Stanford Heart Transplantation Data Revisited: A Real-Time Approach

*Elja Arjas**

Abstract. A form of the discrete time logistic regression model is considered as a means to analyze complicated failure time data. A characteristic property of this approach is that the events recorded in the data are always treated in the model in the order in which they occurred in real time, without first aligning them according to some particular "basic time measurement". Parametric modelling is used throughout. The technique is illustrated by a detailed analysis of the Stanford Heart Transplantation Data. Computational aspects are discussed briefly.

INTRODUCTION

In recent years, the dominant role among regression models of hazard has been played by the semiparametric model of Cox [14] and its extensions (see Andersen and Gill [5] and Andersen and Borgan [4]). An unspecified time-dependent baseline hazard, common to all individuals, is assumed to act multiplicatively on a relative risk function, which is a function of the parameters of interest. Estimation of parameters is based on considering relative risks within risk sets, formed by aligning individuals according to some particular measurement of time, such as age or time since diagnosis. A drawback of such alignment is that it usually changes the natural sequencing of events in real time. In particular, it can destroy the natural martingale structure of the real time counting process models (cf. Sellke and Siegmund [22] and Arjas [6]).

The idea of a multiplicative baseline factor is also present in most fully parametric models of hazard (see Borgan [10] for a general model, Thompson [23], Laird and Olivier [20], Aitkin, Laird and Francis [2] or Tibshirani and Ciampi [24] for piecewise exponential models, and

*Department of Applied Mathematics and Statistics, University of Oulu, Oulu, Finland. This work was completed while the author was visiting the Fred Hutchinson Cancer Research Center in Seattle. It was supported in part by the National Institutes of Health Grant 5R01 GM-28314 and by the Finnish Academy.

Aitkin and Clayton [1] for models having log-linear structure).

We have, in Arjas and Haara [7] and Arjas [6], advocated what might be called a real-time approach to hazard regression, arguing that if a hazard model uses time-dependent covariates, all dependence on time-related quantities can actually be accommodated into such covariates. Then, instead of postulating the existence of a common unspecified multiplicative baseline hazard, a function of some particular measurement of time, one attempts to model how an individual's hazard at a certain (real) time t depends on "the currently prevailing conditions for survival". Frequently, some such conditions are best expressed in terms of conveniently chosen time readings, such as age, time from diagnosis, time from treatment, or indeed, calendar time. Several time readings may be needed simultaneously for a realistic description. Some suitably chosen functions of these readings can then be listed as covariates, among other factors that are thought to be relevant to the individual's survival.

Arjas and Haara [7] showed that, under general conditions, the real time approach leads to likelihood expressions of a common form. This approach has intuitive appeal in that the histories, representing the past and used in the conditioning of the hazards, are always compatible with the actual experiment in the sense of having the events sequenced in the correct order. This makes the results easy to interpret.

On the other hand, the general likelihood formula in [7] has too little structure for immediate statistical application such as parameter estimation. The purpose of this paper is to suggest a concrete way to fill this gap.

In trying to fill the gap, we have incorporated two features that have practical, rather than conceptual or theoretical motivation. First a discrete time parameter is used, which, together with a natural conditional independence assumption, removes all difficulties concerning tied failure times. Second, a logistic regression model with binomial response is the primary statistical tool. This allows us to work within the log-linear family of distributions, and leads to an elegant asymptotic theory and unproblematic numerical routines in ordinary ML-estimation.

In the next section we set up our statistical model. The exact derivation of the corresponding likelihood function and the asymptotic normality of the regression coefficient estimates are deferred respectively to Appendices 1 and 2. We then illustrate the method by considering the well-known Stanford Heart Transplant data. Finally, we make some remarks concerning computation.

THE STATISTICAL MODEL

Choosing some convenient point in real time as the origin, we split the time axis into the unit intervals $(t-1,t]$, $t \geq 1$. As an approxi-

mation, we then think that individuals at risk at the beginning of an interval remain so to the end of it. Consequently, deaths occurring during (t-1,t] are thought of as occurring at t. Similarly, we think of the covariates as remaining fixed during each time interval, with the possible new value always being determined at the beginning of the interval. When the time unit is small, the approximations are unlikely to influence statistical inference a great deal.

The individuals included in the study are indexed by j, $j \geq 1$. We define the risk indicators $Y_j(t-1)$, $j,t \geq 1$, by

$$Y_j(t-1) = \begin{cases} 1 & \text{if individual } j \text{ is at risk during } (t-1,t] \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

and the failure indicators $\Delta N_j(t)$, $j,t \geq 1$, by

$$\Delta N_j(t) = \begin{cases} 1 & \text{if individual } j \text{ at risk during } (t-1,t] \\ & \text{fails during this period} \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

Denote by $R(t-1) = \{j \geq 1 : Y_j(t-1) = 1\}$ the risk set during (t-1,t]. The size of the risk set, $\text{card } R(t-1) = \sum_{j \geq 1} Y_j(t-1)$, is assumed to be finite for all $t \geq 1$.

We do not assume that all individuals are present at time 0; they may enter and leave the risk set many times, and they may also "fail" many times.

Suppose then that, for every individual j and time interval (t-1,t] such that $Y_j(t-1) = 1$, the investigator knows the value of a p-vector $\underset{\sim}{Z}_j(t-1) = (Z_{j1}(t-1), \cdots, Z_{jp}(t-1))$ of the relevant covariates. We shall view the value of $\Delta N_j(t)$ as the outcome of a Bernoulli experiment, where the probability of the event $\{\Delta N_j(t) = 1\}$ depends on $\underset{\sim}{Z}_j(t-1)$. For convenience, we may assume that $\underset{\sim}{Z}_j(t-1) = \underset{\sim}{0}$ whenever $Y_j(t-1) = 0$.

We then assume that the likelihood function corresponding to data up to time t can be expressed as the product

$$L_t^{\beta} = \prod_{s \leq t} \prod_{j \in R(s-1)} P_{\underset{\sim}{\beta}}(\Delta N_j(s) = \Delta n_j(s) \,|\, \underset{\sim}{Z}_j(s-1); Y_j(s-1) = 1) \quad , \tag{2.3}$$

where $\{\Delta n_j(s);\ j \in R(s-1),\ s \leq t\}$ are the observed values of the indicators (2.2). Moreover, we assume that

$$\log \frac{P_{\underset{\sim}{\beta}}(\Delta N_j(s) = 1 \,|\, \underset{\sim}{Z}_j(s-1);\ Y_j(s-1) = 1)}{P_{\underset{\sim}{\beta}}(\Delta N_j(s) = 0 \,|\, \underset{\sim}{Z}_j(s-1);\ Y_j(s-1) = 1)} = \underset{\sim}{\beta}' \underset{\sim}{Z}_j(s-1) \quad . \tag{2.4}$$

The precise sequence of assumptions leading to this logistic regression model for binary data is discussed in Appendix 1. Note the formal similarity of this model and the discrete time model in Cox [14] if one in the latter sets $\lambda_0(t) = 1$. However, setting $\lambda_0(t) = 1$ is a way of

specifying the baseline hazard completely. Here we use a different likelihood than in [14], one that does not suppress $\lambda_0(t)$ by only considering the relative risks of individuals in a risk set, and only at times of failure. Thus (2.3) uses information from all time intervals, whether any failures occur in them or not.

The score function is easily seen to be

$$\frac{\partial}{\partial \beta_i} \log L_{\underset{\sim}{t}}^{\underset{\sim}{\beta}} = \sum_{s \leq t} \sum_j Z_{ji}(s-1)[\Delta N_j(s) - P_{\underset{\sim}{\beta}}(\Delta N_j(s) = 1 \,|\, \underset{\sim}{Z}_j(s-1), Y_j(s-1))] \quad , \tag{2.5}$$

$1 \leq i \leq p$. The equations $\frac{\partial}{\partial \beta_i} \log L_{\underset{\sim}{t}_{max}}^{\underset{\sim}{\beta}} = 0$, $1 \leq i \leq p$, with t_{max} = calendar time of the latest observation, can then be used in the standard way to obtain the ML-estimate $\hat{\underset{\sim}{\beta}}_{t_{max}}$ of the "true" parameter $\underset{\sim}{\beta}_0$. The asymptotic normality of $\hat{\underset{\sim}{\beta}}_t$ as $t \to \infty$ is briefly considered in Appendix 2.

AN EXAMPLE: STANFORD HEART TRANSPLANT DATA REVISITED

The Model. We now apply the above logistic regression model to the Stanford Heart Transplant Data. This famous data set, originally introduced in Clark et al [12], has been since considered and analyzed many times. (See Crowley and Hu [16], Kalbfleisch and Prentice [19]; Section 5.5, and Aitkin, Laird and Francis [2], with discussion, for results and for more references.) Our motive for choosing this data set was that it is widely known from previous analyses, and that the transplantation introduces a set of random and/or time-dependent covariates whose treatment illustrates well our real time approach.

We emphasize that our goal is only to demonstrate the use of a particular statistical technique. Thus, this section should not be taken as one more reanalysis of the Stanford data. Extreme caution should be used in the interpretation of the results, see in particular Gail [18].

We use the data exactly as reported in Crowley and Hu [16]. The follow-up covers 99 patients between September 13, 1967 and March 23, 1974 (4 being deselected because of missing information). The information available about all patients is (j is again the index used to identify the patient):

(1) $T_{BIRTH(j)}$, the date of birth
(2) $T_{ACC(j)}$, the date of acceptance into the program
(3) Previous open heart surgery (1=yes, 0=no)
(4) $T_{EXIT(j)}$, the date last seen
(5) Status on the last day (1=dead, 0=alive)

About transplanted patients there is additionally information about:

(6) $T_{TRANS(j)}$, the date of the transplantation
(7) MM(j), mismatch score, indicating the degree to which donor and recipient are mismatched for tissue type

Crowley and Hu [16] and Kalbfleisch and Prentice [19] applied the semiparametric Cox [14] model on the data, using a single model to accommodate both pretransplant and posttransplant survival. Aitkin, Laird and Francis [2] chose a fully parametric approach and they modelled pretransplant and posttransplant survivals separately.

We follow Aitkin, Laird and Francis in choosing a fully parametric approach, but Crowley and Hu, or Kalbfleisch and Prentice, in that we work with a single model covering both pretransplant and posttransplant survival.

Write for simplicity T_0 for the date September 12, 1967 (the beginning of the follow-up). The variable t always refers to a day in calendar time. We then define the following covariates, some with two alternatives (with and without taking logarithms). A code for each covariate is given in parentheses:

$Z_{j1}(t-1) = 1$

$Z_{j2}(t-1) = \log(t-T_{ACC(j)}+1)$ ("TIME FROM ACC")

$Z_{j3}(t-1) = T_{ACC(j)} - T_0$ ("ACC MONTH")

or $\log(T_{ACC(j)} - T_0 + 1)$

$Z_{j4}(t-1) = T_{ACC(j)} - T_{BIRTH(j)}$ ("ACC AGE")

or $\log(T_{ACC(j)} - T_{BIRTH(j)}$

$Z_{j5}(t-1) = 1$(patient j has had a previous open heart surgery) ("SURGERY")

$Z_{j6}(t-1) = 1(T_{TRANS(j)} \leq t \leq T_{TRANS(j)} + 70)$ ("PHASE 1")

$Z_{j7}(t-1) = Z_{j6}(t-1) \cdot \log(t-T_{TRANS(j)}+1)$ ("TIME FROM TRANS")

$Z_{j8}(t-1) = 1(T_{TRANS(j)} + 70 \leq t \leq T_{TRANS(j)} + 365)$ ("PHASE 2")

$Z_{j9}(t-1) = 1(T_{TRANS(j)} + 365 < t)$ ("PHASE 3")

$$Z_{j10}(t-1) = 1(T_{TRANS(j)} \leq t) \cdot (T_{TRANS(j)} - T_{ACC(j)} + 1) \quad \text{("WAIT")}$$

$$\underline{\text{or}} \quad 1(T_{TRANS(j)} \leq t) \cdot \log(T_{TRANS(j)} - T_{ACC(j)} + 1)$$

$$Z_{j11}(t-1) = 1(T_{TRANS(j)} \leq t) \cdot MM(j) \quad \text{("MISMATCH")}$$

$$Z_{j12}(t-1) = 1(T_{TRANS(j)} \leq t) \cdot (T_{TRANS(j)} - T_0 + 1), \quad \text{("TRANSMONTH")}$$

$$\underline{\text{or}} \quad 1(T_{TRANS(j)} \leq t) \cdot \log(T_{TRANS(j)} - T_0 + 1)$$

$$Z_{j13}(t-1) = 1(T_{TRANS(j)} \leq t) \cdot (T_{TRANS(j)} - T_{BIRTH(j)} + 1) \quad \text{("TRANSAGE")}$$

$$\underline{\text{or}} \quad 1(T_{TRANS(j)} \leq t) \cdot \log(T_{TRANS(j)} - T_{BIRTH(j)} + 1)$$

$$Z_{j14}(t-1) = 1(T_{TRANS(j)} \leq t) \cdot Z_{j5}(t-1) \quad \text{("TRANSSURGERY")}$$

Here $1(\cdot)$ is the indicator of the event inside the parenthesis. Thus covariates $Z_{j6}, \cdots, Z_{j14}$ are only non-zero when $t \geq T_{TRANS(j)}$. In covariates Z_{j4} and Z_{j13} time is expressed in years, in Z_{j3} and Z_{j12} in months (in the latter rounded to the nearest full month). These choices were made purely for convenience.

According to the logistic regression model (2.3) and (2.4), we then assume that, for the j^{th} individual during the t^{th} day (calendar time), the odds ratio between death probability and survival probability is given by

$$\frac{P_{\sim}^{\beta}(j \text{ dies during the } t^{th} \text{ day} \,|\, \underset{\sim}{Z}_j(t-1);\ Y_j(t-1) = 1)}{P_{\sim}^{\beta}(j \text{ survives the } t^{th} \text{ day} \,|\, \underset{\sim}{Z}_j(t-1);\ Y_j(t-1) = 1)} \tag{3.1}$$

$$= e^{\beta_1} \cdot (\text{TIME FROM ACC})^{\beta_2}$$
$$\cdot \exp\{\beta_3 \cdot (\text{ACCMONTH}) + \beta_4 \cdot (\text{ACCAGE}) + \beta_5 \cdot (\text{SURGERY})\}$$
$$\cdot f_j(t;\beta_6,\beta_7,\beta_8,\beta_9)$$
$$\cdot \exp\{\beta_{10} \cdot (\text{WAIT}) + \beta_{11} \cdot (\text{MISMATCH}) + \beta_{12} \cdot (\text{TRANSMONTH})$$
$$+ \beta_{13} \cdot (\text{TRANSAGE}) + \beta_{14} \cdot (\text{TRANSSURGERY})\} \quad ,$$

where

$$f_j(t;\beta_6,\beta_7,\beta_8,\beta_9)$$

$$= \begin{cases} 1 \quad , & \text{if } t < T_{TRANS(j)} \\ e^{\beta_6} \, (\text{TIME FROM TRANS})^{\beta_7}, & \text{if } T_{TRANS(j)} \leq t \leq T_{TRANS(j)} + 70 \\ e^{\beta_8} \quad , & \text{if } T_{TRANS(j)} + 70 < t \leq T_{TRANS(j)} + 365 \\ e^{\beta_9} \quad , & \text{if } T_{TRANS(j)} + 365 < t \quad . \end{cases}$$

We call this the INITIAL MODEL. Note that, in practice, the denominator $P_{\beta}(j \text{ survives the } t^{th} \text{ day} | Z_j(t-1); Y_j(t-1) = 1)$ is very close to one. Therefore the right hand side in (3.1) is "almost" an expression of discrete time hazard.

We then comment on how the time variable t comes up in the model. Consider first the pretransplant survival. Figure 1 describes cumulative hazard from the date of acceptance to the program. (Transplantation itself is treated as a censoring mechanism).

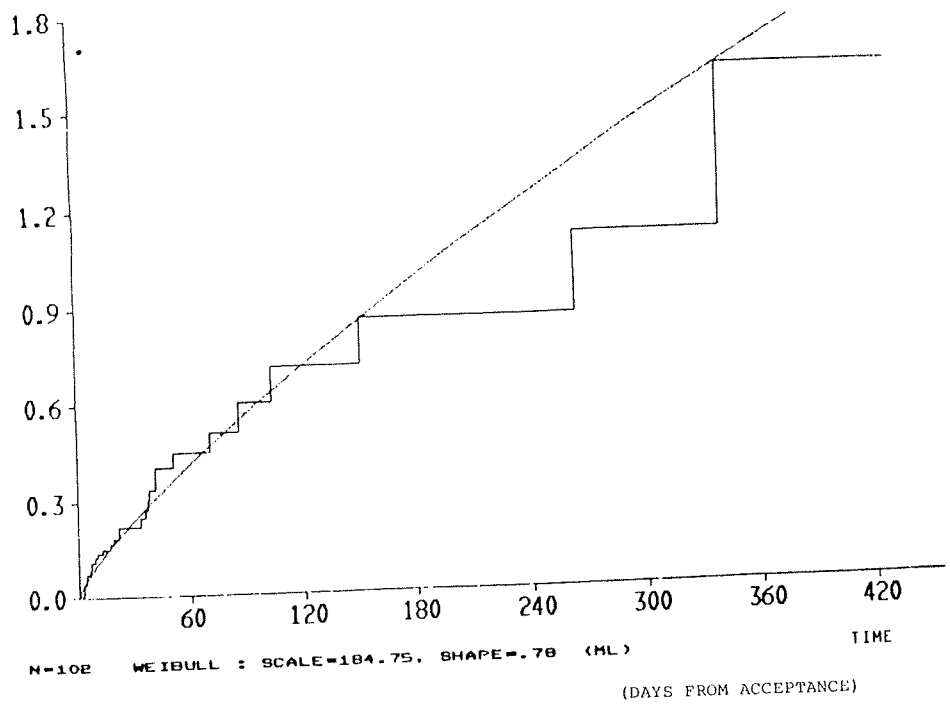

Figure 1. Nelson plot for pretransplant survival times and the fitted Weibull cumulative hazard function.

It is clear from Figure 1 that individuals who have survived long without transplantation, face a smaller risk than those who have only recently been accepted to the program. (This is probably because some of the incoming patients are healthier than others, and survival acts as a selection mechanism, with the healthier patients surviving longer.) Consequently, any reasonable functional expression for hazard should depend on the time from acceptance. We model this dependence by including the term $\beta_2 Z_{j2}(t-1)$ in the linear expression for the log-odds of dying during the t^{th} day. Then the (discrete time) hazard becomes approximately proportional to a power of time from acceptance, in agreement with the Weibull hazard form (cf. Aitkin, Laird, and Francis [2]).

On the other hand, for transplanted individuals, the hazard depends crucially on the time since the transplantation. This is clear from Figure 2, which describes cumulative hazard after the transplantation.

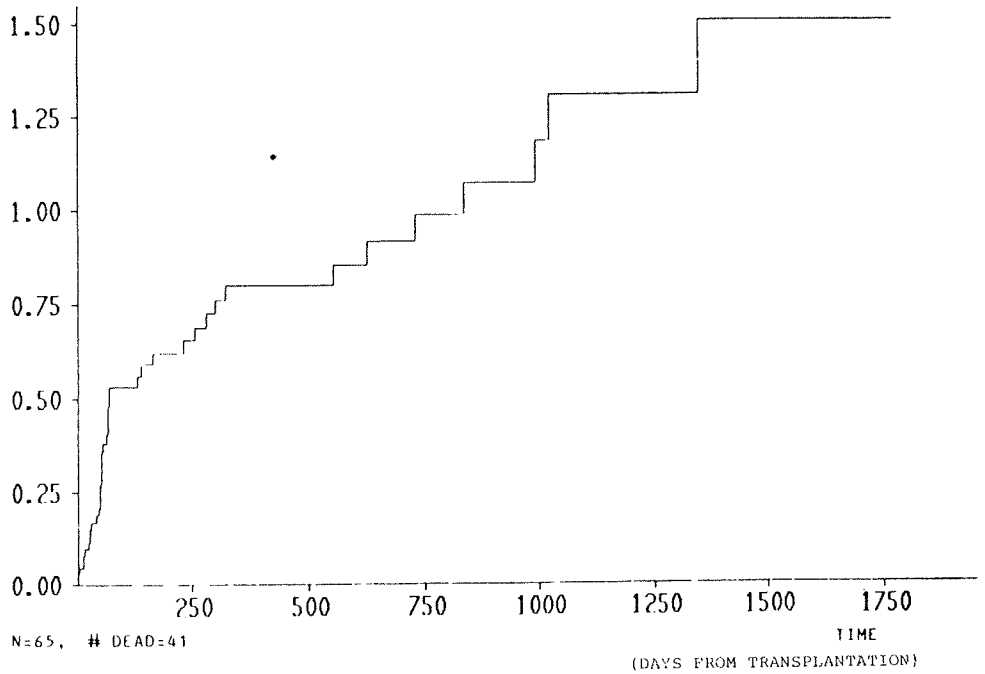

Figure 2. Nelson plot for posttransplant survival times.

There appears to be a very critical period of approximately 70 days after the transplantation ("PHASE 1") after which the hazard abruptly decreases (cf., again, Aitkin, Laird and Francis). In the INITIAL MODEL, we use a Weibull-type hazard model over the 70 day period, and thereafter two intervals of constant hazard ("PHASE 2" and "PHASE 3").

We want the model to describe the potential effect which the transplantation may have on survival. Therefore, it is natural to build it in such a way that, for a patient who is at risk during the t^{th} day and who has been transplanted, the model would also give a value of the hazard corresponding to the alternative where the patient had not been transplanted. But then, by the above argument, it becomes necessary to express the hazard of a transplanted patient as a function which depends on both $(t-T_{ACC(j)})^+$ = time from acceptance and $(t-T_{TRANS(j)})^+$ = time from transplantation. (Compare this with the discussion in Crowley and Storer [17] and Aitkin, Laird and Francis [2]; rejoinder.) By mimicking the behavior found in Figures 1 and 2 by smooth functions, we describe this dependence by including in our model (3.1) the factor $((t-T_{ACC(j)}+1)^+)^{\beta_2} f_j(t;\beta_6,\beta_7,\beta_8,\beta_9)$. Later we are able to simplify this to some extent (see REDUCED MODELS below).

Thus there is some trade-off between our parametric and Cox's semi-parametric approach: We must try to find a reasonable parametric description of the various ways in which "time" influences hazard. But then we are not forced to work with a common baseline hazard, depending on a single time reading, and there is no need to artifically align the observations. We also get a "fuller" form of likelihood to work with (see Appendix 1), and obtain, by a straightforward ML-procedure, estimates for the absolute (instead of only relative) hazard.

We remark that we could have taken the real time approach even further, by employing more calendar time-dependent covariates. For example, we could have used $t-T_{BIRTH(j)}$ instead of $T_{ACC(j)}-T_{BIRTH(j)}$ as an age variable, and $t-T_0$ instead of $T_{ACC(j)}-T_0$ as a variable describing a time trend during the study. On the other hand, the method we use does not require any of the covariates to be a time reading.

Commenting finally on the fixed covariates we see that $\beta_3 Z_{j3}(t-1) + \beta_4 Z_{j4}(t-1) + \beta_5 Z_{j5}(t-1)$ is a patient dependent modification of the pretransplant log-odds $\beta_1 + \beta_2 Z_{j2}(t-1)$ for survival. Recall that the transplantation effect $\beta_6 Z_{j6}(t-1) + \cdots + \beta_{14} Z_{j14}(t-1)$ is zero for $t < T_{TRANS(j)}$. Note also that information which is used in the fixed covariates to describe pretransplant survival, can come up again in a different role, now as part of the transplantation effect.

Empirical Results. We fitted INITIAL MODEL (3.1), trying the covariates Z_{j2}, Z_{j3}, Z_{j10}, Z_{j12} and Z_{j13} both in the logarithmic and non-logarithmic form. We also tried several simpler alternatives, which were obtained from the INITIAL MODEL by dropping some covariates, or replacing covariates by others. Numerical estimates of the regression coefficients, calculated from four different models, are given in Table 1. In REDUCED MODELS, covariates Z_{j6}, Z_{j8} and Z_{j9} corresponding to PHASE 1, PHASE 2 and PHASE 3 were replaced by the single covariate $1(T_{TRANS(j)} \leq t)$. Also, covariates Z_{j5}, Z_{j10}, Z_{j12}, Z_{j13} and Z_{j14} were eliminated.

In order to judge the goodness-of-fit of our models, we used both deviance and a graphical method based on total hazards (generalized residuals). Lack of space does not permit us to consider the graphical method here.

In line with our earlier comment that this section should not be viewed as a reanalysis of the Stanford data, we leave it to the reader to look for similarities and dissimilarities between our results and some earlier analyses. Neither do we comment on the values and significance levels of the regression coefficients in the four models reported. The numerical values of the regression coefficients were very stable from model to model provided that the corresponding covariate remained the same.

TABLE 1.

Parameter estimates from four models (standard errors in parentheses)

PARAMETER	NON-LOGARITHMIC		LOGARITHMIC	
	INITIAL MODEL	REDUCED MODEL*)	INITIAL MODEL	REDUCED MODEL*)
		Pretransplant survival (all patients):		
β_1 INTERCEPT	-4.7395 (.9532)	-4.5576 (.8205)	-3.4796 (1.0240)	-3.1965 (.9139)
β_2 TIME FROM ACC	- .3397 (.1147)	- .3634 (.0973)	- .2853 (.1185)	- .3394 (.0976)
β_3 ACCMONTH	- .0207 (.0090)	- .0172 (.0059)	- .6714 (.2229)	- .6267 (.1615)
β_4 ACCAGE	.0156 (.0174)	.0101 (.0141)	.0376 (.0189)	.0313 (.0141)
β_5 SURGERY	- .0703 (.6148)	---	.0094 (.6363)	---
	Transplantation effect on survival (only transplanted patients):			
γ_7 TIME FROM TRANS	.4376 (.2559)	.4663 (.0997)	.4609 (.2600)	.4727 (.0998)
γ_8 PHASE 1	-1.4866 (1.2679)	-1.7618 (.5170) (Combined estimate)	-1.2083 (1.4147)	-1.7612 (.5181) (Combined estimate)
γ_9 PHASE 2	-1.5259 (.9593)		-1.2149 (1.1074)	
γ_{10} PHASE 3	-1.6182 (.9734)		-1.3559 (1.2098)	
γ_{11} WAIT	1.90×10^{-5} (.0053)	---	- .4543 (.1613)	---
γ_{12} MISMATCH	.4052 (.2989)	.4739 (.2767)	.4779 (.3107)	.4732 (.2772)
γ_{13} TRANSMONTH	.0094 (.0121)	---	.3463 (.7176)	---
γ_{14} TRANSAGE	- .0083 (.0197)	---	- .0102 (.0229)	---
γ_{15} TRANSSURG	- .6164 (.8056)	---	- .5556 (.8093)	---
	Deviance = 892.42	Deviance = 895.11	Deviance = 886.94	Deviance = 890.32

1) In the logarithmic models covariates Z_{j3}, Z_{j10} and Z_{j12} are logarithmic

We now consider survival prognoses that can be obtained from the estimated models. Taking a fictitious patient with given characteristics, we estimate survival distributions corresponding to different waiting times between acceptance and transplantation. In order to do this, we need to consider the time at which a donor heart becomes available as known at the time of acceptance, with no effect on the patient's pretransplant survival.

To take a concrete case, we consider two patients with similar characteristics as in Aitkin, Laird and Francis [2]:

	Patient I	Patient II
ACCMONTH	12	50
ACCAGE	42	55
SURGERY	0	0
MISMATCH	.5	1.5

The options for waiting time are:

(1) transplantation at the time of acceptance (WAIT = 0)

(2) transplantation 50 days after the time of acceptance, if still alive (WAIT = 50)

(3) transplantation 200 days after the time of acceptance, if still alive (WAIT = 200)

(4) no transplantation (WAIT = ∞)

Survival curves were determined on the basis of the fitted models. The curves corresponding to NONLOGARITHMIC INITIAL MODEL are displayed in

Figures 3(a), (b).

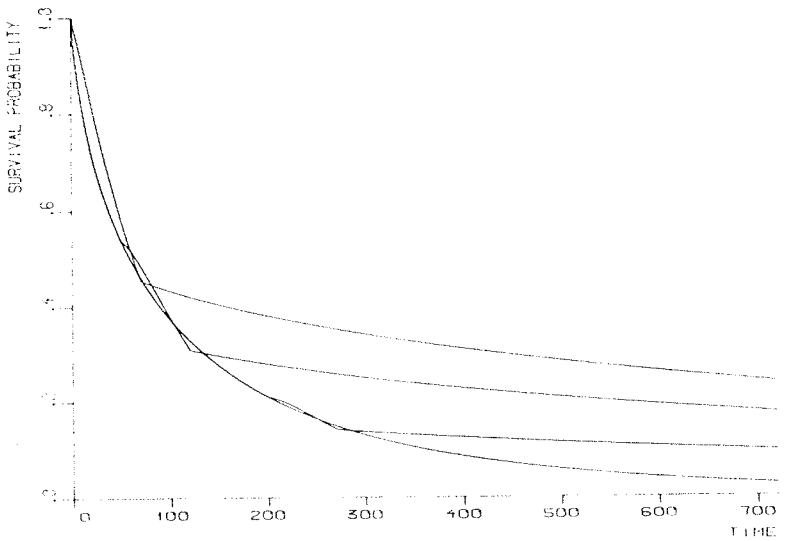

Figure 3(a). Estimated survival probabilities for Patient I (NON-LOGARITHMIC INITIAL MODEL).

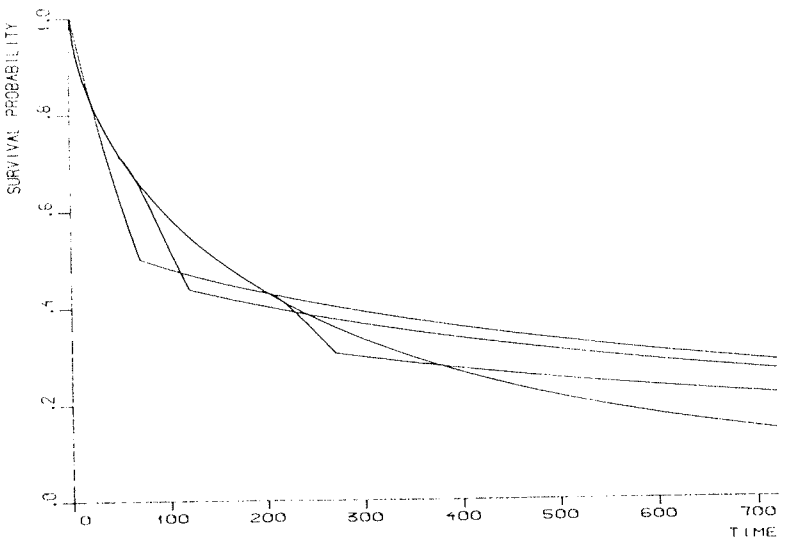

Figure 3(b). Estimated survival probabilities for Patient II (NON-LOGARITHMIC INITIAL MODEL).

For Patient I, all models indicate that transplantation would be beneficial (almost) uniformly over time as compared to no transplantation. The benefit is the bigger the earlier the transplantation. For Patient II, transplantation improves only long-term survival probabilities and the improvements are slight.

For either patient, there was relatively little variation between the prognoses obtained from different models.

NOTES ON COMPUTATION

In a purely technical sense, we have considered a multiple logistic regression model for binary data. In principle, therefore, the parameter estimation could be done by using some well-known statistical package (such as GLIM or BMDP), which includes programs for handling logistic regression. However, there is a practical problem: The data matrix contains the values of $Z_j(t-1)$ and $\Delta N_j(t)$ for each (j,t)-pair such that $Y_j(t-1) = 1$. If the time unit is short (as it preferably should be in order to avoid serious discretization errors), the size of the data matrix becomes prohibitively large. For Stanford data the number of patient days exceeded 30,000. Such a data matrix cannot be stored conveniently into a core of a computer.

A computer program was devised to solve the problem. The key idea in the program is that, instead of computing the covariate values once and for all to be stored in a file, the covariate vectors $Z_j(t-1)$ are determined directly from the data every time they are needed in the iterations. A valuable computational aspect of our approach is that each (j,t)-pair can be treated separately, then only adding a term to the logarithmic likelihood expression. Thus there is no need to form, as in Cox's model, risk sets consisting of individuals with matching baseline hazards, also keeping track on their time-dependent covariates. Finally, as we have remarked earlier, no tie-breaking procedures are required.

A more detailed description of the program, together with examples and experiences about computing, will be reported later.

Acknowledgements. The programming and computing in this work was done by Risto Bloigu and Pekka Kangas. Without their skill and dedication to work, it would have been impossible to apply the method to any reasonable sized problem. My sincerest thanks to them. I am also grateful to Pentti Haara, Suresh Moolgavkar and John Crowley for useful discussions.

APPENDIX I: DETAILS OF THE STATISTICAL MODEL

Let (Ω,F) be a measurable space in which the variables $Y_j(t-1)$, $Z_j(t-1)$ and $\Delta N_j(t)$ are defined. Let F_0 be the σ-field representing "initial information"; usually F_0 is the trivial field. Then the σ-fields F_t and G_{t-1}, $t \geq 1$, defined inductively by

$$G_{t-1} = F_{t-1} \vee \sigma\{R(t-1),\ \{Z_j(t-1);\ j \in R(t-1)\}\} \quad ,$$

$$F_t = G_{t-1} \vee \sigma\{\Delta N_j(t);\ j \in R(t-1)\} \quad ,$$

represent the experimental history registered up to time t, $\mathcal{F}_t$ including and $\mathcal{G}_{t-1}$ excluding the failures during (t-1,t].

Consider a statistical model $\{P^\theta;\theta\in\Theta\}$ for the observation process $(R(s-1), \{\Delta N_j(s), \underset{\sim}{Z}_j(s-1);\ j\in R(s-1)\})_{s\geq 1}$ and a $P\theta$-likelihood which corresponds to data collected up to time t, $t\geq 1$.

Suppose that the parameter θ can be represented in the form $\theta=(\theta_1,\theta_2)$, where θ_1 is the parameter of interest and θ_2 is a nuisance parameter. Typically, we think of θ_1 as parametrizing the conditional distribution of the variables $\Delta N_j(s)$, conditioned on $\mathcal{G}_{s-1}$, and of θ_2 as the parameter associated with the conditional law of the variables $R(s)$ and $\underset{\sim}{Z}_j(s)(j\in R(s))$, given $\mathcal{F}_{s-1}$. The full likelihood corresponding to the observed values $\{r(s-1),\{\Delta n_j(s),\underset{\sim}{z}_j(s-1);\ j\in r(s-1)\};\ s\leq t\}$ can then be expressed as the product of two terms, viz., as

$$\prod_{s\leq t} P^\theta(R(s-1)=r(s-1),\underset{\sim}{Z}_j(s-1)=\underset{\sim}{z}_j(s-1);\ j\in r(s-1)\,|\,\mathcal{F}_{s-1}) \tag{A.1}$$

$$\cdot \prod_{s\leq t} P^\theta(\Delta N_j(s)=\Delta n_j(s);\ j\in r(s-1)\,|\,\mathcal{G}_{s-1}) \quad .$$

Following Cox [15], the second factor can be called a partial likelihood. Ordinary ML-estimation of θ_1, the parameter of interest, can be done by considering that factor alone provided that the following condition holds:

Assumption 1. (i) For each $s\geq 1$, the conditional P^θ-distribution of $\{R(s-1),\{\underset{\sim}{Z}_j(s-1);\ j\in R(s-1)\}\}$, given $\mathcal{F}_{s-1}$, does not depend on θ_1; and (ii) For each $s\geq 1$, the conditional P^θ-distribution of $\{\Delta N_j(s);\ j\in R(s-1)\}$, given $\mathcal{G}_{s-1}$, does not depend on θ_2.

Of course, the validity of Assumption 1 depends on the model $\{P^\theta;\ \theta\in\Theta\}$. Actual verification of this assumption would require that the model were fully specified, including the probability law of the censoring mechanism and possible random covariates. This is usually not done explicitly. However, part (ii) of Assumption 1 becomes obvious if the censoring times and the covariates are fixed, or random but $\mathcal{F}_0$-measurable. More generally, we can consider (ii) to be valid if the censoring is non-informative about θ_1 and the covariates are external (cf. Kalbfleisch and Prentice [19]). For internal covariates more caution is needed: If (i) is not met, also the first factor in (A.1) can depend on θ_1, and then using only the second factor in the maximization is a potential source of bias. Finally, it seems that part (ii) in Assumption 1 can always be met in practice by making a convenient choice for θ_1, the parameter of interest.

For a continuous time version of Assumption 1, see Arjas and Haara [7].

Our next assumption imposes an independence condition between the individuals and simplifies, in particular, the handling of ties.

Assumption 2. For each $s \geq 1$, and $\theta \in \Theta$, the random variables $\{\Delta N_j(s);\ j \geq 1\}$ are conditionally P^θ-independent given $\mathcal{G}_{s-1}$.

This assumption is likely to hold in practice if there are no multiple failures of common cause, or if such failures can occur but the background variable causing the failure can be included as a covariate.

Under Assumptions 1 and 2, the likelihood function (A.1) depends on θ_1 only through the factor

$$\prod_{s \leq t} \prod_{j \in r(s-1)} P^{\theta}(\Delta N_j(s) = \Delta n_j(s) \mid \mathcal{G}_{s-1}) \quad . \tag{A.2}$$

On the other hand, because of Assumption 1 (ii), this expression does not depend on θ_2.

It remains to specify the conditional probabilities in (A.2). Our next assumption guarantees that all relevant information in $\mathcal{G}_{s-1}$, when used as a condition for the probability of $\{\Delta N_j(s) = \Delta n_j(s)\}$, is actually contained in the p-vector $\underset{\sim}{Z}_j(s-1)$ and the indicator $Y_j(s-1)$.

Assumption 3. For all s, $j \geq 1$, and $\theta \in \Theta$, $\Delta N_j(s)$ and $\mathcal{G}_{s-1}$ are conditionally P^θ-independent given $Y_j(s-1)$ and $\underset{\sim}{Z}_j(s-1)$.

As a last step, we specify the conditional probabilities according to the logistic regression model for binomial response (see e.g., Bock [9], Cox [13], Thompson [23] and Plackett [21]). We also change the notation of the parameter, writing $\underset{\sim}{\beta} = (\beta_1, \cdots, \beta_p)'$ instead of θ_1 and $P^{\underset{\sim}{\beta}}$ instead of P^θ.

Assumption 4. For all $t \geq 1$,

$$P^{\underset{\sim}{\beta}}(\Delta N_j(s) = 1 \mid \mathcal{G}_{s-1}) = Y_j(s-1)\, L(\underset{\sim}{\beta}' \underset{\sim}{Z}_j(s-1)) \quad , \tag{A.3}$$

where $L(x) = (1 + \exp(-x))^{-1}$.

As is well-known, an alternative way to (A.3) is to use "log-odds": For (j,t) such that $j \in R(s-1)$,

$$\log \frac{P^{\underset{\sim}{\beta}}(\Delta N_j(s) = 1 \mid \mathcal{G}_{s-1})}{P^{\underset{\sim}{\beta}}(\Delta N_j(s) = 0 \mid \mathcal{G}_{s-1})} = \underset{\sim}{\beta}' \underset{\sim}{Z}_j(s-1) \quad .$$

APPENDIX 2: ASYMPTOTIC NORMALITY

Apart from trivial cases, the exact distribution of the ML-estimate for $\underset{\sim}{\beta}$ is not known. Therefore, asymptotic results are needed to approximate this distribution. Here we only mention the key asymptotic

theorem and a corollary. The exact regularity conditions and the proof of these results can be found in Arjas and Haara [8].

Let β_0 be the fixed "true value" of the parameter and let $\hat{\beta}_t$ be the ML-estimate corresponding to the data on the time interval $[0,t]$. For simplicity we write P in place of P^{β_0}. Denote $I_t(\beta) = -\frac{\partial^2}{\partial\beta^2} \log L_t^{\beta}$.

Theorem (Asymptotic normality of $\hat{\beta}_t$). Suppose that there exists a sequence (a_t) of constants such that

$$\frac{1}{a_t} I_t(\beta_0) \xrightarrow{P} \Sigma \quad \text{as } t \to \infty \quad ,$$

where Σ is positive definite. Then, under further regularity conditions,

$$a_t (\hat{\beta}_t - \beta_0) \xrightarrow{D} N(0, \Sigma^{-1}) \quad \text{as } t \to \infty \quad ,$$

where "D" means convergence in distribution with respect to P.

Corollary. Under the conditions of the Theorem,

$$2(\log L_t^{\hat{\beta}_t} - \log L_t^{\beta_0}) \xrightarrow{D} \chi_p^2 \quad \text{as } t \to \infty \quad .$$

REFERENCES

[1] M. AITKIN and D. CLAYTON. The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. Applied Statistics, 29, (1980), pp. 156-163.

[2] M. AITKIN, N.M. LAIRD and B. FRANCIS. Reanalysis of the Stanford heart transplant data. (With discussion), J. Am. Statist. Assoc., 78, (1983), pp. 264-292.

[3] A. ALBERT and J.A. ANDERSON. On the existence of maximum likelihood estimates in logistic regression models. Biometrika, 71, (1984), pp. 1-10.

[4] P.K. ANDERSEN and Ø. BORGAN. Counting process models for life history data: a review. (To appear), Scandinavian J. of Statist., (1985).

[5] P.K. ANDERSEN and R.D. GILL. Cox's regression model for counting processes: a large sample study. The Annals of Statist., 10, (1982), pp. 1100-1120.

[6] E. ARJAS. Contribution to the discussion on the paper by Andersen and Borgan. (To appear), Scandinavian J. of Statist., (1985).

[7] E. ARJAS and P. HAARA. A marked point process approach to censored failure data ith complicated covariates. Scandinavian J. of Statist., 11, (1984), pp. 193-209.

[8] E. ARJAS and P. HAARA. A logistic regression model for hazard: asymptotic theory. Manuscript in preparation, (1985).

[9] R.D. BOCK. Estimating multinomial response relations. In: Essays in Probability and Statistics (Ed. R.C. Bose et al). University of North Carolina Press, Chapel Hill, (1968).

[10] Ø. BORGAN. Maximum likelihood estimation in parametric counting process models. Scandinavian J. of Statist., 11, (1984), pp. 1-16.

[11] N. BRESLOW. Covariance analysis of censored survival data. Biometrics, 30, (1974), pp. 89-99.

[12] D.A. CLARK, E.B. STINSON, R B. GRIEPP, J.S. SCHOREDER, N.E. SHUMWAY and D.C. HARRISON. Cardiac transplantation in man, VI. Prognosis of patients selected for cardiac transplantation. Annals of Internal Medicine, 75, (1971), pp. 15-21.

[13] D.R. COX. The Analysis of Binary Data. Chapman and Hall, London, (1970).

[14] D.R. COX. Regression models and life tables. (With discussion), J. of the Royal Statist. Society, Ser. B, 74, (1972), pp. 187-220.

[15] D.R. COX. Partial likelihood. Biometrika, 62, (1975), pp. 269-276.

[16] J. CROWLEY and M. HU. Covariance analysis of heart transplant survival data. J. of the Am. Statist. Assoc., 72, (1977), pp. 27-36.

[17] J. CROWLEY and B.E. STORER. Comment on the paper by Aitkin, Laird and Francis. J. of the Am. Statist. Assoc., 78, (1983), pp. 277-281.

[18] M.H. GAIL. Comment on the paper by Aitkin, Laird and Francis. J. of the Am. Statist. Assoc., 78, (1983), pp. 275-277.

[19] J.D. KALBFLEISCH and R.L. PRENTICE. The Statistical Analysis of Failure Time Data. Wiley, New York, (1980).

[20] N.M. LAIRD and D. OLIVIER. Covariance analysis of censored survival data using log-linear analysis techniques. J. of the Am. Statist. Assoc., 75, (1981), pp. 231-240.

[21] R.L. PLACKETT. The Analysis of Categorical Data. Griffin, London, (1981).

[22] T. SELLKE and D. SIEGMUND. Sequential analysis of the proportional hazards model. Biometrika, 70, (1983), pp. 315-326.

[23] W.A. THOMPSON JR. On the treatment of grouped observations in life studies. Biometrics, 35, (1977), pp. 463-470.

[24] R.J. TIBSHIRANI and A. CIAMPI. A family of proportional- and additive-hazards models for survival data. Biometrics, 39, (1983), pp. 141-147.