*Analysis of Censored Data*
IMS Lecture Notes – Monograph Series (1995) Volume 27

# EXPLORATORY METHODS IN SURVIVAL ANALYSIS

J. CROWLEY, M. LEBLANC, R. GENTLEMAN, AND S. SALMON

ABSTRACT. Despite considerable research in the past two decades, much of the practice of survival analysis retains a black box flavor. In this paper we review some of the available exploratory methods for survival data and indicate some directions for future research. Methods discussed include box plots, running median plots, nonparametric estimation of the Cox regression function, and tree-based approaches to forming prognostic groups and building regression functions. The methods are applied to data on patients with multiple myeloma treated on clinical trials conducted by the (US) Southwest Oncology Group, a multi-institutional organization dedicated to finding cures for cancer.

## 1. INTRODUCTION AND NOTATION

Exploratory methods for survival analysis remain underutilized despite having received increasing attention from methodologists over the past decade. There are two aspects of survival analysis which are particularly responsible for the lag in the development and use of exploratory methods. The first is the fact that in most applications most survival data are subject to censoring. This means that for some individuals only partial information is available on their survival time, making many ordinary plotting methods less informative than they would otherwise be. The second reason is that the regression model most used in survival analysis, the Cox proportional hazards model, does not lend itself easily to pictorial representations of the data.

By exploratory methods we shall refer to those methods that attempt to describe the relationship between the response and the covariates of interest, putting very few restrictions upon that relationship. The emphasis of such methods is to use the available data to derive some idea as to the relationships that exist, rather than to test hypotheses that certain relationships hold. These methods are particularly well suited to the situation in which the data of interest have been collected and analyzed for some specific purpose but are still available to explore other questions of interest. The relationships would then be used to suggest directions for further research.

Interest will be focused in two areas, graphical methods and recursive partitioning. The methods will be illustrated using data on patients with multiple myeloma, a malignancy affecting the plasma cells of the blood.

1991 *Mathematics Subject Classification.* Primary 62G07; secondary 62P10 .

*Key words and phrases.* Survival analysis, exploratory methods, graphics, recursive partitioning .

We first provide notation that will be used throughout this paper. Let $X$ be a random variable with survivor function

$$S(t) = P(X > t), 0 \leq t < \infty,$$

and distribution function $F(t) = 1 - S(t)$. We will focus on survival data which are subject only to right-censoring, and for simplicity will assume random right-censoring. Let $C$ denote the random censoring time. The data consist of $T = X \wedge C$, $\Delta = I(T = X)$, and $Z$, where $\wedge$ stands for minimum, $I(\cdot)$ is the indicator function, and $Z$ is some covariate which may be vector valued. Given $Z$, we assume $X$ and $C$ are independent.

The hazard function or age-specific rate of failure given by

$$\lambda(t) = \lim_{h \downarrow 0} \frac{P(t \leq X < t + h \mid X \geq t)}{h}$$

is also of interest, as is the cumulative hazard function

$$\Lambda(t) = \int_0^t \lambda(u) du.$$

## 2. Some Background on Myeloma

Multiple myeloma is a malignant tumor arising from the plasma cell series of immunocytes which are responsible for the production of the antibody-immunoglobulins involved in the immune response. In patients with myeloma, a clone of malignant plasma cells proliferates in the bone marrow and invades bone, causing lytic bone lesions and releasing calcium into the bloodstream. Malignant progenitors also circulate through the blood and thereby are able to colonize marrow and other bony sites. The myeloma cells produce a homogeneous (monoclonal) immunoglobulin which generally has no known antibody function. The tumor also suppresses the production of normal antibody-immunoglobulins via clonal dominance over normal plasma cell precursors, which are suppressed from responding to antigenic stimulation. The monoclonal immunoglobulin or subcomponents of the myeloma protein can be deposited in various tissues in the body and lead to various organ dysfunctions including kidney failure. Other associated findings in myeloma include severe anemia due to bone marrow involvement, bone pain and fractures due to tumor infiltration and calcium loss, and increased susceptibility to infection due to the impairment of the normal antibody response to foreign microbes.

The age-adjusted incidence rate of myeloma in the US is about 5 per 100,000 people per year, is somewhat higher in males than females, and is roughly twice as high in African-Americans as in European-Americans. The incidence in Bombay is about one per 100,000 per year. The etiology is largely a mystery, but radiation, exposures associated with farming, and chronic antigenic stimulation have been implicated. (Riedel, Pottern and Blattner, 1991).

The accepted treatment is with a class of chemicals called alkylating agents, which includes melphalan, and steroid hormones such as prednisone. Median survival after treatment is usually in the range of 24-36 months in various reports, but the course of the disease is extremely variable. Accordingly, there is interest in being able to identify patients who will do well with conventional therapy, and those who will not; more aggressive treatment approaches might be attempted in the latter group, for example. A staging system does exist (Durie and Salmon, 1975), based on a quantification of the number of tumor cells and a classification of kidney function, but improvements may be possible based on blood measurements and pathologic characteristics of the tumor cells. Much of the exploratory analysis we will present is towards this aim of developing an improved staging system, or prognostic grouping, for myeloma patients.

## 3. Graphical Methods

**3.1. Methods for a Single Qualitative Covariate.** With the notation of Section 1, let the observed data for a sample of $n$ individuals be denoted by the triplet $[(t_i, \delta_i, z_i)]_{i=1}^{n}$, and for convenience assume there are no ties among the failure times. If the covariates have no effect on survival, or if their effect is to be ignored, then the estimation of the survivor function $S(t)$ is of interest. The product-limit estimator (Kaplan and Meier, 1958) of $S(t)$ is given by

$$\widehat{S}(t) = \prod_{i:t_i \leq t} \left[1 - \frac{\delta_i}{r_i}\right]^{\delta_i},$$

where $r_i$ is the number of individuals at risk at $t_i$ (the number with observed survival times, censored or not, which are at least $t_i$ ). This reduces to the usual empirical distribution function in the absence of censoring.

Figure 1 gives product-limit estimates for a succession of four clinical trials conducted by the Southwest Oncology Group, a consortium of institutions in the US dedicated to finding improved treatments for all adult cancers. Each of the trials in Figure 1 was a randomized comparison of several treatment strategies (different combinations of alkylating agents and steroids), but the various approaches are ignored in this presentation. The year of initiation of the trial is given by the first two digits of the label in the figure (study 7704 was started in 1977, *etc.*). One conclusion is that we have made little progress in the past two decades in the treatment of patients with multiple myeloma.

The variability of outcomes in patients with multiple myeloma is illustrated in Figure 2A. There was a common treatment arm in all four of these trials (the details of which are not important for our purposes here), and Figure 2A shows the results by trial for patients treated with the same arm in the same research group. While these product-limit plots are widely employed and understood, alternative displays are possible which might have advantages in certain circumstances. For example, box plots are commonly used with uncensored data. Since the features of a box plot can be thought of as functions of the empirical distribution function (sample medians, other quantiles, *etc.*), extensions are readily available using the same functionals of the product-limit estimator $\widehat{S}(t)$ (Gentleman and Crowley, 1991a). Box plots for the four samples in Figure 2A are given in Figure 2B. The median and quartiles form the box in the usual way, and the whiskers extend to the smallest (largest) observed failure time within a distance of 1.5 times the interquartile range. Observations outside the whiskers can be given different symbols to indicate whether they are censored or not. Since survival curve estimates do not typically reach zero due to censoring, the value of the survival curve at the last observation can be given in the plot. If the censoring is so heavy that the upper quartile can not be estimated, the box can be extended to the largest observed failure time and the value of the survival curve at that point can be given. Ad hoc adjustments can be made for the whiskers, but the plots lose meaning when censoring is heavy. As can be seen from Figure 2B, the box plot presentation is especially helpful in displaying the differences in medians and quartiles across the samples. A number of other displays are discussed in Gentleman and Crowley (1992).

### 3.2. Methods for a Single Continuous Covariate.

3.2.1. *Running Quantile Plots.* It has long been known that a few outlying points in a scatterplot can greatly affect the impression that one gets from that plot. When plotting survival data the censored data points can act as outliers and distort the message that the data are trying to provide. Simply using a different plotting character for censored observations cannot remove all of this effect. Figure 3A is a scatterplot of survival and the log of serum $\beta_2$ microglobulin, an enzyme which is elevated by the output of the clone of myeloma cells and also reflects the degree of kidney failure caused by the processing a large quantities of complex proteins. The data for this figure, as for the examples in the remainder of this paper, are from one of the trials shown in Figure 1, Southwest Oncology Group Myeloma Study 8229 (Salmon, Tesh, Crowley *et.al.*, 1990). Censored observations are plotted with open circles, uncensored ones with closed circles. It is difficult to discern trends in this plot, both because of the censoring and because of the inherent variability of the data.

Adding a line which goes through the *center* of the data (in some sense) should provide useful information to the analyst. The usual scatter plot smoothers are running averages or running least squares lines. Neither of these extends particularly well to survival data. Since the data are censored one cannot actually get reliable estimates of the mean; as an alternative running medians have been examined by Doksum and Yandell (1983), building on unpublished work of Beran (1981). The general question of interest is to estimate the conditional distribution of the survival time $X$ given the covariate $Z$,

$$S_z(t) = P(X > t|Z = z),$$

and to plot functionals of these estimators, such as the quantiles. Our approach is to use the $k$ nearest neighbors of a point $z$ to estimate the distribution function, using the product-limit estimator (Gentleman and Crowley, 1991a). Figure 3B shows the same scatterplot of survival and the log of serum $\beta_2$ microglobulin, with the median and two quartile estimates superimposed, using a symmetric neighborhood of about one-third of the data. The tendency for survival to decrease with increasing values of the covariate is more apparent from this plot.

3.2.2. *Nonparametric Estimation of the Regression Function.* The current method of choice for analyzing the effect of covariates upon survival time is the proportional hazards model suggested by Cox (1972). The most general form for the model is

$$\lambda(t, z) = \lambda_0(t)\psi(z),$$

where $\lambda_0(t)$ is an unspecified baseline hazard and $\psi(z)$ is some unspecified function of the covariate. If $\psi(z) = \exp \beta z$ then inference can be made about $\beta$ without specifying $\lambda_0(t)$ . The partial likelihood (Cox, 1975) provides the basis for calculating estimates of $\beta$ and also estimates of the variability of these estimates. Given

$(t_1, \delta_1, z_1), \ldots, (t_n, \delta_n, z_n)$, the risk set at time $x_i$ is denoted by $R_i$ and consists of the indices $j$ such that $x_i \leq x_j$. The partial likelihood is given by

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp \beta z_i}{\sum_{j \in R_i} \exp \beta z_j} \right\}^{\delta_i},$$

and the log partial likelihood is given by

$$l(\beta) = \sum_{i=1}^{n} \delta_i \left\{ \beta z_i - \log \left\{ \sum_{j \in R_i} \exp \beta z_j \right\} \right\}.$$

Note that we have assumed that there are no ties in the survival times. This assumption is for notational convenience only. Approximations such as those suggested by Peto (1972) can easily be extended to all cases given herein.

Several authors have extended this basic proportional model away from the strictly linear case. The extension involves replacing $\exp \beta z$ with $\exp s(z)$, where $s(z)$ is some smooth function. Estimation of $s(\cdot)$ then relies upon the extension of some technique for smooth likelihood based regression to the case of smooth partial likelihood based regression. This usually involves putting some restriction on the form that $s(\cdot)$ may take, since unrestrained maximization results in a function which will not be smooth in the covariate. It seems natural to assume that whatever the effect of the covariate upon survival time that it is at least slowly varying with regard to the covariate.

Tibshirani and Hastie (1987) suggested the method of local likelihood. This method consists of making the assumption that within some neighborhood of $z_i$ the function $s(z_i)$ is linear, so that for $z$ in that neighborhood $s(z) = \alpha_i + \beta_i z$. One then uses the nearest neighbors to estimate $\alpha_i$ and $\beta_i$ and thereby gets an estimate of $s(z_i)$. The function $s(\cdot)$ is then estimated using $s(z_1), \ldots, s(z_n)$ and linear interpolation between the observed values of the covariate. For the proportional hazards model it has long been known that one cannot estimate $\alpha_i$ due the nature of the model. Let $D$ denote the set of indices $i$ of the individuals who failed and let $N_i$ denote the indices of the nearest neighbors of $z_i$. Then it follows that locally the partial likelihood is given by

$$\prod_{j \in D \cap N_i} \frac{\exp(\alpha_i + \beta_i z_j)}{\sum_{k \in R_j \cap N_i} \exp(\alpha_i + \beta_i z_k)} = \prod_{j \in D \cap N_i} \frac{\exp(\beta_i z_j)}{\sum_{k \in R_j \cap N_i} \exp(\beta_i z_k)}$$

and therefore, $\alpha$ is not estimable. Tibshirani and Hastie noted that the estimate of $\beta_i$ corresponds to an estimate of $s'(z_i)$ and suggested that $s(z)$ could be constructed via numerical integration. Estimation of the local parameter $\beta_i$ by the method of local likelihood consists of putting weight $1/k$ on each of the $k$ nearest neighbors of $z_i$. Note that only points close by are used to estimate the value $s(\cdot)$ at any point.

A variation of local likelihood which we have explored (Gentleman and Crowley, 1991b) consists of using an alternating scheme to estimate both the baseline cumulative hazard function and the regression function. The baseline hazard function $\Lambda_0(t)$ is first estimated, using all the data and a preliminary estimate of $\widehat{s}_0(z)$ of $s(z)$,using the estimator proposed by Breslow (1974):

$$\widehat{\Lambda}_{01}(t) = \sum_{i:t_i \leq t} \frac{\delta_i}{\sum_{l \in R_j} \exp \widehat{s}_0(z_l)}. \tag{3.1}$$

Then with $\widehat{\Lambda}_{01}(t)$ considered fixed, the full likelihood is used, locally, to give an updated estimate $\widehat{s}_1(z)$, assuming $s$ is locally linear. The local log likelihood for this step is given by

$$l(\alpha_i, \beta_i) = \sum_{j \in N_i} \left[ \delta_j(\alpha_i + \beta_i z_j) - \widehat{\Lambda}_{01}(z_j) \exp(\alpha_i + \beta_i z_j) \right] .$$

Thus for each value of the covariate a linear approximation to $s$ is made and the appropriate parameters estimated. This estimate $\widehat{s}_1(z)$ is then used in (3.1) to give an updated estimate $\widehat{\Lambda}_{02}(t)$, and the cycle is repeated until convergence. This procedure has the advantage over local partial likelihood of more truly reflecting the proportional hazards model, in that the proportionality is forced to hold globally with respect to the same baseline hazard function.

Figure 4A shows the result of estimating $s(z)$ by local full likelihood for the myeloma data from SWOG 8229 using the covariate serum $\beta_2$ microglobulin, and a neighborhood of about one-third of the data. It can be seen that the regression function is decidedly nonlinear. This nonparametric estimate could be used as a step towards building a multivariable regression function using the additive approach of Hastie and Tibshirani (1986), or it might suggest a transformation of the covariate that would improve the linearity assumption for more conventional Cox regression function fitting. Figure 4B shows the result of fitting a Cox regression using the log of serum $\beta_2$ microglobulin, showing a nearly linear relationship.

3.2.3. *Remarks.* The results of Beran (1981), Owen (1987), and Dabrowska (1987) can be used to establish asymptotic properties of the running quantile plots, but many practical questions remain, among them the choice of the smoothing parameter $k$. If the smoothing parameter is to be estimated via cross validation having a distance measure will be important. A cross-validation scheme would proceed for a given $k$ by deleting each point in turn, estimating the conditional distribution function without that point, then measuring the distance between that distribution and the one-point distribution (extensions to censored points would be required). Minimization over $k$ would then lead to the choice of this smoothing parameter.

Asymptotic results for local partial likelihood should be obtainable using Anderson and Gill (1982), but care would have to be taken in the trade-off between having the neighborhood shrink while the number of points in the neighborhood grows. Results for local full likelihood appear more difficult.

## 4. Tree-Based Methods

Prognostic stratification is often desired as a guide in the choice of medical treatment and clinical trial strategy. Typically, techniques for developing prognostic groups based on survival data use the Cox regression model to identify important predictors, and *ad hoc* methods for using the regression function to divide the predictor space into several regions. While such staging schemes have proven useful, a technique based on recursive partitioning has some advantages. This procedure works in non-linear and synergistic situations, and results in easily described groupings which may be more interpretable to clinicians than scoring rules based on the Cox regression model.

Recursive partitioning algorithms can be very briefly described as follows. A rule is adopted to partition the predictor space into two regions or nodes. The rule is applied recursively to the data until the space has split into a large number of nodes, each containing only a small number of observations; this partitioning can be described by a binary tree. Secondly, there are rules that allow one to prune the tree and choose the "best" pruned subtree.

Several authors have proposed tree-based methods for censored survival data. Gordon and Olshen (1985) and Butler, Gilpin, Gordon and Olshen (1989) modify the Classification and Regression Tree - $CART^{TM}$ algorithm of Breiman, Friedman, Olshen and Stone (1984) to censored survival data by using "distances" between estimated survival curves in place of least squares. The distance measures suggested by Gordon and Olshen are based on $L_p$ and $L_p$ Wasserstein metrics. The "variability" of a node is defined as the minimum distance between the product-limit estimator for the node and any step function. Reduction of within node variability is used to grow the tree. Important aspects of $CART^{TM}$ are adopted, such as initially growing a large tree, and then pruning to obtain a nested sequence of subtrees. Finally, the "best" pruned subtree is chosen by $k$-fold cross-validation. Typically, one would report the product-limit estimator or median survival at the terminal nodes of the best pruned subtree.

Segal (1988) and Ciampi, Thiffault, Nakache and Asselain (1986) propose a different recursive partitioning technique for censored survival data. Instead of splitting based on reducing within node variability, between node splitting is used. This allows one to construct trees using two sample censored data rank statistics, such as the familiar logrank test (Mantel, 1966). While there are advantages to rank based partitioning, there are some problems with the proposed algorithms. Since there is no measure of within node variability, Segal fails to adopt cross-validation to choose the best pruned subtree. Clearly, some method of assessing whether branches describe real structure or are just noise is important. Without such a method, there is a danger of overfitting the model to the data. Ciampi does use Akaike Information Criterion (AIC) to evaluate trees, possibly as an approximation to cross-validation. However, he stops splitting a tree on the basis of significance tests. Breiman *et. al.* note that to stop splitting at fixed split value can lead to poor results; one should initially grow a very large tree and prune back to avoid missing structure. In order to compare the within node distance approach to a between node approach based on the logrank test, we performed a simulation study of the properties of the first split of the predictor space into two regions. The results are given in Section 4.3 below, following some further notation and definitions.

**4.1. $L_p$ Wasserstein and $L_p$ Metrics.** Let $X_1 \sim F_1$ and $X_2 \sim F_2$. Then the $L_p$ Wasserstein distance between $F_1$ and $F_2$ can be expressed (Shorack and Wellner, 1986) as

$$\left[\int_0^1 |F_1^{-1}(u) - F_2^{-1}(u)|^p \, du\right]^{\frac{1}{p}} .$$

The Wasserstein metric focuses on the horizontal distance between proper distribution functions. An adjustment needs to be made for estimated survival curves, which do not reach 0 if the largest observation is censored. For product-limit estimates $\widehat{F}_1$ and $\widehat{F}_2$ of $F_1$ and $F_2$, let

$$\begin{aligned} \lim_{u \to \infty} \widehat{F}_1(u) &= m_1 \leq 1, \\ \lim_{u \to \infty} \widehat{F}_2(u) &= m_2 \leq 1. \end{aligned}$$

Without loss of generality assume $m_1 \leq m_2$. Define $m_3 = F_2^{-1}(m_1)$ and

$$\begin{aligned} \widetilde{F}_2(u) &= \widehat{F}_2(u) \ \text{ if } u < m_3 \\ \widetilde{F}_2(u) &= m_1 \quad \text{ if } u \geq m_3. \end{aligned}$$

The define the distance for censored data based on the $L_p$ Wasserstein metric as

$$\left[\int_0^{m_3} \left|\widehat{F}_1^{-1}(u) - \widetilde{F}_2^{-1}(u)\right|^p du\right]^{\frac{1}{p}}.$$

Note that small differences in the tails of the distributions will have a large effect on the $L_p$ Wasserstein distance.

Now consider the ordinary $L_p$ metric. For $1 \leq p < \infty$ the $L_p$ distance between proper distribution functions $F_1$ and $F_2$ is

$$\left[\int_0^{\infty} |F_1(u) - F_2(u)|^p \, du\right]^{\frac{1}{p}}.$$

This focuses on the vertical distance between distribution functions. With censored data, vertical truncation is used. Let $m_4 = \min\left(\widehat{F}_1^{-1}(m_1), \widehat{F}_2^{-1}(m_2)\right)$. Define the distance based on the ordinary $L_p$ metric to be

$$\left[\int_0^{m_4} \left|\widehat{F}_1(u) - \widehat{F}_2(u)\right|^p du\right]^{\frac{1}{p}}.$$

**4.2. Growing the Tree.** Recall that the data are given by $[(t_i, \delta_i, z_i)]_{i=1}^n$, which in the terminology of recursive partitioning is call the learning sample and is denoted by $\mathcal{L}_n$. Define a measure of node impurity based on a distance between the product-limit estimator of the survival curve for observations at a node $t$, $\widehat{S}_t$, and a step function $\delta_{\widehat{S}_t}$ which has a single jump that minimizes the distance between the any step function with a single jump and the survival function $\widehat{S}_t$. For the $L_1$ metric and uncensored data, $\widehat{\delta}_{\widehat{S}_t}$ is a step function at the median of $\widehat{S}_t$, and for the $L_2$ Wasserstein metric and uncensored data, $\widehat{\delta}_{\widehat{S}_t}$ is a step function at the mean of $\widehat{S}_t$. Then the reduction in impurity at a node $t$ based on the learning sample $\mathcal{L}_n$ is given by

$$G(t) = p(t)d(\widehat{S}_t, \widehat{\delta_{S_t}}) - \left[p(l(t))d(\widehat{S}_{l(t)}, \widehat{\delta_{S_{l(t)}}}) + p(l(t))d(\widehat{S}_{r(t)}, \widehat{\delta_{S_{r(t)}}})\right],$$

where $p(t)$ is the proportion of observations falling into node $t$ and $d(\widehat{S}_t, \widehat{\delta_{S_t}})$ is the $L_p$ Wasserstein or $L_p$ distance between $S_t$ and the step function $\widehat{\delta_{S_t}}$.

A tree is grown on the entire learning sample by partitioning the covariate space recursively into groups based on the $L_p$ Wasserstein or $L_p$ reduction in impurity. There are many types of binary partitions of the space that could be considered. However, the easiest to interpret is a split on a single variable; that is, a split of the form "Is $Z_k \leq c$ ?" for an ordered predictor, or of the form "Is $Z_k \in \mathbf{S}$ ?" for a categorical predictor, where $\mathbf{S}$ is a proper subset of the set of values of

$Z_k$. Since interpretability of the tree structure is important, usually only these simple univariate partitions are considered. Each split is chosen by evaluating the reduction in the impurity splitting statistic for every possible split point for each covariate. The covariate space is split on the variable and the split point that corresponds to the largest splitting statistic. Each of the resulting regions of the covariate space and data corresponding to the regions is represented by a node. The same rule is applied recursively to the resulting groups until there are only a small number of individuals in each node.

The same partitioning scheme is used in Segal (1988) and Ciampi *et. al.* (1986, 1988) except that the logrank test statistic is used to measure the performance of the split instead of the reduction in impurity based on the $L_p$ Wasserstein or $L_p$ metric. Each possible split point for each possible covariate yields a two sample logrank test, and the space is partitioned based on the split point giving the maximum value of the logrank statistic. Figure 5 is a plot of the value of the logrank statistic (in $\chi^2$ form) as a function of covariate value, for several candidate covariates in the myeloma data set described above. Imposed on the plots are the 1% and 5% points of the permutation distribution of the logrank statistic, based on sampling from all possible permutations of the covariate values over the censored survival times. It can be seen that the maximum value of the logrank statistic is for the covariate serum $\beta_2$ microglobulin, at a value of 5.4 nanograms per milliliter; it also appears that more than a chance mechanism is involved.

**4.3. Simulation Study.** A full tree grown by these techniques is likely to be too complex, and some of the branches may only represent noise. Some pruning is necessary, as will be described in subsequent sections. However, for the purposes of this limited simulation study, we will only be concerned with the properties of the splitting statistics used to grow the trees. Among the class of $L_p$ Wasserstein or $L_p$ metrics, the $L_2$ Wasserstein and $L_1$ (W2 and L1 in Figures 7–8) were chosen because of their association with splitting statistics based on squared and absolute error loss for uncensored data; these will be compared with splitting based on the logrank statistic (Lr in Figures 7-8). To keep the study simple only a single split was computed for a range of survival and censoring distributions.

4.3.1. *Method.* We simulated data from four models with several different censoring configurations. The failure time distributions were members of the $G^\rho$ family (Harrington and Fleming, 1982),

$$F(t;\psi,\rho) \equiv P(X \le t;\psi,\rho) = \begin{cases} 1-(1+\rho t\psi)^{-\frac{1}{\rho}} & \text{if} \quad \rho > 0 \\ 1-\exp -t\psi & \text{if} \quad \rho = 0. \end{cases}$$

and censoring times were distributed uniformly on $(0,\gamma)$. Assume $F_j$ and $G_j$ $(j = 1, 2)$ are distribution functions of survival and censoring times. Each data set was produced as follows:

(1) $Z_i = i : i = 1, \cdots, 2m$
(2) $X_i$ generated from $F_i$ if $i \le m$ or $F_2$ if $i > m$
(3) $C_i$ generated from $G_1$ if $i \le m$ or $G_2$ if $i > m$

where $Z_i$ is the covariate value, $X_i$ is the true failure time, and $C_i$ is the censoring time for individual $i$. The survival models and censoring configurations are shown in Table 1, where the percent censoring corresponds to the group with the longer survival times.

Note that the values of the covariate $Z$ were assigned the integers from 1 to $2m$, and that the splitting statistics are invariant under monotone transformations of the covariate. Hence, one way to assess the performance of a split is to consider the number of observations to the left and right of the split. If $F_1$ and $F_2$ are different, a good split would divide the observations simulated above into two groups of about size $m$.

Model $A$ represents a case where the survival distributions are not related to the covariate; all true survival times are exponential random variables with mean 1. Models $B, C$ and $D$ describe simple survival structure such that the median for the failure time distribution $F_1$ is approximately 0.35 and the median for $F_2$ is approximately 0.70 for each model. Hazard functions corresponding to the distributions $F_1$ and $F_2$ for each model are presented in Figure 6. The hazard ratios between the observations is constant or decreasing.

Only one sample size is reported here, for $2m = 100$. The minimum allowed size of a group resulting from a split was 10 observations. This allowed only 81 possible splits for the sample of size 100. One thousand replications were used. The same generated data sets were used for each combination of splitting statistic, survival model and censoring scheme.

### 4.3.2. *Results.*

No Structure. Results from Model $A$ are summarized in Figure 7; they are represented as histograms of the split points on covariate $Z$. $CART^{TM}$ tends to choose splits that send almost all observations to one daughter node if little response structure is present and if the minimum permitted node size is small. This is called end-cut preference by Breiman *et. al.* (1984). In the simulations with uncensored data (column one) and equal 20% censoring (column two), the end-cut preference phenomenon exists for each of the three statistics considered. However, because the number of observations was restricted to be at least 10, Figure 7 shows that the effect is weak. Jesperson (1986) shows that end-cut preference can be a problem with the logrank statistic. He considers splitting on a single covariate for the proportional hazards model in large samples and shows that with high probability the cut point corresponding to the logrank test, for any $\epsilon > 0$, occurs in the $\epsilon$ fraction of the largest or smallest covariate values.

The effect of uneven censoring on the splitting statistics was also investigated. For $Z < 50$ the censoring distribution was $U(0, \gamma_1)$ and for $Z \geq 50$ the censoring distribution was $U(0, \gamma_2)$. The parameters $\gamma_1$ and $\gamma_2$ were chosen so that there was approximately 20% and 50% censoring in the corresponding regions of the covariate space, respectively. The results are presented in the third column of Figure 7. The logrank statistic was not seriously affected by uneven censoring in this example with no structure. Both the $L_1$ and $L_2$ Wasserstein statistics show a striking dependence on the censoring pattern. Since the upper limit of integration in these distance measures depends on where the data run out, heavier censoring in one daughter node tends to increase the value of the statistic. This explains the large proportion of splits near the change in the censoring distribution.

Simple Structure. Figure 8A shows the results of generating data from models $B, C$ and $D$ with no censoring. The split point minus 50 can be thought of as an error in the splitting, since the best split would divide the data into two groups of size 50.

The first column of Figure 8A shows that for uncensored data generated from model $B$ all the statistics detect the structure well. However, in uncensored data from models $C$ and $D$, the efficacy of the splitting statistics is reduced. For the logrank test this result is expected, since in model $B$ the hazard ratio for the distributions $F_1$ and $F_2$ is constant, and for models $C$ and $D$ the hazard ratios are decreasing. The $L_1$ and $L_2$ Wasserstein statistics split more frequently to the right of the optimal split point among the observations with the longer expected lifetime. The skewness is more pronounced for models $C$ and $D$ than for $B$. The higher variability of the observations generated from $F_2$ compared to $F_1$ makes the $L_1$ and $L_2$ Wasserstein statistics more variable for splits among the observations generated from $F_2$, increasing the chance of the maximal splitting statistic occurring among these observations with higher variability. The same phenomenon was noted by Breiman *et. al.* (1984) for the splitting statistic based on squared error loss.

Figure 8B shows histograms for the same models with mild uniform censoring (20%). The row corresponding to the logrank statistic shows that the performance decreases only slightly from the uncensored case. However, the $L_1$ and $L_2$ Wasserstein statistics perform poorly under mild censoring. The $L_1$ statistic splits much more frequently among the observations with the longer expected lifetime and censoring, due to the increased variability of the product-limit estimator near the end of the data. The results are similar for moderate censoring, approximately 50%, as shown in Figure 8C.

Remarks. These limited simulation experiments showed that there is a striking deterioration of the performance of both the $L_1$ and $L_2$ Wasserstein splitting statistics if even mild censoring is introduced. For homogeneous survival distributions and uneven censoring the distribution of the split points depended strongly on the censoring pattern. Therefore, improvements to these splitting statistics are needed before they can be recommended for use in the analysis of censored survival data. The logrank statistic does appear to have promise as a splitting criterion, however, justifying further investigation into pruning algorithms, as discussed below.

**4.4. Pruning the Tree.** Because of the appealing aspects of rank tests and the limited simulation results above, we have extended rank based partitioning to include some of the pruning aspects of the $CART^{TM}$ algorithm (LeBlanc and Crowley, 1993). In $CART^{TM}$, initially a large tree is grown, then the tree is pruned to obtain a nested sequence of subtrees. The "best" subtree is chosen by cross-validation. Since with a rank test there is no intrinsic notion of within node variability, which is key in $CART^{TM}$, pruning and choice of the subtree must be based on between node statistics. Define $G(t)$ to be the "goodness of split" at node $t$. Let $I$ be the set of labels of internal nodes. Then one could evaluate the tree by the split-complexity measure

$$G(T) = \sum_{t \in I} G(t) - \alpha \, |I| \, .$$

The second term is a penalty for the complexity of the tree. Define $r(t)$ and $l(t)$ to be the right and left daughters of $t$. Let $p(t)$ be the proportion of observations that fall into node $t$. This technique is consistent with those procedures using within node variability, since one could define

$$G(t) = p(t)R(t) - p(r(t))R(r(t)) + p(l(t))R(l(t)),$$

where $R(t)$ is the variability of node $t$. In the case of rank trees let $G(t)$ be a two sample rank test; for instance, the two sample partial likelihood ratio, or the logrank test.

As shown in LeBlanc and Crowley (1993), spit-complexity leads to an optimal tree through weakest link cutting for each $\alpha$. The parameter $\alpha$ would typically be between 2 and 4 for test statistics expressed in an approximate $\chi^2$ form, where a value of 4 corresponds roughly to a .05 $p$-value, and a value of 2 is in the spirit of AIC. For a given $\alpha$, $G(T)$ is an over-estimate of the worth of the tree due to the multiple split points considered. A bootstrap bias correction can be applied, in which the overoptimism is estimated by creating a tree for each bootstrap sample, and calculating $G$ for this tree using both the bootstrap and the original sample, then averaging. Applying this correction leads to smaller trees. Other methods of bias correction can be based on cross-validation or permutation sampling.

A fuller implementation of the $CART^{TM}$ algorithm, including cross-validation for the choice of $\alpha$, can be achieved based on (a one-step approximation to) the full nonparametric likelihood, using the Breslow estimator for the cumulative hazard, as in Section (3.2.2). Likelihood deviance then substitutes for least squares as a measure of within node variability. Details can be found in LeBlanc and Crowley (1992).

**4.5. Amalgamation.** The first work on amalgamating nodes with similar survival from distant parts of the final tree, to form a few prognostic groups, was done by Ciampi, Hogg, McKinney and Thiffault (1988). They use a logrank partitioning procedure for growing the tree, then use an algorithm based on minimal logrank tests to combine nodes of the tree into classes. An alternative we have proposed (LeBlanc and Crowley, 1993) is based on defining an ordered categorical variable from some measure (such as median survival or hazard ratio) on the terminal nodes of the tree, then subjecting this variable to a recursive partitioning procedure. While an automatic procedure has some attraction, it is likely that the analyst will want to retain control over which of the terminal nodes should be grouped to form a staging system.

**4.6. The Myeloma Data Set.** We implemented a recursive partitioning scheme on our myeloma data set, using the logrank test for splitting with a minimum node size of 25 observations. The tree was pruned with the split-complexity algorithm, where test statistics were bias-corrected using approximate degrees of freedom calculated by permutation sampling. The parameter $\alpha$ was set to a value of 4 to yield the tree given in Figure 9. Below each split in the tree the logrank test statistic and the approximate degrees of freedom corresponding to the adaptive split are given. Notice that the degrees of freedom decrease as one moves down the branches in the tree because there are fewer potential split points in the lower nodes. Below each terminal node the number of observations and the logarithm of the relative risk compared to the left most node are presented.

The first split on the tree is on serum $\beta_2$ microglobulin (sb2m) at a value of 5.4. The next two splits are both on a derived variable measuring immaturity of cells (iandd) and further splits are on percent plasmablasts (plasmabl), albumin (alb), calcium (cal), age and performance status (swogpf). Based on an ordering of the terminal nodes by the relative risk, four prognostic strata were constructed : I - node 1, 2, and 6, II - nodes 3, 4 and 7, III - nodes 8 and 10, and IV- nodes 5, 9 and

11 (where the nodes are numbered from the left in Figure 9). The survival curves corresponding to the 4 prognostic groups are illustrated in Figure 10A, contrasted with the results based on the Durie-Salmon staging system (Figure 10B).

**4.7. Remarks.** Many questions remain about the properties of the tree-based tools proposed here. The conditions needed for asymptotic results seem to require that the terminal node size approach 0, a condition not generally approximated in practice. Monte Carlo experiments are extremely computer-intensive, and require careful specification of the objectives of the analysis and choice of the experimental conditions. Perhaps the best proof of their usefulness will come with implementation by practitioners, including validation of results with external data sets.

## 5. Concluding Remarks

Several approaches for the exploration of survival data have been presented, and implemented on a set of data on patients with multiple myeloma. It is our hope that the methods presented will prove helpful to the legions of statisticians involved in analyzing censored survival data, and to that end many of the routines have been implemented in the *S* language (Becker, Chambers and Wilks, 1988) and are available from the second author.

## References

[1] Anderson PK and Gill RD: Cox's regression model for counting processes: A large sample study. The Annals of Stat, 10:1100-1120, 1982.

[2] Beran R: Nonparametric regression with randomly censored survival data. Technical Report, Department of Statistics, University of California-Berkeley, 1981.

[3] Becker R, Chambers J and Wilks A: **The New S Language.** Wadsworth International Group, Pacific Grove, California, 1988.

[4] Breiman L, Friedman JH, Olshen RA and Stone CJ: Classification and regression trees. Belmont CA., Wadsworth International Group, 1984.

[5] Breslow N: Covariance analysis of censored survival data. Biometrics 30:89-99, 1974.

[6] Butler JH, Gilpin E, Gordon L and Olshen RA: Tree-structured survival analysis, II. Technical Report, Stanford University, 1989.

[7] Ciampi A, Thiffault J, Nakache J-P and Asselain B: Stratification by stepwise regression, correspondence analysis and recursive partition. Computational Statistics and Data Analysis, 4:185-204, 1986.

[8] Ciampi A, Hogg S, McKinney S and Thiffault J: RECPAM: A computer program for recursive partitioning and amalgamation for censored survival data. Computer Methods and Programs in Biomedicine, 26:239-256, 1988.

[9] Cox DR: Regression models and life-tables (with discussion). J of the Royal Stat Soc, 34:187-220, 1972.

[10] Cox DR: Partial likelihood. Biometrika, 62:269-276, 1975.

[11] Dabrowska DM: Nonparametric quantile regression with censored data. Tech Rep No. 93, 1987.

[12] Doksum KA and Yandell BS: Properties of regression estimates based on censored survival data. In **Fetschrift for Erich L. Lehmann,** PJ, Bickel KA, Doksum and Hodges JL, Eds, 140-156, Wadsworth, 1983.

[13] Durie BGM and Salmon SE: A clinical system for multiple myeloma. Correlation of measured myeloma cell mass with presenting clinical features, response to treatment and survival. Cancer 36:842-854, 1975.

[14] Gentleman R and Crowley J: Graphical methods for censored data. Journal of the American Statistical Association 86:678-683, 1991a.

[15] Gentleman R and Crowley J: Local full likelihood estimation for the proportional hazards model. Biometrics 47:1283-1296, 1991b.

[16] Gentleman R and Crowley J: A graphical approach to the analysis of censored data. Breast Cancer Research and Treatment 22:229-240, 1992.

[17] Gordon L and Olshen RA: Tree-structured survival analysis. CA Treatment Reports, 69:1065-1069, 1985.

[18] Harrington D and Fleming T: A class of rank test procedures for censored survival data. Biometrika, 69:553-556, 1982.

[19] Hastie TJ and Tibshirani RJ: Generalized additive models. Statistical Science 1:297-318, 1986.

[20] Jesperson N: Dichotomizing a continuous covariate in the Cox regression model. Technical Report, University of Copenhagen, 1986.

[21] Kaplan EL and Meier P: Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53:457-481, 1958.

[22] LeBlanc M and Crowley J: Relative risk trees for censored survival data. Biometrics, 48:411-425, 1992.

[23] LeBlanc M and Crowley J: Survival trees by goodness of split. Journal of the American Statistical Association, 88:457-467, 1993.

[24] Mantel, N: Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports, 50:163-170, 1966.

[25] Owen A: Nonparametric conditional estimation. Technical Report No. 25, Laboratory for Computational Stat, Dept. of Stat, Stanford U, USA, 1987.

[26] Peto R: Contribution to the discussion of the paper by Cox DR. J of the Royal Stat Soc, B, 34:205-207, 1972.

[27] Reidel DA, Pottern LM, and Blattner WA: Epidemiolgy of multiple myeloma. In **Neoplastic Diseases of the Blood**, Wiernik P.H., Canellos G.P., Kyle R.A. and Schiffer C.A. eds., pp 347-367, 1991. Churchill Livingstone, New York.

[28] Salmon SE, Tesh D, Crowley J, Saeed S, Finley P, Milder MS, Hutchins LF, Coltman CA Jr., Bonnet JD, Cheson B, Knost JA, Samhouri A, Beckord J and Stock-Novack D: Chemotherapy is superior to sequential hemibody irradiation for remission consolidation in multiple myeloma: A Southwest Oncology Group Study. Journal of Clinical Oncology 8:1575-1584, 1990.

[29] Segal MR: Regression trees for censored data. Biometrics, 44:35-48, 1988.

[30] Shorack G and Wellner J: **Empirical Processes and Applications to Statistics.** John Wiley and Sons, New York, 1986.

[31] Tibshirani RJ and Hastie T: Local likelihood estimation. Journal of the American Statistical Association 82:559-567, 1987.

(J Crowley and M LeBlanc) FRED HUTCHINSON CANCER RESEARCH CENTER, 1124 COLUMBIA STREET, SEATTLE WA 98104

(R Gentleman) DEPARTMENT OF STATISTICS, UNIVERSITY OF AUCKLAND, PRIVATE BAG 92019, AUCKLAND, NEW ZEALAND

(S Salmon) COLLEGE OF MEDICINE, ARIZONA CANCER CENTER, UNIVERSITY OF ARIZONA, TUCSON, ARIZONA 85724

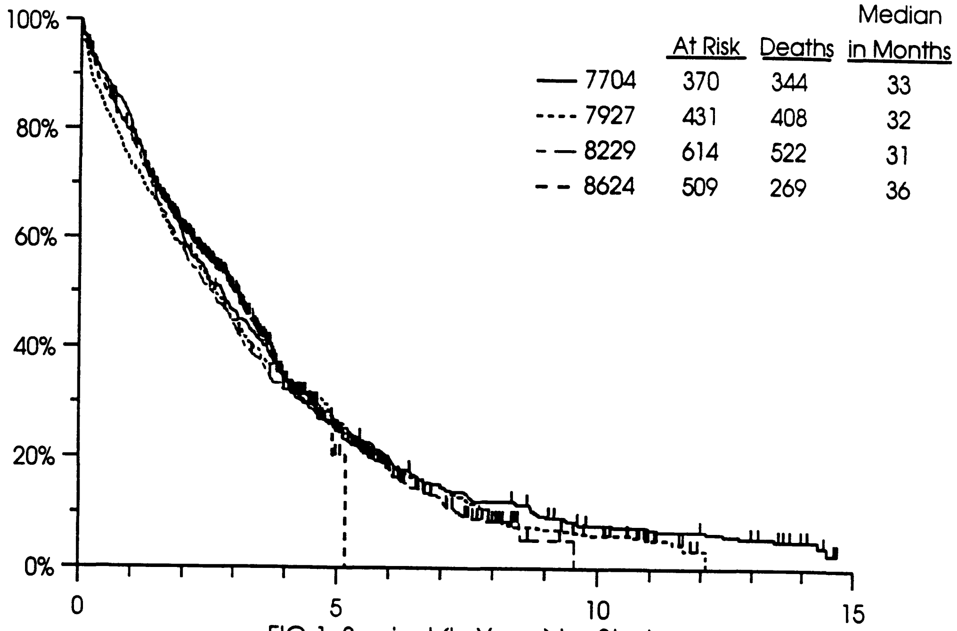

FIG 1: Survival (in Years) by Study

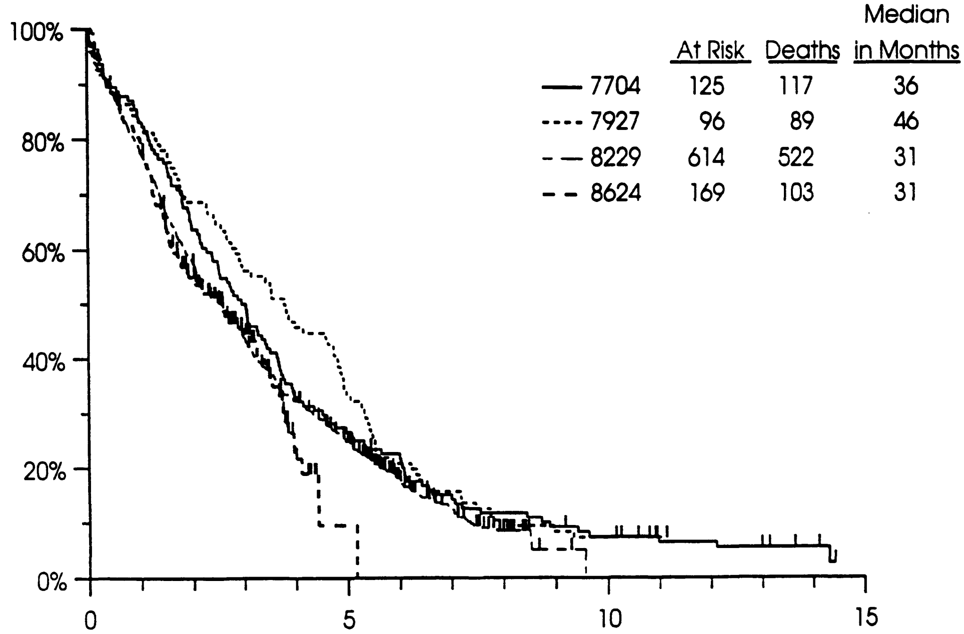

FIG 2A: Survival (in Years) by Study for Same Arm

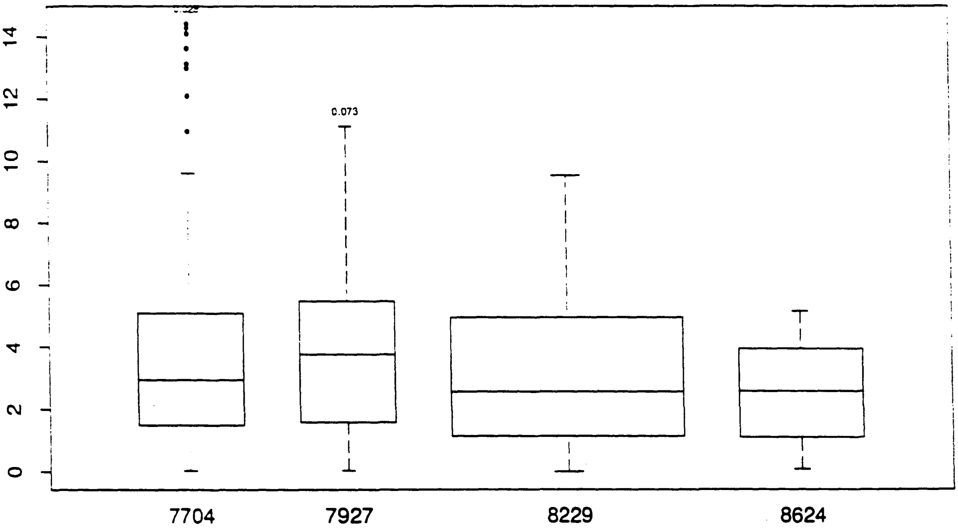

FIG 2B: Box Plots of Survival (in Years) by Study for Same Arm

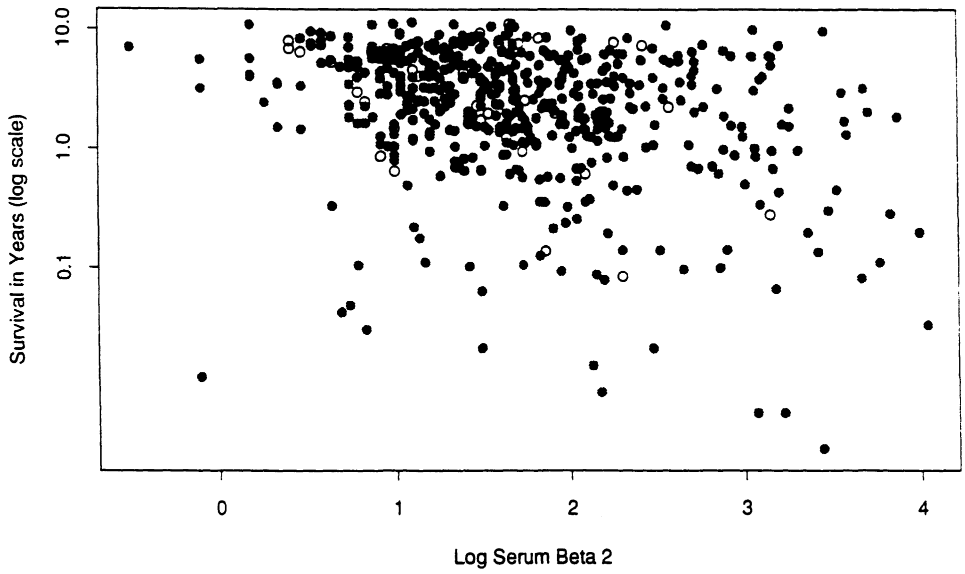

FIG 3A: Scatterplot of Survival Time by the Log of Serum Beta 2 Microglobulin.
Open circles represent censored observations.

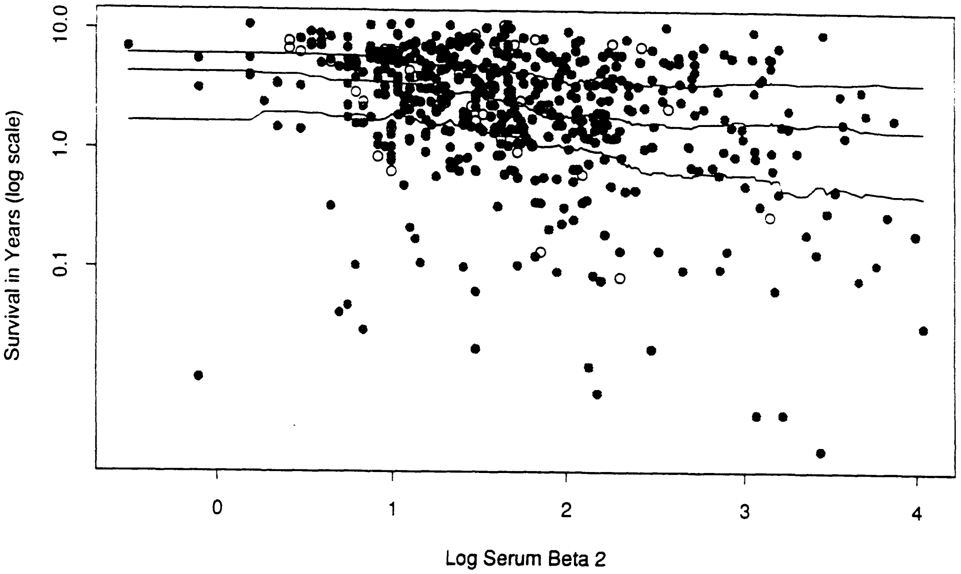

FIG 3B: Scatterplot with Smoothed Quartile Estimators.

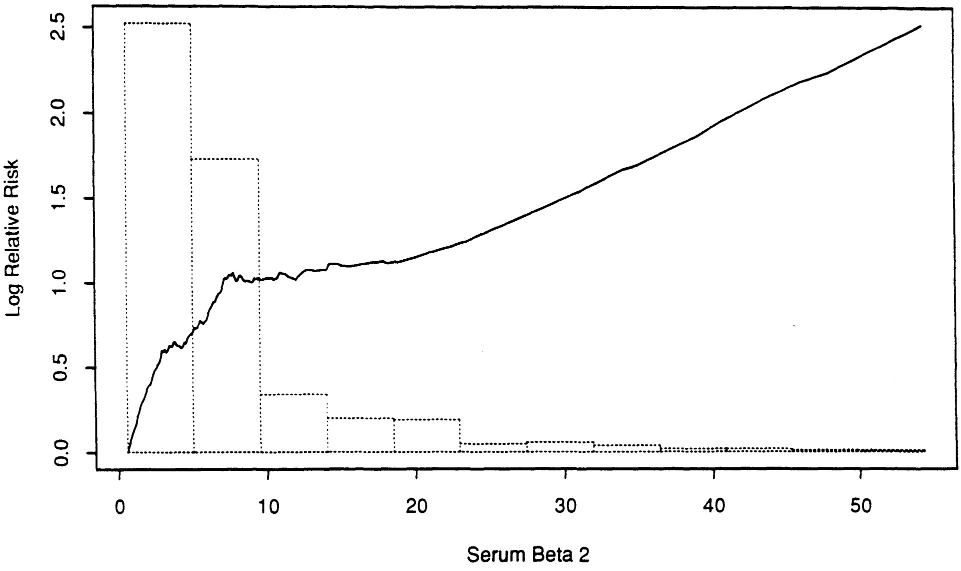

FIG 4A: Local Full Likelihood Estimate of the Log Relative Risk for Serum Beta 2 Microglobulin. Histogram of serum beta 2 is plotted in the background.

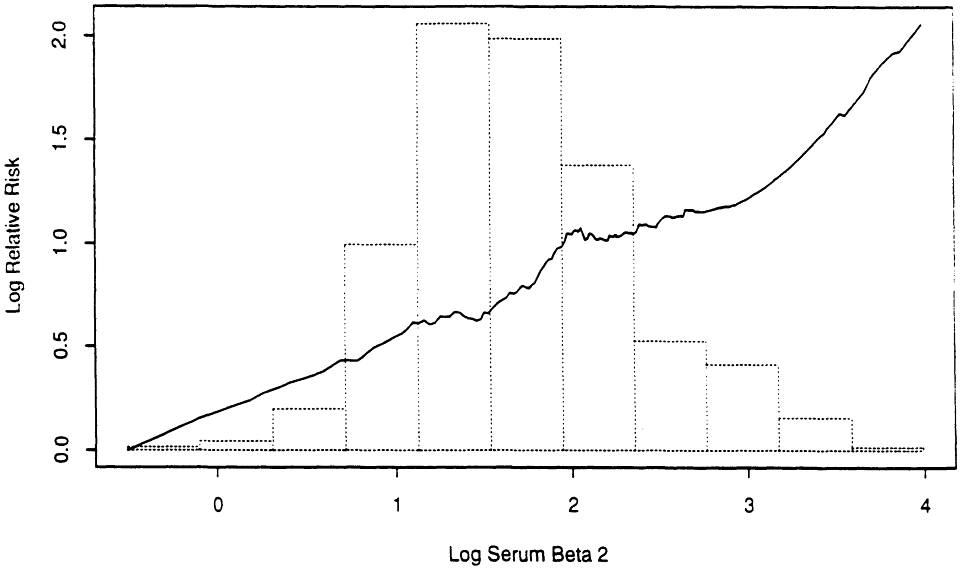

FIG 4B: Local Full Likelihood Estimate of the Log Relative Risk for Log Serum Beta 2 Microglobulin.

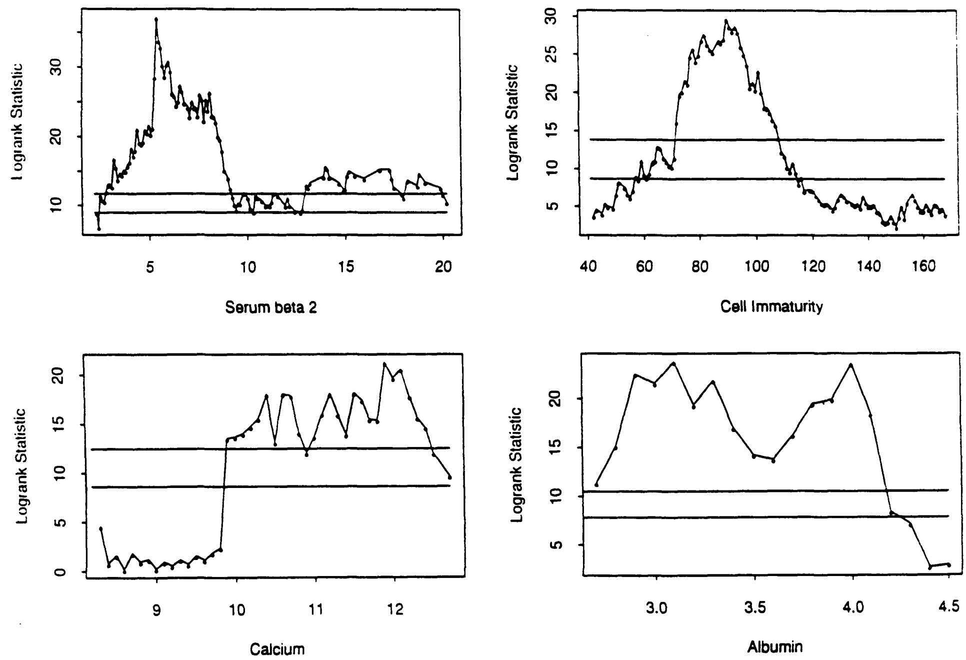

FIG 5: Logrank Statistic as a Function of Several Covariates in the Myeloma Data Set

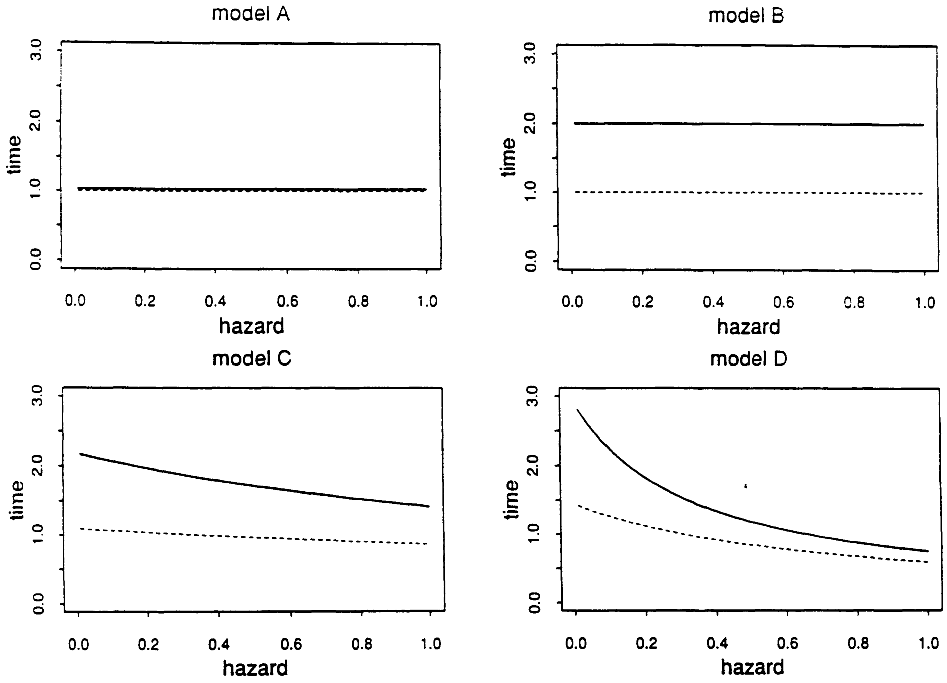

FIG 6: Hazard functions for the simulations. The solid and dashed lines represent the hazard functions for F1 and F2

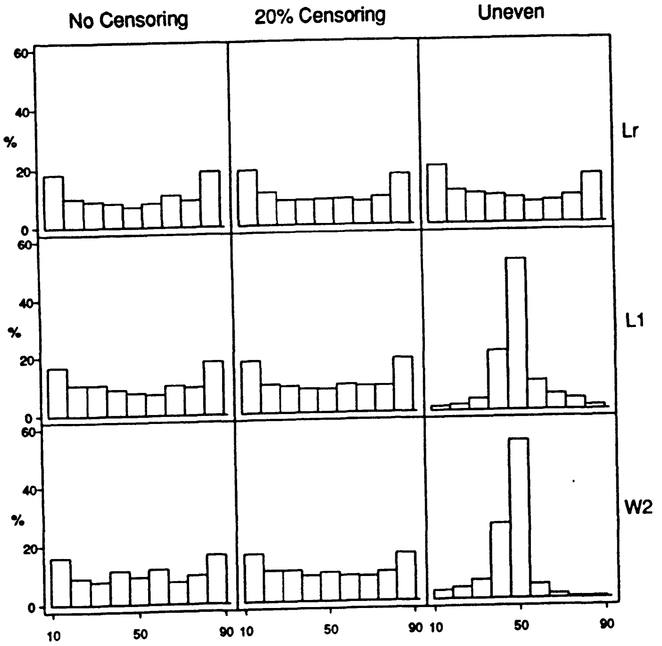

FIG 7: Split Point Frequencies - Model A

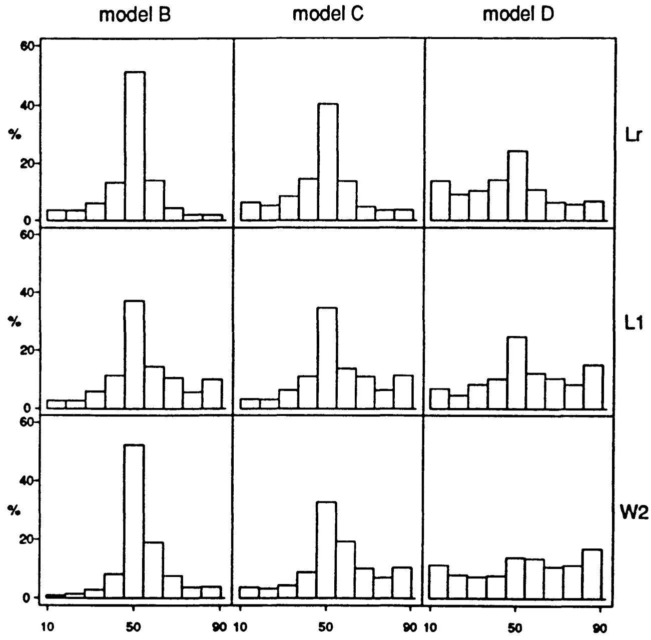

FIG 8A: Split Point Frequencies for Uncensored Data

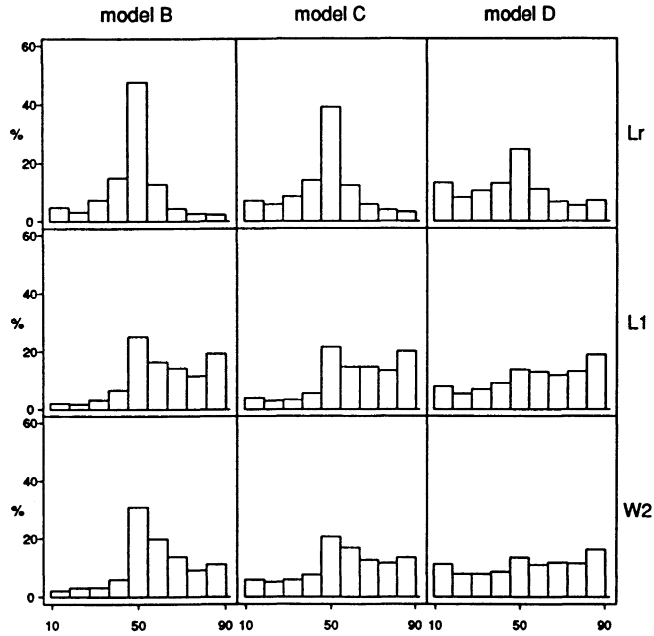

FIG 8B: Split Point Frequencies with 20 % Censoring

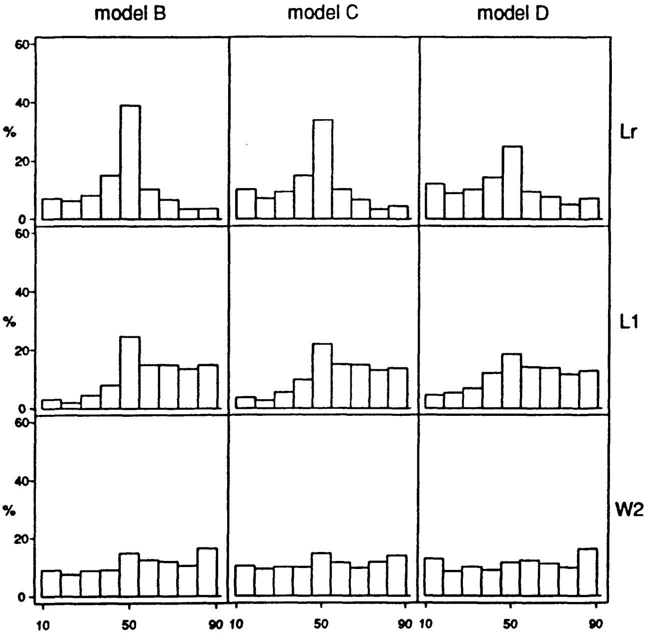

FIG 8C: Split Point Frequencies with 50 % Censoring

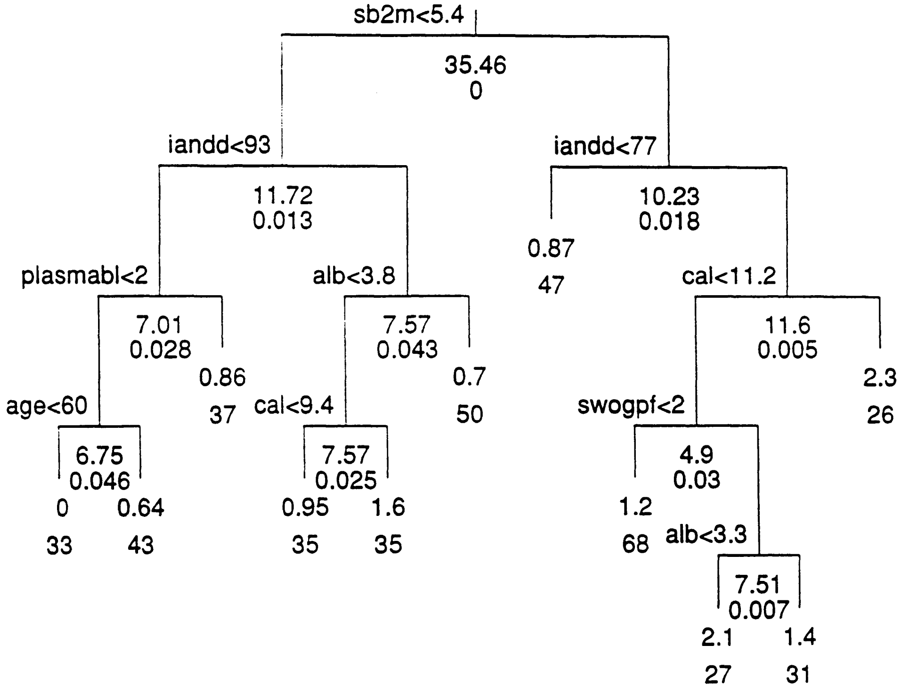

FIG 9: Pruned Survival Tree for the Myeloma Data Set

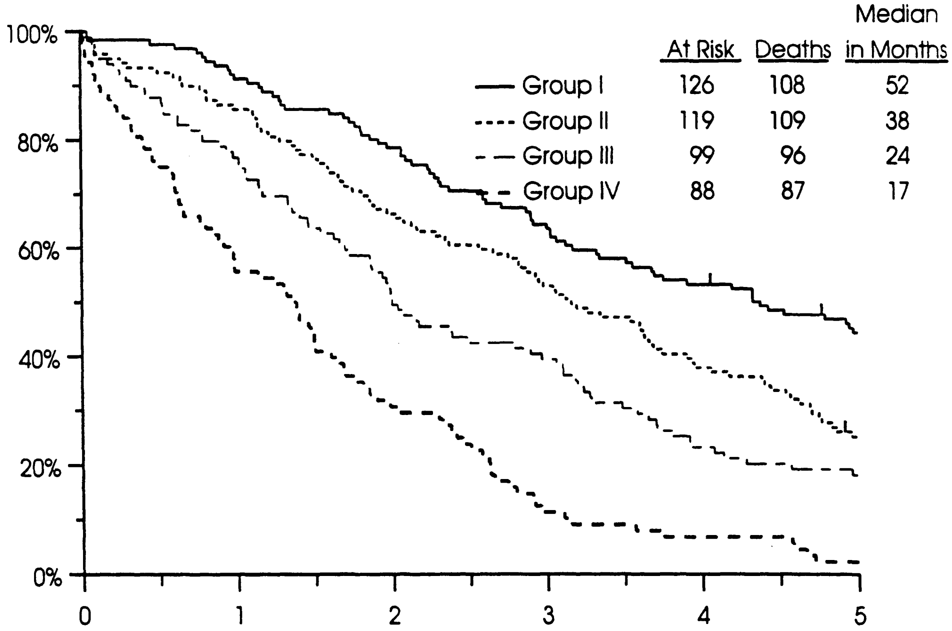

FIG 10A: Survival (in Years) by Amalgamated Nodes

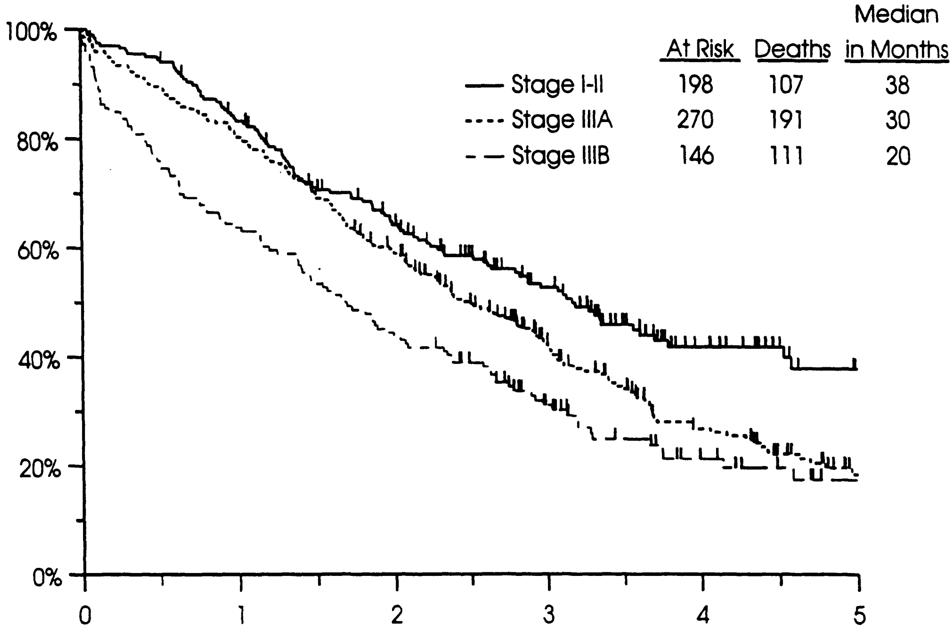

FIG 10B: Survival (in Years) by Durie-Salmon Stage

TABLE 1.

| | | $F_1$ | $F_2$ | Uniform $(0,\gamma)$ | Unequal |
|---|---|---|---|---|---|
| | Model | $\psi_1,\rho_1$ | $\psi_2,\rho_2$ | Censoring | Censoring |
| | | | | 0%, 20%, 50% | |
| No Structure | *A* | 1.000,0.00 | 1.000,0.00 | Yes | Yes |
| Structure | *B* | 2.000, 0.00 | 1.000, 0.00 | Yes | No |
| | *C* | 2.184, 0.25 | 1.092, 0.25 | Yes | No |
| | *D* | 2.888, 1.00 | 1.444, 1.00 | Yes | No |

Survival and Censoring Configurations