Multistate models in survival and event history analysis

Dorota M. Dabrowska

UCLA

November 8, 2011

Research supported by the grant R01 AI067943 from NIAID. The content is solely the responsibility of the author and does not necessarily represent the official views of NIAID or CIBMTR

Outline

1. Survival and event history analysis - basic concepts
2. Competing risk models and multistate models
3. Application to transplant data of CIBMTR

1. Survival and event history analysis - basic concepts

areas of applications:

(i) demography
(ii) engineering (reliability theory)
(iii) medicine
(iv) actuarial science
(v) sociology
(vi) econometrics
(vii) astronomy

The traditional framework of survival and event history analysis deals with experiments in which outcome variable represents “time” till occurence of a single event.

The basic quantity is the survival (or reliability) function, i.e.

$$S(t) = Pr(T > t)$$

The response variable may be continuous or discrete.

Survival functions are often reprepresented using hazard rates or cumulative hazard functions.

Hazard rate α: instantaneous failure rate

$$\alpha(t)h \sim Pr(T \in [t, t+h) | T \geq t) \quad \text{(h - small)}$$

Cumulative hazard function

$$A(t) = \int_0^t \alpha(u)du$$

Common choices of the hazard rate include:

(i) $\alpha(t) \equiv a$ - $T \sim \frac{1}{a}Z$, where Z has exponential distribution with mean 1.

(ii) $\alpha(t)$ - decreasing or increasing, e.g. $\alpha(t) = \gamma t^{\gamma-1}$ (Weibull distribution $T \sim Z^{\frac{1}{\gamma}}$)

(iii) $\alpha(t)$ - U shaped (upwards or downwards)

There is a one-to-one relationship between S and A. For purposes of this talk

$$S(t) = \exp[-A(t)] \quad A(t) = \int_0^t \alpha(u)du$$

If T has continuous distribution then the discrete hazard rate is defined as

$$A(\Delta t) = P(T = t | T \geq t)$$

and

$$S(t) = \prod_{u \leq t} [1 - A(\Delta u)]$$

Sources of missing data

Censoring

(i) Type I (or administrative) censoring: the observable variables are

$$X = \min(T, c) \quad \delta = 1(T \leq c)$$

where c is a fixed time point.

For example, in econometric applications, T may represent income which need not be observable beyond certain level c. In medical applications, c corresponds to termination of a study.

(ii) Random censoring: the observable variables are

$$X = min(T, C) \quad \delta = 1(T \leq C)$$

where C represents time from withdrawal from a study due to reasons unrelated to the study itself.

In medical applications, C represents time till loss-to-follow-up (e.g. the patient may move or die from a cause unrelated to the study).

If a study is conducted during a fixed time period $[0, \tau]$, and patients enter the study at time Y, then their observations are censoring by $C = \tau - Y$.

(iii) Type II censoring: this type of censoring occurs often in reliability and industrial life testing experiments. Instead of observing failures of n items, $T_1, \ldots, T_n$, the observations terminate as soon as $r, r < n$ failures occur, where r is a fixed number.

A more elaborate version of this model corresponds to progressive type II censoring. In this case, $(r_1, n_1), \ldots, (r_k, n_k)$ are fixed numbers. At the time of the r_1-st failure we remove $n_1 - r_1$ items from the surviving $n - r_1$ items. The process continues: at the time of the r_2-nd failure, we remove $n_2 - r_2$ items and the process continues until some fixed number of failures occurs.

In the above examples, observations were censored from the “right”.

(iv) Left censoring arises if the event of interest occurs prior to the beginning of a study and the time of its occurrence is not observed but is known to be smaller than some fixed or random variable C. Thus the observable random variables are

$$X = \max(T, C) \quad \delta = 1(C \leq T)$$

(v) double censoring:
We observe (X, δ), where

$$\begin{aligned} X &= T \quad \text{and} \quad \delta = 1 \qquad \text{if} \quad C_L \leq X \leq C_R \\ &= C_R \quad \text{and} \quad \delta = 0 \qquad \text{if} \quad C_R < X \\ &= C_L \quad \text{and} \quad \delta = -1 \qquad \text{if} \quad X < C_L \end{aligned}$$

where $C_L < C_R$ with probability 1, are fixed or random left and right censoring times.

In epidemiologic applications, age T at onset of a disease may preceed C_L, the age at entrance into the study or it may occur after C_R representing age at the drop-out or termination of the study period.

(vi) current status data

Here observations consist of (X, δ) where X is the time of the examination and δ indicates if the event occurred prior to time X. Thus we observe

$$X \quad \text{and} \quad \delta = 1 \quad \text{if} \quad T \leq X$$
$$X \quad \text{and} \quad \delta = 0 \quad \text{if} \quad T > X$$

The model is different than right-censoring because the failure time T is not observed at all.

Truncation

Left -truncation is a form of biased sampling in which subjects or items are sampled conditionally on an event occurring prior to the entrance of the study. Thus we observe (T, W) where W is the truncation time and T is the failure time. T can be observed provided the truncation time W satisfies $T \geq W$. Very often, observations are also right-censored.

Channing House Data: retirees were followed from the time of entrance in to the retirement home until death. With basic time variable being age at death, the observations are left truncated by W = age at entry into the retirement home. The data are also right-censored by C = age at loss-to-follow-up (termination of the study, move to a different retirement home). (Klein and Moeschberger)

Right truncation arises in astronomic applications: absolute and apparent luminosities of an object in the sky are defined as brightness at a fixed distance and as observed on Earth. In some models, the absolute luminosity and the apparent luminosity are assumed to satisfy $-\log T_{app} = f(x) - \log(T_{abs})$, where f is a known function of x = redshift. Objects in the sky will not be detected by instruments if their apparent luminosity is smaller than a certain lower limit. Thus T_{app} is sampled conditionally on $T_{app} \leq W$.

Double truncation: subjects/objects are sampled conditionally on the failure time T falling within certain bounds: $W_L \leq T \leq W_R$

Regression models

In the regression setting, we also observe covariates and they may be of two types:

(i) time independent or external: variables measured at the time of entrance into the study

(ii) time dependent or internal - variables dependent on the follow-up history

Common choices of regression models with time independent covariates include

(i) Linear transformation models:

$$\log h(T) = \beta^T Z + \varepsilon$$

where ε is an error term with known distribution, h - unknown increasing function, β - a regression coefficient.

Proportional hazard model:

$$\alpha(t|z) = e^{\beta^T z}\alpha_0(t)$$

where α_0 is the baseline hazard rate and β is an unknown regression coefficient. If z_1 and z_2 are distinct levels of covariates then the hazard ratio is constant in time

$$\frac{\alpha(t|z_1)}{\alpha(t|z_2)} = \frac{e^{\beta^T z_1}}{e^{\beta^T z_2}}$$

(ii) Accelerated failure time models:

$$\log T = \beta^T Z + \varepsilon$$

where ε is an error term with unknown distribution

(iii) Additive hazard model

$$\alpha(t|z) = \beta_0(t) + z^T \beta(t) = \beta_0(t) + \sum_{p=1}^{d} z_p \beta_p(t)$$

Multi-state models

The model can be represented as a sequence $(T_n, X_n)_{n \geq 0}$ such that

(i) $T_0 < \ldots < T_m < \ldots$ are ordered times of the occurrence of some events

(ii) X_m is a time-dependent covariate changing at time T_m

Below $X_m = (J_m, Z_m)$, where

(i) $J_m \in \{1, \ldots, r\}$ - type of event occuring at time T_m

(ii) Z_m - covariate dependent on (T_m, J_m) and the sequence $(T_1, J_1, Z_1), \ldots, (T_{m-1}, J_{m-1}, Z_{m-1})$

Competing risk models:

Models for the joint distribution of (T, J), where J is a discrete variable taking on a finite number of values, say $J = \{1, \ldots, r\}$ and representing the type of the event that occurs at time T.

Since J assumes a finite number of values, specification of the joint distribution function $P(T \leq t, J = j)$ is equivalent to specification of the so-called incidence functions

$$F_j(t) = P(T \leq t, J = j) = \int_0^t f_j(u)du$$

The "cause specific" hazard functions

$$\alpha_j(t) = \frac{f_j(t)}{S(t)} = \lim_{h \downarrow 0} \frac{1}{h} P(T \in [t, t+h), J = j | T \geq t)$$

represent the instantaneous risk of failure from cause j.

5. Application to analysis of CIBMTR data

Follow-up data on $n = 1654$ patients who received HLA-identical sibling transplant in first remission during the 1995-2004 period.

Disease: acute lymphoblastic leukemia (ALL) or acute myelogenous leukemia (AML)

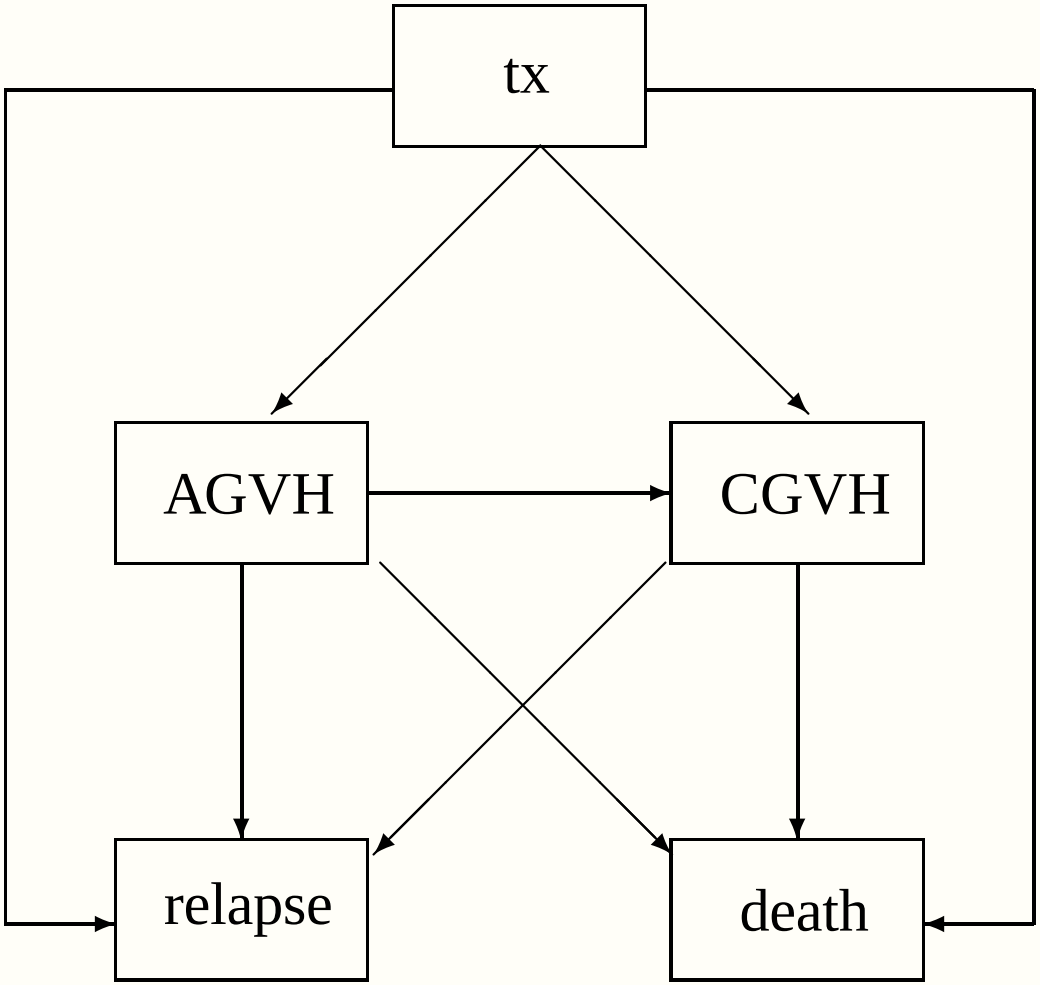
Graft source: bone marrow transplant (BMT) or peripheral blood stem cell transplant (PB).

Age (16-60), FMMATCH - female to male transplant vs other, waiting time for tx.

The two competing causes of transplant failure correspond to death in remission and leukemia relapse.

Other post-transplant complications: Acute or Chronic GVHD

States and possible transitions of the hypothetical model



Observed transition	n	median	range
TX $\rightarrow$ AGVHD	491	.7	4.3
TX $\rightarrow$ CGVHD	372	5.5	106.4
TX $\rightarrow$ relapse	106	5.6	59.4
TX $\rightarrow$ death	179	2.9	131.9
TX $\rightarrow$ censoring	506	56.9	143.8
AGVHD $\rightarrow$ CGVHD	202	4.8	57.4
AGVHD $\rightarrow$ relapse	33	5.2	23.7
AGVHD $\rightarrow$ death	141	2.9	80.3
AGVHD $\rightarrow$ censoring	115	45.7	133.0
CGVHD $\rightarrow$ relapse	27	8.3	98.3
CGVHD $\rightarrow$ death	79	9.8	124.4
CGVHD $\rightarrow$ censoring	266	51.1	144.3
A+CGVHD $\rightarrow$ relapse	25	3.5	53.3
A+CGVHD $\rightarrow$ death	65	5.6	109.3
A+CGVHD $\rightarrow$ censoring	112	56.3	145.2

Table: One-step transition probability matrix

	tx 1	AGVH 2	CGVH 3	A+CGVH $\overline{3}$	rel 4	death 5
tx	0	F_{12}	F_{13}	0	F_{14}	F_{15}
AGVHD	0	0	0	$F_{2\overline{3}}$	F_{24}	F_{25}
CGVHD	0	0	0	0	F_{34}	F_{35}
A+CGVHD	0	0	0	0	$F_{\overline{3}4}$	$F_{\overline{3}5}$
rel	0	0	0	0	1	0
death	0	0	0	0	0	1

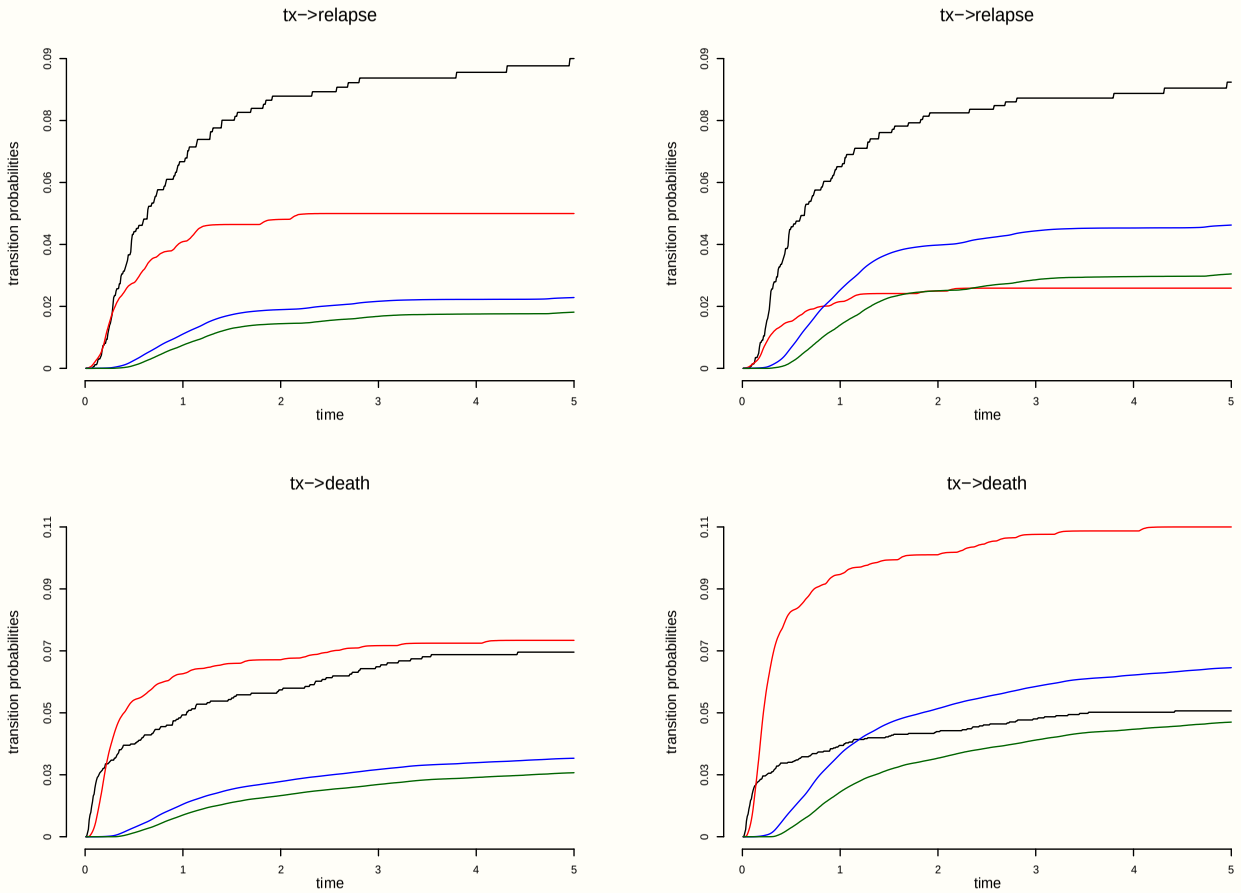
We can also estimate the transition probability matrix. In particular, if T is the time till absorption in $J = e, e =$, relapse or death, then

$$\begin{aligned} H_e(t|z) &= P(T \leq t, J = e|z) = \\ &= F_{1e}(t|z) && \text{TX} \to \text{e} \\ &+ \sum_{k=2}^{3} (F_{1k} \star F_{ke})(t|z) && \text{TX} \to \text{A or C} \to \text{e} \\ &+ (F_{12} \star F_{2\overline{3}} \star F_{\overline{3}e})(t|z) && \text{TX} \to \text{A} \to \text{A + C} \to \text{e} \end{aligned}$$

where for any two subdistribution functions on the positive half-line

$$(F \star F')(t) = \int_0^x F(t-u)F'(du) = \int_0^x F(du)F'(t-u)$$

Transition probabilities left: BMT right PB



To compare covariates we use

$$\Delta_j^F(t|z_1, z_2) = F_j(t|z_1) - F_j(t|z_2), \quad j \in \mathcal{J}_0,$$

and

$$\Delta_j^H(t|z_1, z_2) = H_j(t|z_1) - H_j(t|z_2), j = 4, 5.$$

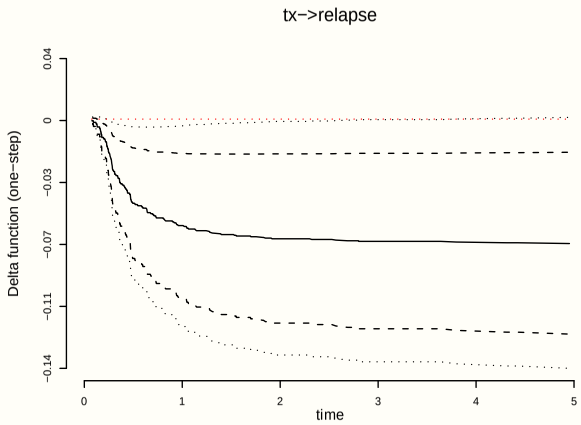
where z_1 and z_2 are two covariate levels. The sample analogues are denoted by $\widehat{\Delta}_j^F$ and $\widehat{\Delta}_j^H$.

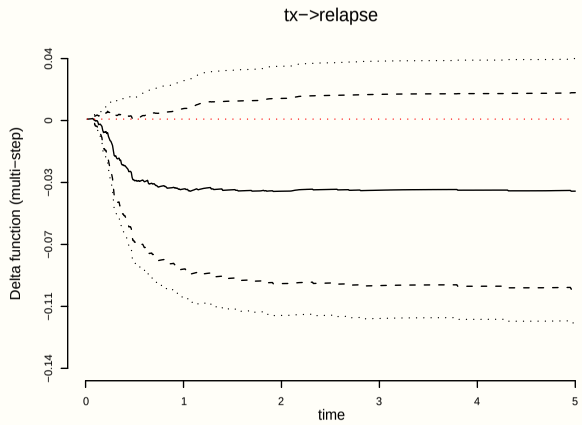
Table: Summary of covariates

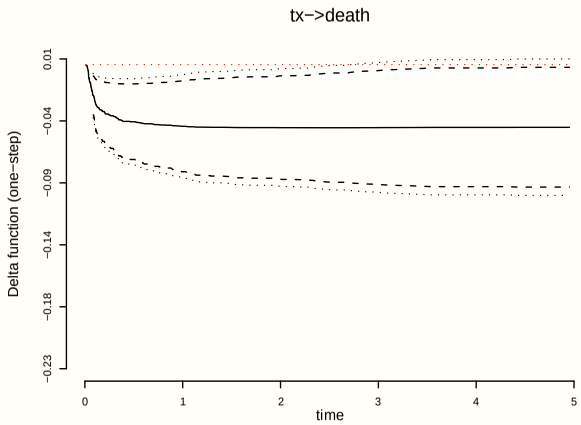
Graft source	n	Disease	n
[BMT]	842	[AML]	1168
PB/PB+BMT	803	ALL	477
Age	**n**	**Sex-Match**	**n**
< 30 (young)	550	FM	441
$[30, 42.5]$	534	[not FM]	1224
> 42.5 (old)	561		

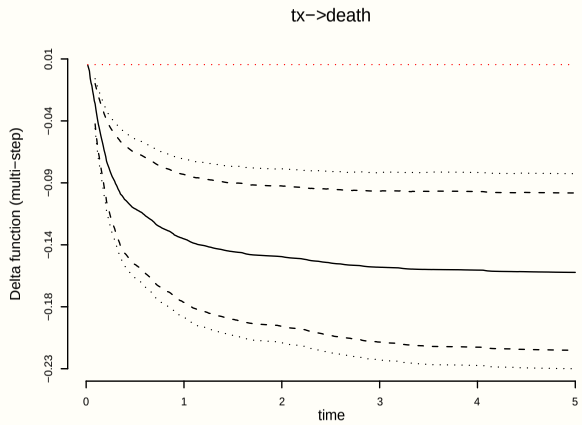
Baseline groups are marked in brackets.
FM represents a female to male transplant

Delta function: age $z_1 \in (16, 29.5]$ vs $z_2 = \in (29.5, 42.5]$

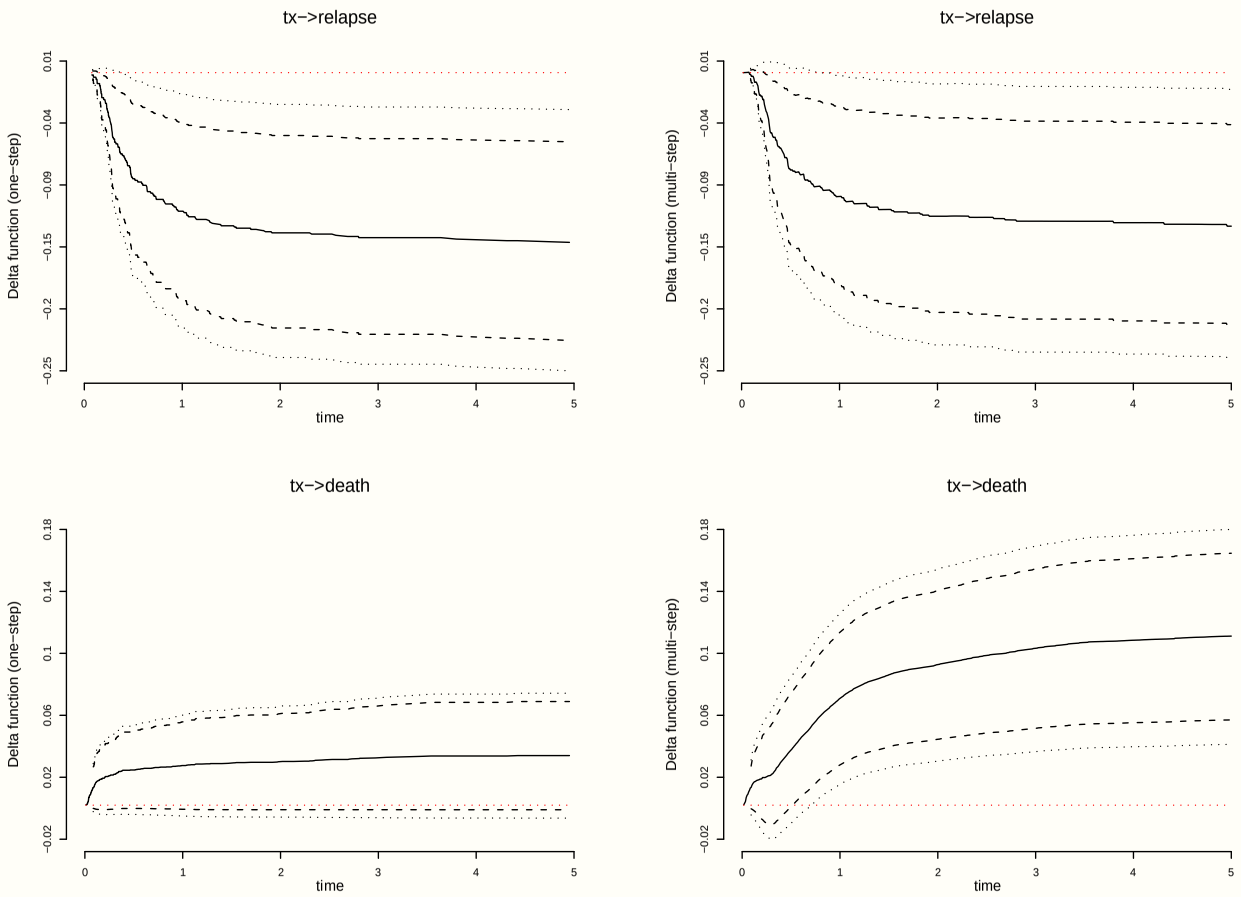




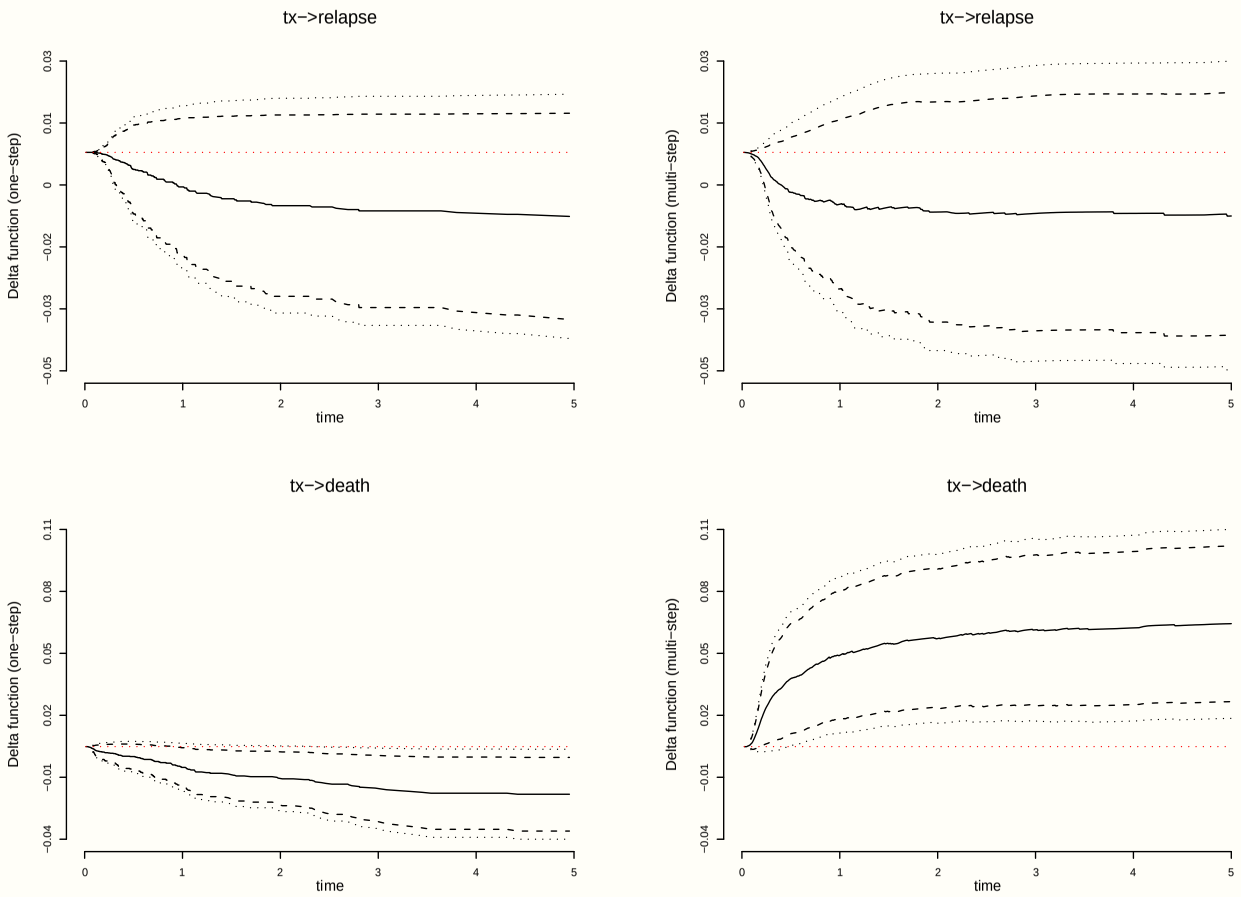




Delta function: age $z_1 \in (29.5, 42.5]$ vs $z_2 = \in (42.5, 60]$



Delta function: PB vs BMT



Delta function: All vs AML