Kaplan-Meier and Cox Proportional Hazards Survival Regression Analysis Illustrated with Immune Deficiency Cohort of Patients after Allogeneic Bone Marrow Transplantation

Ayman A. Mostafa
Department of Mathematics and Statistics, AI
Imam Mohammad Ibn Saud Islamic
University; IMSIU
Email : aasaad@imamu.edu.sa

ABSTRACT

The number of allogeneic and autologous bone marrow transplants continues to grow worldwide. Nevertheless, Bone marrow transplantation (BMT) has become standard therapy for many patients with leukemia, breast cancer, severe immune deficiency, and other diseases; there are risk factors after transplantations. We have obtained dataset of 65,535 patients have been treated with BMT to study such risk factors. The dataset recorded from 1994 to 2005 in the Statistical Center for International Blood and Marrow Transplant Research (CIBMTR). Since the dataset is large, the primary objective of this paper is to introduce the dataset, which will be analyzed in a series of papers using the finest techniques in survival models. The secondary objective of the series will be review the common methods in survival data analysis applied in such useful dataset and then develop some techniques to be suited in such data. In this paper, we discuss the analysis of significant risk factors of severe immune disease after allogeneic BMTs. First, the survival curves were calculated by the Kaplan-Meier method and then the well known Cox model is discussing, that is because it is found that, that model is suited to be applied in immune deficiency in such subset data. When fitting a Cox model, it is vital to assess the assumption of proportionality.

Keywords - Censored Survival data analysis; Cox model; risk; transplants.

1. INTRODUCTION

Survival analysis is a collection of statistical procedures for data analysis where the outcome variable of interest is time until an event occurs. Statistical models and methods for such data and other time-to-event data are extensively used in many fields, including the biomedical sciences, engineering, the environmental sciences, economics, actuarial sciences, management, and social sciences. Survival data is a term used for describing data that measure the time to some event. From a mathematical perspective it is irrelevant whether one is studying time until death or time to any other event and so the term has come to be applied to methods for analyzing "time to event data," we are always interested in the time between two events. For instance one might be studying the age of death (the time from birth until death), survival of immune deficiency patients (the time from diagnosis until death). In biomedical applications, especially in clinical trials, important issue arise when studying "time to event data," that is some individuals are still alive at the end of the study or analysis so the event of interest, namely death, has not occurred. Therefore, survival data analysis is more complicated than the data analysis of other measurements because one often has only partial information regarding the survival time for some individuals. Such partial information "missingness" data is known as "censoring". Duration is censored "if a respondent withdraws from a study for reasons other than the terminating event," or "if a subject does not undergo the event before the end of the observation period." Thus, we know only that they have yet to undergo the event at the time observation ceases. This is known as "right censoring," in that the observed incomplete duration is necessarily less than the unknown full duration until the event of interest. The most common form of partial information arises when a study is stopped before all participants have died. For example, if we assume that a patient has been diagnosed with immune deficiency disorders and that patient decided to enter a clinical trial with Bone Marrow Transplantation (BMT). At that point we might know that patient survived for at least 3.7 years, but have no idea whether she will die a month later or 5 years later. The observation on that patient is said to be (right) censored at 3.7 years.

The goal of a survival analysis might be to describe the survival distribution for a group of individuals. More often clinical trials, epidemiological studies, engineering, finance and social sciences might be interested in factors or "covariates," that influence survival distribution. In such instances, the aim of survival analysis is to estimate the effect of the factors on survival times [1].

In survival analysis, the additive, multiplicative and the class of general additive-multiplicative hazard models provide the three principle frameworks for studying the association between that covariates (X) and the survival time (T). The hazard function, $h(t)$, also called, "the risk," or "intensity," function, of non-negative random variable T, associated with a P-vector of covariates X is

defined as $h(t|X)=f(t|X)/(1-F(t|X))$, where $f(t|X)$ and $F(t|X)$ are the density function and the distribution function, of the random variable T conditioned on the vector of covariates X, respectively. The function $S(t|X)=1-F(t|X)$ is called the survival function. The three functions; $S(t|X)$, $f(t|X)$ and $h(t|X)$ are mathematically equivalent, i.e. if one of them is given, the other two can be derived [2-3].

In studying mortality in a healthy cohort, it is usual to compare their survival (or equivalent hazard) curves to that of the general population. Some form of survival analysis will be required in any cohort study and many of the more complicated designs used in modern epidemiology require quite sophisticated analytical techniques [4]. Several statistical methods have been proposed for modelling survival time data. Under the multiplicative hazard model, or Cox proportional hazards (PH) model [5, 6], proposed modelling the conditional hazard as

$$h(t|X(t))=h_0(t)\times\exp\{\beta'(t)X(t)\}; \quad (1)$$

Under the additive hazard model [7, 8], the hazard function takes the form:

$$h(t|X(t))=h_0(t)+\beta'(t)X(t); \quad (2)$$

The additive regression is an alternative or "supplement," to the Cox model. It results in plots that are informative regarding the effect of covariates on survival, and under the class of general additive-multiplicative models [9]; the hazard function takes the form:

$$h(t|Z)=h_0(t)\times\kappa\{\beta_0'X(t)\}+g\{\alpha_0'(t)R(t)\}. \quad (3)$$

where $h_0(t)$ is "a baseline hazard function," which is common to all observations, $Z=(R',X')'$ is a P-vector of covariates and $(\beta_0',\alpha_0')'=\theta_0$ (say) is a P-vector of unknown regression parameters. The covariate Z can be time-dependent, and $\kappa\{.\}$ and $g\{.\}$ are known link functions. It is obvious that (3) encompasses both models and (1) and (2). Some of common examples of the link functions κ are: $\kappa=\exp(x)$ and $\kappa(x)=1+x$ and those of g are: $g(x)=x$ and $g(x)=\exp(x)$. For $g(x)=\exp(x)$, it is sensible to let $R(t)=1$.

As it is stated above that "the random variable T is often subject to right censoring because certain patients may still be surviving at the end of the study period," the analysis of survival experiments is complicated by "issues of censoring."

In this paper, we focused on "right censored data," since this type of data is most frequently encountered in applications is right censored data. Furthermore, due to the complexity of biological process, it is desirable not to parameterize $h_0(t)$ and therefore, only semi-parametric inference has been used for models (1), (2) and consequently the general model in (3).

The two most distinctive features of the Cox model are proportionality of the hazard rates for different observations and estimation of the regression coefficients based on the partial likelihood. In order to draw semi-parametric inference for model(1), Cox[6] introduced the partial likelihood approach to estimate the regression vector parameter β. Since the proportional hazards assumptions are often violated, the need for more flexible model motivate the introduction of models (2) and (3).

We have obtained a large cohort of dataset comprised of 65, 535 patients have been treated with BMT to study risk factors after different types of transplantations. The dataset recorded in the Statistical Center for International Blood and Marrow Transplant Research (CIBMTR) from 1994 to 2005. Such dataset were collected in the CIBMTR located at the Medical College of Wisconsin. The Center for International Blood and Marrow Transplant Research, or "CIBMTR," collaborates with the global scientific community to advance hematopoietic cell transplantation and cellular therapy research worldwide. A combined research program of the National Marrow Donor Program and the Medical College of Wisconsin, CIBMTR facilitates critical research that has led to increased survival and an enriched quality of life for thousands of patients. As an example, in Saudi Arabia, there is King Faisal Specialist Hosp. & Research Center (NMDP/CIBMTR Research).

Since the dataset is large and complicated due to censoring observations and many diseases which have been treated with both types of well known allogeneic and autologous transplants, the dataset be analyzed in a series of papers using the finest techniques in survival models. In this first article of the series of analyzing the dataset, we will present the advantages from applying model (1) as the basic technique in survival models, including how to produce and interpret survival curves, and how to quantify and test survival differences between two or more groups of patients. Future papers in the series will cover the both models(2), (3) and other methods such as the series of articles introduced by Bradburn et al. [10-13] which contain an excellent survey in four series of papers. Our series of articles will depend on a real dataset from different types of Bone Marrow Transplantations (BMTs); have been collected for this proposal. More detailed accounts of these methods can be found in books written specifically about survival analysis, for example, [14, 15]. We start our methodological considerations with the traditional Cox PH model (1) that specifies the impact of explanatory variables (which factors of X influence) on continuous survival times in a regression to model the hazard (or survival) of patients. the analysis, which will be studied in these papers have not been analyzed reviewed or approved by the CIBMTR, or in any other place.

Therefore, the main primary objective of this paper is to introduce the dataset and then construct the survival curves calculated by the Kaplan-Meier method and then the well known Cox model is discussing, that is because it

is found that, that model is suited to be applied in immune deficiency disease in such subset data. When fitting a Cox model, it is vital to assess the assumption of proportionality.

1.1 Materials and Methods

The data presented here were obtained from the Statistical Center of the Center for International Blood and Marrow Transplant Research (CIBMTR) located at the Medical College of Wisconsin. The analysis has not been reviewed or approved by the CIBMTR. The CIBMTR is comprised of clinical and basic scientists who share data on their transplant patients with CIBMTR Statistical Center. The CIBMTR is partially supported by a Grant, U24-CA76518 from the National Institutes of Health, and by the Health Resources and Services Administration. Between 1994 and 2005 the information is comprised of a large cohort database (65, 535 recipients' patients) of different types of cancer, who has been who received different types from BMTs. BMT is a special therapy for patients with certain cancers or other diseases. In this procedure, a patient is given a high dose of radiation and/or chemotherapy kills the stem cells in the bone marrow which produce white blood cells. New stem cells, taken from a matched donor, are transplanted into the patient to replace their destroyed immune system. Type of "event," in such dataset is the time between the transplantation to the time of death (time from randomization to death in a cancer clinical trial). BMTs are accompanied by serious and life-threatening risks. Furthermore, they are not always an absolute assurance of a cure for the underlying ailment; a disease may recur in the future. There are common diseases that arise after BMT. One of the common diseases that arise after a allogeneic bone marrow transplants is called the graft versus host disease (GVHD), and is due to the rejection of the new tissue by the recipient's immune system. The two types of GVHD are called acute GVHD (aGVHD) and chronic GVHD (cGVHD). As shown in Figure 1, there are different types of such diseases, has been treated with BMTs, in the collected sample. In these series of papers, we investigate the risks of some such diseases on the hazard of death after BMTs under models stated in (1), (2) and (3). Furthermore; many other methods in survival data analysis will be presented using such dataset in Figure 1. In the next series of papers, we briefly review and apply the most important methods in the survival data analysis and apply them in the first disease of our analysis such as Logistic regression [16-18], which is a regression modeling for 0/1 data and how the functional form of the logistic model, how to interpret model coefficients, "odds", and "odds ratio," as well as "risk ratio," and how it differs from odds ratio. The emphasis is on logistic model construction, interpretation, and goodness-of-fit, will be applied. The most important goals of survival data analysis will be applied and clear in these series of articles, because the dataset is very rich with ambiguous features. The improvement in statistical computing and wide accessibility of personal computers led to the rapid development and popularity of nonparametric over parametric procedures. The former required less stringent conditions. But, if the assumptions for parametric methods hold, the resulting estimates have smaller standard errors and are easier to interpret. Nonparametric techniques include the Kaplan-Meier method for estimating the survival function and the Cox proportional hazards model to identify risk factors and to obtain adjusted risk ratios. In cases where the assumption of proportional hazards is not tenable, the data can be stratified and a model fitted with different baseline functions in each stratum. Parametric modeling such as the accelerated failure time model also may be used. Hazard functions for the exponential, Weibull, gamma, Gompertz, lognormal, and log-logistic distributions are described. Since the Kaplan-Meier and the Cox proportional-hazards regression model has achieved widespread use in the analysis of time-to-event data with censoring and covariates, we start to review and apply in this article.

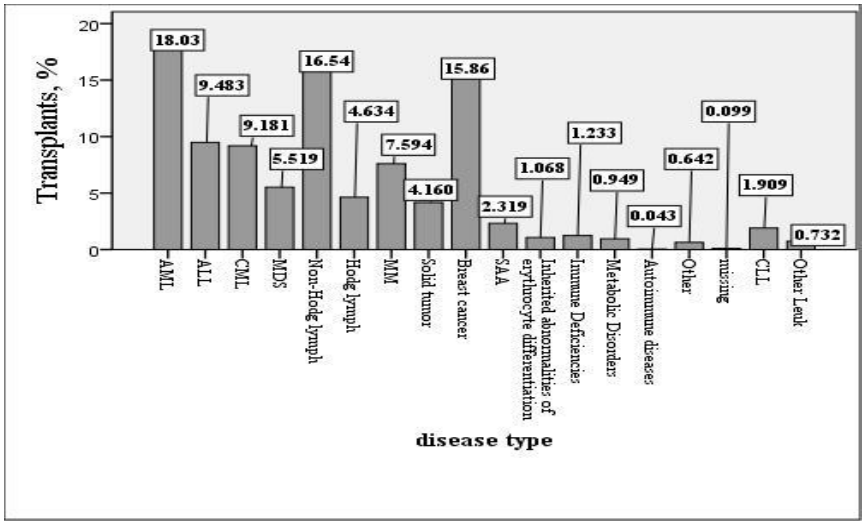

Fig.1 Dataset of 65,535 patients from different types of cancer after BMTs between 1994 and 2005

1.2 Motivate the Cox Model with Immune Deficiency Dataset

Our work in this paper, is motivated by derive model (1) with a real data example as follows. Bone marrow transplantations (BMT) are used as one of the treatments for immune deficiencies patients. BMT using stem cells obtained from a family-related, HLA identical donor (RID) is the optimal treatment for patients with severe combined immune deficiency[19]. The immune deficiencies dataset for our study models (1) collected from the Statistical **C**enter for **I**nternational **B**lood and **M**arrow **T**ransplant **R**esearch (CIBMTR) located at the Medical College of Wisconsin is comprised of a cohort of 808 patients with immune deficiency treated with BMT. The Center for International Blood & Marrow Transplant Research, or CIBMTR, collaborates with the global scientific community to advance hematopoietic cell transplantation and cellular therapy research worldwide. A combined research program of the National Marrow Donor Program and the Medical College of Wisconsin, CIBMTR facilitates critical research that has led to increased survival and an enriched quality of life for thousands of patients.

In the following paper, survival data of 808 (which represent about 1.233% of our sample, as shown in Figure 1) patients with immune deficiency diseases undergoing a BMTs between 1994 and 2005 were collected and will analyzed by univariate analysis and Cox proportional hazard regression model to reveal the risk factors after BMTs. All data were recorded using a standard data form of SPSS (Chicago, IL, United States; version 20.0) and the analyzed of data were created using the R statistical language and environment (R Development Core Team, 2011, www.R-project.org; version 2.14.0). Survival curves were calculated by the Kaplan-Meier method. The log rank test was used to assess differences in survival. Finally; univariate hazard ratios and significant and independent predictors of disease-specific survival and were identified by Cox proportional hazard analysis. The stepwise procedure was set to a threshold of 0.0001.

The collected data accrued the immune deficiencies patients were collected and followed from 1994 until 2005. This collected data will demonstrate the use of models (1) and (2). Table 1 summarizes the response variable t (recorded in months) and several covariates for immune deficiencies patients followed after BMT study. The response variable is the time from transplantation to the last follow-up (in months), which may be right censored. The covariates are the four-level covariate graft type, which is the Hematopoietic stem-cell source ($X_1; X_{1(1)}$:1 = Bone marrow (BM), $X_{1(2)}$: 2 = Peripheral blood stem cell (PB), $X_{1(3)}$:3 = Cord blood, $X_{1(4)}$: 4 = BM + PB), age (X_2), eleven-levels covariate year of transplant (X_3; $X_{3(1)}$:1994, $X_{3(2)}$:1995,…, $X_{3(12)}$:2005), two-level covariate sex ($X_4; X_{4(1)}$:1 = Male, $X_{4(2)}$: 2 = Female), three-level covariate donor type ($X_5; X_{5(1)}$:1 = human leukocyte antigen identical siblings "HLA-id sib," $X_{5(2)}$: 2 = Other relative, $X_{5(3)}$: 3 = Family mismatch, or unrelated donor "URD"). Covariates; X_1, X_3, X_4, and X_5 are all categorical covariates, whereas X_2 is continuous covariate. Continuous variable is summarized as mean, median and range. Categorical variables are described by count and relative frequency of each category (%). X_4 and X_5 have missing variable. Since the missing data fractions for these two covariates are 10% only, we will exclude them from our data analysis (hence; we have $n = 800$ patients). The interrelationships between the outcome (censored and non-censored) and variable over time can lead to bias unless the relationships are well understood. We can see in Figure 2 that the censored and non-censored observations are mixed about 3 to 1 ratio till 45 years. The shape of the plot is controlled by the strong association in these data between age at transplantation and survival time, the fact that survival time is skewed to the right and constraint that patients can be followed for at a final follow-up of 152.83 months (12.7 years). Further longer-term studies are necessary to better evaluate these outcomes. The cloud of points in Figure 2 is densest with the plot truncated ate the maximum length of follow-up. The scatterplot smoothing of a plot such as the one in Figure 2 could be difficult to interpret since censored and non-censored (dead patients) times have been treated equally. That is, the presence of the censored observations in the smoothing process could, make it difficult to visualize the systematic component of the survival times. The simplest statistical distribution with this characteristic is exponential distribution, which is a special case from model (1). After a careful examination of the scatterplot, we arrived at the conclusion that the exponential regression, such as Cox model, might be good setting point to model these data.

Table 1 Summary for dataset of 808 patients for immune deficiencies after BMT*

Completely Observed Variables			Missing Covariates		
X_1 Frequency (%)	BM	511 (63.2)	X_4	Male	604 (74.8)
	PB	104 (12.9)		Female	203 (25.1)
	Cord blood	190 (23.5)	Frequency (%)	Missing	1 (0.1)
X_2 (years)	Mean	3.5	X_5	Autologous	0 (0)
	Median	1.1	Frequency	Other relative	217

	Range	0.022−56.2	(%)		(26.9)
	Std. dev.	5.9		Unrelated	368 (45.5)
	1994	46 (5.7)		Other	2 (0.2)
	1995	60 (7.4)		Missing	7 (0.9)
	1996	70 (8.7)			
	1997	75 (9.3)			
	1998	86 (10.6)			
X_3	1999	64 (7.9)			
	2000	48 (5.9)			
	2001	62 (7.7)			
	2002	64 (7.9)			
	2003	78 (9.7)			
	2004	61 (7.5)			
	2005	94 (11.6)			
t Frequency (%)	Censored**	560 (69.3)			
	dead (non-censored)	248 (30.7)			

* Source: CIBMTR

**Observations that survived to a certain point in time before dropping out from the study for a reason unrelated to the primary outcome of interest (e.g., insufficient time for an observation to meet the event).

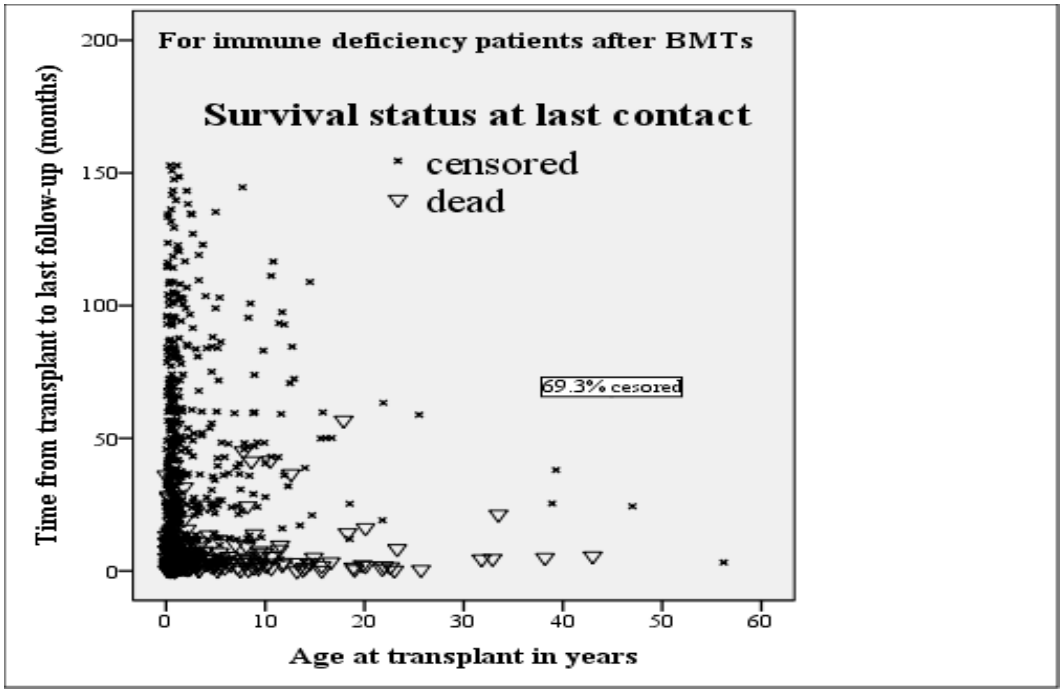

Fig.2 Scatterplot of survival time versus age for severe immune deficiency patients after allogeneic BMTs

1.3 Kaplan-Meier estimates of survival functions

We first study the variables of interest in Table 1 using the Product-limit (PL) curve that estimating the survival rates; this is also called the Kaplan-Meier curve [20]. The Kaplan-Meier can use the information that lies in censored observations efficiently. The result of a Kaplan-Meier analysis is a survival curve like Figure 3. This plot shows time following transplantation on the horizontal axis, and the probability to survive on the vertical axis. The time crosses 66.1% survival at time of 50 months post-transplantations and the time crosses 65% survival at time 67.5 months. Similarly, we can obtain estimates of the 75th percentile of survival time as shown in Figure 3 (8.45 moths). Kaplan-Meier curves can also be used to compare between survival groups, as demonstrated in Figure 4. Basically there are two nonparametric tests available to compare between survival rates in the different groups. The first is the log rank test and the second nonparametric test is the generalized Wilcoxon. A kind of mixture of these two tests is the Tarone-Ware test.

For example, Figure 4 is the output of the test for the null hypothesis that is no difference in the distribution of survival times between males and females groups. We see that survival in the high risk group (females' group) is much worse than in the low risk group (males' group). The three tests (log rank, generalized Wilcoxon and Tarone-Ware test) confirm the importance of the gender in the immune deficiency for survival after transplantations. To visualize the difference in the disease progression rates in the three groups for the covariate X_5, using PL curve Figure 5 displays the graph for the cumulative hazard function for the donor type. For this covariate, the three tests of survival distributions for the different levels of X_5 are high significant (p-value $\ll$ 0.0001). We see a constant gab between the three curves, i.e. the patients with the other relative donor type have constantly higher risk to die following transplantations than patients with the other two types of the source of stem cells.

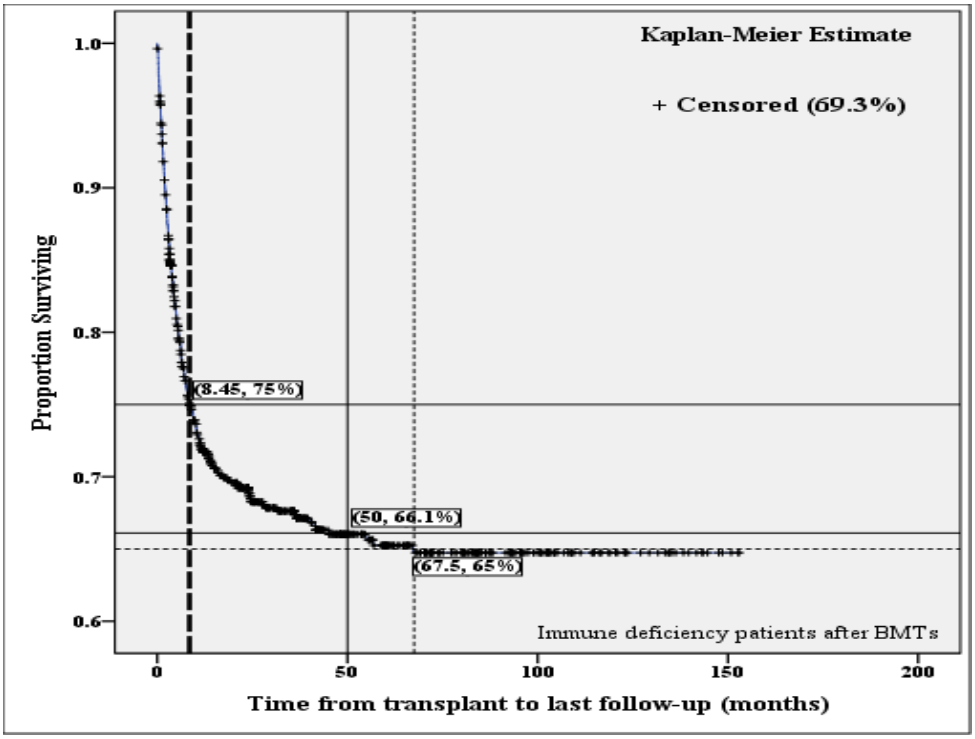

Fig.3 Overall survival (Kaplan-Meier estimate) plots for the immune deficiency following transplantations

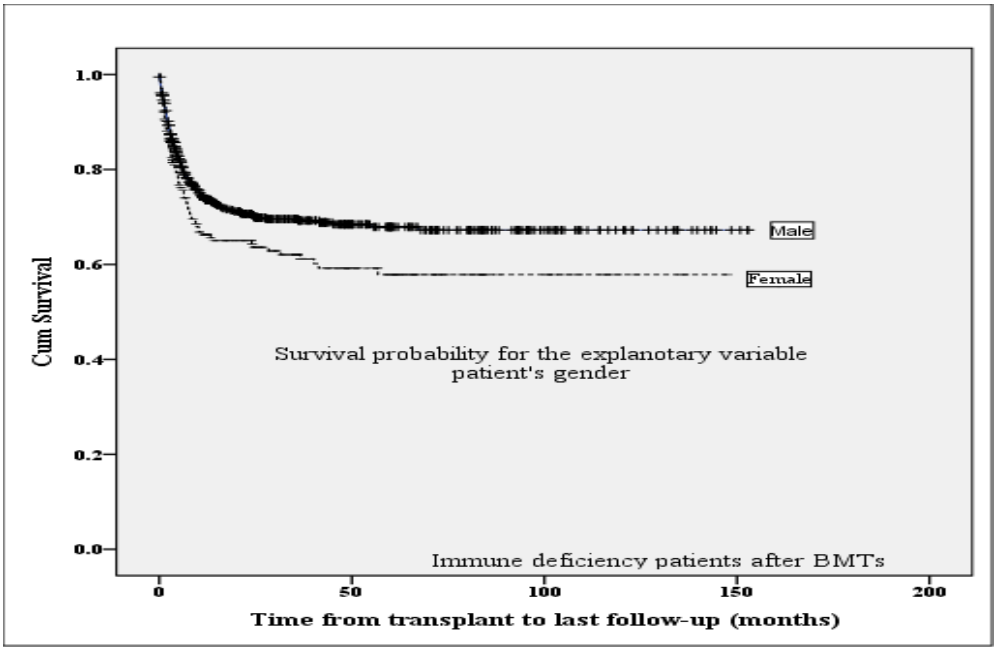

Fig.4 Kaplan-Meier cumulative survival curves for patients with immune deficiency disease following BMTs after 12-year survival (months) stratified by patient's gender

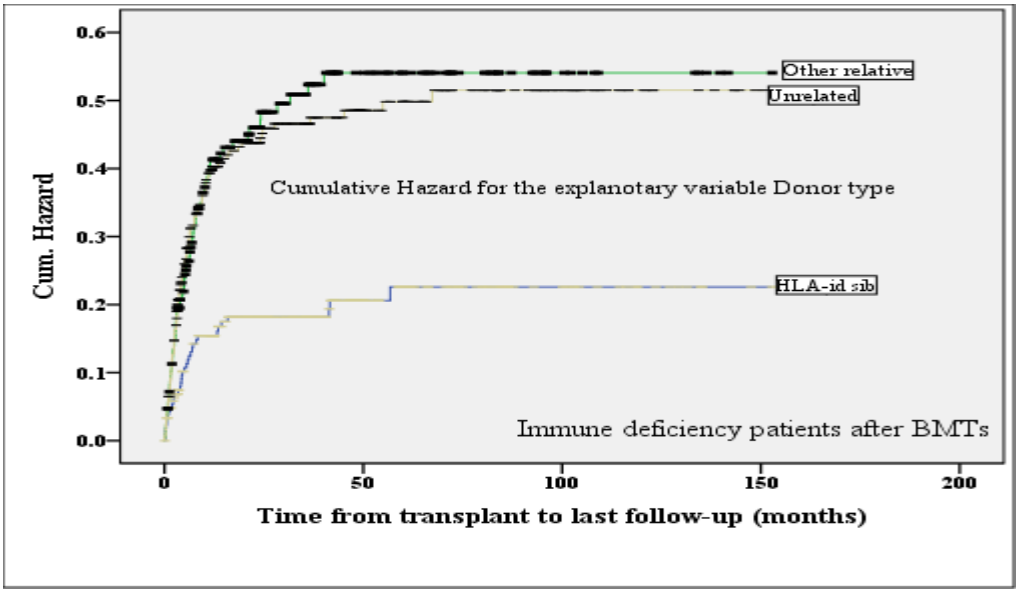

Fig.5 Estimated cumulative hazard function to illustrate HLA-id sib has the lowest risk

2. DISTRIBUTION FREE REGRESSION MODELS: SEMI-PARAMETRIC REGRESSION MODELS

Clinical outcomes, in general, are often described as events: death, stroke, epileptic seizure, multiple sclerosis lesions, recurrence of cancer, disease progression, pain, infection and bacterial/viral eradication, severe toxic adverse effect, resistance to treatment, etc. They may be quantified as time-to-event, counts of events per time interval (rates), their severity grade, or a combination of these. Such data are discrete and require specific modeling structures and methods. While thinking of survival data analysis, potentially the most popular models are life tables and Cox proportional hazard model, life tables seem to be the simplest way to analyze survival time, i.e. time that is measured for each subject from the beginning of the observation period till the defined event occurrence (e.g. an adverse event occurrence, death, full recovery) or till the end of observation period if subject did not experience the event. The idea of this method is to estimate the survival function, $S(t)=P(T \geq t)$, which is the probability of 'survival', (in other words: probability of not experiencing the event) at least to the time point t ϵ T. The cumulative distribution function is specified as well by the survival function $S(t)=1-F(t)$.

Such estimation is being performed at each time point t at which an event occurred. Between the subsequent event occurrences, survival function is assumed to be constant. On the other hand, the hazard function is being estimated, which is interpreted as the conditional probability of the event occurrence at time point t provided that the event had not occurred before. Again, hazard function is being estimated at time points at which an event occurred and is constant between subsequent event occurrences. The hazard function $h(t)$ specifies the instantaneous rate of failure at $T=t$ conditional upon survival to time t and is defined by the limit for $\Delta t \downarrow 0$ of the following

$$h(t)=\lim_{\Delta t \downarrow 0^+}\frac{1}{\Delta t}\Pr(t \leq T < t+\Delta t \mid T \geq t), \qquad (4)$$

be the hazard function. The hazard function, denoted, $h(t)$ gives the instantaneous potential per unit time for the event to occur given that the individual has survived up to time t. In contrast to the survivor function, which focuses on not failing, the hazard function focuses on failing; in other words, the higher the average hazard, the worse the impact on survival. The hazard is a rate, rather than a probability. Thus, the values of $h(t)$, its range is between zero and infinity. Straightforward relationships the distribution, survivor and hazard function should be noticed:

$$\left.\begin{aligned} H(t)=\int_0^t h(u)du, \qquad f(t)=h(t)\exp\left(-\int_0^t h(u)du\right) \\ S(t)=\exp\left(-\int_0^t h(u)du\right), F(t)=1-\exp\left(-\int_0^t h(u)du\right) \end{aligned}\right\} \qquad (5)$$

Eq. (5) shows that each of these functions completely characterizes the distribution of a lifetime.

A proportional hazards model in (1), proposed by D. R. Cox [5], postulates that the failure time T associates with vector of covariates (explanatory variables) X, such as treatment indicators, patient's age at transplant, …, etc., through the flowing function

$$h\left(t \mid X(t)\right)=$$
$$\lim_{\Delta t \downarrow 0^+}\frac{1}{\Delta t}\Pr\left(t \leq T < t+\Delta t \mid T \geq t,\, X(t)\right)=$$
$$h_0(t)\times \exp\left\{\beta' X(t)\right\};\ t \geq 0, \qquad (6)$$

where $h_0(t)$ the unknown and unspecified nonnegative baseline hazard, β is is a P × 1 vector of unknown regression parameters. Note that there is intercept β_0 in model(6), obviously, $h(t \mid X = 0) = h_0(t)$, so that $h_0(t)$ is often called the baseline hazard function. It can be interpreted as the hazard function for the population of subjects with $X_i = 0$. The baseline hazard function in $h_0(t)$ in (6) can take any shape as a function of t. The only requirement is that $h_0(t) > 0$. This is the nonparametric part of the model and $\beta' X(t)$ is the parametric part of the model. Model (6) does not assume any specific statistical distributions, and is therefore called distribution-free or semi-parametric approaches. The Cox's PH model (1) is equivalently characterized as

$$H(t \mid X) = \exp(\beta' X(t)) \int_0^t h_0(u)du = \exp(\beta' X(t)) H_0(t) \quad (7)$$

2.1. Assumptions of the Cox PH model

Though the Cox model is semi-parametric to the extent that no assumptions are made about form of the baseline hazard, $h_0(t)$, there are still a number of important issues which need be assessed before the model results can be safely applied. Assumptions of the Cox PH model can be stated as follows:

[1] The ratio of the hazard function for two individuals with different sets of covariates does not depend on time,

[2] Time is measured on a continuous scale,

[3] Censoring is independent of the event of interest (non-informative censoring). To satisfy this assumption, the design of the underlying study must ensure that the mechanisms giving rise to censoring of individual subjects are not related to the probability of an event occurring,

[4] Sufficient data for inference,

[5] Independence of observations,

[6] Explanatory variables act only on the hazard ratio. They do not affect the baseline hazard.

If one of the above assumptions is violated, the simple Cox model is invalid, the more sophisticated analyses are required [21], such as models (2), (3) and other models, which will adopted in the next issues of papers when our collected data needed to these to draw right conclusions from analysis.

2.2 Interpretation of a Cox PH model

In the following, we introduce the interpretation of the proportional hazards model; this will be useful in the following sections

[1] It is easy to show that under model (6)

$$S(t \mid X(t)) = \{S_0(t)\}^{\exp\{\beta' X(t)\}} \quad (8)$$

where $S(t \mid X)$ is the survival function of the subpopulation with covariate $X(t)$. That is

$$S(t \mid X(t)) = \exp\left\{\int_0^t -h_0(u)du\right\}.$$

[2] Proportional hazards regression model assumes that different groups have similarly shaped hazard functions: for two groups A with covariate $X_i = 1$ and B with covariate $X_j = 0$, there is a constant $\mathrm{HR} = \exp(\beta)$ such that $h_A = HR \times h_B(t)$. Since hazards are chances, this means that the ratio of the hazard functions:

$$\mathrm{HR} = \frac{h_A(t)}{h_B(t)}, \quad \text{for all time points t} \quad (9)$$

Eq. (9) can be interpreted as a relative risk or a hazard ratio, which is a constant over time (so the name of proportional hazards model). Equivalently,

$\log(\mathrm{HR}) = \beta'(X_i - X_j)$ = constant, for all time points t,

In words, the hazard ratio (HR) in Eq. (9) is a ratio of the risk (or hazard) of an event in one group compared with the risk in a comparison group is independent of t. log(HR) in Eq. (10) is the logarithm of the ratio of hazard rates (instantaneous incidence rates) in the exposed to the unexposed at a point in time is constant.

2.2 Brief overview of estimation of β, using Partial Likelihood

In most situations, we are interested in the parameter estimates than the shape of the hazard. The Cox proportional hazard model (6) is well-suited to this goal. From the interpretation of the model in the previous subsection it is obvious that β characterizes the "effect," of X. So β should be the focus of our inference while (6) is "a nuisance parameter". Given a sample of censored survival data, our inferential problems include:

[1] Estimate β and derive its statistical properties,

[2] Testing hypothesis H_0: β = 0,

[3] Diagnostics.

Since the baseline hazard $h_0(t)$ in (6) is left completely unspecified (infinite dimensional), ordinary likelihood methods can't be used to estimate β. In model (6), for the i-th patient; when T_i is a life time to right-censorship, we observe $t_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \le C_i)$; where C_i is the censoring time and I(E) indicates, by the values 1 versus 0, whether or not the event E occurs. Assume that T_i and C_i are independent conditional on X_i. Let (T_i, Δ_i,

X_i) (i = 1,…,n) be independent observations. Andersen and Gill [22] elegantly proved the asymptotic distribution of $\hat{\beta}$ by applying martingale theory in the counting process framework. Gill [23] explained a hint of how Cox's regression model can be extended in many useful ways. Recently, Devarajan and Ebrahimi [24] discussed and develop the Cox model (1) to the generalization of that model to be applied without verification of PH assumption, because a fit of their proposed model provides the both of estimation the model parameters in Cox model and the testing assumption of PH in the same time.

3. APPLICATION OF THE COX'S PROPORTIONAL HAZARDS MODELS TO THE IMMUNE DEFICIENCY PATIENTS AFTER ALLOGENEIC BONE MARROW TRANSPLANTATION

3.1 Preliminary Model (I)

The Cox PH model is the most popular method of examining the effect of explanatory variables on survival. However, it requires the assumption of proportional hazards between strata formed by the combinations of levels of the different explanatory variables. Thus, when fitting a PH model it is vital to assess the assumption of proportionality. There are numerous methodologies in the literature (see for details) for checking the assumption of PHs. The Cox PH model for the hazard of death at time t for the i-th of n individuals in a study can be expressed as Eq. (1) where $\beta' X_i = \sum_{k=1}^{P} X_{ik}\beta_k$, for all i = 1,…, n. We use the initial model with all covariates X_1, X_2, X_3, X_4 and X_5. The full model for the death time for dataset in Table 1 in this situation is modeled as

$$h_i(t \mid X) = h_0(t) \times \sum_{k=1}^{P} X_{ik}\beta_k \text{ ,for all i = 1,…, n} \quad (11)$$

We use i in the above formula to stress that the hazards may differ between individuals and are subject to their covariate values, i.e. $h_i(t \mid X) = h(t \mid X_i)$. Survival, defined as the time from BMT to death or to last contact, was censored on December 31, 2005.

Now we study a Cox regression (6) of time from transplantation to the last follow-up with the constant time covariates specified in order to detect, which covariates in Table 1, is statistically significant , compare outcomes for the categorical covariates and immune reconstitution in a large cohort of patients with severe immune deficiency who receive BMT as follows. The initial fitting of the Cox PH model yielded the following parameter estimates (Table 2). This preliminary model contains the variables $X_1, \ldots, X_5$.

Table 2

(Model I): Estimated Coefficients, hazard ratios, Standard Errors, z-Scores, Two-Tailed p-Value (for the all suspected Effects) for immune deficiency disorders disease after BMT based on Proportional Hazard Model containing graft type, age at transplantation, year at transplantation, sex and donor type $n = 800$.

Parameter	$\hat{\beta}$	exp($\hat{\beta}$)	se($\hat{\beta}$)	Wald z-score	p-value	lower 0.95	upper 0.95
X_1	0.003177	1.003182	0.006649	0.478	0.63280	0.9902	1.016
X_2	0.041030	1.041883	0.007644	5.368	7.98e-08***	1.0264	1.058
X_3	−0.032381	0.968138	0.020173	−1.605	0.10846	0.9306	1.007
X_4	0.399124	1.490519	0.139854	2.854	0.00432**	1.1332	1.961
X_5	0.304038	1.355320	0.060299	5.042	4.60e-07 ***	1.2042	1.525
Likelihood ratio test		= 52.68	on	5	df	p-value =	3.918e-10
Wald test		= 55.82	on	5	df	p-value =	8.865e-11
Score (logrank) test		= 56.76	on	5	df	p-value =	5.659e-11
Statistical significance was defined as:	0	'***'	0.001	'**'	0.01	'*'	0.05

3.2 Model checking – secondary model (II)

The first model has been fitted without considering the best functional form of the continuous variable (age level) and without questioning the underlying assumption of proportional hazard. The fit of this preliminary model was therefore investigated by examining the following residual plots.

[1] The functional form of the continuous variable age at transplantation was investigated by examining Martingale residuals,

[2] Testing the violation of the PH was investigated by examining scaled Schoenfeld residuals,

[3] Score residuals were used to investigate the influence of individual observations.

From the results presented in Table 2, Cox regression analyses revealed a relationship between mortality risk for immune deficiency disease after BMT with age (X_2), sex (X_4) and donor type (X_5). To explore this further, we fit a reduced model that excludes X_1 and X_4. The results are shown in Table 3. Second, as is the case for a linear or generalized linear model, it is desirable to determine whether a fitted Cox regression model adequately describes the dataset in Table 1.

Table 3(Model II): Estimated coefficients, hazard ratios, Standard Errors, z-Scores, Two-Tailed p-Value and for the Main Effects Proportional Hazard Model for immune deficiency disease after BMT.

Parameter	$\hat{\beta}$	exp($\hat{\beta}$)	se($\hat{\beta}$)	Wald z-score	p-value	lower 0.95	upper 0.95
X_2	0.044427	1.045428	0.007559	5.877	4.18e-09 ***	1.030	1.061
$X_{4(2)}$	0.346492	1.414098	0.141087	2.456	0.0141*	1.1332	1.961
$X_{5(2)}$	1.079212	2.942360	0.206330	5.231	1.69e-07 ***	1.964	4.409
$X_{5(3)}$	0.999777	2.717675	0.190431	5.250	1.52e-07 ***	371	3.947
Likelihood ratio test		= 58.43	on	4	df	p-value	6.206e-12
Wald test		= 60.09	on	4	df	p-value	2.775e-12
Score (logrank) test		= 61.1	on	4	df	p-value	1.7e-12
Statistical significance was defined as:	0	'***'	0.001	'**'	0.01	'*'	0.05

X_2: age at transplantation, $X_{4(2)}$: female, $X_{4(1)}$: male, $X_{5(2)}$: dnr is Other relative, $X_{5(3)}$: Unrelated and dnr =$X_{5(1)}$: HLA-id sib.

Next, we go through model diagnosis to confirm whether or not the Cox model (9) does fit the dataset in Table 1, we first eliminate the covariates whose coefficients were not statistically to get Table 3. The sex' patient is coded as indicator variable, sex = $X_{4(2)}$ is corresponds to female, $X_{4(1)}$ corresponding to male and two indicator variables the donor type, $X_{5(2)}$ corresponding to the Other relative, $X_{5(3)}$ corresponding to the Unrelated and $X_{5(1)}$ corresponding to HLA-id sib.

3.3 Assess the Functional Form of Continuous Predictor (age at transplants): Martingale Residuals

We shall use the idea of residuals to examine the best functional form for a given covariate using an assumed Cox model for the remaining covariates. The residual we shall use here, called a martingale residual [25]. If the data are right-censored and all the covariates are fixed at the start of the study the martingale residual may be defined as:

$$\hat{M}_i = \delta_i - \hat{H}_0(t_i) \times \sum_{k=1}^{P} X_{ik}\beta_k = \delta_i - r_i \; ; \; i = 1,2,...,n \qquad (11)$$

where t_i is the time point, when censoring or event occurs, for the i-th patient; $\hat{\beta}_k$ are the estimated covariate coefficients. $\hat{H}_0(t_i)$ is the estimator of the cumulative baseline hazard, δ is 1 or 0 depending on if the event of interest has occurred. The residuals have the property: $\sum_{i=1}^{n} \hat{M}_i = 0$. Also, for large samples the $\hat{M}_i$'s are an uncorrelated sample from a population with a zero mean. The residuals may be interpreted as the difference over time of the number of observed events minus the number of expected events under the assumed Cox model, that is, the martingale residuals are an estimate of the excess number of events seen in the data but not predicted by the model. As in the usual regression formulation, we would like a plot which shows us which, if any, of the observations a response has not well predicted by the fitted model. The martingale residuals are used in this section to check overall model fit, and can be used to determine the functional form of a covariate, (i.e., to check the appropriateness of discretizing or transformed a continuous covariate). A Cox model is fitted with significant covariates and the martingale residuals are plotted along with a LOWESS smooth [26] to reduce the noise level. This plot will reveal if and how the martingale residuals change with increasing values of the covariate that is investigated. If the plot is linear, no transformation is needed. If however, there appears to be a discrete time point where the slope changes, a dichotomized transformation of the covariate may be indicated. Nonlinearity is not an issue for X_4, X_5 because this covariate is dichotomous. For the regression of time to die after BMTs on the age at transplantations, let us examine the plots of martingale residuals against this covariate.

Looking at a martingale residual plot for our data in Figure 6, for the age covariate shows some curvature in the smoothed curve can indicate the need to consider a different functional form for the predictor age at transplantation. The smoothed curve is roughly linear up to about age at transplantation value about 6 years and then levels off. This suggests that

[1] Suppose that the covariate vector age at transplantation coded as an indicator variable. The age is coded as follows:

$$X_2 = \begin{cases} X_{2(1)} = 0, & \text{if age } \le \Theta \\ X_{2(1)} = 1, & \text{if age } > \Theta \end{cases} \qquad (12)$$

The cut-off value (break point) Θ in Eq. (12) is chosen from the values of age in the data set. A profile likelihood may be plotted for each age value in the data set and the Θ value yielding the highest value of the log-likelihood is chosen. Our optimal Cox model is, then,

$$h(t \mid X_{2(1)} = 0,\ X_{2(2)} = 1) = h_0(t) \times \exp(\hat{\beta}_{2(2)}) \qquad (13)$$

A cut-off point 6 years yielded the smallest p-value "1.54e-07" for the covariate and the smallest p-value (log-likelihood =−1535.5 with p-value < 2.2e-16) for the full model. After dichotomizing the age at transplantation covariate, 666 of 798 patients had age = 0 and 132 of 798 patients had age = 1. The original covariate X_2 the model will therefore be substituted by a binary variable $X_{2(1)}$ and $X_{2(2)}$ with cut-off point $\Theta = 6$.

[2] Age at transplantation may be a transform of $\log(X_2)$, X_2^2, or $X_2 \log(X_2)$. We assume that X_2 is independent of X^*. Let $f(X_2)$ be the best function of X_2 to explain its effect on survival. Our optimal Cox model is then,

$$h(t \mid X^*, X_2) = h_0(t) \times \exp(X^* \beta) \times \exp\big(f(x)\big)$$

To find f we fit a Cox model to the data based on X^* and compute the martingale residuals, $\hat{M}_i$, $i = 1, 2, ..., n$. These residuals are plotted against the value of X_2 for the i-th observation. A smoothed fit of the scatter diagram is, typically, used. The smoothed-fitted curve gives an indication of the function f. If the plot is linear, then, no transformation of X_2 is needed. If there appears to be a threshold, then, a transformed covariate version of the covariate is indicated. Details of covariate coefficients for the both models with dichotomous age values and the best function fit our data, which is $f(X_2) = \log(X_2)$ are displayed in Table 4.

Table 4

Estimated Coefficients, hazard ratios, Standard Errors, z-Scores, Two-Tailed p-Value and for the Main Effects Proportional Hazard Model for immune deficiency disorders disease after BMT $n = 798$, with dichotomized age covariate compared with transform the age into log (age).

Parameter	$\hat{\beta}$	exp($\hat{\beta}$)	se ($\hat{\beta}$)	Wald z-score	p-value	lower 0.95	upper 0.95
(age > 6) : $X_{2(2)}$	0.8006	2.2269	0.1526	5.247	1.54e-07 ***	1.651	3.003
$X_{4(2)}$	0.3112	1.3650	0.1405	2.214	0.0268 *	1.036	1.798
$X_{5(2)}$	1.0965	2.9936	0.2071	5.294	1.20e-07 ***	1.995	4.493
$X_{5(3)}$	1.0063	2.7354	0.1904	5.285	1.25e-07 ***	1.883	3.973
Likelihood ratio test		= 58.54	on	4	df	p	5.882e-12
Wald test		= 55.44	on	4	df	p	2.632e-11
Score (log rank) test		= 56.83	on	4	df	p	1.341e-11
log (X_2)	0.25613	1.29193	0.05062	5.060	4.18e-07 ***	1.170	1.427
$X_{4(2)}$	0.31886	1.37556	0.14119	2.258	0.0239 *	1.043	1.814
$X_{5(2)}$	1.16057	3.19176	0.2072	5.523	3.33e-08 ***	2.114	4.818
$X_{5(3)}$	0.94299	2.56763	0.1901	4.979	6.40e-07 ***	1.771	3.722
Likelihood ratio test		= 59.67	on	4	df	p	3.397e-12
Wald test		= 53.97	on	4	df	p	5.327e-11
Score (log rank) test		= 55.07	on	4	df	p	3.137e-11
Statistical significance was defined as:	0	'***'	0.001	'**'	0.01	'*'	0.05

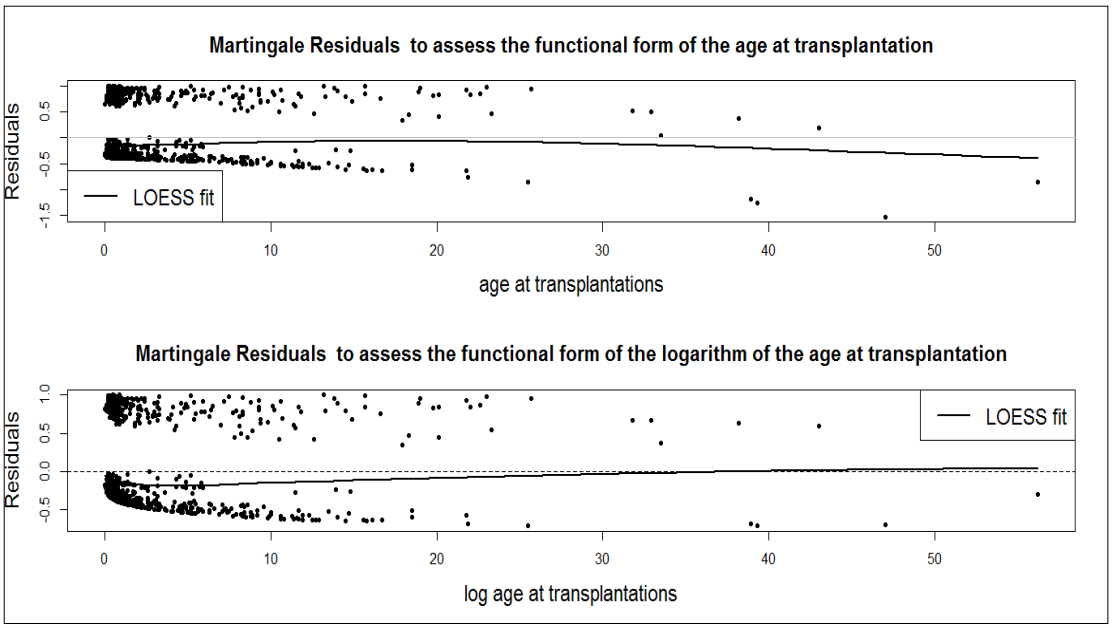

Fig.6 Martingale residuals to check the functional form of the continuous variable age produced by local linear regression (using the lowess function): Although there is no strong evidence of nonlinearity for the covariate age, the log(age) appears more suitable than the covariate age

3.4 Testing violation of the PH assumption

There are several ways to formally test PH assumptions, leading to approximately the same results. The PH assumption was examined Kaplan-Meier survival estimates in graphical format are useful in preliminary identification of proportional hazards for levels of categorical variables taken individually. Kaplan-Meier curves, as in Figures 4 or 5, estimate the survival or cumulative hazard function, respectively, that summarizes the survival data. A test of proportional hazards assumption was not significant, as illustrated by the fact that the Kaplan-Meier survival curves did not cross.

Such curves work best for time fixed covariates with few levels. If the predictor satisfies the proportional hazard assumption then the graph of the $\log(-\log(S(t))$ versus the logarithm of survival time should result in parallel lines. In Figure 4, for example, the proportional hazards hold for patient's sex, since there is constant divergence, and the cumulative hazards may not even cross. While we can use such plots to judge upon the PH assumption in univariable models, it is not straightforward to apply these plots to multivariable models, as the cumulative hazard may be confounded by other variables.

Test and graphical diagnostics for the proportional hazards assumption may be based on the scaled Schoenfeld residuals. The global goodness-of fit test for Cox PH models, proposed by Schoenfeld [27] partial residual, that has power to detect the insufficiency of covariates in describing the relative risks and the assumption of PH, was applied to the fitted model. For each covariate, a Schoenfeld residual can be calculated for each case that was not censored. A plot of these residuals against time should be "approximately flat" if the PH assumption holds. We illustrate a check on the proportionality assumption for the immune deficiency dataset using a Schoenfeld residual plot with a smooth curve fit to these residuals [28].

There is strong evidence for proportionality as shown by the large global test statistic. The idea behind the PH test is that if the PH assumption is satisfied, then the residuals should not be correlated with survival time (or ranked survival time). On the other hand, if the residuals tend to be positive for subjects who become events at a relatively early time and negative for subjects who become events at a relatively late time (or vice versa), then there is evidence that the hazard ratio is not constant over time (i.e., PH assumption is violated). In Figure 5, Schoenfeld residual plot for the significant covariates has been obtained. Following Grambsch and Therneau [28] and Therneau and Grambsch [29], Systematic convergence from a horizontal line are indicative of proportional hazards for all covariates occurred for all time t, (i.e., time-independent covariates is supported for the data).

Furthermore, Table 5 presents a Schoenfeld residual analysis to test the proportional hazards assumption, which revealed strong evidence that the proportional hazards assumption was valid and cannot violated by these data (χ^2 - test , p-value for slope is = 0.795 $\gg$ 0.05).

Table 5
Test statistics on the PH assumption based on Schoenfeld residuals.

Parameter	rho	chisq	p
X_2	0.0780	0.9979	0.318
$X_{4(2)}$	0.0215	0.1154	0.734
$X_{5(2)}$	0.0515	0.6456	0.422
$X_{5(3)}$	0.0177	0.0746	0.785
GLOBAL	NA	1.6759	0.795

Since T is a continuous random variable, then the p-th quantile (or 100p-th percentile, $0 < p < 1$), t_p is the point satisfying the equation $P(T \le t_p) = 1 - S(t) = p$. To find the p-th quantile solve the equation $S(t) = 1 - p$ for t. Hence, the median of T is the 50-th percentile, $Med(T) = t_{0.5}$. We may ask what is the probability that the lifetime of a patient with immune deficiency disease will survive is more than 50 months after BMTs ?. From Figure ???, we can see that, it is expected that 68.5% from patients with immune deficiency disease will be expected to survive more that 50 months after a BMT. The survival rate among the 798 recipients of allogeneic Bone Marrow Transplantation with autologous was 67.5% in the average, with a follow-up of 67.5 months.

3.5 Deviance Residuals

In this section, we shall consider the problem of examining a model for outliers, after a final proportional hazards model has been fit to the data. As in the usual regression formulation, we would like a plot which shows us which, if any, of the observations a response has not well predicted by the fitted model.

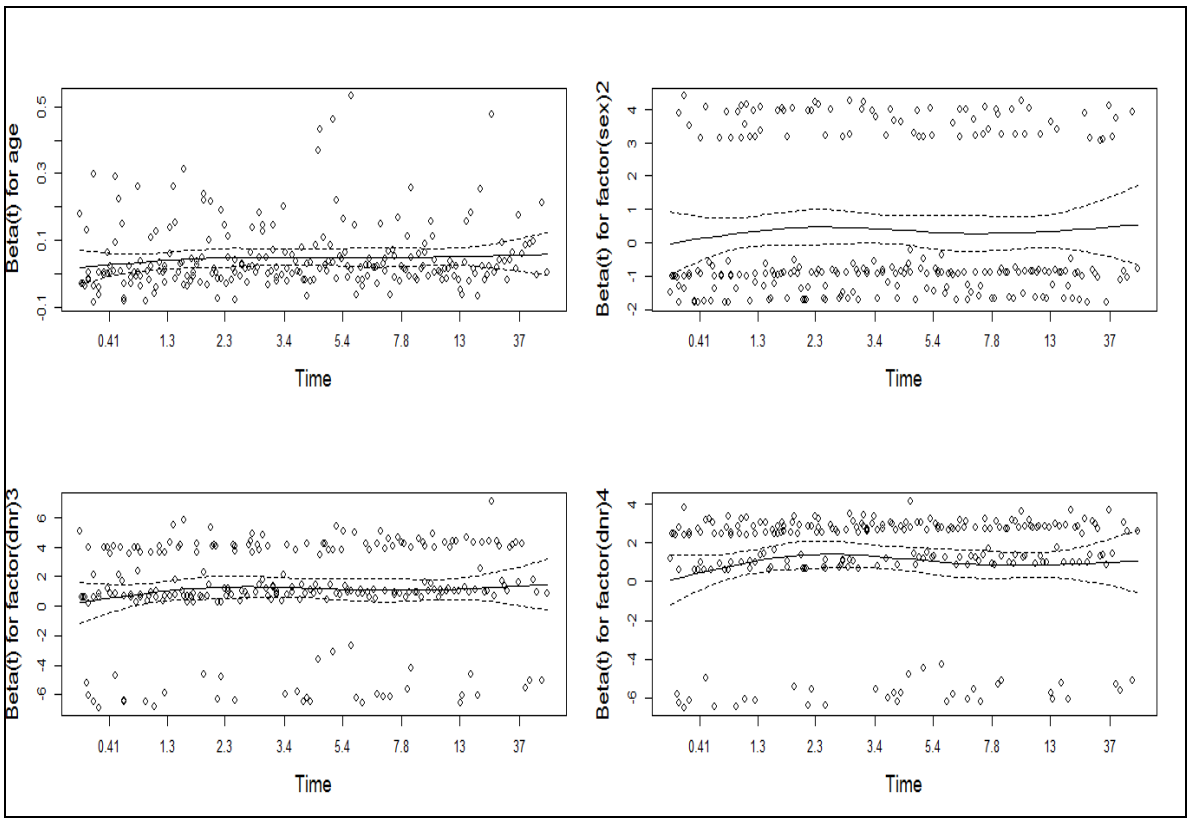

Fig.7 Plots scale Schoenfeld residual for the effect of significant factors on immune deficiency mortality after transplantations plotted against ordered survival time in months, "the fitted curve slopes are slightly horizontally, the slope of the curve, indicate a possible validation of the proportional hazards assumption". The solid reference line at zero is a smoothing-spline fit to the plot, with the broken lines representing a ±2-standard-error band around the fit.

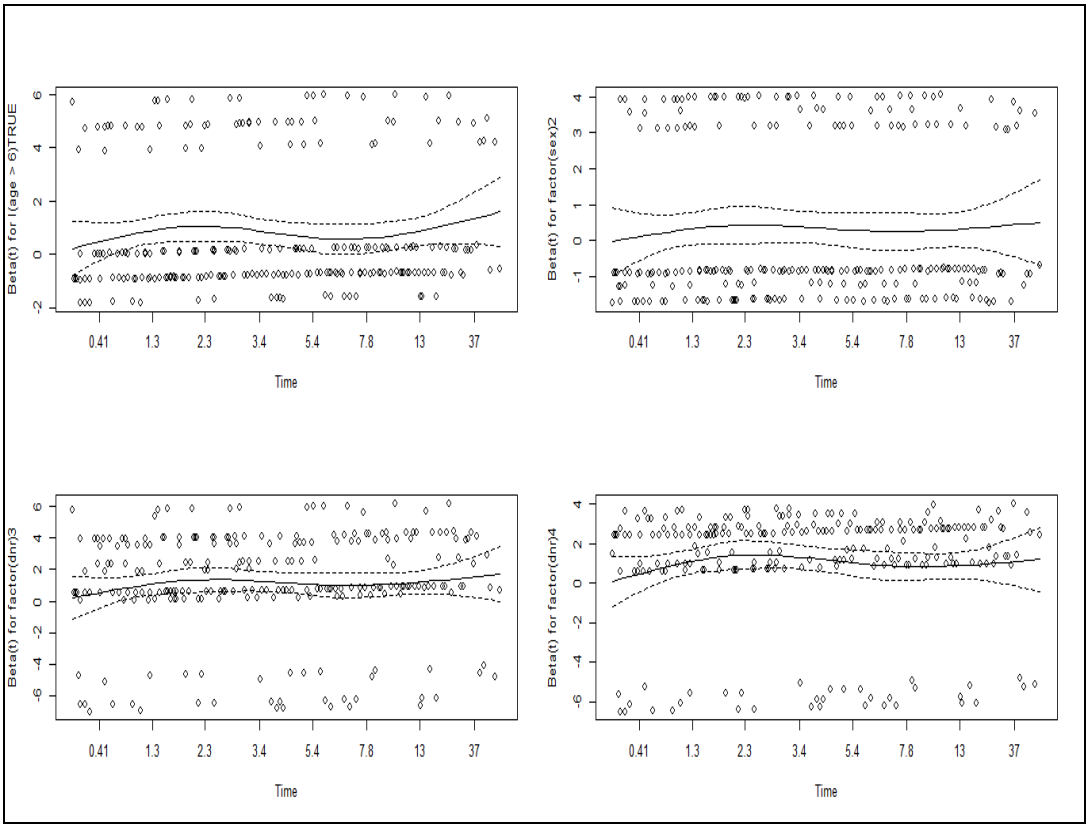

Fig. 8 Plots scale Schoenfeld residual for the effect of significant factors on immune deficiency mortality after transplantations plotted against ordered survival time in months, "the fitted curve slopes are slightly horizontally, the slope of the curve, indicate a possible validation of the proportional hazards assumption". The solid reference line at zero is a smoothing-spline fit to the plot, with the broken lines representing a ±2-standard-error band around the fit.

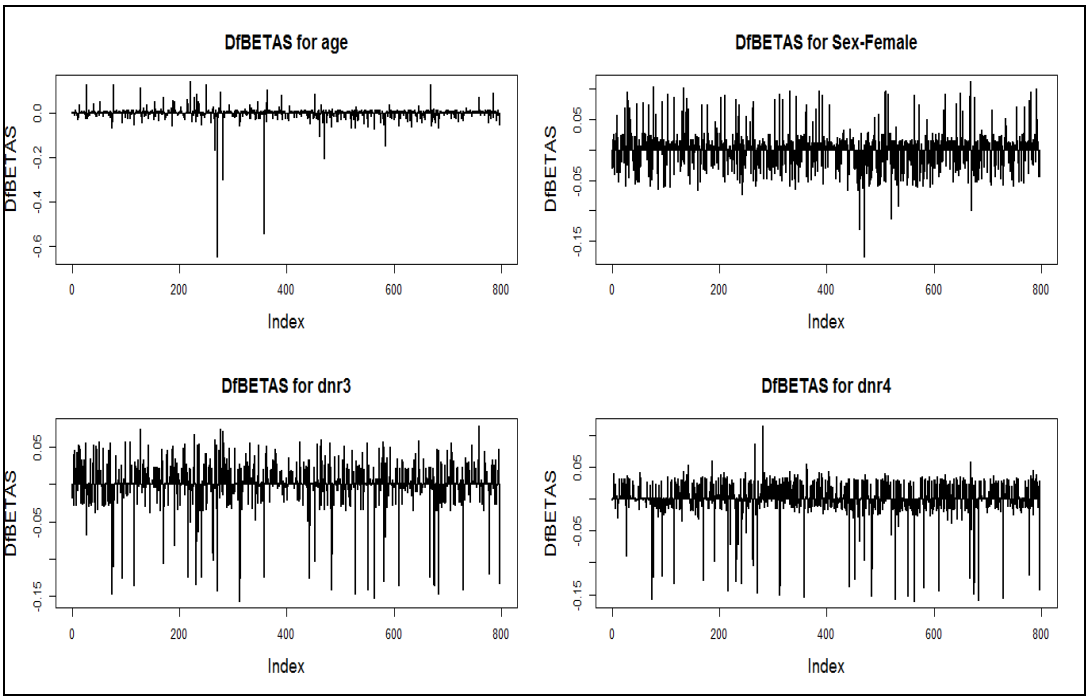

Fig. 9 Plots the DfBeta values for each significant covariate against patient ID to detect influential observations (delta-betas).

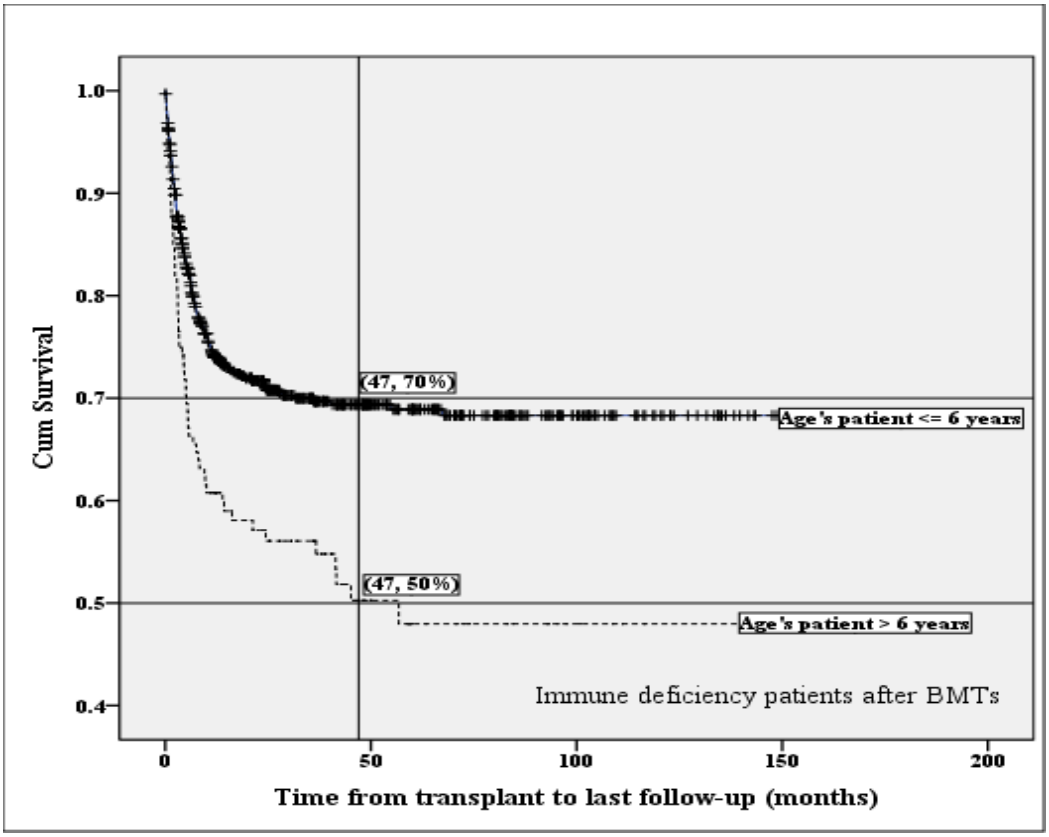

Fig.10 Estimated survival function to illustrate the difference between the survival of two cohort of patients with immune deficiency following BMTs

3.6 Conclusions with interpretation of Results

As I noted earlier, the Cox model uses a proportional hazards (PH) specification. When we do a backward elimination, we get the same result, hence, we study the model that only includes the three significant variables age, sex and dnr in Tables 3 and 4. As a result, the outputs in those tables conclude the following important features about the immune deficiency after a BMT:

[1] The factor and percentage changes in the hazard ratio can be calculated for every significant factor X_2, X_4 and X_5. We can exponentiate the coefficients to obtain hazard ratios. The exponential coefficients are interpretable as multiplicative effects on the hazard. You can then use these hazard ratios to calculate the factor change or percentage change in the baseline hazard associated with a one unit increase in a covariate. For example, holding the other significant covariates constant, an additional year of age raises the monthly hazard of death after BMT by a factor of $\exp(0.041030) = 1.041883$ on average that is; by 4.1 percent (i.e. we can confirm that the clinical trial has a lower success rate the greater the recipient's age).

[2] We can interpret the sign and statistical significance of the coefficients β's. The sign of the coefficient indicates how a covariate affects the hazard rate. Thus, a positive coefficient increases the hazard rate and, therefore, reduces the expected duration post-transplantations. A negative coefficient decreases the hazard rate and, therefore, increases the expected duration. The statistical significance of the coefficient indicates whether these changes in the expected duration will be statistically significant or not. After adjustment for confounders, the covariate age (Hazard Ratio [HR] 1.045428; 95 %-confidence of interval [CI] 1.03–1.061; p-value for $\beta_2 \ll 0.0001$). These associations were stronger with increasing the age at transplantation.

[3] For the dichotomous (or binary) variable "X_4: sex," the death rate for group 1 (male) had the lowest risk of death after BMTs (HR: 1.37556, 95% CI: 1.043–1.814; p-value for $\beta_{4(2)} < 0.05$) compared with group 1 (female), holding age, donor type are constant.

[4] For the categorical variable "X_5: donor type," the death rate for the group "HLA-id sib", had the lowest risk of death after BMTs ($HR_{(2,1)}$: 3.19176, 95% CI: 2.114–4.818; p-value for $\beta_{5(2)} \ll 0.0001$ and $HR_{(3,1)}$: 2.56763, 95% CI: 1.771–3.722; p-value for $\beta_{5(3)} \ll 0.0001$).

The standard asymptotic likelihood inference tests, the Wald, score, and likelihood ratio tests, are also valid under the Cox partial likelihood to test hypotheses about β. Wald statistics are based on the asymptotic normality of the estimated regression coefficients. Likelihood ratio statistics are based on the log likelihood ratio for two

nested models. Score statistics are based on the asymptotic normal distribution of the score function. See, for example, Kleinbaum and Klein ([3]; Ch. 2), for a detailed discussion of these tests. In this study, the test statistics the likelihood-ratio, Wald and score tests are asymptotically equivalent tests of the null hypothesis that all of the β's are zero are in close agreement, and the hypothesis is soundly rejected. As a final conclusion, the hazard function for the data in Table 1 can be estimated from one of the following Eqs.

$$\hat{h}(t \mid X) = \hat{h}_0(t) \times \exp\{0.044427X_{2} + 0.346492X_{4(2)} + 1.079212X_{5(2)} + 0.9997777X_{5(3)}\} \quad (15)$$

$$\hat{h}(t \mid X) = \hat{h}_0(t) \times \exp\{0.8006X_{2(2)} + 0.3112X_{4(2)} + 1.0965X_{5(2)} + 1.0063X_{5(3)}\} \quad (16)$$

$$\hat{h}(t \mid X) = \hat{h}_0(t) \times \exp\{0.2561 \times \log(X_{2}) + 0.31886X_{4(2)} + 1.16057X_{5(2)} + 0.94299X_{5(3)}\} \quad (17)$$

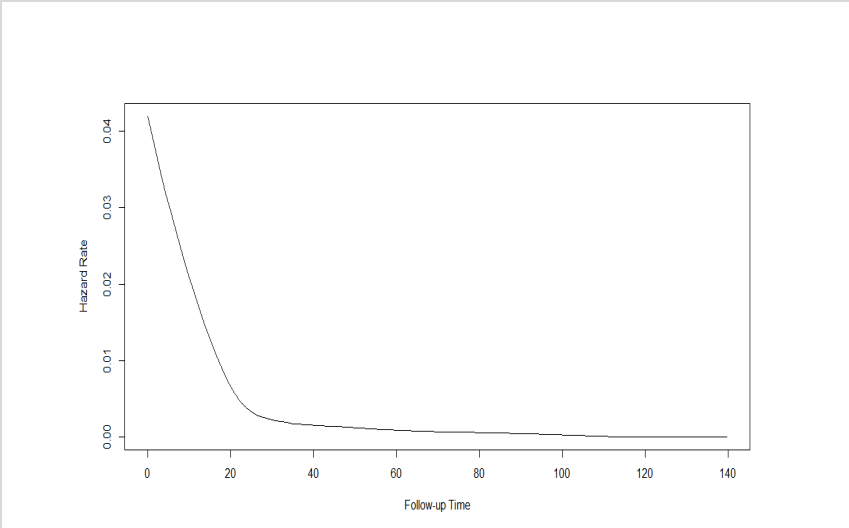

Fig.11 Smoothed estimation for $\hat{h}_0(t)$ curve using kernel-based method for the immune deficiency data after BMTs

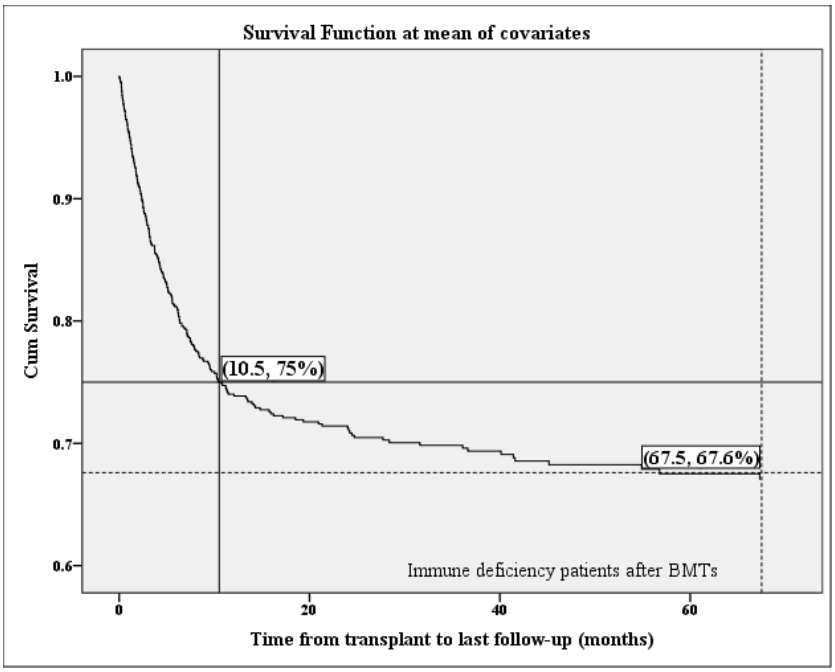

Fig.12 Estimated survival function for the immune deficiency data after BMTs

REFERENCES

[1] M. Bhattacharyya, "Analyzing survival times based on the proportional hazard regression model," International Journal of Pure and Applied Mathematics, vol. 46, no. 4, pp. 325-332, 2008.

[2] J. F. Lawless, Statistical Models and Methods for Lifetime Data, John Wiley and Sons Inc., New York, NY, USA, 2nd edition, 2003.

[3] D. G. Kleinbaum and M. Klein, Survival Analysis: A Self-Learning Text, Springer-Verlag, New York, YN, USA, 2nd edition, 2005.

[4] P. D. Sasieni, Survival Analysis, Handbook of Epidemiology, Springer Verlag, New York, YN, USA, pp. 693-728, 2005.

[5] D.R.Cox, "Regression models and life tables (with discussion)," Journal of the Royal Statistical Society, Series **B**, no. 34, pp. 187-220, 1972.

[6] D.R.Cox, "Partial likelihood," Biometrika, vol. 62, no. 2, pp. 269-276, 1975.

[7] O. O. Aalen, "A linear regression model for the analysis of life times," Statistics in Medicine, vol. 8, no. 8, pp. 907–925, 1989.

[8] D. Y. Lin and Z. Ying, "Semiparametric analysis of the additive risk model," Biometrika, vol. 81, no. 1, pp. 61–71, 1994.

[9] D. Y. Lin and Z. Ying, "Semiparametric analysis of general additive-multiplicative hazard models for counting processes," The Annals of Statistics, vol. 23, no. 5, pp. 1712–1734, 1995.

[10] M. J. Bradburn, T. G. Clark, S. B. Love, D. G. Altman, "Survival analysis part I: Basic concepts and first analysis," British Journal of Cancer, vol. 89, pp. 232-238.

[11] M. J. Bradburn, T. G. Clark, S. B. Love, D. G. Altman, "Survival analysis part II: Multivariate data analysis-an introduction to concepts and methods," British Journal of Cancer, vol. 89, pp. 431-436.

[12] M. J. Bradburn, T. G. Clark, S. B. Love, D. G. Altman, "Survival analysis part III: Multivariate data analysis-Choosing a model and assessing its adequacy and fit," British Journal of Cancer, vol. 89, pp. 605-611.

[13] M. J. Bradburn, T. G. Clark, S. B. Love, D. G. Altman, "Survival analysis part IV: Further concepts and methods in survival analysis," British Journal of Cancer, vol. 89, pp. 781-786.

[14] D. Collett, Modelling Survival Data in Medical Research, Chapman and Hall, London, UK, 1994.

[15] D. W. Hosmer, S. Lemeshow, and S. May, Applied Survival Analysis: Regression Modelling of Time to Event Data, John Wiley and Sons Inc., New York, NY, USA, 2nd edition, 2008.

[16] D. W. Hosmer and S. Lemeshow, Applied Logistic Regression, John Wiley and Sons Inc., New York, NY, USA, 1989.

[17] E. Vittinghoff, D. V. Glidden, S. C. Shiboski and C. E. McCulloch, Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models, Springer Verlag, New York, YN, USA, 2nd edition, 2012.

[18] T. Brenn and E. Arnesen, "Selecting risk factors: a comparison of discriminant analysis, logistic regression and Cox's regression model using data from the Tromsø heart study," Statistics in Medicine, vol. 4, no. 4, pp. 413–423, 1985.

[19] E. Grunebaum, E. Mazzolari, F. Porta, D. Daller et al., "Bone marrow transplantation for severe combined immune deficiency," American Medical Association, vol. 295, no. 5, pp. 508-518, 2006.

[20] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," Journal of the American Statistical Association, vol. 53, no. 282, pp. 457-481, 1958.

[21] I. Persson, "Essays on the assumption of proportional hazards in Cox regression," Doctoral Dissertation, Acta Universitatis Upsaliensis, 2002. Available from www.diva-portal.org (Accessed on July 13, 2013).

[22] P. K. Andersen and R. D. Gill, "Cox's regression model for counting processes: A large sample study," Annals of Statistics, vol. 10, no. 4, pp. 1100-1120, 1982.

[23] R.D.Gill, "Understanding Cox's regression model: A martingale approach," Journal of the American Statistical Association, vol. 79, no. 386, pp. 441-447, 1984.

[24] K. Devarajan, and N. Ebrahimi, "A semi-parametric generalization of the Cox proportional hazards regression model: Inference and applications," Computational Statistics and Data Analysis, vol. 55, no. 1, pp. 667-676, 2011.

[25] J.P.Klein and M. L. Moeschberger, Survival Analysis: Techniques for Censored and Truncated Data, Springer Verlag, New York, YN, USA, 2nd edition, 2003.

[26] W. S. Cleveland, "Robust locally weighted regression and smoothing scatter plots," Journal of the American Statistical Association, vol. 74, no. 368, pp. 829-836, 1979.

[27] D.Schoenfeld, "Partial Residuals for the proportional hazards regression model," Biometrika, vol. 69, no. 1, pp. 239-241, 1982.

[28] P.M.Grambsch and T.M. Therneau, "Proportional hazards tests and diagnostics based on weighted residuals," Biometrika, vol. 81, no. 3, pp. 515-526, 1994.

[29] T. M. Therneau and P. M. Grambsch, Modeling Survival Data: Extending the Cox Model. Springer-Verlag, New York, YN, USA, 2000.