Universiteit Utrecht

Prediction for Transition Probabilities in Multi-State Models

Bachelor Thesis Mathematics
Utrecht University

Author: Marion Snijders

Supervisor: Cristian Spitoni

January 18, 2017

Contents

Introduction

In many medical settings, people are interested in analysing data and making predictions about the future with this information. The precautions one has to take in drawing conclusions is illustrated in a story mentioned in [6]: *a centenarian on his 100th birthday [was] proclaiming that he was looking forward to many more years ahead because "I read the obituaries every day, and you almost never see someone over 100 listed there".* In this thesis we focus on prediction in a model in which we deal with incomplete data. We want to predict the chance of some event happening to a selected patient, given certain information about this patient.

By using a model we simplify and categorize reality to be able to understand a part of the process. In this thesis this will be done by choosing a Markov chain model with a finite number of states. We take (part of) the life cycle of a patient by describing that patient being in a certain state and making transitions from one state to another state when an event (or condition) of interest happens. We call this a multi-state model, which will be explained in more detail in this thesis.

In a simplified model, the two-state model alive-death illustrated in figure 1 where all patients start at the same point and all information is available, one could easily estimate a probability of surviving by the proportion of people still alive in the sample. In a medical setting this is typically not the case; patients leave the study too early, data at the start is missing or the study stops before a patient experiences the event that we are interested in, etc.. This is a problem of 'censored data'. Because of these complications we need an ad-hoc model, in this thesis we choose the multi-state Markov model with hazard based estimation.

As an example of the theory of these multi-state models we will use a dataset from the European Society for Blood and Marrow Transplantation, ebmt4, available through the R-package mstate. [4] This example will be used throughout this thesis as an illustration of the various concepts that are introduced.

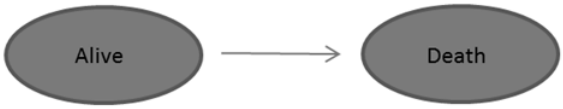

Figure 1: two-state model

Aim and structure of the thesis

The problem of censoring causes estimating transition probabilities to be interesting and complicated. The main focus of this thesis will be on developing theory for a hazard based multi-state model, a useful estimator of the transition probabilities and how to combine this with extra information available of an individual.

We start Chapter 1 with an introduction of the concept of a multi-state model, together with a more detailed introduction of the ebmt4 dataset example. In Section 1.1 we go more into the mathematical background by introducing some random variables together with the notion of filtrations, counting processes and the definition of a Markov process. In Section 1.2 we write about the hazard rate and related variables in the context of multi-state models. In Chapter 2 we focus on estimators. As we aim to estimate the transition probability, we explain this estimator in Section 2.3. As a preliminary we have to introduce two other estimators in Section 2.1 and Section 2.2, respectively the Kaplan-Meier and the Nelson-Aalen estimators. As a last step we want to be able to take the influence of various physical characteristics of a patient into account, by using the Cox model in Section 2.4. We conclude with a discussion of the possibilities and the flaws of the chosen model. The programming code used to produce pictures and data can be found in Appendix 1.

Chapter 1

Multi-state Models

The model we will use to analyze data in this thesis is a multi-state model, in which we assume the data to be a finite state stochastic process. This model is built up out of a finite number of states and some possible ways to make a transition between them. The most basic form of this model is the "alive-death" model mentioned in the introduction. Another possibility is a 'competing risks' model, this is a multi-state model where only one transition is made out of the initial state, but there is more than one possible absorbing state. This is illustrated in figure 1.1. We won't build separate theory for the competing risk or the alive-death model, because it is a variation of the more general multi-state model. Many examples of the practical use of a multi-state model can be found, detailed examples are given in [1]. A few examples in different fields of research are listed below:

- Medical: Many disease processes can be modelled, in such a way that different complications or different levels of a disease, sometimes including recovery, are chosen as the different states a patient could be in. The ebmt4 dataset is a good example of this concept, we will look into it in further detail below.
- Animal behaviour: In a certain situation one can model the different possible reactions of animals and the consequences of that behaviour as different states.
- Technical: In an industrial setting, you could be interested in the different kind of possible failures a machine can have. Which of those will occur first? And if a machine can still be productive after a first failure and reparation, what would be the second type of failure to come up?
- Economic: If a company changes something for their employees, what will their reaction be? This could be modelled in a competing risk model.

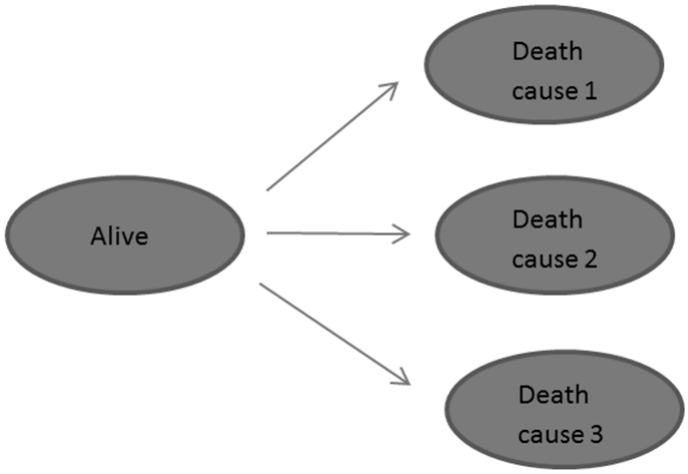

Figure 1.1: competing risk model

In multi-state models the first state is called the initial state. People enter a study for example by having a surgery or by being born. States from which one can move to another state are transient states. An absorbing state is, as the name suggests, a final state where making a transition to move out is not possible. The state 'death' is an example of this state.

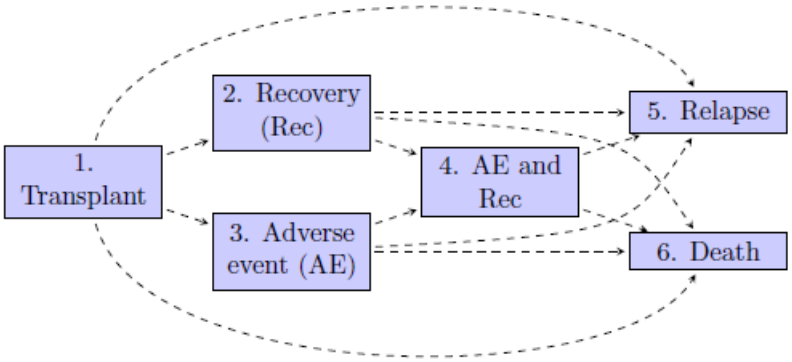

Figure 1.2: multi-state model for dataset ebmt4, source: [8]

In our main example the initial state is entered by patients having a transplantation. After this treatment they move to the other states in figure 1.2. Because arrows going both ways would complicate the computations and

analysis later on, we solve this problem by creating the new state 'recovery and adverse event'. If the patient experiences a relapse and dies as a consequence, we register this only as going into the absorbing state 'relapse'. To get an idea what the ebmt4 dataset looks like, we show the first few lines of it in figure 1.3. The first 11 columns are about tracking down when a patient makes a transition into another state. The columns with 0/1 entries are about data (not) being censored. Censored data is non-complete data, we will speak about that in further detail in Section 1.3. The last columns are other data that might be relevant for estimation, those will be taken into account in the Cox model in Section 2.4. We will not get into the technical details of arranging such a dataset in a form that could be used for calculating, neither will we get into the details of the programming language R. Although these are used to produce examples and graphics for the thesis, they are more of a practical matter, explained well in [9] and [8]. Another practical matter is the way of writing down the transition times. In a clock-reset setting, the time is set to zero every time a patient steps to a new state. In the clock -forward setting we just register the transition times in relation to the starting time at the initial state. For simplicity reasons we will use the second approach in this text.

	id	rec	rec.s	ae	ae.s	recae	recae.s	rel	rel.s	srv	srv.s	year	agecl
1	1	22	1	995	0	995	0	995	0	995	0	1995-1998	20-40
2	2	29	1	12	1	29	1	422	1	579	1	1995-1998	20-40
3	3	1264	0	27	1	1264	0	1264	0	1264	0	1995-1998	20-40
4	4	50	1	42	1	50	1	84	1	117	1	1995-1998	20-40
5	5	22	1	1133	0	1133	0	114	1	1133	0	1995-1998	>40
6	6	33	1	27	1	33	1	1427	0	1427	0	1995-1998	20-40

	proph	match
1	no	no gender mismatch
2	no	no gender mismatch
3	no	no gender mismatch
4	no	gender mismatch
5	no	gender mismatch
6	no	no gender mismatch

Figure 1.3: printed head of dataset ebmt4

1.1 Mathematical Notions and Notations

Here we introduce some mathematical theory we will need for our multi-state model. We fix the continuous time interval on which an experiment is done as $\mathcal{T} = [0, \tau]$. We take the stochastic process $\{X(t), t\}$. This can be viewed as a time-indexed collection of random variables, registering the state a patient is in at time t. These random variables are defined on the probability space $(\Omega, \mathcal{F}, P)$, with the sample space $\Omega = \{1, ..., k\}$ the k possible states of the model, $\mathcal{F}$ being the filtration of all possible history paths and P the probability measure. We explain these notions in a little more detail below:

Sample space: This is a set of possible outcomes of the experiment, so these are the values that the random variable can give as an output.

Filtration and σ-algebra's: To briefly explain the use of the word filtration; by this we mean an increasing family of sub-σ-algebra's, indexed by time. This is a family of subsets of Ω including the empty set, which is closed under complement and closed under countable unions and intersections. The increasing property can be written as

$$\text{if } s \leq t, \mathcal{F}_s \subseteq \mathcal{F}_t \tag{1.1}$$

This property is important in this context, because we view this filtration as the history of the process. A set $\mathcal{F}_t$ contains all possible paths with only the events for times $s \leq t$ fixed. One could think of this as a description of the past and present ($s \leq t$) with an open future ($s > t$).

Probability measure: This is a real valued function which assigns probabilities to the various outcomes of the experiment. We assume the reader is familiar with this concept.

Counting process: A useful mathematical approach to these multi-state models is to look at them as a counting process, $\mathbf{N} = (N_1, ...N_k)$. We define $S = 1, ..., n$ the different numbered states of the model. The N_i represent the n different states in the model, registering every patient jumping out of this state. We have a stochastic process $\{N_i(t), t\}$ with t and N_i nonnegative, but now with the sample space defined by $\Omega = \{1, 2, ...\}$, $\mathcal{F}$ again being the filtration of possible history paths. We assume that no two patients make a transition at the same time, so every process N_i is piecewise constant, non-decreasing and only jumps with $+1$ at a time. In this thesis we assume that we will work with a finite number of states, patients and transitions, so our counting processes will be finite. This holds for $N_\cdot = \sum_{i=1}^{k} N_i$ as well, so we look at this total process of all states together as a counting process as well. Later on we will refer to this last counting process as $N(t)$. We summarize some of the properties mentioned above:

$$N_i(t) \geq 0 \text{ and } t \geq 0 \tag{1.2}$$

$$N(t) \in \mathbb{N} \tag{1.3}$$

$$N(t) \leq N(s) \text{ if } s \leq t \tag{1.4}$$

Cadlag is an abbreviation of 'continu à droite, limité à gauche'. A function is cadlag if the right limit exists for all t and is given by $\lim_{s \downarrow t} f(s) = f(t)$, and

the left limit $\lim_{s\uparrow t} f(s)$ exists, but is not necessarily equal to $f(t)$. This is the case for our functions, as they jump to a new value at the exact times t_i and are continuous at all other times with no jumps. In the diagram below (figure 1.4) this is visualised.

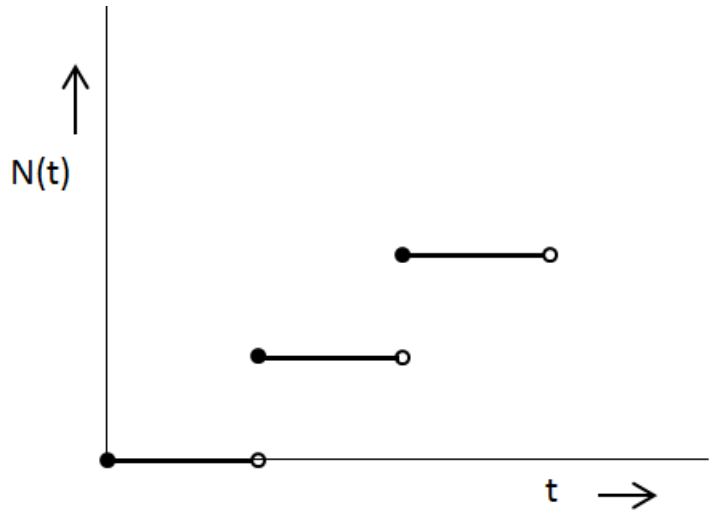

Figure 1.4: Visualisation of process N(t), cadlag

Markov process: We make the assumption we have a Markov model: the future depends on the history only through the present. This will be useful when we start estimation in the model. To write this property down more formally, we denote the history of the process until time s as a filtration $\mathcal{F}_{s-}$ in the probability space of a stochastic process $X(t)$ as defined above. We define the transition probability of going from state g to h in the interval $(u, v]$ with the patient history until time u $\mathcal{F}_{u-}$

$$P_{gh}(u, v) = P(X(v) = h | X(u) = g, \mathcal{F}_{u-}) \tag{1.5}$$

A process $X(t)$ is Markov if the history had no influence on the probability above:

$$P(X(v) = h | X(u) = g, \mathcal{F}_{u-}) = P(X(v) = h | X(u) = g) \text{ for all } \mathcal{F}_{u-} \tag{1.6}$$

The probabilities in the two equations above are *transition probabilities*. We could also look at *occupation probabilities* $P(X(t) = c)$, but we focus on the first because this is the most flexible approach in the practical analysis later on.

Markov multi-state process Combining the two notions above, we arrive at the model we will use in this thesis; a counting process viewed as a multi-state model in which we assume the process to be Markov.

1.2 Hazard rates in Multi-state Models

As we want to make calculations in our model , we will introduce some basic notations. First, we look at them in the basic alive-death model, afterwards I will explain some modifications for our more complicated multi-state example. We look in a more detailed way at patients making transitions and introduce the concept of a hazard rate. As we will see that we can look at the multi-state model as a hazard based model, this will be an important concept when we start with estimation. We define another random variable T, being the time of the alive-death transition in the basic model. This random variable is defined on a probability space $(\Omega, \mathcal{F}, P)$ with $\Omega = [0, \tau]$, $\mathcal{F}$ the family of all history paths, and P the probability measure. The survival function is defined as a probability function

$$S(t) = P(T > t). \tag{1.7}$$

Naturally, the failure function is given by

$$F(t) = 1 - S(t) = P(T \leq t). \tag{1.8}$$

Both functions are not measured directly in data. More easily accessible from the data is the hazard rate function

$$\alpha(t) = \lim_{\Delta t \to 0} \frac{P(T \leq t + \Delta t | T \geq t)}{\Delta t}, \tag{1.9}$$

which can be thought of as the instantaneous probability of going to the next state per time unit. There is a relation between $S(t)$ and $\alpha(t)$, shown in the next few steps. As we know that $P(A|B) = \frac{P(A \cap B)}{P(B)}$, we rewrite the numerator of α as

$$\alpha(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t)}{P(T \geq t)\Delta t} \tag{1.10}$$

and define f as the probability density function of F. In this way we can substitute:

$$\alpha(t) = \frac{f(t)}{S(t)}. \tag{1.11}$$

Because $S'(t) = (1 - F(t))' = -f(t)$ we can now derive

$$\alpha(t) = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)} = \frac{d}{dt} - ln(S(t)). \tag{1.12}$$

Rewriting this to a function of S we get

$$S(t) = e^{-\int_0^t \alpha(u)du}. \tag{1.13}$$

Although from a mathematical point of view the function $S(t)$ could feel like a logical starting point, in applications much more emphasis is on $\alpha(t)$, as this is the accessible measure in the survival data. We will go deeper into this subject in the next chapter. We define a cumulative hazard rate

$$A(t) = \int_0^t \alpha(s)ds. \tag{1.14}$$

As we step to a more complicated multi-state model, we will use indices to explain from and to which state the transition is made. When writing α_{ij} we mean the hazard rate of the transition from state i to state j. If we have a situation where from a state i different transitions $j, ..., K$ are possible, it is not possible to find a uniquely determined joint distribution of those transitions without further assumptions. For explanation on this statement I refer to [1]. We will not get into the details of this theory, but just focus on the separate transitions, which we can estimate. We can still use a survival function for a state i in the same interpretation as before, but modified to the situation where more than one hazard cause is possible:

$$S_i(t) = e^{-\sum_{n=j}^{K} A_{in}}. \tag{1.15}$$

This formula can now be interpreted as the probability to not have made any transition at time t. If we want to say something about a specific cause we can use the cumulative incidence function,

$$I_{ij}(t) = \int_0^t \alpha_{ij}(s)S_i(s)ds, \tag{1.16}$$

the probability of making a transition from state i to j before t. The last two formulas we will introduce are that of the intensity process λ

$$\lambda_{ij}(t) = Y_i(t)\alpha_{ij}(t), \tag{1.17}$$

and the cumulative intensity process Λ

$$\Lambda(t) = \int_0^t \lambda(s)ds, \tag{1.18}$$

with $Y_i(t)$ being the number of patients at risk in state i at time t. To keep these formulas clear of confusion; α is a proportion of patients making a transition, λ is a measure for the actual number of patients making a transition. Most variables of the above formula's are only known at the time they happen, but $Y_i(t)$ is known "just before" t. We write this as $t-$ and mention that $Y(t-) = Y(t)$, which makes Y a predictable process.

1.3 Censoring

As mentioned before, a main issue which makes the analysis of survival data difficult, is the problem of censoring. In a basic alive-death model with full

data information, we would estimate the probability of surviving by the proportion of the people still alive in the database. In all kinds of situations, data cannot be fully observed. We define two random variables, the event time of an individual as T and the censoring time as C, both on the interval $\mathcal{T} = [0, \theta]$. In the next examples, in which we distinguish two main kinds of censoring, we show what the actually observed time $\tilde{T}$ would be.

- Right censoring: The event of interest has not happened yet. In the embt4 example this could be a patient still being alive at the end of the study. The observed transition time is given by $\tilde{T} = \min(T, C)$.
- Left censoring: The starting point of an individual is unknown. This could be the case if one wants patients to be in the initial state if an infection occurs, but as the consequences of this infections are only noticed at a later time, the precise starting time will be never known. We have $\tilde{T} = \max(T, C)$ with in this case T the starting point and C the time of entering the study.

If one wants to be explicit about the censoring data, we can write $(\tilde{X}_i, D_i)$. If $D_i = 1$, we actually observed $X_i = \tilde{X}_i$. If $D_i = 0$ the data was censored; we only know that $X_i > \tilde{X}_i$. Other types of censoring can be seen as a combination of the two stated above. In this thesis we will focus on the right censoring, which is present in many datasets.

To be able to make calculations, we later on need an assumption of independent right censoring. In other words, we assume that the probability distributions of C and T, as defined at the start of this paragraph, are independent.

$$T \perp C \implies \text{for all } t \text{ and } c \; P(T \geq t, C \geq c) = P(T \geq t)P(C \geq c) \tag{1.19}$$

If we define $G = P(C > t)$ we could write

$$P(\tilde{T} > t) = S(t)G(t) \tag{1.20}$$

In this assumption, we want the uncensored data to be representative for the dataset if there had been no censoring. This is the case if the censoring distribution is independent of the survival time distribution. If the data is censored because the study has ended, the independence assumptions is often safely made. If patients step out of the study by themselves we have to take more caution. If, for instance, people who know they will die in a short timespan go home to spend their last time with their family, this is not a case of independent censoring. The dataset that remains is not representative for the whole dataset, because people with a relatively high chance of dying left the study. This leads to a bias in our estimators later on. Another case where caution is needed is when an competing event has occurred. This means that

the event of interest cannot happen any more. We could be interested in the occurrence of a specific cancer type, but some patients will die of other causes. If we treat these cases as censored, this is certainly not a case of independent censoring. We assume that the patients that remain have the same chance of an event happening as the whole population including the censored data. This means that the censored data must have the same survival distribution as the remaining data. As the competing event prevents the event of interest of happening, the distribution is clearly not the same.

To be clear about the notation of time in our database we look at figure 1.5 and 1.6. The closed dots are deaths, the open dots a censoring. Here we see what could happen with censoring due to follow up, which means the study has ended. In the database the data is not stored by calendar time, but by the time since transplantation.

The last thing we want to add here is that we will assume that the data will come in three components $(\tilde{t}, d, \mathbf{z})$. Here we define $\tilde{t}$ the observed transition time, d the censoring indicator and $\mathbf{z}$ a vector of covariates. The covariate vector contains extra information, which could for example be age, length, other diseases which a patient has. In the last section of Chapter 2 we explain a model in which these are taken into account.

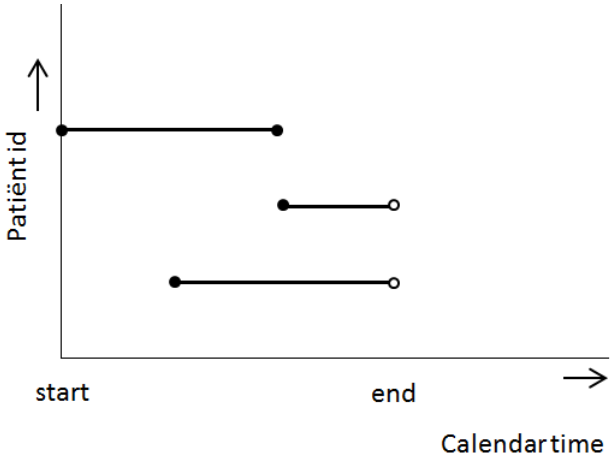

Figure 1.5: visualisation censoring, calendar time

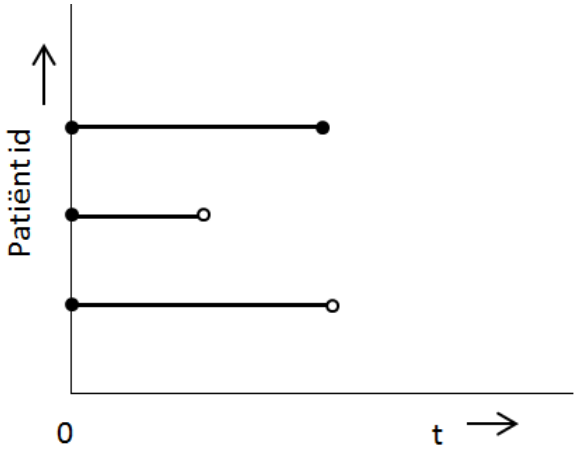

Figure 1.6: visualisation censoring, time since transplantation

Chapter 2

Estimation of Transition Probabilities

In this chapter we will focus on estimating transition probabilities. In the case of complete data we could estimate this probability by computing the empirical average, but as we have censored data we would overestimate the hazard rate if we do so. The estimator we will use to compute transition probabilities is the Aalen-Johanson estimator. Explaining this estimator and its basic properties are the main goals of this chapter. To guide the reader into the subject we will start with explaining the Kaplan-Meier and Nelson-Aalen estimators. The Kaplan-Meier estimator helps the reader to develop some understanding of the subject and we will use it to introduce the product integral. In Section 2.3 we will see that this first estimator is a special case of the Aalen-Johanson estimator. The Nelson-Aalen estimator is needed in the Aalen-Johanson estimator, so we explain this estimator before starting with the last one as well. As we want to take other information than the transition times about the patient into account as well, we conclude the chapter with an explanation of the Cox-model. We combine the non-parametric multi-state model with this semi-parametric model to write a final overview of estimation in our example of ebmt4 data.

2.1 Kaplan-Meier Estimator

This first estimator estimates $S(t)$, which can be used in the basic alive-death model of figure 1. This concept is generalised to the Aalen-Johansen estimator for multi-state models in Section 2.3. We take T as a random variable registering the transition times $t_i \in \mathbb{R}$ of an individual. To explain the

estimator we assume an ordered list of transition times $0 < t_1 < t_2 < ... < t_N$ from a dataset. From a probabilistic point of view, we can split the formula $S(t)$ into a product:

$$S(t_j) = P(T \geq t_j) = P(T \geq t_j | T \geq t_{j-1}) P(T \geq t_{j-1}) \tag{2.1}$$

$$S(t_j) \approx (1 - \alpha(t_j)) S(t_{j-1}) \tag{2.2}$$

Because we make the independent censoring assumption, we assume that the uncensored part of the data is representative for the original dataset of patients. This means we can easily estimate the probability of failing as a proportion of transitions made in relation to the total of patients at risk at that time. We define d_j as the number of observed events at time t_j, which will be one in this set-up of discrete time and the assumed absence of ties. The number of uncensored patients at risk just before t_j is defined as n_j. We combine this with the repeated use of the recurrence relation of S in 2.2 and arrive at:

$$\hat{S}(t_j) = \prod_{i=1}^{i=j} (1 - \frac{d_i}{n_i}) \tag{2.3}$$

If there are many events in the chosen time interval, this estimator will come closer to a product-integral, which uses a continuous distribution. This approach is used later in the Aalen-Johanson estimator. We state the definition of this product-integral as given in [1]. We will not need the matrix notation in this context yet as we work with scalar values now, but still write it in this form because it will be useful later on.

Definition 2.1.1 (product-integral)**.** Let $\mathbf{X}(t)$, $t \in \mathcal{T}$, be a $p \times p$ matrix of cadlag functions of locally bounded variation. We define

$$\mathbf{Y} = \prod(\mathbf{I} + d\mathbf{X}), \tag{2.4}$$

the product-integral of $\mathbf{X}$ over intervals of the form $[0, t]$, $t \in \mathcal{T}$, as the following $p \times p$ matrix function:

$$\mathbf{Y}(t) = \prod_{s \in [0,t]} (\mathbf{I} + \mathbf{X}(ds)) = \lim_{max|t_i - t_{i-1}| \to 0} \prod (\mathbf{I} + \mathbf{X}(t_i) - \mathbf{X}(t_{i-1})), \tag{2.5}$$

where $0 = t_0 < t_1 < ... < t_n = t$ is a partition of $[0, t]$ and the matrix product is taken in its natural order from left to right. In the leftmost term of the product, $\mathbf{X}(0)$ must to be replaced by $\mathbf{X}(0-) = \mathbf{0}$ because the left endpoint 0 is included in the interval $[0, t]$.

This product integral was introduced by Volterra as the unique solution Y of the equation

$$Y(t) = \mathbf{1} + \int_{s \in]0,t]} Y(s-) X(dt_n). \tag{2.6}$$

In our model $Y(t)$ is our probability function and $X(dt)$ is representing the hazard rate. More on the topic of this equation and the appliance of product integration is found in [3].

In case of a continuous distribution of $A(t)$ we could write the Kaplan Meier estimator as:

$$\hat{S}(t) = \prod_{s \leq t} (1 - d\hat{A}(s)) \tag{2.7}$$

If we assume A a jump function, the times s when there is a contribution to the outcome of the estimator are when the function $\hat{A}$ makes a jump. In other cases, when $d\hat{A} = 0$, this factor can be ignored in multiplication. The estimator is still a finite product of scalar values, as in Section 2.3. The two formulas for the Kaplan-Meier estimator bring us to a next question; how can we estimate $A(t)$? We go into this subject in the next paragraph.

2.2 Nelson-Aalen Estimator

The Nelson-Aalen estimator is estimating $A(t)$. We need this estimator again because of censored data in our dataset. If we take the empirical average we overestimate the failure rate because some of the drop-out times are due to censoring, not to dying. First we will start with a more intuitive definition of $\hat{A}$, afterwards writing this in a more formal and suitable way to proceed with the analysis of the Aalen-Johansen estimator.

If we are interested in the cumulative hazard rate, we want to sum up all deaths and compare them in a ratio with the total number of patients. Because of censoring this cannot be done directly, after every transition the ratio 'transition out of state/total number of patients in the state' has to be recomputed. We use the same notation as with the first way of writing down the Kaplan-Meier estimator. This leads to the estimator

$$\hat{A}(t_j) = \sum_{i=1}^{j} \frac{d_i}{n_i}. \tag{2.8}$$

In the same way as the Kaplan-Meier estimator approaches a continuous distribution in case of many jumps in a given time interval, we can write the Nelson-Aalen estimator in another form. We defined $Y_h(t)$ and $N_h(t)$ as the number of patients at risk in state h and the jumps made out of state h. This leads to

$$\hat{A}_h(t) = \int_0^t \frac{dN_h(s)}{Y_h(s)}, \tag{2.9}$$

which still comes down to equation 2.8 if we look at it from a jump-function perspective. The notation of formula 2.9 is used in proofs of several properties

of the Nelson-Aalen estimator. These proofs involve martingales and compensators, for this theory I refer to [1].

We define consistency here to state the second property of the estimator:

Definition 2.2.1 (consistency)**.** An estimator $\hat{\theta}_n$ of θ with sample size n is *consistent* if

$$P(|\hat{\theta} - \theta| > \epsilon) \to 0 \text{ as } n \to \infty. \tag{2.10}$$

In this definition we want the estimator to converge to the real value of the parameter as we increase the sample size to infinity. Consistency implies asymptotically unbiasedness. [5] The Nelson-Aalen estimator as defined above is consistent.

2.3 Aalen-Johansen Estimator

In this paragraph we arrive at estimating transition probabilities $p_{ij}(s,t)$ from state i to state j. The difference between $\alpha_{ij}(t)$ and $p_{ij}(s,t)$ is that the first is an instantaneous hazard rate and the second the chance of a transition in a given time interval $(s,t]$ These probabilities form the transition probability matrix $\mathbf{P}$ with p_{ij} being the element on the i'th row and j'th column. Analogue we define $\mathbf{A}$ to be the matrix with on the i'th row and j'th column the element A_{ij}. In figure 2.1 we write all possible transitions down in matrix notation assigning them a number, NA indicating that no transition is possible.

	to					
from	Tx	Rec	AE	Rec+AE	Rel	Death
Tx	NA	1	2	NA	3	4
Rec	NA	NA	NA	5	6	7
AE	NA	NA	NA	8	9	10
Rec+AE	NA	NA	NA	NA	11	12
Rel	NA	NA	NA	NA	NA	NA
Death	NA	NA	NA	NA	NA	NA

Figure 2.1: transition matrix of dataset ebmt4

An example of the information that a reader can extract here is that in our model the transition from 'adverse event' to recovery cannot be made, but that the transition 'adverse event' to 'adverse event and recovery' is possible and has got the number 8 assigned to it. To understand how the Aalen-Johansen estimator works, we take the same approach as with the Kaplan-Meier estimator.

We start with a small example to create some understanding and will give the formal definition of the Aalen-Johansen estimator afterwards. For the sake of

simplicity we take a situation with only two transition times here. We want to emphasize this is a discrete approach to create understanding, the final estimator of $\mathbf{P}$ will be defined for continuous time. A doctor or patient could be interested in $P(X(4) = 6|X(2) = 3)$. In words, the chance of being dead at $t = 4$, if the patient is in the state 'adverse event' at $t = 2$. This could be calculated by adding the probabilities of the time paths that lead from 'AE' to 'Death'.

- $t = 3$ no transition, $t = 4$ transition to death
- $t = 3$ recovery, $t = 4$ death
- $t = 3$ transition to death, $t = 4$ no transition

giving

$$P[X(4) = 6|X(2) = 3] = \tag{2.11}$$

$$P[X(3) = 3|X(2) = 3]P[X(4) = 6|X(3) = 3]+ \tag{2.12}$$

$$P[X(3) = 4|X(2) = 3]P[X(4) = 6|X(2) = 4]+ \tag{2.13}$$

$$P[X(4) = 6|X(2) = 3]P[X(4) = 6|X(2) = 6] \tag{2.14}$$

Now we take two matrices for $t = 3$ and $t = 4$, with the transition numbers changed for the instantaneous transition probabilities and on the diagonal the chance of not making a transition at that time. If we multiply these matrices to get $\mathbf{P}$ and look at the calculation made for $p_{3,6}$ we get a computation similar to 2.16. This is an intuitive way of understanding the logic of the chosen definition of the Aalen-Johanson estimator in the next paragraph. From here we take a continuous approach again.

We define $A_{ij}(t) = \int_0^t \alpha_{ij} ds$ for all $i \neq j$, with $\alpha_{ii}(t) = -\sum_{i \neq j} A_{ij}$. Then the transition probability matrix is given by

$$\mathbf{P}(s,t) = \prod_{(s,t]} (\mathbf{I} + d\mathbf{A}(u)) \tag{2.15}$$

This suggests to estimate $\mathbf{P}$ by

$$\hat{\mathbf{P}}(s,t) = \prod_{(s,t]} (\mathbf{I} + d\hat{\mathbf{A}}(u)) \tag{2.16}$$

with the elements of $\hat{\mathbf{A}}$ the Nelson-Aalen estimator $\hat{A}_{ij}$. Just as the product integral notation of the Kaplan-Meier estimator still came down to a finite product of scalar values, the Aalen-Johansen estimator comes down to a finite product of matrices. At times when there is no transition in the model, the matrix $d\mathbf{A}$ is one with all elements estimated zero. This comes down to

multiplying with the identity matrix, which has no effect on the outcome of $\hat{\mathbf{P}}$. Only the transition times have to be taken into account while estimating $\mathbf{P}$.

We stated above that the Nelson-Aalen estimator is consistent, which can be carried through to the Aalen-Johansen estimator. For proofs, we refer to [1] again. We will state a possible formula (Greenwood) to calculate the covariances of the Aalen-Johansen estimator beneath, for which more details can be found in the reference as well. Another way of computing is via the Aalen covariance formula, but these are approximate equal to each other in practice. The proofs of these formulas and statements are beyond the scope of this thesis and because of the same reason cannot give full information on the notation and meaning of this formula here. The benefit of the formula beneath, which is why we state it here, is that it is a closed formula, so the covariance can be computed directly without the need to simulate.

$$\widehat{cov}(\hat{\mathbf{P}}(s,t)) = \int_s^t \hat{\mathbf{P}}(u,t)^\top \otimes \hat{\mathbf{P}}(s,u-)\widehat{cov}(d\hat{\mathbf{A}}(u))\hat{\mathbf{P}}(u,t) \otimes \hat{\mathbf{P}}(s,u-)^\top \quad (2.17)$$

2.4 Cox Proportional Hazard Model

Until this point we have been focussing on estimating the transition probabilities in the multi-state model. Another interesting question is how other factors, for example age or sex, from now on called covariate values, influence the hazard rates. Given the data as a vector $(\tilde{t}, d, \mathbf{z})$ as introduced in Section 1.3, we could ask for $P(X(t) = c|X(t) = b, \mathbf{z})$, or the Aalen-Johansen estimator now being $\hat{\mathbf{P}}(s,t) = \prod_{(s,t]}(\mathbf{I} + d\hat{\mathbf{A}}(u, \mathbf{z}))$. We will look at this question through the Cox proportional hazard model. This is a semi-parametric model because we assume the baseline hazard rate $\alpha_0(t)$ as non-parametric (not making any assumptions on the form of the function) and the influence of the specific covariates multiplicative and parametric in an exponential model. We denote the values of the covariates of an individual X_i in a vector $\mathbf{z}_i = (z_{i1}, ..., z_{in})$ and define a vector with scalar values who influence the effect of the covariates on the hazard rate: $\beta = (\beta_1, ..., \beta_n)$ This brings us to the Cox Model:

$$\alpha_i(t) = \alpha_0(t) \exp(\mathbf{z_i}\beta^T) = \alpha_0(t) \exp(z_{i1}\beta_1 + z_{i2}\beta_2 + ... + z_{in}\beta_n) \quad (2.18)$$

This model is called a proportional hazard model because we can calculate the influence of the covariates in relation to another covariate. As we will see further down, we do not need to specify or estimate the baseline hazard to do this, which is an advantage of the model. First we will propose a way to

estimate β. Note that this model assumes that the influence of the covariates are not varying over time. We take a discrete approach again by defining the event-times $0 < t_1 < t_2 < ... < t_N$ and denote the set of index numbers of patients at risk just before time t by R_t. Now to explain how we estimate β through a maximum likelihood estimate, we state that at time t_i an individual makes a transition. Given that there is a transition, the conditional probability of this to be the specific individual X_j is

$$\frac{\alpha_j}{\sum_{k \in R_{t_i}} \alpha_k} = \frac{\alpha_0 \exp(\mathbf{z_j}\beta^T)}{\sum_{k \in R} \alpha_0 \exp(\mathbf{z_k}\beta^T)}. \tag{2.19}$$

In this formula, the baseline hazard cancels out. Cox proposed to take the product of these probabilities and maximize this so called 'partial likelihood' to find an estimate for β.

$$L(\beta) = \prod_{i=1}^{N} \left(\frac{\exp(\mathbf{z_j}\beta^T)}{\sum_{k \in R_{t_i}} \exp(\mathbf{z_k}\beta^T)}\right)^{C_j} \tag{2.20}$$

The model is chosen in such a way that the influence of the covariates is not changing over time. We use this fact in the interpretation of the proportional hazards. In a comparison we want to fix all but one element of the vector Z_i and Z_j, with the only different element in these vector being Z_{ik} and Z_{jk}. Assuming we have estimated the vector β, we calculate the proportion of these two hazard rates by

$$\frac{\alpha_i}{\alpha_j} = \frac{\alpha_0 \exp(\mathbf{z_i}\beta^T)}{\alpha_0 \exp(\mathbf{z_j}\beta^T)} = \frac{\exp(\mathbf{z_i}\beta^T)}{\exp(\mathbf{z_j}\beta^T)} = \exp(\beta(Z_{ik} - Z_{jk})). \tag{2.21}$$

The outcome of this formula is the factor that the hazard rate will be multiplied with, every time that the covariate value Z_{ik} jumps a step of $Z_{ik} - Z_{jk}$.

2.5 A Practical Example of Estimation using R

To illustrate the theory of estimation and the partial likelihood formula above, we will conclude this chapter with a further discussion of our example. We start with a subgroup analysis. Estimation just after having a transplant is sometimes not very informative. If we use dynamic prediction, we could asses the situation again after a 100 days. Now, the prospects for a patient still being in state 1, or a patient already in state 4 could be very different. We take another factor into account, for example the difference between patients in age class $20 - 40$ and age class > 40. To generate this picture, we first selected the subgroups with the covariance factors we are interested in and perform estimation on this group(see appendix 1). This yields to the diagrams in figure 2.2.

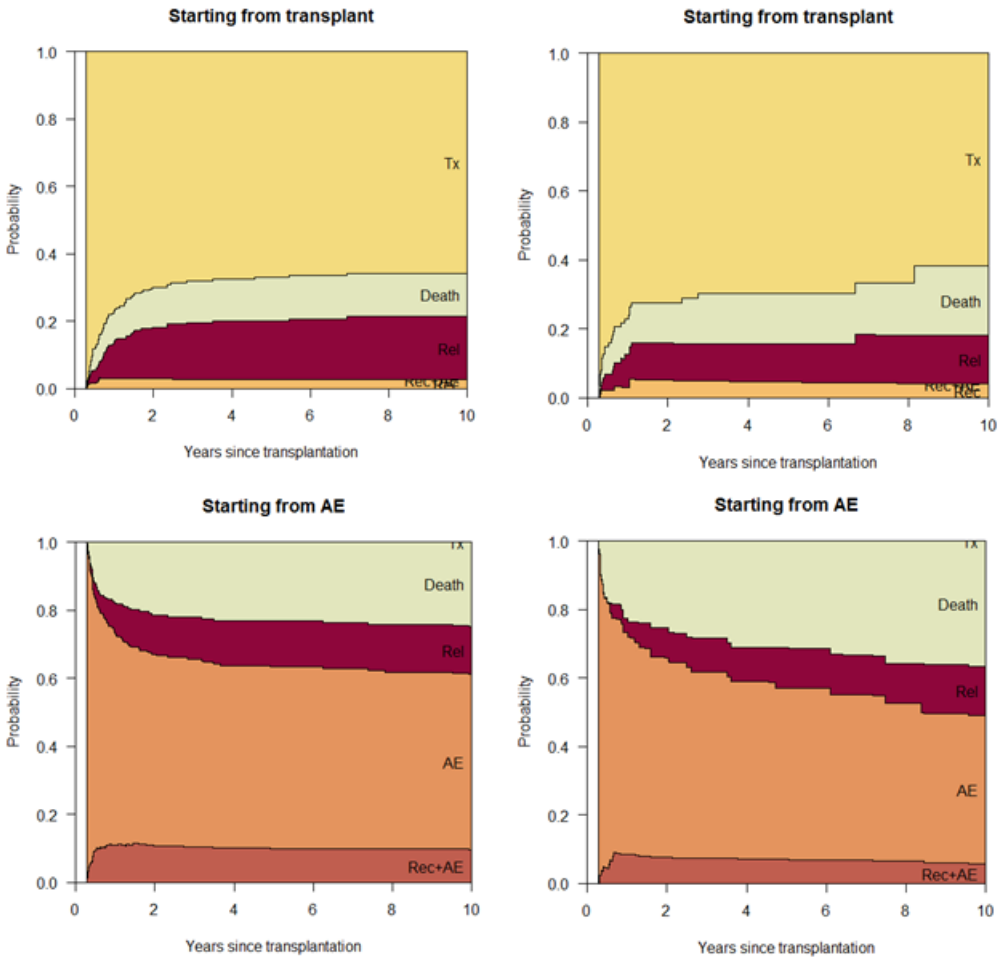

Figure 2.2: Estimation in ebmt4, age class $20-40$ on the left, age class >40 on the right

We can see, as expected, that people in age class 20-40 have a better perspective, as well as people not having had any transitions. A problem in this analysis is that by selecting a subgroup the dataset gets smaller, which is not desirable. The next step we take is the Cox model, in this estimation we take all data into account. We can ask the programming language R to estimate the covariate values in the Cox model for our ebmt4 data set and show part of the output in figure 2.3. We concentrate on the age subgroups again, the first twelve lines representing the transitions for patients with age $20 - 40$, the next twelve for patients > 40. We show part of the output, which contains, amongst other things, the standard error and p-value, marking the significant values in the table with stars. Positive regression factors cause the hazard rate to be higher. Negative ones are of a protective nature, causing the hazard rate to be lower. This table can serve as an indication of which parameters are interesting to further investigate. We can see that for example the regression coefficients of transition 8 are quite similar, the predictions will be much alike. For transition 7 and 12, the prediction for different age classes will be interesting, because the regression coefficients, and therefore the prediction, differ a lot here.

To continue our investigation of the influence of age we plot the perspective of these patients, being in state 1 after 100 days again, in figure 2.4. We see the estimation is more smooth than in figure 2.2, which is caused by the use of the whole dataset instead of selecting a subgroup at the start of estimating.

```
              coef exp(coef)  se(coef)      z Pr(>|z|)
agecl1.1   0.049112  1.050338  0.088876  0.553 0.580543
agecl1.2   0.123340  1.131268  0.082832  1.489 0.136481
agecl1.3  -0.093608  0.910639  0.232175 -0.403 0.686816
agecl1.4   0.766029  2.151207  0.228562  3.352 0.000804 ***
agecl1.5   0.292378  1.339610  0.187504  1.559 0.118922
agecl1.6  -0.255416  0.774594  0.223056 -1.145 0.252178
agecl1.7   0.150255  1.162130  0.490927  0.306 0.759557
agecl1.8  -0.393161  0.674920  0.116296 -3.381 0.000723 ***
agecl1.9   0.172498  1.188269  0.366786  0.470 0.638144
agecl1.10  0.237579  1.268176  0.205262  1.157 0.247090
agecl1.11  0.414041  1.512919  0.250179  1.655 0.097929 .
agecl1.12  0.759524  2.137260  0.272286  2.789 0.005280 **
agecl2.1   0.199439  1.220718  0.102441  1.947 0.051550 .
agecl2.2   0.067313  1.069630  0.100945  0.667 0.504883
agecl2.3  -0.232173  0.792809  0.322286 -0.720 0.471281
agecl2.4   0.934224  2.545238  0.264412  3.533 0.000411 ***
agecl2.5   0.470068  1.600102  0.205277  2.290 0.022026 *
agecl2.6  -0.100720  0.904186  0.264126 -0.381 0.702957
agecl2.7   1.464513  4.325437  0.481106  3.044 0.002334 **
agecl2.8  -0.327631  0.720629  0.142436 -2.300 0.021437 *
agecl2.9   0.422855  1.526313  0.432526  0.978 0.328252
agecl2.10  0.494656  1.639934  0.236828  2.089 0.036738 *
agecl2.11  0.256183  1.291989  0.304150  0.842 0.399626
agecl2.12  1.336744  3.806630  0.287013  4.657 3.20e-06 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 2.3: Estimation in ebmt4 covariates in Cox model

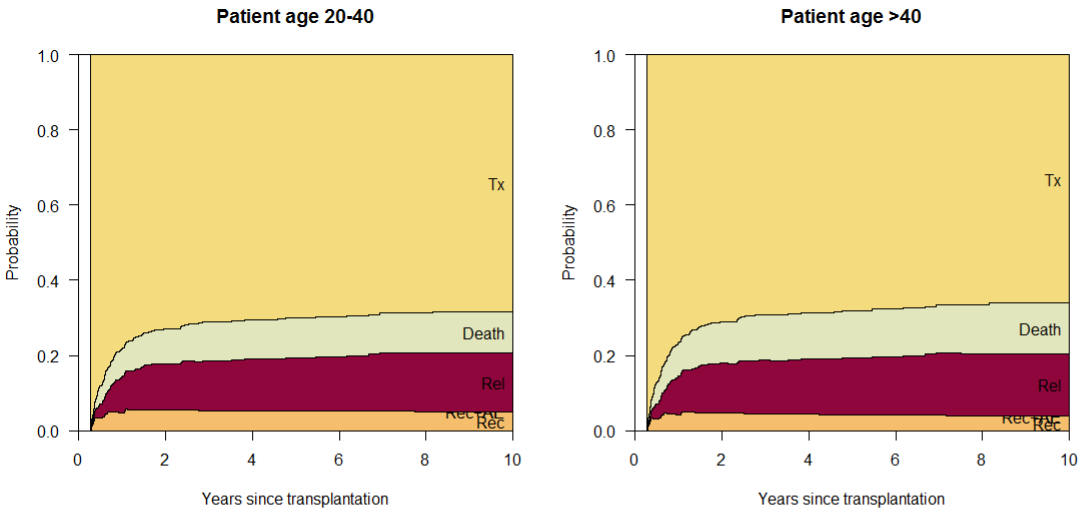

Figure 2.4: Estimation in ebmt4 using the Cox model

Chapter 3

Discussion

In this thesis we have seen a hazard based multi-state model which can be used in varying practical situations with incomplete data. We can make predictions taking time and other covariates into account, which makes it very useful for getting insight in, for example, the course of a disease. The reader should keep in mind that this model is still a simplification of reality, and some assumptions are made. Sometimes these are doubtful for the application we want to model, so we want to advance or change the model. A big assumption is made by stating our model is Markov. For example, we don't take into account the history of where and how long a patient has been in earlier states, but this might be very relevant for the prediction of diseases. A starting point to learn more about non-Markov multi-state models is [7]. Another assumption we made was the use of a semi-parametric Cox model. We assume that the influence of these covariates are proportional in a specific way and that this influence doesn't change over time. If, on the other hand, the influence changes a lot over time, we have to take another approach. In [2] the Cox-model is studied in more detail and with time dependent covariates as well. Another problem that might occur is the long time of calculation of the covariance of the Aalen-Johansen estimator. Solutions for this problem have been found, of which information can be found in, amongst others, [1]. We conclude that the model of this thesis is a good starting point in estimating transition probabilities with censored data, with a broad range of refinement and possibilities to investigate from here.

Acknowledgements

I would like to thank the EBMT organisation for providing the data used in this thesis. I would like to thank Cristian Spitoni for guiding me to and through the interesting subject of this thesis en its enormous amount of literature available. I thank Erna, for all her useful remarks, even though she is not a mathematician at all. And mostly Thomas, for keeping up with my mood and feelings, which were varying between a broad range of positive and negative ones in the process of writing this thesis.

Appendix 1

```
library("mstate")
data("ebmt4")
head(ebmt4)
ebmt <- ebmt4
tmat <- transMat(x = list(c(2, 3, 5, 6), c(4, 5, 6),
 c(4, 5, 6), c(5, 6), c(), c()),
names = c("Tx", "Rec", "AE", "Rec+AE", "Rel", "Death"))
print(tmat)

msebmt <- msprep(data = ebmt, trans = tmat, time = c(NA, "rec", "ae",
 "recae", "rel", "srv"), status = c(NA, "rec.s", "ae.s", "recae.s",
 "rel.s", "srv.s"), keep = c("match", "proph", "year", "agecl"))

covs <- c("match", "proph", "year", "agecl")
msebmt <- expand.covs(msebmt, covs, longnames = FALSE)
msebmt[, c("Tstart", "Tstop", "time")] <- msebmt[, c("Tstart","Tstop",
 "time")]/365.25

msebmt0 <- msebmt[msebmt$agecl == "20-40", -c(15:48, 61:84)]
msebmt1 <- msebmt[msebmt$agecl == ">40", -c(15:48, 61:84)]

c0 <- coxph(Surv(Tstart, Tstop, status) ~ strata(trans), data = msebmt0,
 method = "breslow")
msf0a <- msfit(object = c0, vartype = "aalen", trans = tmat)

c1 <- coxph(Surv(Tstart, Tstop, status) ~ strata(trans),
 data = msebmt1, method = "breslow")
msf1a <- msfit(object = c1, vartype = "aalen", trans = tmat)

library("colorspace")
statecols <- heat_hcl(6, c = c(80, 30), l = c(30, 90),
 power = c(1/5, 2))[c(6, 5, 3, 4, 2, 1)]
```

```
pt100a <- probtrans(msf0a, predt = 100/365.25, method = "aalen")
plot(pt100a, ord = c(2, 4, 3, 5, 6, 1), xlab = "Years since transplantation",
 main = "Starting from transplant", xlim = c(0, 10), las = 1,
 type = "filled", col = statecols[ord])

pt100b <- probtrans(msf1a, predt = 100/365.25, method = "aalen")
plot(pt100b, ord = c(2, 4, 3, 5, 6, 1), xlab = "Years since transplantation",
main = "Starting from transplant", xlim = c(0, 10), las = 1,
 type = "filled", col = statecols[ord])

plot(pt100a, from = 3, ord = c(2, 4, 3, 5, 6, 1),
 xlab = "Years since transplantation",
 main = "Starting from AE", xlim = c(0, 10), las = 1,
 type = "filled", col = statecols[ord])
plot(pt100b, from = 3, ord = c(2, 4, 3, 5, 6, 1),
 xlab = "Years since transplantation",
 main = "Starting from AE", xlim = c(0, 10), las = 1,
 type = "filled", col = statecols[ord])

cfull <- coxph(Surv(Tstart, Tstop, status) ~ match.1 +
 match.2 + match.3 + match.4 + match.5 + match.6 + match.7 +
 match.8 + match.9 + match.10 + match.11 + match.12 +
 proph.1 + proph.2 + proph.3 + proph.4 + proph.5 + proph.6 +
 proph.7 + proph.8 + proph.9 + proph.10 + proph.11 +
 proph.12 + year1.1 + year1.2 + year1.3 + year1.4 +
 year1.5 + year1.6 + year1.7 + year1.8 + year1.9 + year1.10 +
 year1.11 + year1.12 + year2.1 + year2.2 + year2.3 +
 year2.4 + year2.5 + year2.6 + year2.7 + year2.8 + year2.9 +
 year2.10 + year2.11 + year2.12 + agecl1.1 + agecl1.2 +
 agecl1.3 + agecl1.4 + agecl1.5 + agecl1.6 + agecl1.7 +
 agecl1.8 + agecl1.9 + agecl1.10 + agecl1.11 + agecl1.12 +
 agecl2.1 + agecl2.2 + agecl2.3 + agecl2.4 + agecl2.5 +
 agecl2.6 + agecl2.7 + agecl2.8 + agecl2.9 + agecl2.10 +
 agecl2.11 + agecl2.12 + strata(trans), data = msebmt,
 method = "breslow")


whA <- which(msebmt$agecl == "20-40")
patA <- msebmt[rep(whA[1],12),9:12]
patA$trans <- 1:12
attr(patA, "trans") <- tmat
patA <- expand.covs(patA, covs, longnames = FALSE)
patA$strata <- patA$trans
msfA <- msfit(cfull, patA, trans = tmat)
ptA <- probtrans(msfA, predt = 100/365.25)
plot(ptA, ord = c(2,4,3,5,6,1), main = "Patient age 20-40", las = 1,
```

```
 xlab = "Years since transplantation", xlim = c(0,10),
 type = "filled", col = statecols[ord])

whB <- which(msebmt$agecl == ">40")
patB <- msebmt[rep(whB[1],12),9:12]
patB$trans <- 1:12
attr(patB, "trans") <- tmat
patB <- expand.covs(patB, covs, longnames = FALSE)
patB$strata <- patB$trans
msfB <- msfit(cfull, patB, trans = tmat)
ptB <- probtrans(msfB, predt = 100/365.25)
plot(ptB, ord = c(2,4,3,5,6,1), main = "Patient age >40", las = 1,
 xlab = "Years since transplantation", xlim = c(0,10),
 type = "filled", col = statecols[ord])
```

Bibliography

[1] P.K. Andersen, O. Borgan, R.D. Gill and N. Keiding. *Statistical Models Based on Counting Processes*. Springer-Verlag, New York, 1992.

[2] P.K. Andersen and R.D. Gill, Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annuals of Statistics*, volume 10, no. 4, 1100-1120, 1982.

[3] R. D. Gill and S. Johansen, A Survey of Product-Integration with a View Towards Application in Survival Analysis. *The Annuals of Statistics*, volume 18, no. 4, 1501-1555, 1990.

[4] H. Putter, L. de Wreedea and M. Fiocco. *mstate: Data Preparation, Estimation and Prediction in Multi-State Models.* Available on https://cran.r-project.org/, package 'mstate' version 0.2.10, 2016.

[5] J. A. Rice.*Mathematical Statistics and Data Analysis.* Third edition, Books/Cole cengage learning, Belmont, 2007.

[6] T. Therneau, C. Crowson and E. Atkinson. *Multi-state models and competing risks.* 2016, available at https://cran.r-project.org/web/packages/survival/vignettes/compete.pdf.

[7] J. de Una-Alvarez. Recent Developments in Censored, Non-Markov Multi-State Models. *Combining Soft Computing and Stats. Methods.* AISC 77, p173-179, Springer-Verlag, Berlijn, 2010.

[8] L. de Wreede, M. Fiocco and H, Putter. mstate: An R Package for the Analysis of Competing Risks and Multi-State Models 1994. *Journal of statistical software*, volume 38, issue 7, 2011.

[9] *The Comprehensive R Archive Network*, https://cran.r-project.org/.