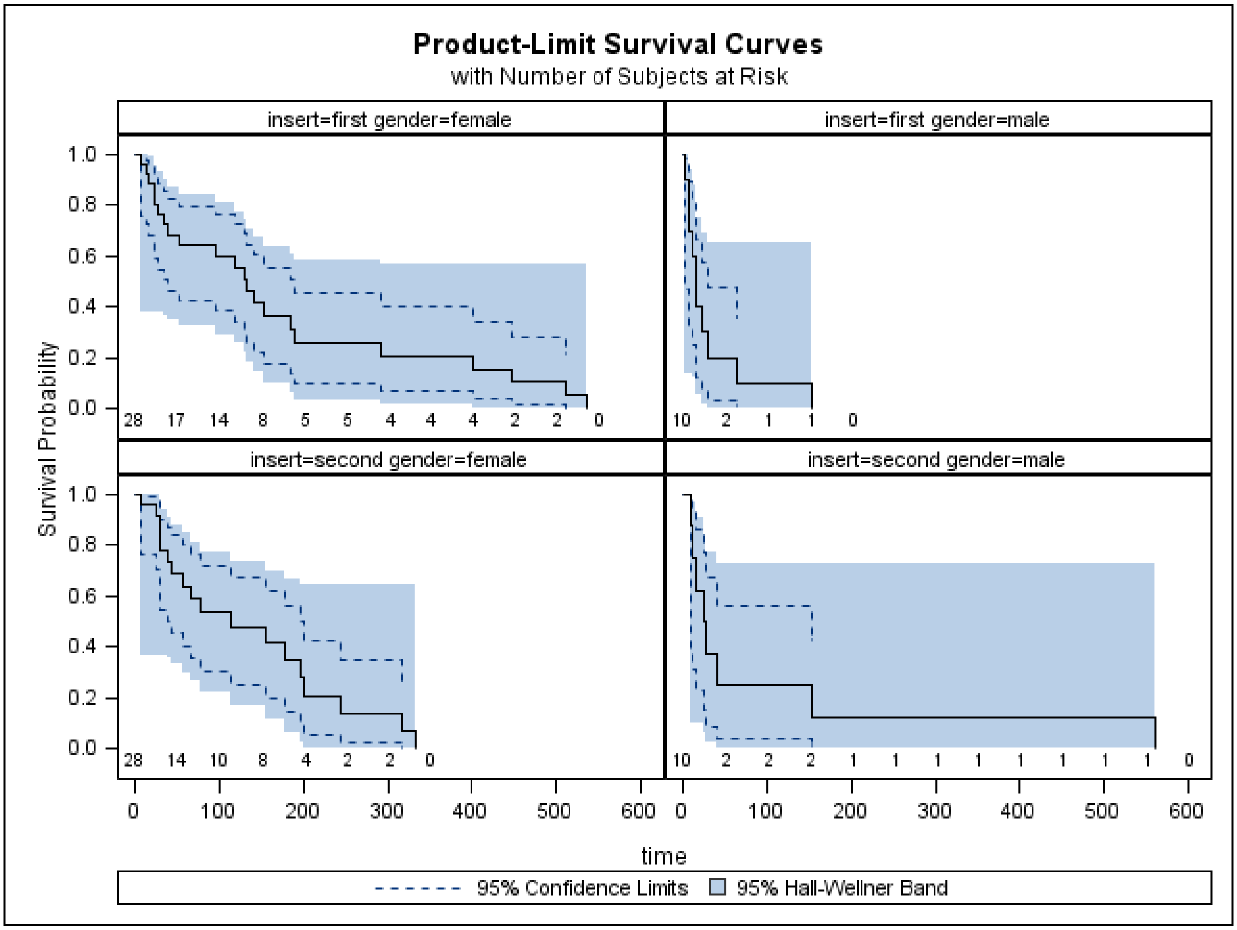
McGilchrist & Aisbett (1991) describe a study in 38 kidney dialysis patients where the time in days to infection at the catheter insertion point was recorded. Each patient (*i*) has two times. After the first insertion of the catheter, the time to infection T_{i1} if observed is recorded. If the catheter is removed for any reason other than infection, T_{i1} is considered right censored at the removal time U_{i1} . If infection occurs, the catheter is removed, the patient is treated and cleared of the infection and then after some time, a second catheter is inserted. The second time to infection T_{i2} , measured from the time of second insertion, is either observed or right censored if the catheter removed at time U_{i2} for any reason other than infection.

The sample data are $\{(X_{i1}, \delta_{i1}, X_{i2}, \delta_{i2}, \mathbf{z}_i) : 1 \leq i \leq n\}$ where $X_{ij} = \min(T_{ij}, U_{ij})$, $\delta_{ij} = [T_{ij} \leq U_{ij}]$ j=1,2 and δ_{ij} denotes the event indicator. Censoring is assumed independent of infection times. The data set KIDNEY has two records per patient with variables TIME, FAIL and covariates AGE (=average age between the insertions of the catheter), and GENDER. PATIENT and INSERT identify patient and the two infection times. Nonparametric methods use the accumulating count of events up to time *t*, $N_j(t)=\sum_{i=1}^{n}[X_{ij}\le t,\delta_{ij}=1]$ and the number at risk at time *t*, $Y_j(t)=\sum_{i=1}^{n}[X_{ij}\ge t]$.

a. ESTIMATION OF SURVIVAL CURVES

The following syntax will produce the product-limit estimates of infection time by insertion for females and males. Use of formats when applicable makes the output display more readable.

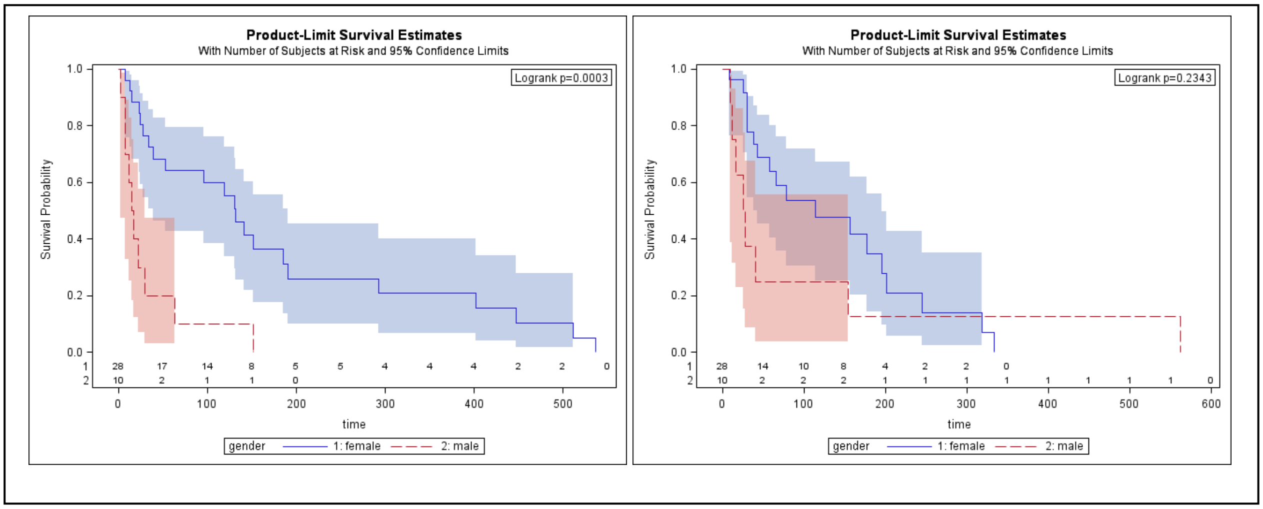
```
proc format;
value gender 0='male' 1='female';
value insert 1='first' 2='second';
run;

ods graphics on;
proc lifetest data=kidney
      plots=survival(nocensor cb=hw cl strata=panel atrisk=0 to 600 by 50);
strata insert gender;
time time*fail(0);
format gender gender. insert insert.;
run;
ods graphics off;
```



The NOCENSOR option suppresses the display of censored times (with the symbol +); the CB=HW option displays the simultaneous Hall-Wellner 95% confidence band for the survival curves; CL displays the lower and upper limits of the pointwise 95% confidence interval, and at the foot of each plot the ATRISK option shows the number of patients at risk of infection at specified times. STRATA=panel requests display of the four individual plots in a 2×2 panel, instead of the default overlay. For the first catheter insertion it appears that females have a longer infection-free duration than males. Stratified by INSERT the default logrank test for comparing the time to infection distributions for males and females is requested by

```
strata insert/group=gender test=logrank;
```

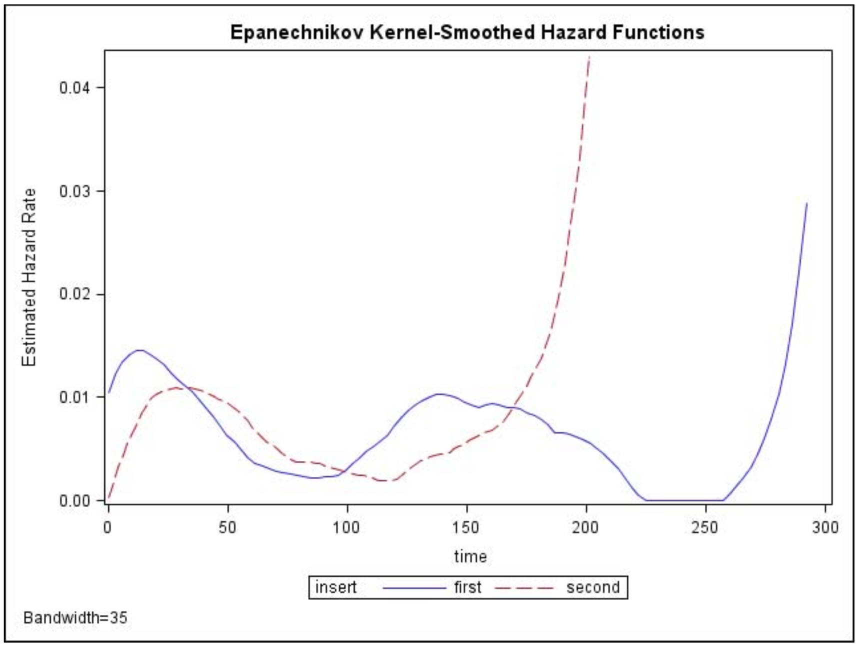
The test is significant (p=.0009). However, when comparisons are made separately at each catheter insertion this significance is seen in the figure below for the first insertion (left panel) and not the second insertion (right panel) (p=.2343). Pointwise 95% confidence intervals are also displayed.



b. CUMULATIVE HAZARD AND KERNEL-SMOOTHED HAZARD

The option `method=pl nelson` in the LIFTEST statement adds to the default table of Kaplan-Meier estimates the Nelson-Aalen estimate of the cumulative hazard for each stratum specified in the strata statement. With strata defined by INSERT, the estimate is $\hat{H}_j(t) = \int_0^t \{Y_j(u)\}^{-1} dN_j(u)$, j=1, 2. It could be used to obtain an estimator of the survival curve $\hat{S}_j(t) = \exp(-\hat{H}_j(t))$. In addition, the PLOTS option in the syntax below produces an estimate of the kernel-smoothed hazard function using the Epanechnikov kernel and bandwidth of 35 days. Optimal selection of a bandwidth by minimizing the mean integrated squared error was not feasible with this small data set. The risk of infection appears to first increase and then taper off, but another increase is seen for the second catheter insertion. The sharp rise towards the end is somewhat typical in this context. However, another choice of kernel or bandwidth might depict a different pattern. Larger bandwidth produces smoother curves.

```
ods graphics on;
proc lifetest data=kidney method=pl nelson
                  plots(only)=hazard(kernel=e bw=35);
strata insert;
time time*fail(0);
format insert insert.;
run;
ods graphics off;
```



II. PARAMETRIC MODELS-ACCELERATED FAILURE TIME MODEL

Procedures LIFEREG and RELIABILITY can be used for inference from survival data that have a combination of left, right and interval censored observations. The *accelerated failure time* (AFT) model is specified by $\log T = \mu + \sigma\varepsilon$ with location and scale parameters μ, σ, respectively. Covariate effects are modeled by $\mu = \mathbf{z}'\beta_1$, and additionally if plausible heteroscedasticity by $\log\sigma = \mathbf{z}'\beta_2$ By specifying a distribution for the random variable ε, independent of **z**, one induces a distribution on T. Estimation of parameters (β_1, β_2) is via maximum likelihood. Survival distributions within the AFT class are the exponential, Weibull, lognormal and loglogistic. All distributions have the functional form $S(t) = S_0((t/\alpha)^\gamma)$ where $\sigma = \gamma^{-1}, \mu = \log\alpha$, $\alpha > 0, \gamma > 0$, and S_0 is a known survival distribution

SAS also allows the generalized gamma (GG) distribution which has an additional shape parameter. Here ε has the one-parameter log-gamma distribution with shape parameter k>0, i.e., S_0 is the gamma survival distribution with shape parameter k. A re-parameterization suggested by Prentice (1974), recasts the GG in the AFT form $\log T = \mathbf{z}'\beta_1 + \sigma_0 Z$ where $k = \delta^{-2}$, $\sigma_0 = \sigma\delta$ and the distribution of Z is defined for all $\delta \neq 0$. In the limit as $\delta \to 0$, Z converges to the standard normal. SAS calls δ the *shape* and σ_0 the *scale* of the GG. Defined in this way, GG returns three special cases: with δ=0 the log normal; with δ=1 the Weibull; with δ=1 and σ_0 =1 the exponential. Testing of these restrictions within the parent GG is valid under maximum likelihood (ML) via for example, likelihood ratio and Lagrangian multiplier (score) tests.

a. FITTING PARAMETRIC MODELS

Initially we assume the within-patient times (T_{i1}, T_{i2}) are independent, making our sample comprise of 76 individual catheter insertions. The dist=gamma option requests fitting the GG to the model with covariates age and gender (Table 1, column 1). Wald tests produced by default indicate that age is not significant (p=.57), but gender is strongly significant (p<.0001). The positive estimated β-coefficient for female gender shows that female dialysis patients have a longer infection-free time compared to male patients.

```
proc lifereg data=kidney;
class gender;
model time*fail(0)=age gender/dist=gamma;
format gender gender.;
run;
```

Table 1: Summary of results of fitting parametric AFT models to infection times

	Maximum likelihood estimate (standard error)				
Parameter	**GG**	**Lognormal**	**Weibull**	**Exponential**	**Loglogistic**
Intercept	3.4188(0.5322)	3.4490(0.4939)	4.2916(0.5505)	4.4025(0.4971)	3.4052(0.4636)
AGE	–0.0054(0.0097)	–0.0054(0.0097)	–0.0042(0.0103)	–0.0046(0.0094)	–0.0073(0.0093)
GENDER-female	1.3830(0.3283)	1.3269(0.3261)	0.9655(0.3248)	0.8853(0.2871)	1.5526(0.3265)
Scale	1.1863(0.1085)	1.1847(0.1077)	1.1031(0.1035)	1(fixed)	0.6793(0.0720)
Shape	–0.0473(0.3164)	0(fixed)	1(fixed)	1(fixed)	na
–2 log L	197.032	197.053	206.197	207.348	198.532
BIC	218.686	214.375	223.520	220.340	215.855
LM test p-value	na	.887	<.0001	Shape <.001 Scale .316	na

The log-normal, Weibull and exponential models can be fitted directly by changing the dist= option (e.g., dist=lnormal) or by restricting the GG model's shape and scale parameters. Then Lagrangian multiplier (LM) 1-degree of freedom chi-square tests are produced.

For lognormal: $H_0 : \delta = 0$. Use `dist=gamma noshape1 shape1=0;`
For Weibull: $H_0 : \delta = 1$. Use `dist=gamma noshape1 shape1=1;`
For exponential: $H_0 : \delta = 1, \sigma_0 = 1$. Use `dist=gamma noshape1 shape1=1 noscale scale=1;`
For GG vs exponential, SAS does not produce a joint test 2 df LM test. However, in all situations except for comparing GG to loglogistic one can perform a likelihood ratio test (LRT). The LM test and LRT are asymptotically equivalent. In this example, compared to the GG model the simpler lognormal model is acceptable. It also has the lowest BIC.

b. ESTIMATION OF PERCENTILES

In the AFT model $\log T = \mathbf{z}'\beta + \sigma\varepsilon$ for a specified covariate profile **z** the 100(1–*p*)-th percentile t_p of the event time T is obtained from $t_p = \exp(\mathbf{z}'\beta + \sigma w_p)$ where w_p is the corresponding percentile of ε. Although statistics computed via the output statement from the fitted model in LIFEREG may be used for this purpose, an easier approach is to use PROC RELIABILITY. Suppose we want estimates and 95% confidence intervals for the 25th, 50th and 75th percentiles for females age=44 years and males age= 49 years. These are approximately the median ages in the data set. We add these two profiles to the data set kidney:

```
data covar;
input gender age @@;
datalines;
1 44 0 49
;
run;
```

```
data kidney2;
set covar(in=one) kidney;
if one then control=1;
else control=0;
run;
```

The same lognormal model statistics of Table 1 column 2 are obtained using the syntax below. The OBSTATS options produce the desired estimates. The ODS statements are added to permit some editing of the output data set ModObstats. The control= option reduces observation-wise calculations to the six records in ModObstats for the control variable value=1 only. Results are shown in Table 2.

```
ods select ModObstats;
proc reliability data=kidney2;
class gender;
distribution lognormal;
model time*fail(0)=age gender/obstats(quantiles=.25 .50 .75 control=control);
format gender gender.;
run;
```

Table 2: Estimates of percentiles in lognormal model

Age	Gender	p	Estimate	95% Lower CL	95% Upper CL
44	female	0.25	44.2	30.9	63.3
44	female	0.50	98.3	69.6	138.9
44	female	0.75	218.7	148.3	322.5
49	male	0.25	10.9	6.2	19.2
49	male	0.50	24.2	13.9	42.0
49	male	0.75	53.7	30.2	95.4

The LOGSCALE statement in RELIABILITY permits modeling heteroscedasticity in the scale parameter σ. For example `logscale age gender;` models $\log\sigma = \beta_{20} + \beta_{21} Age + \beta_{22}[Gender = female]$ resulting in a 6-parameter model. It turns out that (β_{21}, β_{22}) are not significant indicating that our simpler model is adequate. Generally, when specifying the covariates for (μ, σ) one should consider exclusion restrictions where at least one covariate present in $\mu = \mathbf{z}_1'\beta_1$ is excluded in $\log\sigma = \mathbf{z}_2'\beta_2$, and vice versa. This could ensure stability in ML estimates and estimated standard errors. Exclusions restrictions are informed by the subject matter rather than statistical considerations.

c. JOINT MODELING OF INFECTION TIMES

Consider a joint model for the infection times (T_{i1}, T_{i2}) allowing correlation between them. Create one record for both infection times, transform to the log scale $y_{ij}^* = \log T_{ij}$ and create a variable UB_{ij} (upperbound) as $UB_{ij} = \log X_{ij}$ if $\delta_{ij} = 0$ (censored); $UB_{ij} = \log(X_{ij} + c)$ if $\delta_{ij} = 1$ (infection time) where c is an arbitrary positive constant. Our model is $y_{ij}^* = \mathbf{z}_{ij}'\beta + u_{ij}$ with (u_{i1}, u_{i2}) ~Normal $(\mathbf{0}, \rho, \sigma_1, \sigma_2)$, but the observed analysis variable is:

$$y_{ij} = \begin{cases} UB_{ij} & \text{if } y_{ij}^* \geq UB_{ij} \\ y_{i1}^* & \text{if } y_{ij}^* < UB_{ij} \end{cases}.$$

The following syntax sets up the data set BIVAR with one record per patient.

```
proc sort data=kidney; by patient; run;
data bivar;
merge kidney(keep=insert patient time fail age gender where=(insert=1))
      kidney(keep=insert patient time fail where=(insert=2)
                   rename=(time=time2 fail=fail2));
by patient;
drop insert;
ub=log(time+10); ub2=log(time2+10); /*c=10*/
lgtime=log(time);  lgtime2=log(time2);
if fail=0 then ub=lgtime;
if fail2=0 then ub2=lgtime2;
run;
```

We use the same covariates (AGE, GENDER) in the two-equation model although generally covariate specification should consider exclusion restrictions. PROC QLIM estimates the joint model by maximum likelihood (Table 3).

```
proc qlim data=bivar;
class gender;
format gender gender.;
endogenous lgtime~censored(UB=UB);
endogenous lgtime2~censored(UB=UB2);
model lgtime=age gender;
model lgtime2=age gender;
run;
```

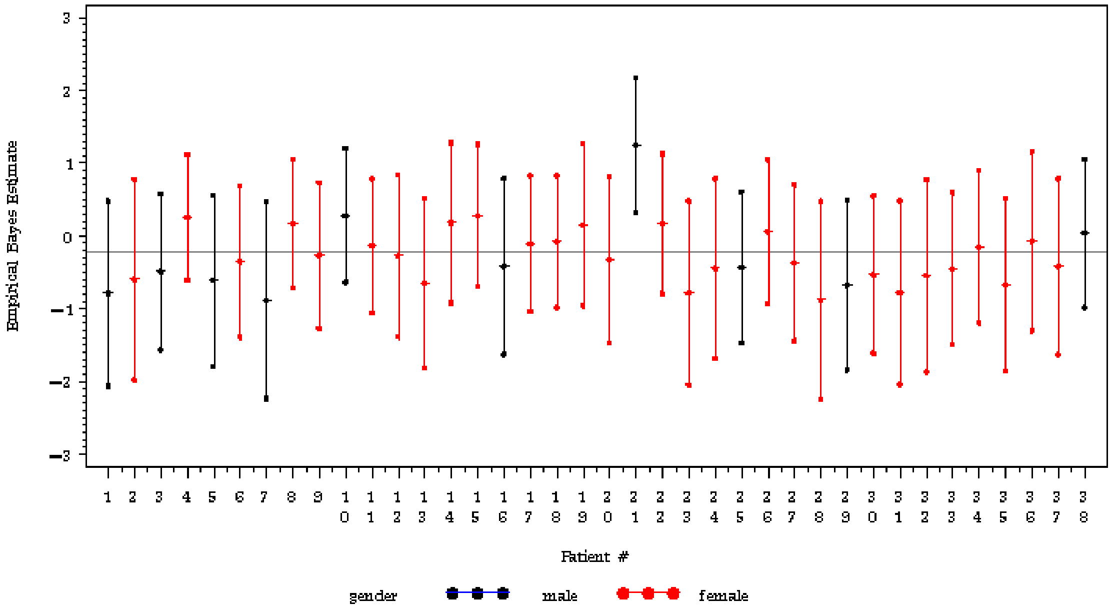
Table 3. Parameter estimates for joint model for infection times

Parameter		DF	Estimate	Standard Error (Hessian)	t Value	Approx Pr > |t|	Standard Error (QML)
lgtime.Intercept		1	3.411765	0.702542	4.86	<.0001	0.532027
lgtime.age		1	–0.013135	0.013567	–0.97	0.3330	0.011349
lgtime.gender	female	1	1.743168	0.453233	3.85	0.0001	0.434777
lgtime.gender	male	0	0	.	.	.	.
_Sigma.lgtime		1	1.205189	0.150094	8.03	<.0001	0.117485
lgtime2.Intercept		1	3.481873	0.662375	5.26	<.0001	0.649470
lgtime2.age		1	0.004995	0.013432	0.37	0.7100	0.013230
lgtime2.gender	female	1	0.866471	0.457612	1.89	0.0583	0.506512
lgtime2.gender	male	0	0	.	.	.	.
_Sigma.lgtime2		1	1.102781	0.148801	7.41	<.0001	0.131952
_Rho		1	0.144550	0.181132	0.80	0.4249	0.209544

The Wald test for no correlation $H_0 : \rho = 0$ is not significant. The gender effect is strong for the first insertion but weak for the second supporting what we had seen in our nonparametric analysis. Standard errors are obtained from the Hessian matrix based on the second derivative of the log-likelihood. Optionally, with `covest=qml` added to the PROC QLIM statement, quasi-maximum likelihood (QML) standard errors are obtained as a 'sandwich' of the Hessian and outer product (OP) matrices. QML standard errors are shown in the last column of Table 3. To obtain standard errors from OP only use `covest=op`. Although asymptotically equivalent, in finite samples the results from the three methods could differ.

d. FRAILTY MODEL

Another approach to incorporating correlation between the infection times is through a shared frailty ν_i, a random effect, via $\log T_{ij} = \mathbf{z}'_{ij}\beta + \nu_i + \sigma\varepsilon_{ij}$. Under the assumption that (T_{i1}, T_{i2}) are conditionally independent given $(\nu_i, \mathbf{z}_{i1}, \mathbf{z}_{i2})$, ML estimation of the marginal model based on the data $\{(X_{ij}, \delta_{ij}, \mathbf{z}_{ij}) : j = 1,2, 1 \le i \le n\}$ can be carried out under assumed parametric distributions on $(\nu_i, \varepsilon_{ij})$. Currently there is no direct SAS procedure to carry out the computations. However, informed by LIFEREG for suitable starting values of the model's parameters, PROC NLMIXED can be used to optimize the marginal likelihood. Post estimation provides empirical Bayes (EB) estimates of ν_i, ie, $E(\nu_i \mid data)$. The plot below shows the EB estimates and 95% confidence limits for the 38 patients in the sample for the Weibull model with $\nu_i \sim N(-\frac{1}{2}\sigma_\nu^2, \sigma_\nu^2)$. The frailty effect is strong with a LRT p-value<.01, but the effect appears to be influenced by patient #21.



III. SEMIPARAMETRIC MODEL-PROPORTIONAL HAZARDS MODEL

The workhorse of survival analysis for over three decades, the proportional hazards model (PHM) assumes $h(t \mid \mathbf{z}) = h_0(t)\exp(\mathbf{z}'\beta)$ where h_0 is an unspecified baseline hazard function and the parameter β is unknown. For time-invariant covariates, $S(t \mid \mathbf{z}) = S_0(t)^{\exp(\mathbf{z}'\beta)}$ where S_0 is the survival function corresponding to h_0. Given two covariate profiles $\mathbf{z}_1, \mathbf{z}_2$ the hazard ratio $h(t \mid \mathbf{z}_1) / h(t \mid \mathbf{z}_2) = \exp((\mathbf{z}_1 - \mathbf{z}_2)'\beta)$ is constant in time. The *stratified* PHM given by $h_k(t \mid \mathbf{z}) = h_{0k}(t)\exp(\mathbf{z}'\beta)$ maintains the proportional hazards assumption in each stratum k for a K-level stratification factor. For example, survival data from a multicenter clinical trial are often analyzed with center as the stratifying variable.

In addition to analysis based on the traditional PHM, enhancements to PROC PHREG allow for several additional data structures. These include time-dependent covariates, multiple failure times, recurrent events, and delayed entry or left censoring. Although many covariates of interest are assessed at t=0, for example, age at entry, gender, race, comorbid conditions, baseline clinical measurements, we may have some covariates measured during the period of follow-up making them time-dependent. Intermediate events that may occur during follow-up could influence occurrence of the primary event of interest. Multiple events of different types or recurrences of the same event are typical in longitudinal studies or in data structures that are clustered (e.g., animals within the same litter). A unified approach to analysis of event history data has been explicated (Anderson *et al,* 1993) based on the theory of multivariate counting processes.

Suppose there are K event types. Let $N_k(t)$ denote the number of type k events that have occurred by time t; $Y_k(t)$ denotes the number of individuals at risk for the type k event just before t; and $\mathbf{z}(t)$ the covariate history observed just prior to t. Conditional on the prior history (denoted by $\Im_{t-}$) the *multiplicative intensity model* (MIM) is $E(dN_k(t)\,|\,\Im_{t-}) = Y_k(t)\alpha_k(t\,|\,\mathbf{z}(t))dt$ where $\alpha_k(t\,|\,\mathbf{z}(t)) = \alpha_{k0}(t)\exp(\mathbf{z}'(t)\beta_k)$. For a single event type, the MIM reduces to the previously described PHM. To harness the power of PHREG to fit the MIM, some preliminary data processing may be required to structure the event history and covariate data appropriately to permit the correct evaluation of the at-risk sets.

a. FITTING THE PHM

Consider again the two times to infection since insertion of the catheter in 38 dialysis patients. The following syntax fits the PHM to both infection times, $h_k(t\,|\,\mathbf{z}) = h_{0k}(t)\exp(\beta_{1k}AGE + \beta_{2k}GENDER)$ where INSERT=k. Because $(\beta_{1k}, \beta_{2k}, k = 1,2)$ may be correlated a robust covariance is requested by the COVSANDWICH (AGGREGATE) option and all standard errors used in subsequent inference will use this covariance. The class statement uses GLM coding. Results of maximum partial likelihood estimation are in Table 4. We notice that the effect of gender is strong for the first infection time, with a lower infection rate among female patients compared to male patients. The comparison for the second infection time is not significant. These conclusions are in line with our previous nonparametric and parametric analyses.

```
proc phreg data=kidney covsandwich(aggregate);
id patient;
class gender insert/param=glm;
strata insert;
model time*fail(0)=age*insert gender*insert;
format gender gender. insert insert.;
run;
```

Table 4: Parameter estimates in PHM model for infection times

Parameter			DF	Parameter Estimate	Standard Error	StdErr Ratio	Chi-Square	p-value
age*insert	first		1	0.00964	0.01115	0.901	0.7467	0.3875
age*insert	second		1	–0.00332	0.01064	0.751	0.0971	0.7553
gender*insert	female	first	1	–1.38599	0.44621	1.062	9.6479	0.0019
gender*insert	female	second	1	–0.54276	0.60239	1.316	0.8118	0.3676
gender*insert	male	first	0	0	.	.	.	.
gender*insert	male	second	0	0	.	.	.	.

Because we have used GLM coding in the class statement, all contrast statements shown below must be provided in the comparative style: female vs male. For computing hazard ratios (HR) and 95% confidence intervals use:

```
contrast 'HR female vs male at INSERT=1' gender*insert 1 0 -1/estimate=exp;
contrast 'HR female vs male at INSERT=2' gender*insert 0 1 0 -1/estimate=exp;
```

Contrast	Estimate	Standard Error	95% Confidence Limits		p-value
HR female vs male at INSERT=1	0.2501	0.1116	0.1043	0.5996	0.0019
HR female vs male at INSERT=2	0.5811	0.3501	0.1785	1.8925	0.3676

A forthcoming enhancement to the HAZARDRATIO statement would give the same results from a single statement: `hazardratio "Gender effect" gender/cl=wald;` The label is optional.

Because the gender effect is dissimilar for the two catheter insertions, a test of equality of the gender effect is unwarranted. However, for illustration this test of equality $H_0: \beta_{21} = \beta_{22}$ is obtained from

```
contrast "Same Gender Effect" gender*insert 1 -1 -1 1;
```

The resulting Wald test is barely significant (p=0.038).

ILLUSTRATIVE EXAMPLE 2

Data set BMT contains follow up data on 137 patients who underwent a bone marrow transplant for treatment of acute leukemia (Klein & Moeschberger, 1997). These data have been analyzed extensively to meet different objectives using different strategies. We focus here on two events death/relapse combined and the event of platelet recovery when a patient's platelets return to normal levels. It is an important indicator of prognosis of survival. Initially following surgery all patients have depressed platelet count. Subsequently, in 120 patients recovery to normal levels was observed (PRI=1). TRETP is the recovery time. For the other 17 patients without platelet recovery (PRI=0), TRETP is set to missing. TFREEST denotes the time to death/relapse which was observed in 83 patients (DFI=1), 67 of whom had platelet recovery. Event times are in days from transplant. TFREEST is censored (DFI=0) if the event death/relapse has not occurred at the end of follow-up.

b. FITTING A PHM WITH TIME-DEPENDENT COVARIATES

In the analysis of TFREEST we will create a multiple record file to handle the time-dependent status of platelet recovery. Details of SAS code to create the long file BMT_LG is given in Gardiner, Luo & Lin (2008). All patients begin at TSTART=0. For a patient who had platelet recovery we create two records one of each time interval (0, TRETP] and (TRETP, TFREEST]. For the first record define TSTOP=TRETP, PLSTATUS=0, STATUS=0 and STRATUM='01'. For the second record define TSTART=TRETP, TSTOP=TFREEST, PLSTATUS=1, STATUS=DFI and STRATUM='12'.

For a patient who did not have platelet recovery we create a single record: TSTOP=TFREEST, PLSTATUS=0, STATUS=DFI and STRATUM='02'. PLSTATUS defines the platelet recovery status just prior to TSTOP and STATUS indicates whether or not death/relapse occurred at TSTOP. All time-invariant covariates are retained on each record. For this illustration we consider disease group (DGROUP) only. The variable STRATUM is created for convenience. It can be used to verify counts of events and censored values.

```
proc format;
value dgroup 1='ALL' 2='AML low risk' 3='AML high risk';
value plstatus 0='before' 1='after';
run;
```

Consider estimation of the PHM $h(t\,|\,\mathbf{z}(t)) = h_0(t)\exp(\mathbf{z}'(t)\beta)$ for the risk of death/relapse. With dummy variables AML_H and AML_L for AML-high risk and AML-low risk the linear predictor $\mathbf{z}'(t)\beta$ is defined as: $\beta_1 AML_H + \beta_2 AML_L + \beta_3 PLSTATUS(t) + \beta_4 AML_H \times PLSTATUS(t) + \beta_5 AML_L \times PLSTATUS(t)$ which is $\beta_1 AML_H + \beta_2 AML_L$ for t <TRETP, and $(\beta_1 + \beta_4)AML_H + (\beta_2 + \beta_5)AML_L + \beta_3$ for t ≥TRETP.

Estimation of the β-parameters is carried out by maximum partial likelihood estimation: the counting process style input must be used in the model statement to create the appropriate risk sets at each death/relapse time.

```
proc phreg data=bmt_lg;
class  plstatus(ref='before') dgroup(ref='ALL')/param=ref;
model (tstart, tstop)*status(0)=dgroup|plstatus/rl;
hazardratio dgroup/diff=ref cl=wald;
format dgroup dgroup. plstatus plstatus.;
run;
```

Table 5: Parameter estimates from PHM for time to death/relapse

Parameter			DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
dgroup	AML high risk		1	0.83069	0.69324	1.4358	0.2308
dgroup	AML low risk		1	1.05334	0.71832	2.1503	0.1425
plstatus	after		1	–0.45715	0.62896	0.5283	0.4673
plstatus*dgroup	after	AML high risk	1	–0.52620	0.75215	0.4894	0.4842
plstatus*dgroup	after	AML low risk	1	–1.85157	0.78769	5.5254	0.0187

The HAZARDRATIO statement is needed to produce the estimates of hazard ratios and 95% confidence intervals (Table 6). They are not produced by default because the dgroup|plstatus specification in the model is viewed as containing an interaction of dgroup with plstatus. The option DIFF=ref requests hazard ratios for the two AML disease groups with ALL as referent, and CL=wald gives the 95% confidence limits.

With the aforementioned parameterization, in the first row of Table 6 the point estimate is obtained from Table 5: $\exp(\hat{\beta}_1 + \hat{\beta}_4)$=exp(0.83069–0.52620)=1.356. Compared to the ALL patient group, the AML low risk group has improved prognosis for survival after platelet recovery (p=0.012). The same results can be obtained from contrast statements including p-values.

Table 6: Hazard Ratios for Disease Group

Description	Point Estimate	95% Wald Confidence Limits	
dgroup AML high risk vs ALL At plstatus=after	1.356	0.765	2.403
dgroup AML low risk vs ALL At plstatus=after	0.450	0.241	0.840
dgroup AML high risk vs ALL At plstatus=before	2.295	0.590	8.930
dgroup AML low risk vs ALL At plstatus=before	2.867	0.701	11.719

c. PLOTTING SURVIVAL CURVES

The PLOTS=survival option in the PROC PHREG statement produce graphs of estimated survival curves at specified covariate profiles. Consider the six profiles defined by DGROUP and PLSTATUS output to the data set COVAR.

```
proc sort data=bmt_lg out=covar(keep=dgroup plstatus) nodupkey;
by dgroup plstatus;
format dgroup dgroup. plstatus plstatus.;
run;
```

The BASELINE statement and its options produce the Nelson-Aalen estimator of the survival function at a fixed profile $\mathbf{z}_0$ using: $\hat{S}(t\,|\,\mathbf{z}_0) = \exp\left(-H_0(t,\hat{\beta})\exp(\mathbf{z}_0'\hat{\beta})\right)$ where $H_0(t,\hat{\beta}) = \int_0^t \{S^{(0)}(u,\hat{\beta})\}^{-1} dN(u)$, $S^{(0)}(t,\hat{\beta}) = \sum_{i=1}^n Y_i(t)\exp(\mathbf{z}'(t)\hat{\beta})$ and $N(t)$ is the counting process for death/relapse events in the sample. Survival curves derived from a PHM with time-dependent covariates should be interpreted with caution. For example, the relationship $S(t\,|\,\mathbf{z}(t)) = \exp\left(-\int_0^t h_0(u)\exp(\mathbf{z}'(u)\beta)du\right)$ holds under the assumption of strict exogeneity of the accumulating covariate process $t \to \mathbf{z}(t)$. By strict exogeneity we mean that $t \to \mathbf{z}(t)$ evolves as $P[\mathbf{z}(t+\Delta t)\,|\,T \geq t+\Delta t, \mathbf{z}(t)] = P[\mathbf{z}(t+\Delta t)\,|\,\mathbf{z}(t)]$. See Lancaster (1990) for further discussion.

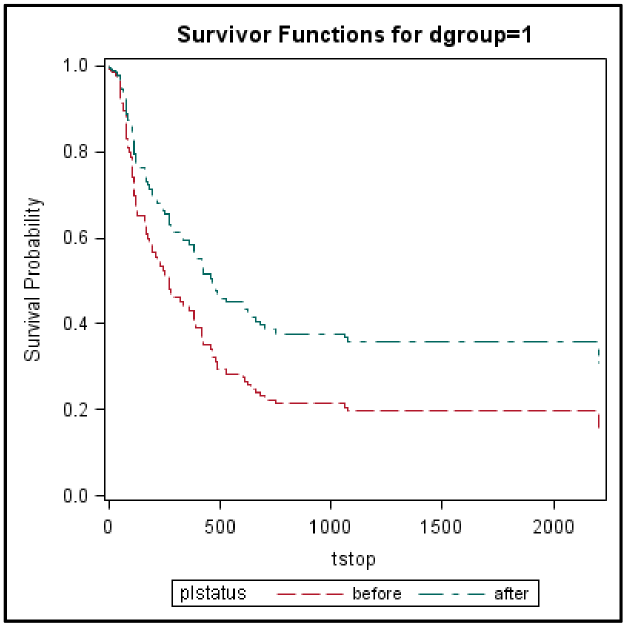
The following syntax will produce the plots shown next.

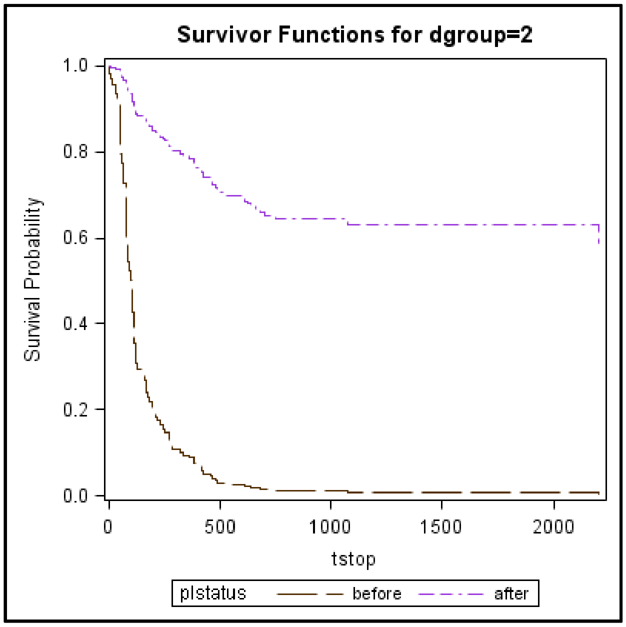
```
ods graphics on/width=4in height=4in;
proc phreg data=bmt_lg plots(overlay=group)=survival;
class  plstatus(ref='before') dgroup(ref='ALL')/param=ref;
model (tstart, tstop)*status(0)=dgroup|plstatus;
baseline covariates=covar out=surv survival=survival/method=ch group=dgroup
                                                    rowid=plstatus;
format dgroup dgroup. plstatus plstatus.;
run;
ods graphics off;
```

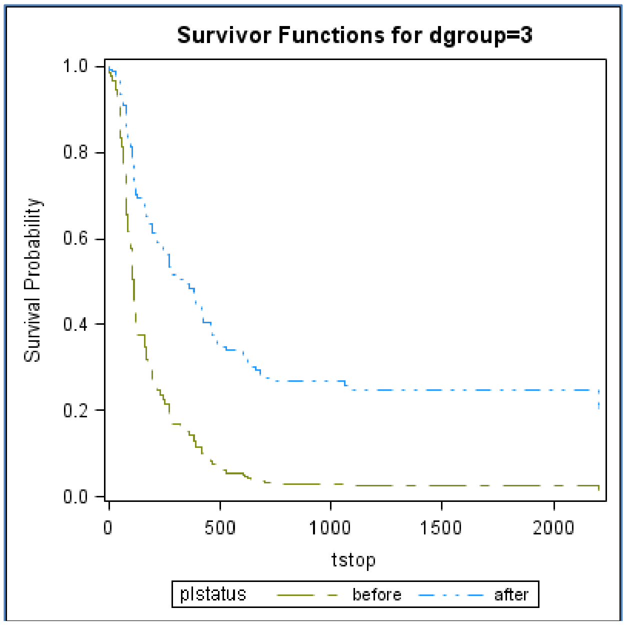
Six curves are displayed in three panels. The GROUP= option collates the curves by disease group, and ROWID appropriately labels the curves. Further modification of the plots (e.g., changing colors, title, line type, axes and legends) would need more manipulation of the output graphics file through PROC TEMPLATE, or some editing of the plot with the ODS graphics editor. See *Statistical Graphics using ODS in SAS/STAT® User's Guide.* An alternative is to use the SURV output dataset to plot the curves by PROC GPLOT.

d. MULTI-STATE MODELS

Although not discussed in detail here, the same technique of expanding the data set appropriately could be used in other settings including analyses of recurrent events, multiple failure times and competing risks models. For example, if platelet recovery is viewed as an intermediate event along with the terminal event death/relapse, another expansion of the data set BMT_LG would place all 137 patients at risk of each event, together with 120 records for the post-recovery transition to the terminal event. The data set will have 394 records. This is a three-state model with transitions 0→1 (platelet recovery), 0→2 (death/relapse without platelet recovery), and 1→2. The multiplicative intensity model $\alpha_{hj}(t\,|\,\mathbf{z}(t)) = \alpha_{hj0}(t)\exp(\mathbf{z}'(t)\beta_{hj})$ with stratum-specific covariates and hj denoting the $h \to j$ transition can be analyzed using PHREG. See Gardiner *et al*, (2008).







Disease–free survival plots

dgroup=1: ALL
dgroup=2: AML-low risk
dgroup=3: AML-high risk
plstatus: Platelet recovery status (time-dependent).

Within each disease group the survival plots are for two "what if" scenarios: (i) for a patient without platelet recovery (lower curve), (ii) for another patient with platelet recovery (upper curve). All curves are estimated at the same grid of event times (76 distinct times for 83 events). Plots are computed directly from the formula

$$\hat{S}(t\,|\,\mathbf{z}_0) = \exp\left(-H_0(t,\hat{\beta})\exp(\mathbf{z}_0'\hat{\beta})\right)$$

for a fixed profile $\mathbf{z}_0$.

IV. BAYESIAN ANALYSES

Frequentist analyses are based on the distribution of the data **y** that leads to a likelihood function L(**θ**; **y**) with parameters **θ** considered as fixed constants. It is the distribution of **y** that provides a basis for statistical inference on the unknown **θ**. We cannot make probabilistic statements about **θ**. Rather, the distributional properties of its estimator $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y})$ such as consistency, asymptotic normality are used to make inferences about **θ**. For example, the classical 100(1−α)% confidence interval for a 1-dimensional parameter θ with confidence limits $\mathrm{LCL}(\hat{\theta}), \mathrm{UCL}(\hat{\theta})$ and $P[\mathrm{LCL}(\hat{\theta}) < \theta < \mathrm{UCL}(\hat{\theta})] = 1 - \alpha$ for all values of θ is a probability statement about the confidence limits and not the parameter θ.

The Bayesian paradigm on the other hand places a distribution on **θ**, the prior distribution π(**θ**) which expresses our degree of belief in **θ**, and when combined with L(**θ**; **y**) gives the posterior distribution π(**θ**|**y**) of **θ** given **y**. Because Bayes' theorem gives π(**θ**|**y**) ∝ L(**θ**; **y**) π(**θ**) the term Bayesian analysis is applied to inferences drawn from the posterior distribution. For instance, the aforementioned frequentist confidence interval for θ can be replaced by a probability statement $P[a < \theta < b] = \int_a^b \pi(u \mid \mathbf{y})du$ based on the posterior distribution. If this probability is (1−α) we call (*a*, *b*) a 100(1−α)% credible interval for θ.

A closed-form expression for π(**θ**|**y**) can only be derived in a relatively few cases. Therefore, a general approach is to simulate π(**θ**|**y**) by drawing samples $\{\boldsymbol{\theta}^{(b)} : 1 \le b \le B\}$ and use them for inference. For example, the posterior mean is calculated as $\bar{\boldsymbol{\theta}} = B^{-1}\sum_{b=1}^{B} \boldsymbol{\theta}^{(b)}$ and an equal-tail 95% credible interval for a one dimensional θ is the interval between the 2.5-th and 97.5-th percentiles of the sample. The theory underlying the simulation approach is the Markov Chain Monte Carlo (MCMC) method that constructs a Markov chain whose stationary distribution is the posterior distribution. The process of drawing samples from the posterior distribution is based on Metropolis-Hastings algorithms or its variants (e.g., Gibbs sampler). The MCMC procedure designed to analyze Bayesian models fuels the capability of LIFEREG and PHREG to provide a Bayes solution to several survival models.

The BAYES statement in both LIFEREG and PHREG invokes the Bayes engine. For most analyses none of the myriad of options in the BAYES statement needs to be explicitly specified. However, a diligent investigation of the results should be undertaken to ascertain convergence of the underlying Markov Chain to its stationary distribution and whether the samples from the posterior exhibit dependencies. Several useful diagnostics and plots are produced by default if ODS Graphics is enabled with the PLOTS request. Finally, the posterior sample can be saved in a data set with the OUTPOST option for additional analyses. For quantities of interest such as the hazard ratio and percentiles of the survival curve that can be expressed as a function *g*(**θ**), the posterior sample $\{g(\boldsymbol{\theta}^{(b)}) : 1 \le b \le B\}$ is used to describe summary statistics for *g*(**θ**).

The following are standard MCMC options in the BAYES statement (default in parenthesis).

SEED= sets the random number generator for simulating the Markov chain samples (time of day).
NBI= # burn-in iterations discarded before the samples are saved (2000).
NMC=# iterations after burn-in (10000)
THIN=*k* retains one in every *k* samples after burn-in (*k*=1)

The initial values $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_K^{(0)})$ are arbitrary (can be set by INITIAL=). One iteration of the Gibbs sampler produces $\boldsymbol{\theta}^{(1)} = (\theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_K^{(1)})$ based on component-by-component random draws from conditional distributions: $\theta_1^{(1)}$ drawn from $\pi(\theta_1 \mid \theta_2^{(0)},\ldots,\theta_K^{(0)},\mathbf{y})$, $\theta_2^{(1)}$ drawn from $\pi(\theta_2 \mid \theta_1^{(1)},\theta_3^{(0)}\ldots,\theta_K^{(0)},\mathbf{y})$,…, $\theta_K^{(1)}$ drawn from $\pi(\theta_K \mid \theta_1^{(1)},\theta_2^{(1)}\ldots,\theta_{K-1}^{(1)},\mathbf{y})$. After B iterations this leads to the chain $\{\boldsymbol{\theta}^{(b)} : 1 \le b \le B\}$.

a. BAYESIAN ANALYSIS WITH LIFEREG

Consider the model $\log T_i = \mathbf{z}_i'\beta + \sigma\varepsilon_i$ for the time to death/relapse (TFREEST) in bone marrow transplant patients with disease group (DGROUP) as covariate. For a Bayes analysis, a prior distribution on $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ is specified through COEFFPRIOR and for σ through SCALEPRIOR. The following syntax fits a Weibull model with a normal prior $\boldsymbol{\beta} \sim N(0, 10^6\mathbf{I}_3)$, and gamma prior $\sigma \sim G(10^{-4}, 10^{-4})$. Several options need not be explicitly stated as they are the defaults. After a burn-in of 4000, one-half of the 20000 samples is retained. The data set BMT is the single record per patient file (137 patients). The order=freq option is used to preserve the previously used parameterization with the ALL group as referent.

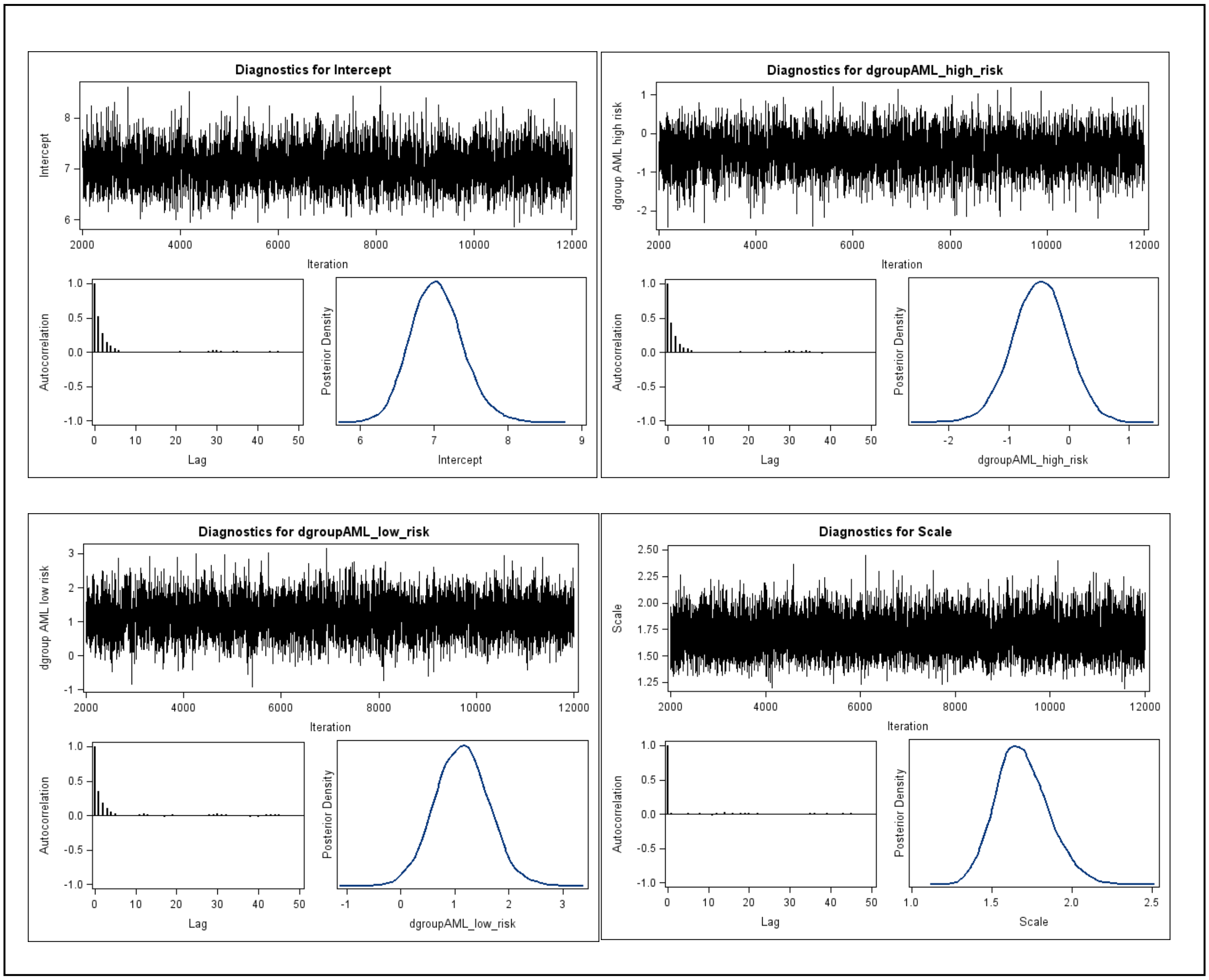
```
ods graphics off;
proc lifereg data=bmt order=freq;
class dgroup;
format dgroup dgroup.;
model tfreest*dfi(0)=dgroup/dist=weibull;
bayes seed=3538623 outpost=post_w nbi=4000 nmc=20000 thin=2
      coeffprior=normal(var=1E6)scaleprior=gamma(shape=1E-4, iscale=1E-4);
run;
ods graphics close;
```

Trace, autocorrelation and density plots are produced for each of the four parameters $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \sigma)$. It is imperative that these (and other diagnostics) be examined before any conclusions are drawn from the simulated posterior samples $\{\boldsymbol{\theta}^{(b)} : 1 \le b \le B\}$. The results shown on the next page are almost perfect. The trace plot show excellent mixing, the autocorrelation decreases to near zero, and the density is bell-shaped. The trace plots are centered near their respective posterior mean and traverse the posterior space with small fluctuations. For the intercept β_0 which corresponds to the ALL group, the trace plot is centered near the posterior mean of 7.0. Samples in both tails are covered. These results exhibit convergence of the Markov chain to its stationary distribution. The Geweke test (not shown) produced by default, compares the posterior mean from the early part (first 10%) of the Markov chain to posterior mean from the latter part (last 50%). There are no differences for each of the parameters.

Table 7 reports the simple statistics, percentiles, credible intervals, and high probability density (HPD) intervals for each of the parameters based on the posterior sample of 10000. Because the priors used are non-informative, the mean, standard deviation and credible interval should be fairly close to the corresponding maximum likelihood estimates (estimate, standard error, 95% CI).

Table 7. Posterior Statistics for parameters from Bayes analysis of the Weibull model

				Percentiles			Posterior Intervals			
Parameter	N	Mean	Standard Deviation	25%	50%	75%	Equal-Tail Interval		HPD Interval	
ALL	10000	7.0373	0.3506	6.7969	7.0253	7.2601	6.3857	7.7702	6.3528	7.7304
AML low risk	10000	1.1346	0.4947	0.8006	1.1330	1.4620	0.1684	2.1239	0.1479	2.0888
AML high risk	10000	–0.4810	0.4517	–0.7835	–0.4756	–0.1777	–1.3781	0.3898	–1.3442	0.4197
Scale	10000	1.6934	0.1621	1.5804	1.6826	1.7967	1.4026	2.0383	1.3750	2.0048



b. ESTIMATION OF PERCENTILES

For the Weibull, the *p-th* percentile is $t_p = \exp(\mathbf{z}'\beta + \sigma w_p)$ where $w_p = \log(-\log(1-p))$. A Bayes estimate is constructed from the posterior samples $t_p^{(b)} = \exp(\mathbf{z}'\beta^{(b)} + \sigma^{(b)} w_p), b = 1,\ldots,B$. Results are shown for the median in Table 8, right hand side panel with corresponding nonparametric and MLE estimates for comparison. The results are obtained by processing the OUTPOST=post_w data set. Similar, but not necessarily identical results can be derived using PROC MCMC using its MONITOR option.

Table 8: Estimate of Median disease-free survival (in days)

	Nonparametric			**MLE (Weibull)**			**Bayes (Weibull)**		
	Median	**95% Confidence Interval**		**Median**	**95% Confidence Interval**		**Posterior Mean**	**Equal-Tail Credible Interval**	
ALL	418	192	…	590.58	307.42	1134.53	650.88	317.45	1256.91
AML low risk	2204	641	…	1810.64	951.49	3445.55	2023.39	1010.18	3898.97
AML high risk	183	113	390	376.73	214.80	660.72	395.38	211.05	682.01

Likewise, disease-free survival at t days can be estimated from $S(t\,|\,\mathbf{z},\theta^{(b)}) = \exp(-\exp(-y^{(b)}))$ b= 1,…,B where $y^{(b)} = (\log t - \mathbf{z}'\beta^{(b)}) / \sigma^{(b)}$. For an example see SAS/STAT®: The MCMC Procedure.

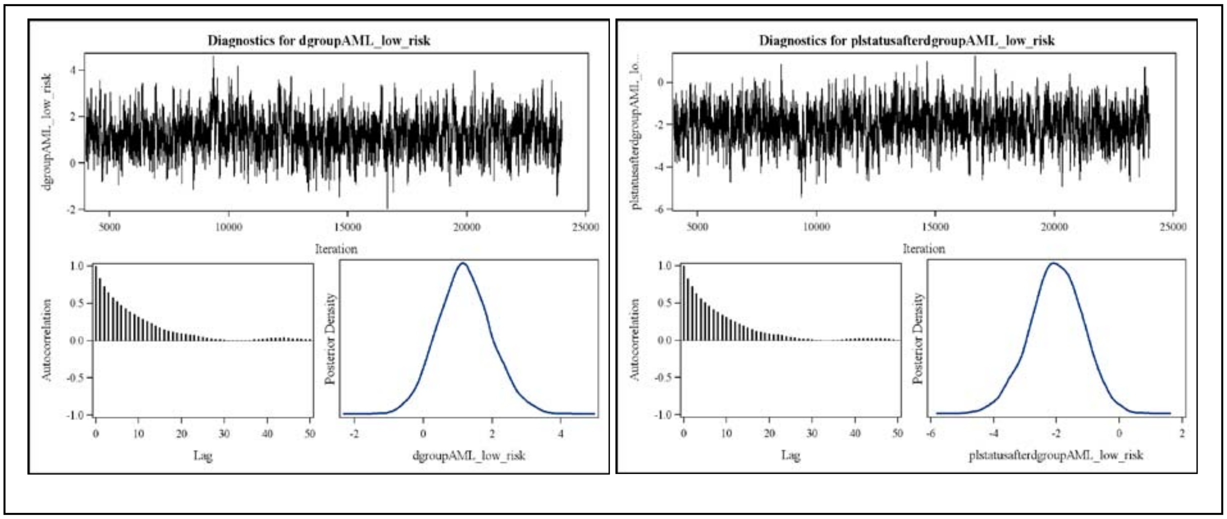
c. BAYESIAN ANALYSIS IN PHREG

There are two approaches to a Bayes analysis of the PHM $h(t\,|\,\mathbf{z}(t)) = h_0(t)\exp(\mathbf{z}'(t)\beta)$ .The first is based on the partial likelihood L($\boldsymbol{\beta}$; $\mathbf{y}$) combined with a prior $\pi(\boldsymbol{\beta})$ which produces the posterior $\pi(\boldsymbol{\beta}\,|\,\mathbf{y})$. The baseline hazard $h_0(t)$ is left unspecified. Inference is made from samples $\{\boldsymbol{\beta}^{(b)} : 1 \le b \le B\}$ drawn from $\pi(\boldsymbol{\beta}\,|\,\mathbf{y})$. Thus the partial likelihood is treated as a likelihood function just as in the previous analysis. The second approach discussed later parameterizes $h_0(t)$ by a finite-dimensional parameter $\boldsymbol{\lambda}$, $h_0(t) = h_0(t,\boldsymbol{\lambda})$ producing a full likelihood L($\boldsymbol{\theta}$; $\mathbf{y}$), where $\boldsymbol{\theta}$=($\boldsymbol{\beta}$,$\boldsymbol{\lambda}$).

Returning to our analysis of the time to death/relapse among bone marrow transplant patients we consider the PHM with disease group DGROUP, the time-dependent indicator PLSTATUS(*t*) of return of platelets to normal levels, and their interaction. This is a 5 parameter model. The extended file BMT_LG is used. The BAYES statement invokes the analysis. Options are the same as in LIFEREG. We only need to specify a prior for $\boldsymbol{\beta}$ which is taken here as $\boldsymbol{\beta} \sim N(0,10^6\mathbf{I}_5)$.

```
ods graphics on;
proc phreg data=bmt_LG;
class plstatus(ref='before') dgroup(ref='ALL')/param=ref;
model (tstart, tstop)*status(0)=dgroup|plstatus/ ties=breslow;
format dgroup dgroup. plstatus plstatus.;
bayes seed=4112010 outpost=postsample nbi=4000 nmc=20000 thin=2
           coeffprior=normal(var=1E6);
hazardratio dgroup/diff=ref cl=wald;
run;
ods graphics off;
```

Before drawing inferences from the posterior sample, we should examine the trace, autocorrelation and density plots for each parameter to be content that the underlying chain has converged. The plots for the two parameters involving the AML low risk group shown below suggest that the mixing in the chain is acceptable, although we notice long correlation times. Plots for the 3 other parameters (not shown) are very similar.



The HAZARDRATIO statement delivers the Bayes solution corresponding to the previous classical ML analysis in Table 6. These results (Table 9) can also be derived from the OUTPOST=postsample data set. We can also use postsample to assess the posterior probability that the HR for AML low risk vs ALL after platelet recovery is <1. The probability is over 99%.

Table 9: Hazard Ratios for Disease Group (10000 samples)

Description	Mean	Std Dev	Quantiles			95% Equal-Tail Interval		95% HPD Interval	
			25%	50%	75%				
AML high risk vs ALL At plstatus=after	1.411	0.423	1.111	1.351	1.647	0.768	2.392	0.691	2.241
AML low risk vs ALL At plstatus=after	0.470	0.153	0.361	0.449	0.556	0.234	0.827	0.208	0.773
AML high risk vs ALL At plstatus=before	3.711	3.827	1.625	2.674	4.420	0.649	13.191	0.294	10.046
AML low risk vs ALL At plstatus=before	4.631	4.863	1.961	3.264	5.542	0.757	17.112	0.275	12.778

d. SURVIVAL CURVES (FROM BAYES ANALYSIS)

Disease-free survival at t days for a specified covariate profile $\mathbf{z}_0$ is estimated from the posterior sample $S(t\,|\,\mathbf{z}_0,\beta^{(b)}) = \exp(-H_0(t,\beta^{(b)})\exp(\mathbf{z}_0'\beta^{(b)}))$ b= 1,…,B. This approach is similar to the classical estimates $\hat{S}(t\,|\,\mathbf{z}_0) = \exp\left(-H_0(t,\hat{\beta})\exp(\mathbf{z}_0'\hat{\beta})\right)$ where $\hat{\beta}$ denotes is the maximum partial likelihood estimator.

We use the same COVAR data set with six profiles defined by disease groups and platelet recovery status. The BASELINE statement requests a data set SURV_BAYES be formed to contain the output. With SURVIVAL=_ALL_ we obtain at each event time the posterior mean, standard error, equal-tailed credible interval limits, and HPD interval limits. For the purpose of plotting the survival curves we can use SGPLOT or GPLOT. However, when fully operational, under ODS graphics the PLOTS option in the PHREG statement together with additional options in the BASELINE statement would also yield the desired results.

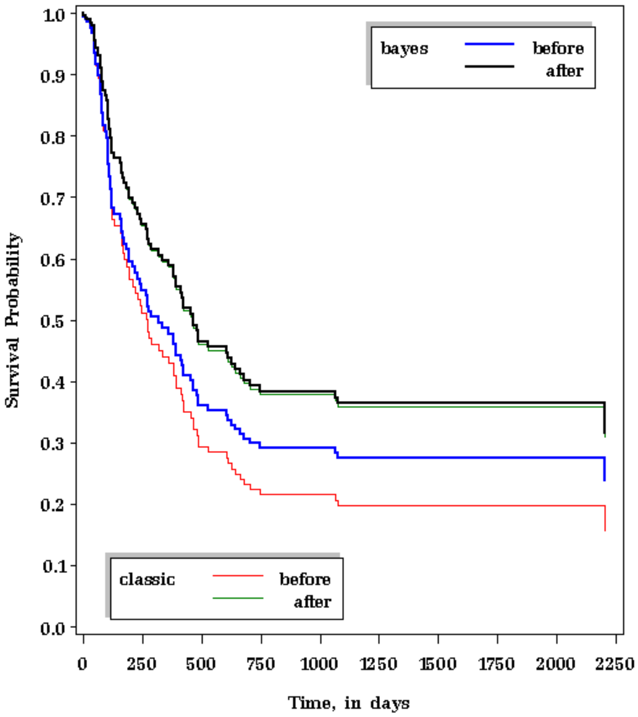
```
proc phreg data=bmt_LG ;
class plstatus(ref='before') dgroup(ref='ALL')/param=ref;
model (tstart, tstop)*status(0)=dgroup|plstatus;
format dgroup dgroup. plstatus plstatus.;
bayes seed=5808208 outpost=postsample nbi=5000 nmc=25000 thin=2
            coeffprior=normal(var=1E6);
baseline covariates=covar out=surv_bayes survival=_ALL_;
run;
```

The plots shown next are obtained by combining into one data set the survival estimates from the classical analysis with the posterior means from the Bayes analysis. We use GPLOT (with two plot statements) to exploit various options for axes, colors, legends etc.

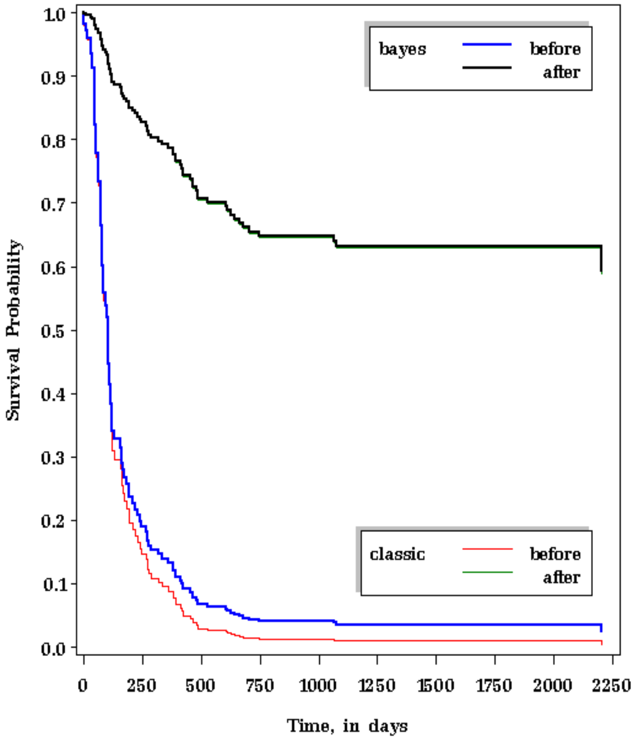
The following syntax plots the Bayes estimates and pointwise HPD bands for the DGROUP=1 (ALL patients). The plot is not shown.

```
ods graphics on;
proc sgplot data=surv_bayes(where=(dgroup=1));
    band x=tStop lower=lowerHPDSurvival  upper=upperHPDSurvival /
             group=plstatus modelname="Survival" transparency=.8;
   step x=tStop y=Survival / group=plstatus name="Survival";
   title "DISEASE GROUP = ALL";
   run;
ods graphics off;
```

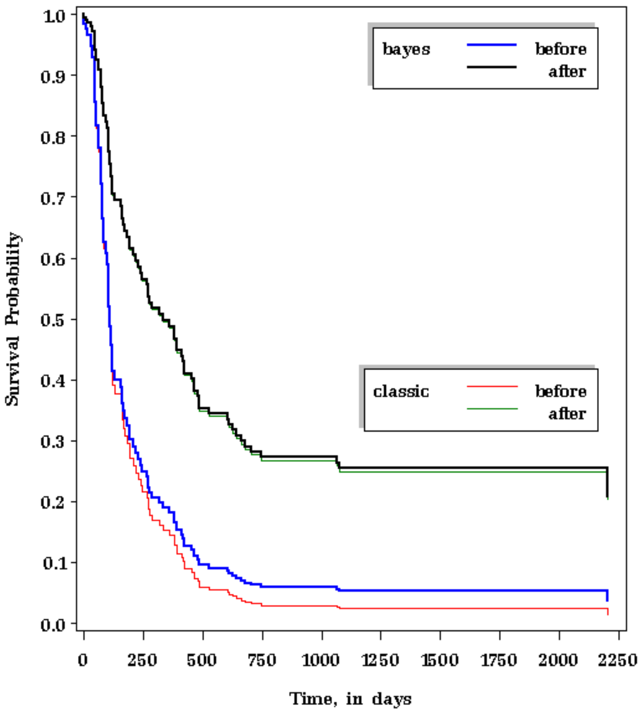
Disease Group ALL



Disease Group AML Low Risk



Disease Group AML High Risk



Disease/relapse-free estimates of survival

Plots for the classical analysis are obtained from the estimates $\hat{S}(t \mid \mathbf{z}_0) = \exp\left(-H_0(t,\hat{\beta})\exp(\mathbf{z}_0'\hat{\beta})\right)$ where $\hat{\beta}$ denotes the maximum partial likelihood estimator.

Plots for the Bayes analysis are derived from the posterior samples
$S(t \mid \mathbf{z}_0, \beta^{(b)}) = \exp(-H_0(t,\beta^{(b)})\exp(\mathbf{z}_0' \, \beta^{(b)}))$
$1 \le b \le B$ which are obtained from $\{\boldsymbol{\beta}^{(b)} : 1 \le b \le B\}$.

All calculations are made at a fixed profile $\mathbf{z}_0$ and at the same grid of event times. The two sets of estimates track each other, especially for after platelet recovery.

e. PIECEWISE CONSTANT HAZARD

The second approach to a Bayes analysis includes a parameterization of the baseline hazard $h_0(t)$ in the PHM as a piecewise constant function. Let $0 = a_0 < a_1 < .... < a_{J-1} < a_J = \infty$ denote a partition of the time axis into *J*-intervals $[a_{j-1}, a_j), j = 1,...,J$. The piecewise constant hazard is $h_0(t,\boldsymbol{\lambda}) = \sum_{j=1}^{J} \lambda_j [a_{j-1} \le t < a_j]$, with parameters $\boldsymbol{\lambda} = (\lambda_1,...,\lambda_J)$, $\lambda_j > 0$ for all *j*. An alternative parameterization uses log-hazards $\boldsymbol{\alpha} = (\alpha_1,...,\alpha_J), \alpha_j = \log \lambda_j$. Because the PHM is parametric in $\boldsymbol{\theta} = (\boldsymbol{\lambda},\boldsymbol{\beta})$ or $\boldsymbol{\theta} = (\boldsymbol{\alpha},\boldsymbol{\beta})$, the likelihood $\mathrm{L}(\boldsymbol{\theta}; \mathbf{y})$ can be constructed for the observed data **y**. Together with a specified prior $\pi(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$ we obtain the posterior $\pi(\boldsymbol{\theta} \mid \mathbf{y}) \propto \mathrm{L}(\boldsymbol{\theta}; \mathbf{y})\ \pi(\boldsymbol{\theta})$. The basis for inference is the sample $\{\boldsymbol{\theta}^{(b)} : 1 \le b \le B\}$ drawn from this distribution using the Gibbs sampler. The MLE of $\boldsymbol{\theta}$ obtained by maximizing $\mathrm{L}(\boldsymbol{\theta}; \mathbf{y})$ are produced which serve as the default initial values for the sampler.

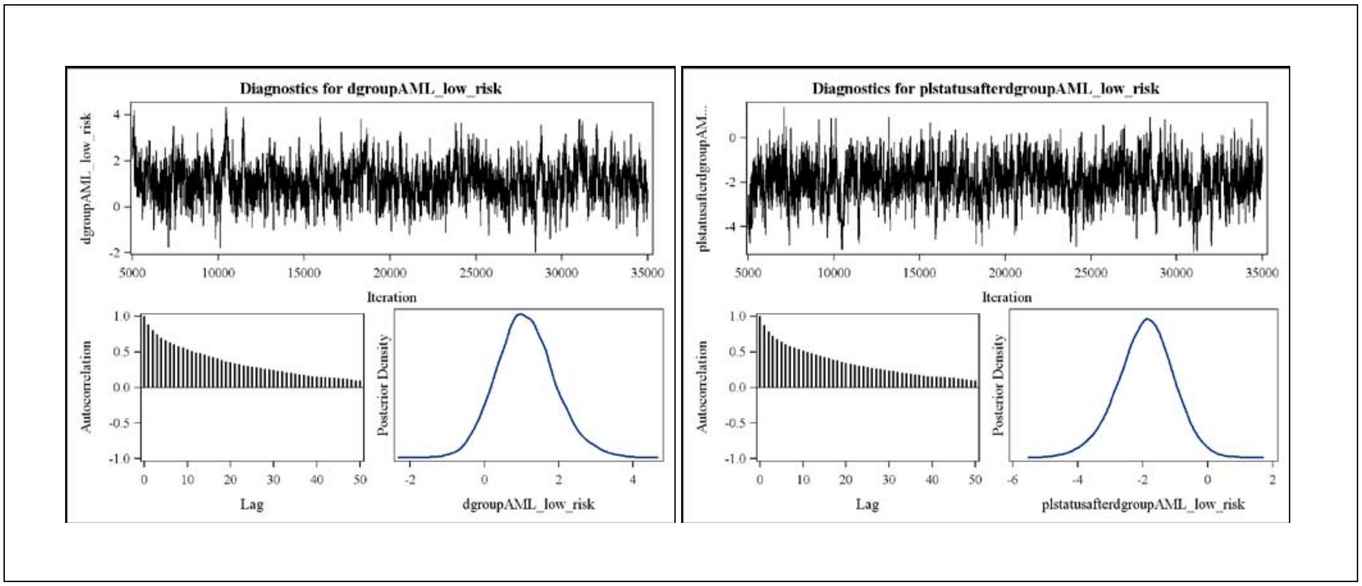
Simply adding PIECEWISE alone to the Bayes statement triggers the following: (i) log-hazard parameterization (ii) *J*=8 intervals (iii) uniform prior $\pi(\alpha_j) \propto 1$ for all *j*. This is the same as an improper prior on λ_j, that is, $\pi(\lambda_j) \propto \lambda_j^{-1}$. Interval cut-points are chosen by default to have approximately an equal number of events in each interval. Of course, all of these can be changed by options. The total number of events in the BMT data set is 83. By default 8 intervals are constructed to have about 10-11 events in each interval. Increasing the number of intervals could produce unstable estimates of $\boldsymbol{\lambda}$. Too few intervals could lead to poor fit. To obtain a feasible solution for $\boldsymbol{\lambda}$ the intervals must have at least one event. After trial and error, we use *J*=12. The following syntax specifies independent normal priors for $(\boldsymbol{\alpha},\boldsymbol{\beta})$. Correlation times are still large, but Geweke diagnostics are quite good. Results are in Table 10.

```
ods graphics on;
proc phreg data=bmt_LG;
class plstatus(ref='before') dgroup(ref='ALL')/param=ref;
model (tstart, tstop)*status(0)=dgroup|plstatus;
format dgroup dgroup. plstatus plstatus.;
bayes seed=4122010 outpost=postsample nbi=5000 nmc=30000 thin=2
  coeffprior=normal(var=1E6)
piecewise=loghazard(Ninterval=12 prior=normal(var=1e6));
run;
ods graphics off;
```

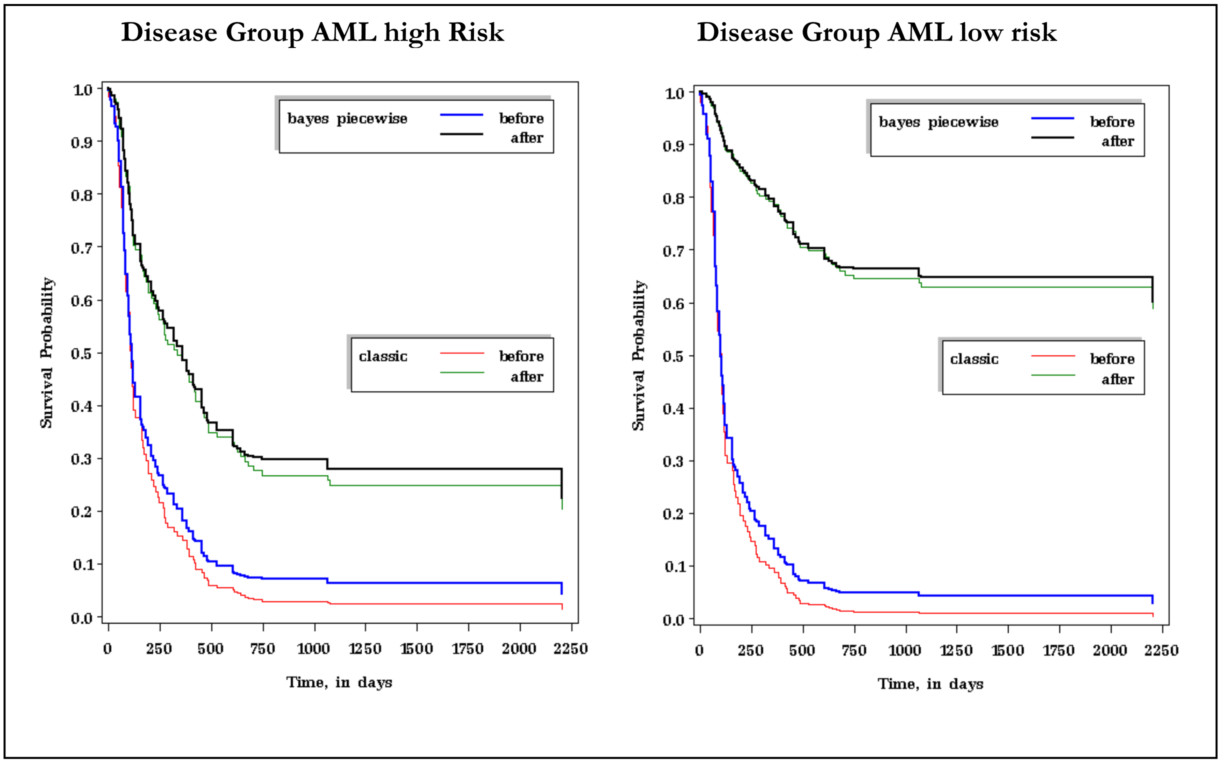
Table 10: Piecewise constant hazard model

	Maximum likelihood estimates				**Bayes estimates**			
Parameter	**Estimate**	**Standard Error**	**95% Confidence Limits**		**Posterior Mean**	**Standard Deviation**	**95% HPD Interval**	
AML: high risk	0.8414	0.6925	–0.5158	2.1986	0.8947	0.7604	–0.5284	2.4939
AML: low risk	1.0552	0.7158	–0.3477	2.4582	1.1162	0.7902	–0.4504	2.6721
PLSTATUS: after recovery	–0.4296	0.6301	–1.6646	0.8054	–0.3246	0.6861	–1.6085	1.0542
PLSTATUS × AML high risk	–0.5548	0.7515	–2.0277	0.9181	–0.5987	0.8205	–2.3587	0.9024
PLSTATUS × AML low risk	–1.8651	0.7852	–3.4040	–0.3261	–1.9215	0.8614	–3.7017	–0.2770

Diagnostic plots are shown below for 2 of the 5 regression parameters. They could be compared with the corresponding plots shown earlier for the Cox model on page 18.



A BASELINE statement is used to save the Bayes estimates of the survival curves and other optional quantities. Depicted below are curves for the AML low risk and AML high risk groups, paralleling the corresponding plots shown on page 20. The patterns are very similar, but with slightly more separation between estimates from the Bayes and classical analyses.



DATA SETS

The two data sets KIDNEY and BMT used in this paper are widely circulated via the world-wide-web. We used the original sources McGilchrist & Aisbett (1991) and Klein & Moeschberger (1997).

ACKNOWLEDGEMENTS

I wish to thank Ying So, Gordon Johnston and Jan Chvosta of the SAS Institute who have provided valuable feedback on my numerous questions and comments over many years. It is a pleasure to have learnt from them the many intricacies, nuances and versatility of SAS procedures. All errors and omissions in this paper are solely mine. This research was supported in part by the Agency for Healthcare Research & Quality under Grant 1R01 HS14206.

REFERENCES

Aalen O, Borgan O, Gjessing H. *Survival and Event History Analysis*. New York: Springer-Verlag; 2008.

Allison PA. *Survival Analysis using the SAS System--A Practical Guide.* Cary, NC: SAS Institute, Inc; 1995.

Andersen PK, Borgan O, Gill RD, Keiding N. *Statistical Models Based on Counting Processes*. New York: Springer-Verlag; 1993.

Box-Steffensmeier JM, Jones BS. *Event History Modeling--A Guide for Social Scientists*. Cambridge: Cambridge University Press; 2004.

Collett D. *Modelling Survival Data for Medical Research, 2nd edition*. London, UK: Chapman-Hall; 2003.

Cook RJ, Lawless JF. *The Statistical Analysis of Recurrent Events*. New York: Springer-Verlag; 2007.

Gardiner JC, Liu, L, Luo Z. Analyzing Multiple Failure Time Data Using SAS® Software. *Computational Methods in Biomedical Research.* Editors: R. Khattree and DN. Naik. Chapter 6, 153-188. Chapman-Hall; 2008.

Gardiner JC, Luo Z. Survival Analysis. *Encyclopedia of Epidemiology.* S. Boslaugh. Editor. Sage, 2008; 1019-1024.

Heckman JJ, Singer B, eds. *Longitudinal Analysis of Labor Market Data*. Cambridge, UK: Cambridge University Press; 1985.

Hosmer DW, Lemeshow S. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. New York: John Wiley & Sons; 1999.

Hougaard P. *Analysis of Multivariate Survival Data*. New York: Springer-Verlag; 2000.

Ibrahim JG, Chen M-H, Sinha D. *Bayesian Survival Analysis*. New York: Springer-Verlag; 2001.

Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer Verlag; 1997.

Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. New York: John Wiley & Sons; 1980.

Lancaster T. *The Econometric Analysis of Transition Data*. Cambridge, UK: Cambridge University Press; 1990.

Lawless JF. *Statistical Models and Methods for Lifetime Data, 2nd Edition*. Hoboken: John Wiley & Sons; 2003.

Marubini E, Valsecchi MG. *Analysing Survival Data from Clinical Trials and Observational Studies*. Chichester, England: John Wiley & Sons, Inc; 1995.

McGilchrist CA, Aisbett CW. Regression with frailty in survival analysis. *Biometrics.* 1991;47:461-466.

Nelson W. *Applied Lifetime Data Analysis*. New York: John Wiley & Sons; 1982.

Thernau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model.* New York: Springer-Verlag; 2000.

CONTACT INFORMATION

Joseph C. Gardiner
Division of Biostatistics, Department of Epidemiology
Michigan State University, East Lansing MI 48824
jgardiner@epi.msu.edu