Asian European Journal of Probability and Statistics
Vol. 1, No-1, (2024), pp. 13-27
© Permanent Blue, India

A revisit on the survival analysis of two popular data sets via nonparametric and semi-parametric methods: New results with discussion

Indranil Ghosh[a], Giana Maldari[b] and Jonathan Richards [c]

[a] *Department of Mathematics and Statistics, University of North Carolina, Wilmington, NC, USA* [b]*Department of Mathematics and Statistics, University of North Carolina, Wilmington, NC, USA* [c] *Department of Mathematics and Statistics, University of North Carolina, Wilmington, NC, USA*

ARTICLE HISTORY
Compiled June 14, 2024

Received 11 December 2023; Accepted 18 April 2024

Abstract
In this article, we revisit the survival analysis of two different well-known data sets– (i) Stanford Heart Transplant (SHT) data, and (ii) AIDS data via nonparametric and semi-parametric methods. The novelty of this current work is manifold. For the STH data, we compare the performance of the survival analysis is considered via semi-parametric methods (this is new), and also explore the survival rates of patients between recipients and non-recipients of a heart transplant to illustrate the benefits of a heart transplant. On the other hand, we performed a survival analysis of the AIDS data via four different estimation strategies and make a comparison study. Such an extensive study has not been done earlier for this AIDS data to the best of the knowledge of the authors.

KEYWORDS
Survival Analysis; Life-table; hazard Model; Stanford heart plant data; Kaplan-Meier curve; Cox Regression; Semiparametric methods; AIDS data.

1. Introduction

The applications of life-table method can be found in several different aspects of human life including, but not limited to pharmaceutical industries, insurance risk modeling among others. A majority of survival analysis methods focus on right censoring since it occurs far more frequently than left censoring. The incompleteness of data makes the conventional statistical methods inappropriate. The analysis of survival data can be considered by any of the following three approaches—(a) nonparametric, (b)parametric, and/or (c) semiparametric. Recently, there has been some studies in which researchers have advocated the strategy of online updating of nonparametric survival estimator and nonparametric survival test, see Xue et al. (2020) and the references cited therein. However, the proposed strategy is more suitable to deal with

CONTACT Author[a]. Email: ghoshi@uncw.edu

massive amounts of data. Dzinza and Ngwira (2022) made a comparison study between parametric and Cox regression using HIV/AIDS survival data from a retrospective study in Malawi and established that parametric models may perform equally well as the Cox regression. For a rigorous in depth overview on the topic of recent statistical methods for survival analysis, we refer an enthusiastic reader a special issue on this topic by the Japenese Journal of Statistics and Data Science (2021). In the next, we discuss a brief history of the Stanford heart transplantation data and associated methodology The Stanford Heart Transplantation Program, for details, see Clark et al. (1971). Noticeably, the main objective in Turnbull et al. (1974) was to assess the effect on survival of transplantation, assuming that the underlying patient population is homogeneous. However, in Brown et al. (1973) the influence of a number of concomitant variables was analyzed via pairwise correlations. However, this particular analysis of heart transplant survival data was first conducted by Crowley and Hu (1977). It is considered to be one of the first survival analyses pertaining to heart transplants to ever be conducted. The entire Stanford Heart Transplantation Program commenced in the fall of 1967. Since then, over 900 transplants on 850+ patients have been performed and closely analyzed and monitored as a part of this program. Of the 103 patients enrolled in the Crowley and Hu (1977) study, 69 received transplants and four patients were declared fit enough to receive any further transplants. Of those four, two were lost during the follow-up process and the other two died. Patients were enrolled in this particular heart transplant program between 1967 and 1974. Techniques examined include the Cox and Breslow methods, in which the simultaneous effect of several covariates which is denoted by V_n for the purpose of this paper. The hazard function is the conditional probability of an event occurring within a set interval divided by the width of the interval was also utilized, along with the exponential model and the survival function. Based on the exponential model, each patient was assumed to have a constant hazard variable that will be further explained later in the paper. Participants in this strictly observational study were permitted to be included only once it was determined that they were unlikely to respond to any type of other methods of therapy. Personal and familial consent were also required. We will be discussing the data obtained and notation used in the Crowley and Hu (1977) paper, several techniques and methods that were introduced, and the final results of the analysis. In particular we will discuss several one sample nonparametric methods for estimating the associated survivorship function related to this program, including but not limited to, the Kaplan Meier estimator, Life-table (Actuarial Estimator) etc.The major objective of this paper is to further examine the benefits of heart transplants based on the data presented in Crowley and Hu (1977). Next, we provide some useful background and history of the Stanford heart transplantation data analysis which will help the readers to find the reason as to why we are interested in this study. A quick search on `Google Scholar` survival analysis of Stanford heart transplant data resulted in 75200 references which justifies the fact that there is a need to re-analyse this data set (in various pertinent perspectives). A non-exhaustive list of such pertinent references can be cited as follows. Aitkin et al. (1983) re-analyzed the data by modeling survival time as a function of patient covariates and transplant status, and compare the results obtained using various parametric representations for survival time, including Weibull, lognormal, and piecewise exponential distributions. Pretransplant and post-transplant survival are considered separately and the effect of transplantation on survival is examined by comparison of the separate hazard functions. Storer and Crowley (1985) in a follow up paper provided some comments on the statistical problems regarding larger societal concerns with heart transplantation. Dag et al. (2017) discussed the prediction regarding $1-, 5-, 9-$ year patient's graft survival following a heart transplantation surgery via the adapation of analytical models that are based on four powerful classification algorithms (i.e., decision trees, artificial neural networks, support vector machine and logistic regression). Mancicni et al. (2021) discussed the rate of improvement regarding survival after heart transplantation despite increasing complexity. Moayedi et al. (2019) discussed and opined that overall survival doesn't differ between men and women after cardiac transplantation. Women who survive to heart transplantation appear to have lower risk features than male recipient but receive hearts from higher risk donors. In this paper, we did not include the gender factor which can be taken up in a future article.

The rest of the paper is organized as follows. In Section 2, we discuss briefly the idea on survival hazard rate model, and the survival function. In Section 3, we discuss in detail the survival data from Crowley and Hu (1977) with some useful notations and terminology used in this paper. In Section 4, we discuss the methodology used to analyze the heart transplant data. In Section 5, we provide two real-data application—one with the expanded analysis for the Stanford heart transplantation data and another for the survival of the AIDS data. Finally, some closing remarks are presented in Section 6.

2. Survival and Hazard Model

An alternative characterization of the distribution of T is given by the hazard function, or instantaneous rate of occurrence of the event, defined as

$$\text{lim}_{dt\to 0}\frac{P\left((t<T)<t+dt|\,(T>t)\right)}{dt}=\lambda(t) \tag{1}$$

The numerator of this expression is the conditional probability that the event will occur in the interval $[t, t+dt]$ given that it has not occurred before, and the denominator is the width of the interval. Dividing one by the other, the rate of event occurrence per unit of time is obtained. Taking the limit as the width of the interval goes down to zero, an instantaneous rate of occurrence is obtained.

The conditional probability in the numerator may be written as the ratio of the joint probability that T is in the interval [t,t+dt] and $T \geq t$ to the probability of the condition $T > t$. The former may be written as $f(t)dt$ for small dt, while the latter is $S(t)$ by definition. Dividing by dt and passing to the limit gives the useful result

$$\lambda(t)=\frac{f(t)}{S(t)}, \tag{2}$$

which is a standard definition of the hazard function. Equivalently, the rate of occurrence of the event at duration t equals the density of events at t, divided by the probability of surviving to that duration without experiencing the event. Note from the previous equation that f(t) is the derivative of S(t).

$$\lambda(t)=-\frac{d}{dt}\log S(t). \tag{3}$$

In this analysis, the integrated version of the hazard function is used:

$$\lambda(t) = \int_0^t \lambda(s)ds. \tag{4}$$

2.1. Survival Function

The survival function has the following standard form

$$S(t) = P\left(T > \ t\right) = \int_t^\infty f(x)dx, \tag{5}$$

which generally gives us the probability that participants would be alive just before the duration of (t) or it can also do the opposite and give us the probability that an event of interest has not occurred during the time frame of variable (t). For the purposes of this analysis, the curve is represented by:

$$F(t) = \exp\left\{-\lambda(t)\right\}. \tag{6}$$

3. Stanford Heart Data: Preliminaries

For this data provided by Crowly and Hu (1977), notations are used systematically throughout the paper. P represents the total number of participants in the study, in this case $P = 103$; where as H equals the number of patients who actually received a new heart, $H = 69$; C represents the total number of patients who were enrolled and did not go through with the transplant, in this case $C = 34$. Among the people who have received a transplant, let H represent the uncensored data and H^c represent the censored data. As for the people who did not receive the heart transplantation (C), let w be equal to the people in this sample that had uncensored data, with C^c people with censored data. For this sample $h = 45$ and $c = 30$.

The patients acceptance date is represented by T_1 while T_2 represents the patients last day being observed. With this, the survival time of patients without receiving a heart can be obtained as T_2-T_1. The day of transplant, for the patients that received a new heart, is represented by T_3, where $T_1 \leq T_3 \leq T_2$. Waiting time for patients between acceptance and transplant date can be found by $W = T_3{-}T_1$. There were also a number of covariates included in the analysis of each patient. They are represented by V_0,V_1,V_2, all the way up to V_8. Variables considered included transplant status, age at transplant, and three separate tissue types. T_5 in the data set represents the numerical measure of closeness of alignment between donor and recipient tissue. The survival variables $Y_1, Y_2, ...Y_m$ are right censored by fixed constants $t_1, t_2, ...t_m$, if the sample consists of ordered pairs (Z_i, δ_i), for $i = 1, 2, ...m)$, where for each i $\rightarrow$ $Z_i = min(Y_i, t_i)$. Z is measured by the minimum value between the time until an event happens, and the time until the end of the study. The data can be represented in a more simple way when using it in SAS and/or R programming.

In this sample, the data is right censored, meaning a patient has either died before the study ended through a different cause, or they withdrew from the study before the ending date, or the study has ended and the patient is still alive. In this sample, many patients didn't come back for follow up check up and their records were lost along the way of the study. The censored data is represented by

$$\begin{aligned} \delta_i &= 1 \quad \text{if} \quad Y_i \leq t_i \quad \text{(uncensored)} \\ &= 0 \quad \text{if} \quad Y_i \ > \ t_i \quad \text{(censored)}. \end{aligned} \tag{7}$$

4. Methodology

4.1. Actuarial Table

Survival analysis can be examined using different formulas when observing data sets based on sample size, truncation, and censored data. When the random variables $Y_1, Y_2 ... Y_m$ are independent, evenly distributed and censored, with survival function S and probability density function f, we use a method called the cohort life table method. The m patients alive are recorded of their survival time or time to censor, within a fixed disjoint interval.

$$\eta_j = [u_{j-1}, u_j), \ j = 1, 2, 3, ...k+1 \text{ such that } u_0 = 0 \text{ and } u_{k+1} = \infty \tag{8}$$

From here, the different notations for the j-th interval $\eta_j = [u_{j-1}, u_j)$ can be written:
N_j is number at risk in η_j,
D_j is the number of deaths or observed failures in η_j,
W_j is the number censored in η_j.

With these definitions, we know that $N_1 = n$, meaning the whole sample is at risk in the $j-1$-th interval. The life table chart shows the survival rate of a person in a given interval, which is also called the conditional probability structure(Smith 2002). These probabilities across η_j are represented by:

$$p_j = P\left(Y > u_j | Y > u_{j-1}\right) = \frac{S(u_j)}{S(u_{j-1})}, \tag{9}$$

where $S(u_j) = p_1 p_2 p_3 ... p_j$. To estimate p_j, a binomial estimator is used, is given by

$$\widehat{p_j} = 1 - \frac{\text{number of dying in } I_j}{\text{number with the potential to die in } I_j} \tag{10}$$

or equivalently,

$$\hat{p_j} = 1 - \frac{D_j}{N_j'}. \tag{11}$$

For the actuarial estimate of p_j, the effective number at risk, N_j' needs to be defined so that the number at risk in a given interval can be assumed to be censored uniformly and is given by

$$N_j' = N_j - \frac{1}{2} W_j. \tag{12}$$

With the data and Eqs. (8)-(12), estimates can be derived for the life table. To further complete the table, the standard error of each estimate should be found, by the use of Greenwood's (1926) formula. This formula is also used in the Kaplan Meier product limit estimator. The standard error of the life table estimate(s) are given by

$$S.E(\hat{S}(u_j) = \hat{S}(u_j)\sqrt{\sum_{i=1}^{j} \frac{\hat{q}_i}{\hat{p}_i N_i'}},\ j = 1, 2, 3, ...k + 1, \tag{13}$$

where $\hat{q}_i = 1 - \hat{p}_i$.

With this information, a life-table can be constructed and analyzed to see the survival rate of patients who received a new heart over a certain amount of time measured in days (given in the Appendix). Next, We will consider several different strategies.

4.2. Kaplan-Meier Estimator

The Kaplan-Meier product limit (PL) method is a special case of the lifetable technique. It estimates the probability of surviving longer than a given time t, i.e., $S(t)$. The estimate is the product of a series of estimated conditional probabilities. For example, the probability of surviving longer than k years is estimated as,

$$\widehat{P}\,(T > k) = \widehat{S}(k) = \prod_{i=1}^{m} p_i,$$

where p_m denotes the proportion of patients surviving the m-th year after they have survived $m - 1$ years. An important assumption of the Kaplan-Meier method is that the probability of a censored observation is independent of the actual survival time. The Kaplan-Meier method is commonly used by medical researchers and epidemiologists. Recently, it was applied to health economics by Fenn et al. (1995). The authors advocate the use of survival analysis, particularly the Kaplan-Meier method in the economic evaluation of cost-effectiveness of treatment where censored cost data are present. Censored data arise when the course of treatment extends beyond the end of the clinical trial period and when patients withdraw from the trial for reasons unconnected with the treatment under study. The Kaplan-Meier estimator is probably the most popular approach. It may be justified from several perspectives which are listed below:

- product limit estimator
- likelihood justification
- redistribute to the right estimator

This estimates the survival function S, that has right censored data. The difference between this estimator and the actuarial method, is that the endpoints are not fixed and are created by the spaces between the observed data points. Equation [7] is used for this method to define δ_i, to show which data points are censored or not. Intervals are created by dividing $\big(0, Z_{(n)}\big)$ into disjoint intervals

$$I_j = (Z_{(j-1)}, Z_{(j)}],\ j = 1, 2, 3, ...n \,\text{such that}\, Z_0 = 0. \tag{14}$$

The risk set at time u denoted by $R(u)$ indicates the set of subjects who are still alive and observed at time u^{-}, which is a time right before u. Next, revised/redefined notations can be written in terms of the PL-estimator. We begin our discussion by introducing the following notations and terminologies listed below:

- N_j the number at risk is number of elements in $R(Z_{(j)})$;
- D_j is the number of deaths observed in $Z_{(j)}$ (0 or 1);
- p_j is the conditional probability $P\left(\text{surviving through} \quad I_j|\text{alive at start of} \quad I_j\right)$.

Next, note that when u is fixed and there are ties within the sample, the right censored data is denoted by

$$\left(Z'_{(1)}, \delta'_{(1)}), (Z'_{(2)}, \delta'_{(2)}), ..., (Z'_{(k)}, \delta'_{(k)}\right), \tag{15}$$

consequently the Kaplan-Meier product-limit estimator of S will be

$$\hat{S}(u) = \prod_{j:Z'_{(j)} \leq u} \left(1 - \frac{D_j}{N_j}\right)^{\delta'_{(j)}}. \tag{16}$$

The standard error of the survival estimator can be found by using Greenwood(1926)'s formula . We describe it below. Notice that we can rewrite Eq. (16) as

$$\hat{S}(u) = \prod_{j:Z'_{(j)} \leq u} \left(1 - \widehat{\theta}_j\right)^{\delta'_{(j)}}, \tag{17}$$

where $\widehat{\theta}_j = \frac{D_j}{N_j}$. Next, since $\widehat{\theta}_j$ are basically binomial proportions, we may apply standard likelihood theory to establish that $\widehat{\theta}_j$ is approximately normal with mean θ_j and $var(\widehat{\theta}_j) \equiv \frac{\widehat{\theta}_j(1-\widehat{\theta}_j)}{N_j}$. Also, without loss of generality, we can say that for large samples, $\widehat{\theta}_j$ are independent. Consequently, we can estimate its variance using the `delta method`. Next, we consider the `delta method` in brief (and associated Greenwood's formula to compute the estimated variance (and subsequently the standard error) of the survival function) as follows:

- `delta method`: If X is normal with mean μ and variance σ^2, then for a one-to-one function g, $g(X)$ is approximately normally distributed with mean $g(\mu)$ and variance $[g'(\mu)]^2\sigma^2$.
- Next, instead of dealing with $\hat{S}(u)$ directly, we consider the logarithm of it, given by

$$\log\left(\hat{S}(u)\right) = \sum_{j:Z'_{(j)} \leq u} \log\left(1 - \widehat{\theta}_j\right).$$

Thus, by approximate independence of the $\widehat{\theta}_j$'s

$$\begin{aligned}
var\left(\log\left[\hat{S}(u)\right]\right) &= \sum_{j:Z'_{(j)}\leq u} var\left(\log\left(1-\widehat{\theta}_j\right)\right) \\
&= \sum_{j:Z'_{(j)}\leq u} \left(\frac{1}{1-\widehat{\theta}_j}\right)^2 var\left(\widehat{\theta}_j\right) \\
&= \sum_{j:Z'_{(j)}\leq u} \left(\frac{1}{1-\widehat{\theta}_j}\right)^2 \frac{\widehat{\theta}_j\left(1-\widehat{\theta}_j\right)}{N_j} \\
&= \sum_{j:Z'_{(j)}\leq u} \frac{\widehat{\theta}_j}{\left(1-\widehat{\theta}_j\right)N_j} \\
&= \sum_{j:Z'_{(j)}\leq u} \frac{D_j}{\left(N_j-D_j\right)N_j}.
\end{aligned}$$

Next, since $\hat{S}(u) = \exp\left[\log\left(\hat{S}(u)\right)\right]$, we can write

$$\begin{aligned}
var\left(\hat{S}(u)\right) &= \left[\hat{S}(u)\right]^2 var\left(\log\left[\hat{S}(u)\right]\right) \\
&= \left[\hat{S}(u)\right]^2 \left\{\sum_{j:Z'_{(j)}\leq u} \frac{D_j}{\left(N_j-D_j\right)N_j}\right\}.
\end{aligned}$$

Consequently, the estimated standard error of $\hat{S}(u)$ will be $\sqrt{\widehat{var\left(\hat{S}(u)\right)}}$. The Kaplan-Meier curves can be obtained either in SAS (`PROC LIFETEST`), or in `R` (`Survival` package) and also in `Mathematica` software as well. We have utilized `SAS` to analyze the data and subsequently all other associated computations in this paper.

4.3. Cox Regression

Cox regression, also known as proportional hazards regression, is a type of model assumes a parametric form for the effects of the explanatory variables, but allows an unspecified form for the underlying survivor function. In public health research, it is often of interest to know whether a certain personal characteristic is related to the occurrence of a certain health-related event. For example, are cigarette smoking, elevated cholesterol value, and family history of heart disease related to the development of cardiovascular disease? In this case, cigarette smoking, cholesterol value, and family history of heart disease are referred to as risk factors (or in a clinical setting, prognostic factors), or covariates. In most public health studies, data on many risk factors are collected and therefore the identification of the most significant risk factors becomes an important task. In addition to examining individually each variable's relationship to the length of disease-free time, survival, or remission, multivariate regression analysis is necessary to control for confounding factors. Cox's (9) regression model has been the most widely used method in survival data analysis regardless of whether the survival time is discrete or continuous and whether there is censoring. Though the Cox model is introduced initially in the framework of proportional hazards, the model easily can be extended to cases of non-proportional hazard functions. Some examples are models that include covariates that are time dependent (i.e. their values change over the follow-up period) and those that have multiple events for some individuals. How do we know if the assumption of proportional hazards holds? An easy way to check the hazards assumption is to plot the cumulative hazard functions.

It is used in the analysis of survival data to explain the effect of explanatory variables on hazard rates. This model can be written as:

$$h(t|z) = h_0(t) \exp\left(z^T \beta\right), \tag{18}$$

where $h(t)$ is the expected hazard time and $h_0(t)$ is the baseline function for the model, z is a $m \times 1$ vector of covariates such as treatment indicators, and β is a $m \times 1$ vector of regression coefficients. Obviously, $h\left(t|z=0\right) = h_0(t)$. Therefore, $h_0(t)$ is often called the baseline hazard function. It may be interpreted as the hazard function for the population of subjects with $z = 0$. The baseline hazard function in Eq. (15) can take any shape as a function of t. This is the nonparametric part of the model and z is the parametric part of the model. Consequently, Coxs' proportional hazards model is a semi parametric model.

This model can be used to calculate the likelihood ratio, which finds a goodness of fit between the null hypothesis and the alternative hypothesis. Cox regression examines the probability and effects between two or more groups within a sample. If the proportional hazards assumption does not hold, one remedy is to stratify the data into subgroups and apply the model for each stratum. Note that the hazards are non-proportional because the baseline hazards may be different between strata. A drawback of this approach is that the estimate of the effect of the stratifying variable can not be calculated. Non-proportional hazards also occur when some covariates are time dependent. For example, status of cigarette smoking may change during the study period and therefore is time dependent. The setup of the partial likelihood functions of the time-independent covariates is still applicable. In this case, the covariates are indexed by time. The risk for survival is allowed to vary with time. The model takes into account the recent information about the covariates, though it does not model the changes that occurred in the variable over the observation period. This is in contrast to the time-independent model, which only includes baseline information.

5. Real data application

5.1. Expanded Analysis of the Heart transplantation data

Throughout this analysis, it was found that waiting time was not a key variable and the efficacy of surgery was not strengthened by the variable of calendar time. Further analysis into these variables justified this point because of the strength of all of the other variables(age at transplant, previous surgery, etc.) when expanded analysis was conducted to include participants who did not receive a heart transplant. Instead, the mismatch score, in terms of HL-A typing, was a more valuable variable because it served as an indicator that typing did in fact reduce the chances of acute rejection and therefore increasing the odds of survival. The risk factor of age at acceptance could also serve as a significant variable in this study. In this analysis, it was found that age at acceptance only showed promise as a significant addition to the model because of the accompanying variable of age at transplant as a post-transplant risk and the correlation between the two. The variables are analyzed in two separate models.

The first model explores the results of the explanatory variables age at acceptance (dependent variable) and transplant status (censoring variable). With this model it was found that the risk of age at acceptance as a pre-transplant risk factor increases dramatically ($p = 0.0289$) Also, based on this model there is a slight decrease in transplant risk status with ($p = 0.8261$, therefore not significant)

The second model surveyed three separate explanatory variables; transplant status, transplant age and mismatch score. (Four participants did not have mismatch values and were therefore excluded) The variables transplant age and mismatch score are both time dependent. Results concluded that transplant age was statistically significant ($p = 0.0143$), meaning participants who received heart transplants at younger ages experienced longer lives. Also, mismatch score was found to have minimal effect on survival ($p = 0.1121$).

5.2. Survival data analysis of the AIDS data

Our example deals with the ribavirin clinical trial in AIDS patients given Table 1 of Lipsitz and Parzen (1996). This data set contains $N = 36$ eligible patients, each having a maximum of nj = 3 blood samples. The observed response for the kt-h blood sample from patient j is the minimum of number of days to virus positivity and the censoring time. Next, let us consider some useful preliminaries before considering a specific model for the data.

If T_{ik} be the failure time for the k-th member of the cluster i, with $i = 1, 2, \cdots, N$; $k = 1, 2, \cdots, n_i$. Assuming that T_{ik} follows a Cox proportional hazard model (see, Cox (1972), then the hazard function for T_{ik} at time $t > 0$, conditional on the $p \times 1$ covariate vector at time t, $Z_{ik}(t)$ will be

$$\lambda\left(t|Z_{ik}(t)\right) = \lambda_0(t) \exp\left(\beta Z_{ik}(t))\right),$$

where $\lambda_0(t)$ is an arbitrary hazard function and β is a $p \times 1$ vector of regression coefficients. Wei et al. (1989) had fitted separate treatment effects for each of weeks 4, 8, and 12; that is, for the k-th blood sample from patient j, they assumed the baseline hazard function for T_{ik} to be of the following form

$$\lambda\left(t|Z_{ik}(t)\right) = \lambda_{k0}(t) \exp\left(\beta_{k1}\delta\left[TRT_j = 2\right] + \beta_{k2}\delta\left[TRT_j = 3\right]\right),$$

where β_{k1} is the log-relative risk for low dose versus placebo at time $k1$, and β_{k2} the log-relative risk for high dose versus placebo at time $k2$, and δ is an indicator function. For the null hypothesis $H_0 : \beta_{11} = \beta_{12} = \beta_{13}$; the associated p-value = 0.2638, therefore we fail to reject the null hypothesis. A similar conclusion was independently obtained in Wei et al. (1989). Consequently, for simplicity, we consider the following model:

Table 1. Survival analysis of the AIDS data

Methods of estimation	$\widehat{\beta_1}$		$\widehat{\beta_2}$		$\widehat{S_t}$	
	Estimate	SE	Estimate	SE	Estimate	SE
Kaplan Meier	-0.8539	0.3112	-0.3926	0.2783	0.1385	0.3629
Bootstrap	-0.7543	0.1816	0.3598	-0.2938	0.1782	0.1563
Jackknife	-0.7647	0.3783	1.932	-0.3124	0.2346	0.1745
K-step repeated Jackknife	-0.7549	0.2541	-0.2842	0.1933	0.5127	0.1689

$$\lambda\left(t|Z_{ik}(t)\right) = \lambda_{k0}(t)\exp\left(\beta_1\delta\left[TRT_j=2\right] + \beta_2\delta\left[TRT_j=3\right]\right),$$

which assumes a common baseline hazard $\lambda_{k0}(t) = \lambda_0(t)$. Next, to estimate β_1, β_2 and the associated survival function, we consider the Kaplan Meier, the Bootstrap, the Jackknife and the K-repeated jackknifing estimation strategies. For pertinent details each of the last three estimation methods, see Adewara and Mbata (2014) and the references cited therein. From the above Table 1, one may observe the following:

- In comparison among the four different methods, the bootstrap procedure appears to be most efficient (in the sense of minimum value of the MSE for all the estimands. However, we can not make a general comment that among these methods adopted in survival analysis, Bootstrap will always perform uniformly better.
- The K-repeated jackknife procedure appears to be performing second best for this particular data.
- It must be mentioned that estimates under the Jackknife method has been independently obtained by Lipsitz and Parzen (1996) but only for the population quantities β_1, β_2. Our estimates are very close to what they have obtained.

6. Conclusion

In this paper we studied in more details the Stanford Heart Transplant survival data via several nonparametric and semi-parametric methods. Based on the analysis done in this paper, it can be concluded that after one year, the survival rate of patients that received heart transplants decreased significantly. Based on the Kaplan-Meier test, the results were equally as unimpressive, as the average lifespan for individuals receiving a heart transplant was approximately a year and a half. Both the life-table and product-limit estimator graphs agreed with each other by showing survival rates that fluctuated between 0.2 and 0.4 over 1000 days, after which the survival rate begins to stabilize and the hazard rate generally decreases. These results can be attributed to a number of circumstances which were stated earlier. Extenuating circumstances do have to be taken into consideration, such as whether the patient had received any type of transplant prior to this study. It was also found that younger patients experienced longer survival rates through transplantation once a suitable heart was found for them. To conclude, the transplant program that begin in 1967 had many positive impacts on the survival rates of patients enrolled in the study. Even though the rates weren't long term, the results were positive for younger patients. It is safe to say that the survey made in this article are far from complete. Survival analysis with incomplete patient data will involve a more complex approach and associated methodologies will involve a careful analysis under both the parametric as well as nonparametric set-up. We plan to take up this matter in a separate article. In addition, in this paper, as a separate study, we have also performed a survival analysis due to Lipsitz and Parzen (1996) via various estimation methods. It is observed that among the four different methods adopted, the bootstrap appears to be most efficient. However, we can not recommend that the bootstrap method will be uniformly better in other situations.

Acknowledgement

The authors would like to thank the anonymous reviewers and the Editors on an earlier version of the manuscript which has significantly improved the current version.

References

[1] Xue, Y., Wang, H., Yan, J. and Schifano, E.D., 2020. An online updating approach for testing the proportional hazards assumption with streams of survival data. Biometrics, 76(1), pp.171-182.

[2] Dzinza, R. and Ngwira, A., 2022. Comparing parametric and Cox regression models using HIV/AIDS survival data from a retrospective study in Ntcheu district in Malawi. Journal of Public Health Research, 11(3), p.22799036221125328.

[3] Ke, S., Fang, Q., Lan, J., Qiao, N., Zhang, X., Xie, C., & Fan, Y., 2023. Survival times of HIV/AIDS in different AIDS Diagnostic and Treatment Guidelines from 2006 to 2020 in Liuzhou, China. BMC Public Health, 23(1), 1745.

[4] Clark, D.A., Stinson, E.B., Griepp, R.B., Schroeder, J.S., Shumway, N.E., and Harrison, D.C., 1971. Cardiac Transplantation in Man. VI. Prognosis of Patients Selected for Cardiac Transplantation. Annals of Internal Medicine, 75, 15-21.

[5] Turnbull, B.W., Brown, B.W., Jr., and Hu, M., 1974. Survivorship Analysis of Heart Transplantation Data. Journal of the American Statistical Association, 69, 74-80.

[6] Mantel, N., and Byar, D.P., 1974. Evaluation of Response-Time Data Involving Transient States: An Illustration Using Heart-Transplant Data. Journal of the American Statistical Association, 69, 81-86.

[7] Brown, B.W., Jr., Hollander, M., and Korwar, R.M., 1973. Nonparametric Tests of Independence for Censored data with Applications to Heart Transplant Studies. Paper presented at the Florida State University Conference on Reliability and Biometry.

[8] Crowley, J., and Hu, M., 1977. Covariance Analysis of Heart Transplant Survival Data. Journal of the American Statistical Association, 72, 27-36.

[9] Aitkin, M., Laird, N., & Francis, B., 1983. A reanalysis of the Stanford heart transplant data. Journal of the American Statistical Association, 78(382), 264-274.

[10] Crowley, J. and Storer, B.E., 1983. A reanalysis of the Stanford heart transplant data: comment. Journal of the American Statistical Association, 78(382), pp.277-281.

[11] Dag, A., Oztekin, A., Yucel, A., Bulur, S. and Megahed, F.M., 2017. Predicting heart transplantation outcomes through data analytics. Decision Support Systems, 94, pp.42-52.

[12] Mancini, D., Gibson, G. T., & Rangasamy, S., 2021. Improving survival after heart transplantation despite increasing complexity. European Heart Journal, 42(48), 4944-4946.

[13] Moayedi, Y., Fan, C. P. S., Cherikh, W. S., Stehlik, J., Teuteberg, J. J., Ross, H. J., & Khush, K. K., 2019. Survival outcomes after heart transplantation: does recipient sex matter?. Circulation: Heart Failure, 12(10), e006218.

[14] Lipsitz, S. R., & Parzen, M., 1996. A jackknife estimator of variance for Cox regression for correlated survival data. Biometrics, 291–298.

[15] Smith, J. P., 2002. Analysis of Failure and Survival Data.Chapman & Hall/CRC.

[16] Greenwood, M., 1926. The Natural Duration of Cancer. Reports of Public Health and Related Subjects, 33, 1-26. HMSO, London.

[17] Fenn, P., McGuire, A., Phillips, V., Backhouse, M., & Jones, D., 1995. The analysis of censored treatment cost data in economic evaluation. Medical Care, 33(8), 851–863.

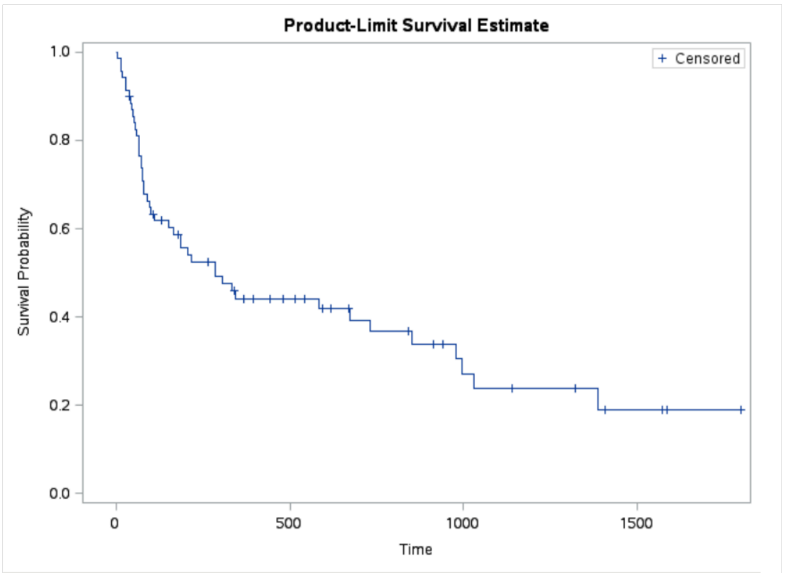
[18] Kaplan, E.L., and Meier, P., 1958. Non parametric estimation from incomplete observations. Journal of the American Statistical Association, 53, 448–457.

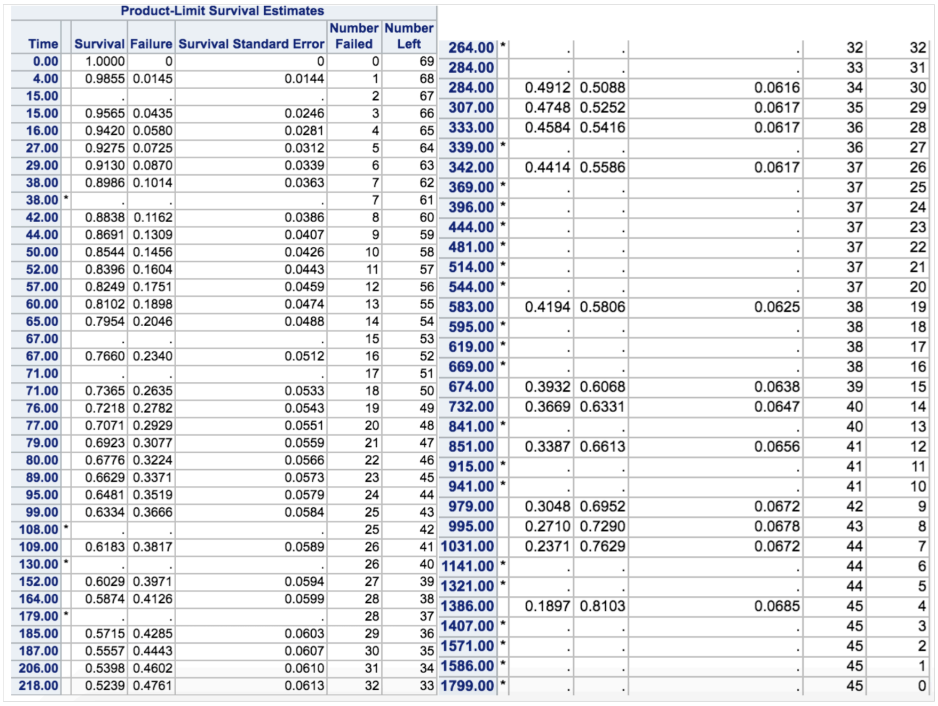
[19] Cox, D. R., 1972. Regression models and life tables (with discussion). Journal of the Royal Statistical Society, Series B, 34, 187–220.

[20] Wei, J., Hou, J., Su, B., Jiang, T., Guo, C., Wang, W., Zhang, Y., Chang, B., Wu, H. and Zhang, T. (2020). The prevalence of Frascati-criteria-based HIV-associated neurocognitive disorder (HAND) in HIV-infected adults: a systematic review and meta-analysis. Frontiers in neurology, 11, p.581346

[21] Adewara, J.A., and Mbata, U.A. (2014). Survival Estimation Using Bootstrap, Jackknife and K-Repeated Jackknife Methods. Journal of Modern Applied Statistical Methods: Vol. 13 : Iss. 2 , Article 15. DOI: 10.22237/jmasm/1414815240.

Appendix

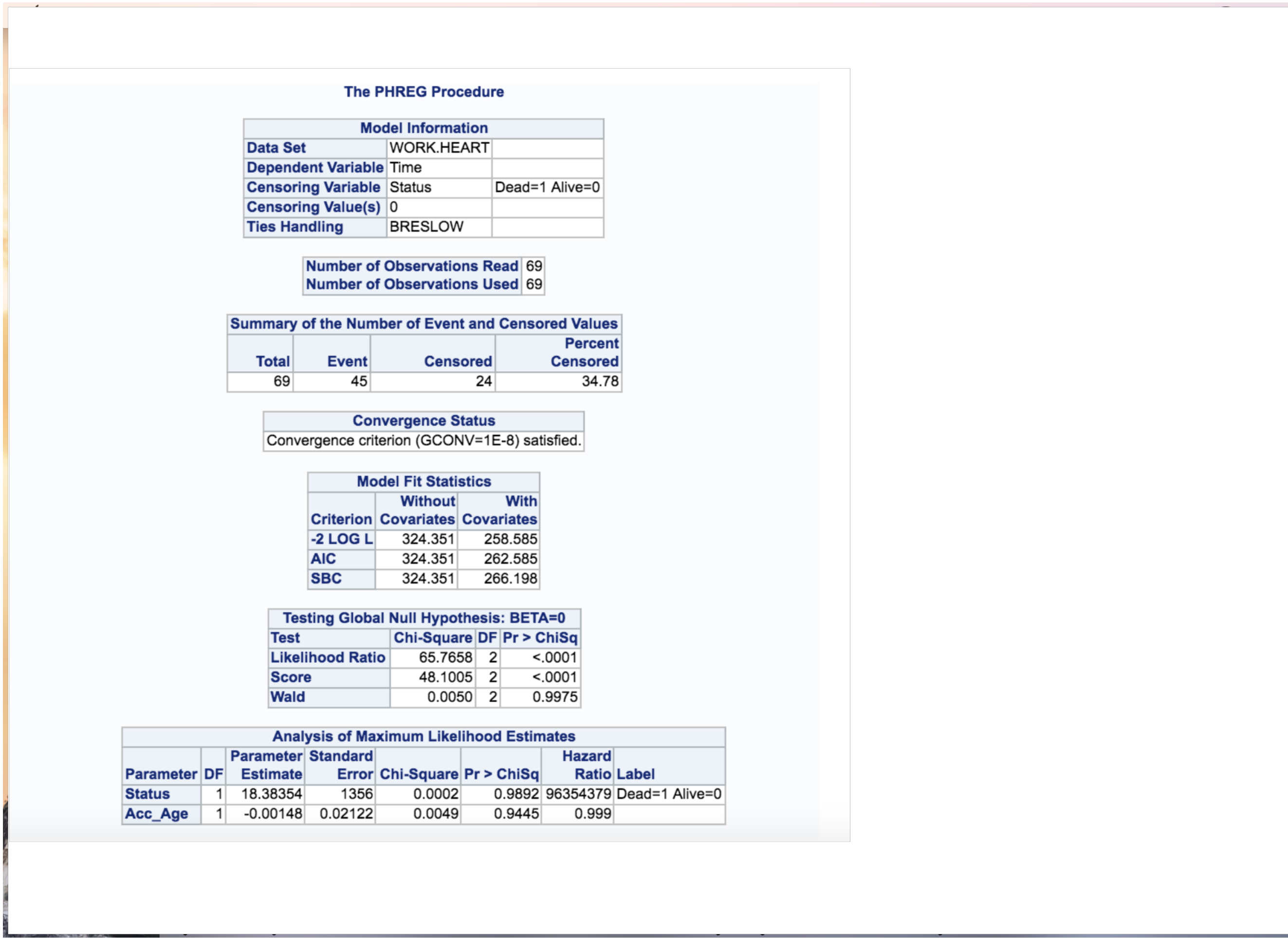




Product-Limit Survival Estimates

Time		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.00		1.0000	0	0	0	69
4.00		0.9855	0.0145	0.0144	1	68
15.00		.	.	.	2	67
15.00		0.9565	0.0435	0.0246	3	66
16.00		0.9420	0.0580	0.0281	4	65
27.00		0.9275	0.0725	0.0312	5	64
29.00		0.9130	0.0870	0.0339	6	63
38.00		0.8986	0.1014	0.0363	7	62
38.00	*	.	.	.	7	61
42.00		0.8838	0.1162	0.0386	8	60
44.00		0.8691	0.1309	0.0407	9	59
50.00		0.8544	0.1456	0.0426	10	58
52.00		0.8396	0.1604	0.0443	11	57
57.00		0.8249	0.1751	0.0459	12	56
60.00		0.8102	0.1898	0.0474	13	55
65.00		0.7954	0.2046	0.0488	14	54
67.00		.	.	.	15	53
67.00		0.7660	0.2340	0.0512	16	52
71.00		.	.	.	17	51
71.00		0.7365	0.2635	0.0533	18	50
76.00		0.7218	0.2782	0.0543	19	49
77.00		0.7071	0.2929	0.0551	20	48
79.00		0.6923	0.3077	0.0559	21	47
80.00		0.6776	0.3224	0.0566	22	46
89.00		0.6629	0.3371	0.0573	23	45
95.00		0.6481	0.3519	0.0579	24	44
99.00		0.6334	0.3666	0.0584	25	43
108.00	*	.	.	.	25	42
109.00		0.6183	0.3817	0.0589	26	41
130.00	*	.	.	.	26	40
152.00		0.6029	0.3971	0.0594	27	39
164.00		0.5874	0.4126	0.0599	28	38
179.00	*	.	.	.	28	37
185.00		0.5715	0.4285	0.0603	29	36
187.00		0.5557	0.4443	0.0607	30	35
206.00		0.5398	0.4602	0.0610	31	34
218.00		0.5239	0.4761	0.0613	32	33
264.00	*	.	.	.	32	32
284.00		.	.	.	33	31
284.00		0.4912	0.5088	0.0616	34	30
307.00		0.4748	0.5252	0.0617	35	29
333.00		0.4584	0.5416	0.0617	36	28
339.00	*	.	.	.	36	27
342.00		0.4414	0.5586	0.0617	37	26
369.00	*	.	.	.	37	25
396.00	*	.	.	.	37	24
444.00	*	.	.	.	37	23
481.00	*	.	.	.	37	22
514.00	*	.	.	.	37	21
544.00	*	.	.	.	37	20
583.00		0.4194	0.5806	0.0625	38	19
595.00	*	.	.	.	38	18
619.00	*	.	.	.	38	17
669.00	*	.	.	.	38	16
674.00		0.3932	0.6068	0.0638	39	15
732.00		0.3669	0.6331	0.0647	40	14
841.00	*	.	.	.	40	13
851.00		0.3387	0.6613	0.0656	41	12
915.00	*	.	.	.	41	11
941.00	*	.	.	.	41	10
979.00		0.3048	0.6952	0.0672	42	9
995.00		0.2710	0.7290	0.0678	43	8
1031.00		0.2371	0.7629	0.0672	44	7
1141.00	*	.	.	.	44	6
1321.00	*	.	.	.	44	5
1386.00		0.1897	0.8103	0.0685	45	4
1407.00	*	.	.	.	45	3
1571.00	*	.	.	.	45	2
1586.00	*	.	.	.	45	1
1799.00	*	.	.	.	45	0

Figure 1. Brief summary of the SAS output related to product limit survival estimates for the Stanford heart transplantation data along wit the Kaplan-Meier estimate.



The PHREG Procedure

Model Information		
Data Set	WORK.HEART	
Dependent Variable	Time	
Censoring Variable	Status	Dead=1 Alive=0
Censoring Value(s)	0	
Ties Handling	BRESLOW	

Number of Observations Read	69
Number of Observations Used	69

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
69	45	24	34.78

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	324.351	258.585
AIC	324.351	262.585
SBC	324.351	266.198

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	65.7658	2	<.0001
Score	48.1005	2	<.0001
Wald	0.0050	2	0.9975

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
Status	1	18.38354	1356	0.0002	0.9892	96354379	Dead=1 Alive=0
Acc_Age	1	-0.00148	0.02122	0.0049	0.9445	0.999	

Figure 2. Brief summary of the SAS output for the Cox-Ph regression.