UNIVERSITA' DEGLI STUDI DI MILANO-BICOCCA
Corso di Dottorato in Epidemiologia e Biostatistica
XXVII Ciclo - III anno

DYNAMIC PREDICTION IN SURVIVAL ANALYSIS WITH BINARY NON-REVERSIBLE TIME-DEPENDENT TREATMENT INDICATOR

Davide Paolo Bernasconi

Tutor: Dr.ssa Laura Antolini

Supervisore: Prof.ssa Maria Grazia Valsecchi

Anno Accademico 2014/2015

"If everything is coming your way,
chances are you're in the wrong lane."

Luke Duke

Acknowledgements

With these few but sincere words, I would like to express my greatest appreciation and gratitude to Prof. Maria Grazia Valsecchi and Prof. Stefania Galimberti, for giving me the chance to join this PhD course and to work in their extraordinary research group.

A very special thanks goes to my tutor, Dr. Laura Antolini, for the brilliant intuitions, the involving enthusiasm and the endless patience she dedicated to my work.

Thanks to Anita, Emanuela, Ilia, Jessica, Maria Chiara, Matteo, Paola, Roberta, and all the wonderful people who have been working at the Center of Biostatistics for Clinical Epidemiology in Monza during the last four years: thank you all for the great time we had together, the scientific discussions, and for your friendship.

I also wish to thank Prof. Hein Putter from the Leiden University Medical Center, for the opportunity I have been given to spend a nice research period working in his Department.

Mostly, thanks to Marina and my family for their love.

Table of contents

1 Introduction

The survival experience of two groups defined at baseline with a categorization fixed in time, for instance according to treatment, is generally summarized non-parametrically by the Kaplan-Meier (KM) estimator of the cumulative survival probability (1958). The interpretation of the obtained curves is clear: they estimate over time the survival probability of being free from failure that a patient would expect if assigned to each treatment group/covariate level. In some situations, however, the group allocation is dynamic in time allowing patients to switch from one group to another one. This may occur when there is a treatment switching during the follow-up, where patients either: i) start from an initial treatment and continue it, ii) start from an initial treatment and switch to an alternative one after some time, iii) start directly from the alternative treatment since the beginning of the follow-up**.** In this context, the treatment received in time is a **binary non reversible time-dependent (BNRTD) covariate** which starts, for instance, from a value of 0 (initial treatment) and may switch to 1 (alternative treatment) at some time point. A BNRTD variable can also be defined by the achievement of some intermediate event in event-free patients at the beginning of the follow-up, such as response to treatment. A typical example of BNRTD treatment is the chemotherapy vs stem-cell transplant treatments in Acute Lymphoblastic Leukemia (ALL). Each patient is treated with chemotherapy since the starting point (e.g. first complete remission) and, after a certain amount of time (waiting time), he/she could receive transplant. In this case, the treatment switching can depend only on something external to the disease process (e.g. the chance of having a donor), but an "intention-to-treat analysis" cannot be performed due to the absence of randomization (Brand et al. 2013). On the other hand, a "treatment-performed" analysis by classifying patients on the ground of the BNRTD level at the end of the follow-up would not lead to a meaningful interpretation of the obtained curves (Beyersmann et al. 2008, Snapinn et al. 2005, Dubin et al. 2001, Webster 2011). This approach has been adopted and discussed several times in the medical literature, involving epidemiologic studies on cardiovascular diseases (Gail 1972, Mathew et al. 2001), respiratory diseases (Suissa 2003), AIDS (Glesby and Hoover 1996), cancer (Giobbie-Hurder 2013) and even sociologic studies (Rothman et al. 1992, Sylvestre et al. 2006, Hanley et al. 2006). The prevalence and the impact on the conclusions of this type of analysis in the published literature were investigated by Van Walraven et al. (2004).

Let us consider the case of chemotherapy vs transplant example. The BNRTD variable level at the end of the follow-up results from the competing action between the potential survival time the patient would have under chemotherapy-only and the potential waiting time to transplant. Thus,

patients are classified according to the "treatment performed" classification into the chemotherapy group or in the transplant group depending on whether the potential survival time under chemotherapy-only is lower or equal than the potential waiting time or not. As a consequence, the observed survival time of a patient in the transplanted group will be necessarily greater than his/her potential waiting time, and the observed failure time of a patient in the non-transplanted group will be necessarily shorter or equal than his/her latent waiting time. This is commonly referred in the literature as "immortal time bias" on the failure time data observed for the transplant group and "mortal time bias" on the failure time data observed for the chemotherapy group. To remove the aforementioned intuitive bias of the data, a "rescaling approach" of the failure times was proposed. For a patient who switched to the alternative treatment (or who experienced the intermediate event), the starting point is his/her waiting time to treatment switch/intermediate event; for a patient who did not switch the treatment/experience the intermediate event, the starting point is the median of the observed waiting times among patients of the other group. Patients who did not switch treatment (experience the intermediate event) and who developed failure before the median waiting time are excluded from the analysis. The KM estimator is applied according to the "treatment performed" classification after the "rescaling" of the failure times. An example of application is the study by Arnold et al. (2003) where the aim was to investigate the effect of the occurrence of heart failure on overall survival. This approach would seem an intuitive way for removing the "immortal time bias" from the original data by recalculating the failure time directly from the waiting time, and then by compensating this in the group of patients who did not switch the treatment (experience the intermediate event), by recalculating their failure time from the median of the waiting times. As noted by Snapinn et al. 2005, however, the "mortal time bias" in this latter group is still intuitively present since the original observed failure time will be necessarily shorter or equal than his/her latent waiting time. In addition, the theoretical meaning of the obtained estimates is not clearly defined in the literature.

The theoretical survival probabilities to gain insights into the BNRTD treatment impact are in fact not well established. Alternative applications of the KM estimator were however proposed, mostly relying on a heuristic bases and without specifying the theoretical quantity to estimate.

The "**landmark** classification" approach was motivated by the need of comparing responders to non-responders to some treatment, where the achievement of first response is the BNRTD variable. Anderson et al. 1983 proposed to classify patients into two groups according to the achievement or not of response at some fixed time LM in the follow-up called "landmark" (LM), i.e. according to the BNRTD variable level at the LM time. The starting point for the calculation of the survival time

for all patients is the LM ("landmark rescaling") and patients who developed failure before the LM are excluded from the analysis regardless of the BNRTD variable value at the LM time. In practical applications the LM is usually chosen relatively close to the beginning of the follow-up. The KM estimator is applied according to the LM classification measuring time from LM. The obtained curves are interpreted as estimates of the survival probability from failure conditional on being free from failure at the LM time and according to the BNRTD classification at the LM time. Curves can be graphically displayed starting from LM but on the original time axis (i.e. from time 0) leaving blank the part of the graph between 0 and LM to denote the conditioning.

This method was extensively used in applied literature even when the BNRTD variable is an intervention, in particular in the context of cardiovascular disease, comparing for instance the survival rate of patients receiving heart transplant and controls. A nice review can be found in the work of Dafni et al (2011). This approach requires the set-up of the LM time which may be a controversial choice that relies on a trade-off between empirical and clinical/interpretative considerations. Let us consider for instance the transplant versus chemotherapy example. From an empirical point of view, the LM specification should be early in time to keep under control the fraction of patients who are excluded from the analysis, and forward in time to achieve a satisfactory sample size of the two groups. The LM set up is also directly involved in the clinical interpretation of the theoretical meaning of the estimates, since the pragmatic classification is performed at the LM time and the dynamic evolution of the BNRTD variable after the LM time is somehow ignored.

In 1984, **Simon and Makuch** proposed a further method to describe the survival experience of two groups defined by a BNRTD variable. The approach is also known in literature as "extended Kaplan-Meier" method (Snapinn et al. 2005). The basic idea consists in estimating the hazard involved in the KM formula in groups of patients defined according to BNRTD variable dynamically updated in time. In the transplant versus chemotherapy example, patients contribute to the hazard estimate of the chemotherapy group up to their waiting time, and from that point on they start to contribute to the hazard estimate of the transplant group. This, in other words, means that the transplant the survival curve is obtained from transplanted patients considering the waiting time to transplant as the (delayed) entry time, and the chemotherapy survival curve is obtained from all patients considering the waiting time to transplant as a cause of censoring. This approach is consistent with the rationale of the Mantel-Byar test which relies on the hazards of failure dynamically updated in time (i.e. accounting for treatment switch) (1974). If at the beginning of the follow-up there are no patients having the BNRTD variable equal to 1, the initial estimate of the

hazard needs to be necessarily obtained in two groups of patients defined by the BNRTD variable value at some initial landmark (ILM) time. All failure times are consistently measured from ILM and patients with failure time lower than the ILM are excluded from the analysis. Again, the probabilistic meaning of the obtained curves was not clarified in the literature. Quoting Beyersmann et al. (2005), "these [curves] cannot have an interpretation as survival probabilities". Despite the simplicity of the Simon and Makuch (SM) estimator and its frequent use in medical literature, there is still no evidence of what this curves represent. In addition, the dynamic update in time of the risk set for transplanted patients intrinsically assumes that the hazard under transplant does not depend of the time since transplant. This assumption is not realistic for the action of an aggressive treatment such as transplant which causes an initial peak of hazard.

In practical applications it is of interest to estimate the survival experience of patients conditional on being failure-free at some LM time points during the follow-up, regardless of whether or not we are in the presence of a BNRTD treatment. This approach in the literature is known as "**dynamic prediction**". Roughly speaking it answer questions such as:
<<Let us consider a patient failure free at a given time point, what is his/her expectation of surviving failure free forward in time?>>.
Unlikely in the aforementioned LM and SM methods, these new LM time points are chosen depending on clinical considerations and can be even very far from the beginning of the follow-up.
The type of questions posed in the dynamic prediction framework arise naturally after the basic one, posed at the beginning of the follow-up, relative to the "**fixed prediction"**: <<Let us consider a patient failure free at the beginning of the follow-up, what is his/her expectation of surviving failure free forward in time?>>.

When studying the dynamic of disease evolution in the presence of BNRTD treatment, such as transplant, it comes natural to investigate the role of the **waiting time to transplant** on the velocity of development of the disease. One could think to address this issue simply using a hazard based regression model, such as the Cox model which is the most popular tool to assess the relative impact of covariates accounting also for their time-dependent nature and for possible time-varying effects. The impact of treatment could be assessed by setting the waiting time as the (delayed) entry for transplanted patients. This would imply, however, assuming markovianity, i.e. the hazard of failure under transplant in time does not depend of the waiting time to transplant. One could think of overcoming this assumption including in the model also the waiting time as a time-dependent covariate. Doing so, however, would induce to observe a spurious effect of the waiting time which

would be indeed a mixture including the real effect of the waiting time on the velocity of disease development and the effect of the time past transplant (sojourn time).

A further aspect that drives to the need of regression models, is to account for further covariates on top of the BNRTD treatment in estimating the survival function. Especially from the point of view of a newly diagnosed patient, it comes natural to focus on **profile-specific prediction** accounting for individual characteristics to target treatment options. To address the issue of profile-specific survival curves, the Cox model presents two strong limitations: i) curves cannot be obtained in the presence of time-varying effects and/or time-dependent covariates (Therneau, 2000), ii) curves will assume a non-attractive steps-in-time shape, due to the nature of the baseline hazard estimator (Julien and Hanley, 2008).

Standard fully parametric models, such as Weibull model, represent a possible alternative. Their reliability, however, depends on the correct specification of the underlying density distribution of time. As a consequence, in non-standard situations, one should move towards flexible modeling solutions which do not require assuming a well-defined density distribution of time, and enable to account for time-varying effects and time-dependent covariates. These models could be hard to handle computationally and to interpret clinically, but they could be suitable to provide profile specific survival prediction curves.

The **Hanley-Miettinen (HM)** model (2009) is a fully parametric model based on the instantaneous hazard, in analogy with the Cox model. It consists in using the standard logistic regression model to estimate the hazard function over time as follows. An expanded dataset is created by splitting the observed survival time of each subject into a number of time-units with the pertinent updated covariate values, including treatment. The number of events in every single interval is thought as a Binomial random variable which is used as the response variable of the logistic model where time is included as a covariate with possible interaction with treatment. To overcome the computational issues arising from this kind of approach, due to the potentially huge size of the expanded dataset, Hanley and Miettinen reviewed and extended a method developed by Mantel (1973) which is based on a random sampling of the individual time-units which is accounted by a suitable offset term in the model.

The semi-parametric **"landmark regression model"** (Van Houwelingen 2007, Van Houwelingen and Putter, 2008, 2012) was specifically conceived to address the issue of dynamic prediction of the cumulative incidence at LM times up to a fixed-width time window. It is based on a particular application of the Cox model as described in the following. A dens grid of LMs is set and for each

LM a data set is created using subjects event free at LM who are artificially censored at the end of the time-window. Treatment classification and the value of further covariates are frozen at each LM. Once data sets are stacked together, a single Cox model is fitted using LM both as entry time and as a covariate and including an interaction between treatment and LM. Since we use the LM acronym to denote the non-parametric "landmark" method, to avoid ambiguities we refer to this regression model as the "LMR" model.

This thesis was motivated from clinical questions arising in a multicenter study on pediatric patients affected by Philadelphia chromosome-positive (Ph+) Acute Lymphoblastic Leukemia (ALL). Patients initially received chemotherapy with possible treatment switch to additional stem cell transplant (SCT) (Aricò et al. 2010).

The main aims are reviewing and developing methods to:

1. describe non-parametrically the survival experience according to a BNRTD treatment resorting to fixed and dynamic prediction functions;
2. use regression models to assess the impact of the waiting time to treatment switch on the hazard of failure after treatment switch through interpretable parameters;
3. use regression models to develop profile-specific predictions in the presence of a BNRTD treatment.

Concerning the methodological aspects, the thesis is organized in two parts: the first deals with aim 1, the second deals with aims 2 and 3.

The goal of **part one** is to clarify the theoretical quantities estimated by the LM and SM methods. After discussing the limitations of these methods, in terms of both interpretation and estimation characteristics, I present a novel approach based on counterfactual questions and dynamic prediction, and an estimation method based on measuring time since transplant (clock-back scale) for transplanted patients. The validity of the proposed method is discussed theoretically and checked through a simulation protocol.

In **part two,** I show how to properly estimate the effect of waiting time to transplant on the velocity of disease development using a hazard-based standard regression model, such as the Cox model (aim 2). Moreover, I review the HM and LMR regression models and discuss their use to make profile-specific dynamic predictions on the motivating ALL data (aim 3).

On both aims 2 and 3, the approach of measuring time in the original scale (clock-forward) is contrasted to that of measuring time in the clock back scale. The impact of this choice is discussed theoretically and checked through a simulation protocol. Specifically, a simulation protocol is set to:

i) compare the clock-back and the clock-forward scale to assess the impact on the hazard function of the waiting time variable using the Cox, HM and LMR models,
ii) compare the performance of HM and LMR models with both the time-scales to provide treatment-specific dynamic predictions.

Concerning the organization in chapters, the thesis is structured as follows:

- In **Chapter 2**, the context of application is described, setting the notation and describing the motivating example on transplant vs chemotherapy data. The distribution of the waiting time to transplant and the hazard functions under the two treatments is investigated, resorting also to multistate nonparametric estimators to evaluate state probabilities in time.

Part I consists in 4 chapters:

- In **Chapter 3** the LM and SM estimators are reviewed and the meaning of the corresponding theoretical quantities is identified. In addition, the application of these methods to the motivating example are shown.
- In **Chapter 4** a modification of the SM approach for transplant is proposed to address the issue of semi-Markovian data (Putter et al. 2007).
- In **Chapter 5** a novel LM approach to answer questions on dynamic prediction is proposed.
- In **Chapter 6** a simulation protocol is set up to investigate the estimators' performance.

Part II consists in 5 chapters:

- In **Chapter 7** I discuss the issue of assessing the impact of waiting time to transplant on the hazard of failure after transplant using regression models; specifically I talk about the choice of the proper time scale to fit the purpose.
- In **Chapter 8** I define the fixed width conditional incidence failure function which represents the quantity to estimate to develop (profile-specific) dynamic prediction curves.
- In **Chapter 9** I review the two flexible models to make dynamic predictions: the Hanley-Miettinen (HM) model and the landmark regression (LMR) model.

- In **Chapter 10** I apply the methods discussed, both for estimating the effect of waiting time and for dynamic predictions, on the ALL data. I compare the predictive ability of the two flexible models using the time-dependent Brier score.
- In **Chapter 11** a simulation protocol is set up to investigate the performance of the methods discussed both for the assessment of the effect of waiting time and for dynamic prediction.

- **Chapter 12** contains a discussion of the issues and findings presented.

2 Context and notation

2.1 Motivating Example: Chemotherapy vs stem-cell transplant in pediatric leukemia

The data used for illustration are relative to a multicenter study (Aricò et al. 2010) on a cohort of 542 children, aged less than 18 years, affected by Philadelphia chromosome-positive Acute Lymphoblastic Leukemia (ALL). All enrolled patients were in first complete remission and were treated with chemotherapy only (217) or with chemotherapy plus hematopoietic stem-cell transplantation (325) performed at different points in time. The end point considered was disease free survival (DFS), defined as the time from complete remission until relapse or death during remission, whichever comes first. Other baseline covariates included in the analysis were: sex, age at diagnosis, leukocyte count at diagnosis, early response according to bone-marrow result or peripheral-blood result. A graphical representation of the multistate structure of the application is provided in figure 1, which is consistent with an *illness-death model* where the intermediate state is an intervention rather than an intermediate event monitored during the disease course.

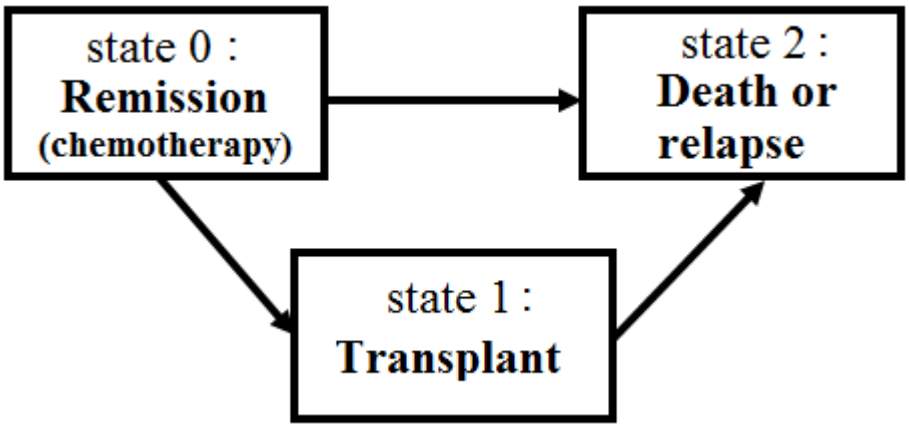


Figure 1. Graphical representation of transitions between states. A total of 325 (60.0%) patients out of 542 were observed to be switched in time to transplant (from state 0 to state 1). Among the 325 patients who were observed to enter state 1, 152 were right censored while 173 developed failure (from state 1 to state 2). A total of 217 (40.0%) patients out of 542 were either right censored while under chemotherapy (55) or developed failure under chemotherapy (162, from state 0 to state 2).

2.2 Notation

Let us define the following random variables which have an explicit reference to this motivating example:

- T_{dh} is the latent time to failure from remission under chemotherapy,

- W is the latent time from remission to switch to transplant (i.e. waiting time),
- $\mathrm{E} = I(W \le T_{ch})$ is the BNRTD final treatment indicator status,
- T_{ptr} is the latent time to failure from transplant,
- $T_{tr} = W + T_{ptr}$ is the latent time to failure from remission going through transplant,
- T is the time to failure,
- C is the censoring time.

Time variables are referred as latent since, besides the limitation of observation due to standard censoring, their observability depends on the joint behavior of T_{ch} and W. Specifically, the latent failure time T_{ch} is observable only if $W > T_{ch}$ $(\mathrm{E} = 0)$, and in this case $T = T_{ch}$. The latent waiting time W and the subsequent latent failure time T_{tr} are observable only if $W \le T_{ch}$ $(\mathrm{E} = 1)$, and in this case $T = T_{tr}$. This overall leads to the time to failure

$$T = (1-\mathrm{E}) \cdot T_{ch} + \mathrm{E} \cdot T_{tr} \cdot \qquad (1)$$

Although all patients are potentially eligible to transplant, the feasibility of transplant depends mainly on the availability of a matched related donor. Thus, we can hypothesize that treatment assignment, as well as the waiting time to transplant, are mainly due to genetic chance. This implies that: i) the latent times T_{ch} and W are independent, which also implies independence between the latent times T_{ch} and T_{tr}, and ii) the distribution of the latent waiting time to transplant W reflects the genetic chance process of the donor availability.

With reference to each patient $i=1,\ldots,n$ I set the following notation:

- t_i is the observed time from remission to failure (either under chemotherapy or under transplant) or censoring;
- δ_i is the observed event indicator, where $\delta_i = 1$ if t_i is a failure time, while $\delta_i = 0$ if t_i is a censoring time;
- ε_i is the observed treatment indicator at the end of the follow-up, where $\varepsilon_i = 1$ if the patient i was transplanted before t_i, while $\varepsilon_i = 0$ if the patient i either failed or was censored under chemotherapy;
- w_i is the latent waiting time to transplant, which is observed only if $\varepsilon_i = 1$.

The vector of observed data $(t_i;\delta_i;\varepsilon_i;\varepsilon_i \cdot w_i)\ \ i=1,...,n$ is the sampling realization of $(Z;\Delta;\mathrm{E};\mathrm{E}\cdot W)$ where $Z=\min(C;T)$, where T was defined in (1), $\Delta = I(T \le C)$, $\mathrm{E} = I(W<T)$. In addition, I denote by

- $t_{(j)}$ *(j=1,...,K)* the *K* observed ordered distinct failure times from the achievement of remission.
- $t^{s}_{(j)}$ *(j=1,...,K^s)* the *K*^{s} observed ordered distinct failure times from transplant, where the superscript "s" stands for "sojourn" time in the transplant state. These times pertain only to the subsample of patients who were observed to switch treatment.

2.3 Preliminary analyses of the application example data

The distribution of the waiting time to transplant is commonly heuristically investigated resorting only to the observed waiting times, that in my application are available on the 325 patients who were observed switching treatment. The corresponding cumulative distribution function represented in figure 2 (thin line) is an the estimate of $P\left(W \le t \mid (W<C)\cap(W<T_{ch})\right)$. It reaches 1 at the maximum observed waiting time of 1297 days (the minimum was 48 days). This behavior depends on the joint action of: i) donor availability process, ii) failure development under chemotherapy and iii) censoring process. It is worth of note that *W* can also be investigated resorting directly to the distribution of the latent waiting time by estimating $P(W \le t)$. The latent time *W* is in fact also known to be greater than the observed time provided by the 217 patients who were either censored (55) or who developed failure before transplant (162). This is accounted in the KM estimator represented in figure 2 (thick line) by considering failure before transplant (i.e. under chemotherapy) as an additional independent censoring on the latent waiting time *W*.

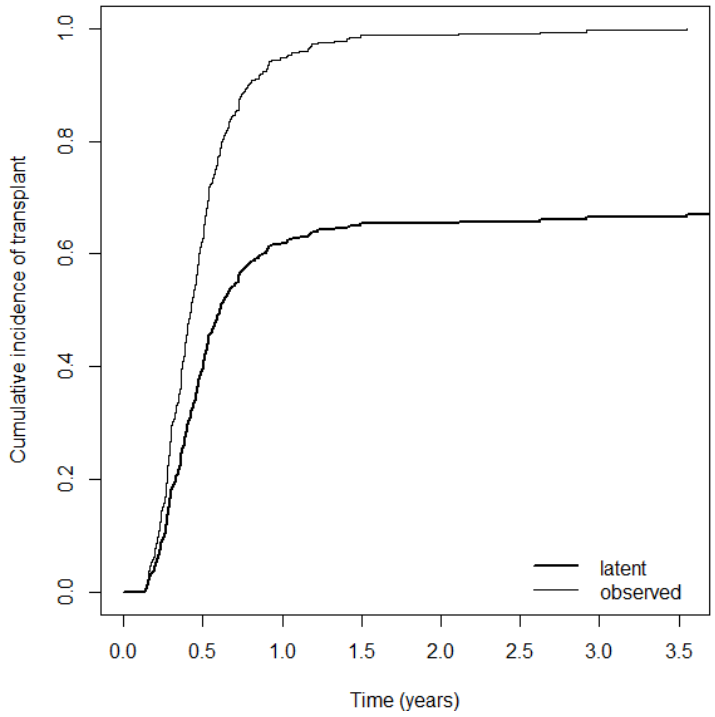


Figure 2. Cumulative incidence of observed (thin line, 325 patients) and latent (thick line, 542 patients) waiting time to transplant.

As expected, the distribution function of the latent waiting time W lies always below the distribution of the observed times since it includes also the partial contribution of the latent waiting time to transplant of patients who failed or were censored under chemotherapy. It provides a suitable summary of the latent donor availability process in time.

When analyzing data generated by a multi-state process such as in figure 1, it is important to check whether the Markov property is satisfied (Meira-Machado et al. 2006, Rodrigeuz-Girondo and de Una-Alvarez 2012), and more generally to choose an appropriate time scale and modelling for the hazard of failure after intermediate events (Iacobelli and Carstensen 2013). In general, a multistate process is markovian if the future of a patient depends only on the present state and not on the previous states and on transition times between them. In the example, under this assumption we would expect that the hazard of failure at time t (from remission) for a patient transplanted at $w<t$ (and thus with a sojourn time $s=t-w$) is equal to the hazard of failure at time t for a patient transplanted at a different $w^*<t$ (and thus with a sojourn time $s^*=t-w^*$). This property however is not plausible in the ALL data, while it has been shown that the failure rate increases sharply soon after transplant and then it starts to decrease (Aricò et al. 2010). Thus the hazard after transplant is expected to depend on the sojourn time (semi-Markov assumption) and could possibly depend also

on the waiting time (extended semi-Markov assumption, Putter et al. 2007, Meira-Machado et al. 2009). This has been investigated as follows. The hazard function was estimated separately in subgroups of patients with homogeneous waiting times, for instance defined from the quartiles of the distribution of the observed waiting time (106, 154 and 211 days) leading to 4 groups of patients (median waiting times equal to 84.5, 132, 177, 266 days respectively). The corresponding hazard estimates according to the sojourn time scale ("clock reset" or "clock back" scale Putter et al. 2007, Iacobelli and Carstensen 2013) are represented in figure 2. Since the hazard functions are almost overlapping, we conclude that the influence of the waiting time is negligible. This means that the hazard of failure at time t of a patient with a waiting time $w<t$ and a sojourn time s (i.e. s time units after transplant) in the transplant state ($t=s+w$) equals the hazard at t^* of a patient with a different waiting time $w^*< t^*$ and the same sojourn time s ($t^*=s+w^*$). The absence of a possible effect of waiting time on the hazard under transplant was also confirmed by the analysis of hazard models, namely through the score test of a Cox model fitted on the transplanted patients measuring time on the clock back scale (p-value=0.677). This approach is more extensively discussed in section 10.2. Thus, the data reasonably satisfy the semi-Markov assumption, which in terms of random variables means independence between W and T_{ptr}.

The hazard of failure under chemotherapy is estimated from remission on the whole sample, provided that the subjects who were switched to transplant are considered as censored at their switching time. This estimate is also represented in figure 3 (solid line) and is a monotone function decreasing less steeply than the hazard after transplant, but starting with a lower value. It can be noticed that, at any time, the treatment switching implies a jump in the hazard. For instance, a patient under chemotherapy at 6 months after remission experiences an hazard rate of about 0.44 per year, but if he/she would switch to transplant, his/her hazard would jump to 0.72 per year, as shown by the hazard of the third group. However, after some sojourn time in the transplanted state, he/she gains an advantage with respect to chemotherapy. For instance, at 2 years from remission, i.e. at 1.5 years from transplant, he/she experiences a hazard of about 0.14 which is lower than the hazard he/she would have experienced under chemotherapy (about 0.28).

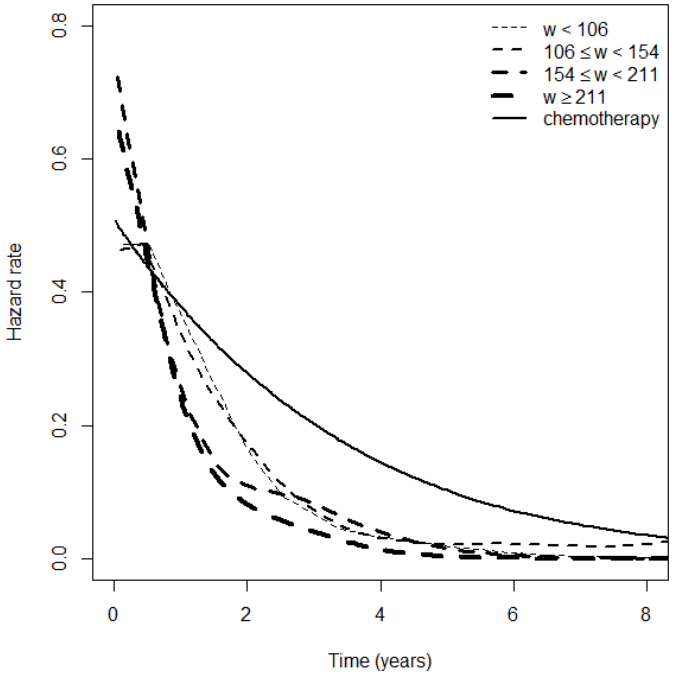


Figure 3. Smoothed non-parametric hazard of failure under chemotherapy (solid line) and in four subgroups of transplanted patients (dashed lines) with waiting times in intervals [48, 106) , [106, 154) , [154, 211) , [211, 1297] with corresponding medians equal to 84.5, 132, 177, 266 (days). The hazard functions under transplant are plotted from time since transplant, while the hazard function under chemotherapy is plotted in the original scale from remission. The thickness of the dashed lines corresponds to the quartiles of the observed waiting time distribution, where thicker lines correspond to increasing waiting times. The analysis was carried out using the "bshazard" R package by Rebora et al. (2014).

The joint action of the hazard of the latent T_{dh} (figure 3, solid line), the genetic chance that drives patients to be transplanted (figure 2, thick line), and the hazard of failure after transplant (figure 3, dashed lines) implies that at each time point any patient will be in one of four conditions: failure under transplant, failure free under transplant, failure under chemotherapy, failure free under chemotherapy. The pragmatic description of the observed probabilities in time can be obtained by the non-parametric Aalen-Johansen estimators in the multi-state model theory by the so called "stacked" prediction probabilities (figure 4, Putter et al. 2007, Meira-Machado et al. 2009, Andersen and Keiding 2002). These standard estimators are valid even in non-markov settings (Datta and Satten 2001, Glidden 2002).

Specifically, the probabilities of the four conditions mentioned before are obtained as follows: $\hat{P}(W \le t; W \le T_{ch}; T_{tr} \le t)$ = dotted-0 = failure under transplant, $\hat{P}(W \le t; W \le T_{ch}; T_{tr} > t)$ = dashed-dotted = failure free under transplant, solid-dashed = $\hat{P}(W > T_{dh}; T_{dh} \le t)$ = failure under chemotherapy, 1-solid = $\hat{P}(W > t; T_{ch} > t)$ = failure free under chemotherapy.

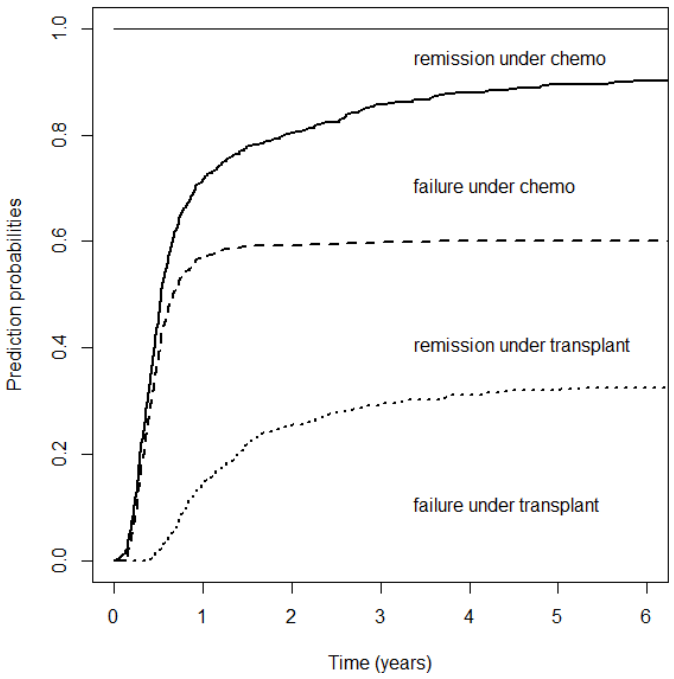


Figure 4. Stacked prediction probabilities at the beginning of the follow-up. Each state probability at time *t* is obtained by difference between the curves: dotted-0 = $\hat{P}(W \le t; W \le T_{ch}; T_{tr} \le t)$ = failure under transplant, dashed-dotted = $\hat{P}(W \le t; W \le T_{ch}; T_{tr} > t)$ = failure free under transplant, solid-dashed = $\hat{P}(W > T_{dh}; T_{dh} \le t)$ = failure under chemotherapy, 1-solid = $\hat{P}(W > t; T_{ch} > t)$ = failure free under chemotherapy.

Let us consider the time *t=1* (year). The value 0.568 of dashed curve estimates the probability $P(W \le t; W \le T_{ch})$ of being observed to be switched to transplant up to *t*, and similarly the value 0.432 of the distance between 1 and the dashed curve estimates the probability $1 - P(W \le t; W \le T_{ch})$ of not being switched to transplant up to *t* . These two probabilities are defined regardless of being event free or not at *t*, thus they can be subsequently split into two parts depending on the development or not of the end point. Specifically, the 0.568 of dashed curve is decomposed into the estimate of the probability $P(W \le t; W \le T_{ch}; T_{tr} > t)$ of being transplanted and event free at *t* (dashed – dotted = 0.568 – 0.146 = 0.422), and the estimate of the probability $P(W \le t; W \le T_{ch}; T_{tr} \le t)$ of failure under transplant up to *t* (dotted – 0 = 0.146 – 0 = 0.146). Similarly, the value obtained by the distance between 1 and the dashed curve at *t* (1 – 0.568 = 0.432) is decomposed into the estimate of the probability $P(W > t; T_{ch} > t)$ of being event free and under chemotherapy at *t* (1-solid = 1 – 0.714 = 0.286), and the estimate of the probability $\hat{P}(W > T_{dh}; T_{dh} \le t)$ of failure under chemotherapy up to *t* (solid – dashed = 0.714 – 0.568 = 0.146).

The estimation of the state probabilities provides a description of the survival experience and the BNRTD treatment administration. Thus, it refers to a joint distribution between treatment allocation and survival status, which does not allow us to get insights into the chance of survival conditional on the BNRTD treatment.

Part I

Non-Parametric Methods

In the first part of this thesis, my aim is to fill the lack of clarity about the established theoretical quantities to describe the survival experience in the presence of a BNRTD variable. First, I critically identify the quantities corresponding to the KM-type estimators used in the LM and SM approaches. Second, motivated from the chemotherapy vs stem-cell transplant application, I propose a novel approach based on counterfactual questions, which states that the survival experience under chemotherapy can be summarized resorting to the latent time of failure under chemotherapy-only. Concerning the survival experience under transplant, the novel approach states that it can be summarized resorting to the latent failure time under transplant:

i) as if transplant was administered at the beginning of the follow-up; or

ii) as if transplant was administered at a given LM time conditional on being failure free at LM.

3 Review of the Kaplan-Meier-like estimators and identification of the corresponding theoretical quantities

3.1 Landmark approach

The method is based on the classification of the patients according to the BNRTD status at a given LM time. In the application example the LM is set at 6 months relying on clinical considerations.

3.1.1 Transplant

The estimate is calculated on the sub-sample of 193 subjects at risk, i.e. satisfying the following three conditions at LM time: waiting time observed (i.e. treatment switching occurred), failure not observed, censoring not observed. The application of the KM formula leads to the estimator

$$\hat{S}_{tr}^{LM}(t) = \prod_{LM \le t_{(j)} \le t} \left(1 - \frac{\sum_{i=1}^{n} \varepsilon_i \cdot I(\varepsilon_i w_i \le LM) \cdot I(t_i = t_{(j)}) \cdot \delta_i}{\sum_{i=1}^{n} \varepsilon_i \cdot I(\varepsilon_i w_i \le LM) \cdot I(t_i \ge t_{(j)})} \right) \qquad t \ge LM \qquad (2)$$

where $\varepsilon_i \cdot I(\varepsilon_i w_i \le LM)$ specifies the fixed selection of patients transplanted up to LM. The corresponding curve is displayed in figure 3 (dashed line).

To clarify the meaning of the KM estimator (2), the selection of the sample on which it applies needs to be analysed. Let us consider the case of no censoring. The estimation relies on patients who were observed to be transplanted up to LM and failure free, i.e. satisfying the condition

$(W \le LM) \cap (W \le T_{ch}) \cap (T_{tr} > LM)$ (3)

Given the independence between T_{ch} and T_{tr}, the sampling data satisfying (3) are a random sub-sample of the (unobservable) sample of subjects satisfying the condition $(W \le LM) \cap (T_{tr} > LM)$. In the presence of random censoring a further random selection is applied on (3), i.e. $(C > LM)$. The hazard of failure estimated by (2) at time $t = t_{(j)}$ $(t \ge LM)$ corresponds theoretically to

$\lim_{\Delta t \to 0} P[t < T_{tr} < t + \Delta t \mid (W \le LM) \cap (T_{tr} > t) \cap (C > t)] / \Delta t$

where from the independence between *W* and *C* , T_{tr} and *C*, we can write

$\lim_{\Delta t \to 0} P[(t < T_{tr} < t + \Delta t) \mid (W \le LM) \cap (T_{tr} > t)] / \Delta t$ (4)

This is indeed a mixture of hazards of failure since it involves history of patients who switched to transplant up to LM. The theoretical survival function corresponding to (4) is

$$P[T_{tr} > t \mid (W \le LM) \cap (T_{tr} > LM)] \qquad (5)$$

It is worth noting that this quantity does not refer to the survival function of a single transplanted patient, because it involves a mixture of hazards of patients with different waiting times up to LM, but it answers the following question: << Let us consider a group of patients whose waiting time to transplant is lower than LM with values depending on donor availability. For those who are still failure free at LM, what is their expectation of surviving failure free forward in time?>>.

3.1.2 Chemotherapy

The estimate is calculated on the sub-sample of 300 subjects at risk, i.e. satisfying the following three conditions at LM time: waiting time not observed (i.e. treatment switching not occurred), failure not observed, censoring not observed. The application of the KM formula leads to the estimator

$$\hat{S}_{ch}^{LM}(t) = \prod_{LM \le t_{(j)} \le t} \left(1 - \frac{\sum_{i=1}^{n} (1-\varepsilon_i) \cdot I(\varepsilon_i w_i > LM) \cdot I(t_i = t_{(j)}) \cdot \delta_i}{\sum_{i=1}^{n} (1-\varepsilon_i) \cdot I(\varepsilon_i w_i > LM) \cdot I(t_i \ge t_{(j)})} \right) \qquad t \ge LM \quad (6)$$

where $(1-\varepsilon_i) \cdot I(\varepsilon_i w_i > LM)$ specifies the fixed selection of patients not transplanted up to LM, which includes both patients transplanted after LM and patients never observed switching to transplant. The corresponding curve is displayed in figure 5 (solid line).

To clarify the meaning of the KM estimator (6), the selection of the sample on which the formula is applied needs to be analysed. Let us consider the case of no censoring. The estimation relies on patients who were not observed to be transplanted up to LM and who are failure free, i.e. satisfying the condition

$$(W > LM) \cap (T_{ch} > LM) \qquad (7)$$

Given the independence between T_{ch} and W, the sampling data satisfying (7) are a random sub-sample of the (unobservable) sample of subjects satisfying the condition $(T_{ch} > LM)$. In the presence of random censoring a further random selection is applied on (7), i.e. $(C > LM)$. The hazard of failure estimated by (6) at time $t = t_{(j)}$ $(t \ge LM)$ corresponds theoretically to

$$\lim_{\Delta t \to 0} P[t < T < t + \Delta t \mid (T > t) \cap (C > t)] / \Delta t$$

where T instead of T_{ch}, is used to represent the possible treatment switching after LM. From the independence between T and C we can write

$$\lim_{\Delta t \to 0} P[(t < T < t + \Delta t) \mid (T > t)] / \Delta t \qquad (8)$$

This is indeed a mixture of hazards of failure since it involves the history of patients who kept on being under chemotherapy and patients who switched to transplant after LM. The theoretical survival function corresponding to (8) is

$$P(T > t \mid T_{ch} > LM) \qquad (9)$$

It is worth noting that this quantity does not refer to the survival function of a single patient, because it involves a mixture of hazards of patients with different treatment histories after LM but it answers the following question: <<Let us consider a group of patients whose only treatment option up to LM is chemotherapy. For those who are still failure free at LM, what is their expectation of surviving failure free forward in time considering that they could also receive transplant in the future depending on donor availability?>>.

3.1.3 Graphical comparison of transplant vs chemotherapy

The contrast between the LM-based theoretical questions is fair since both are built on conditional probabilities of being failure free at LM and the treatment status frozen at LM. The former conditioning enlightens the dynamic flavor of this approach. The heterogeneity of the waiting times prior to LM in (5) and the heterogeneity of the possible waiting times after LM in (9) imply that both quantities are averages.

Observing the graph in figure 5, we can state that, on average, patients surviving relapse-free and without receiving transplant until LM (6 months) have a slightly better prognosis in the early time (within two years after LM) than patients transplanted up to LM. However, in the long term, patients transplanted at an early time (before LM) have an advantage in survival.

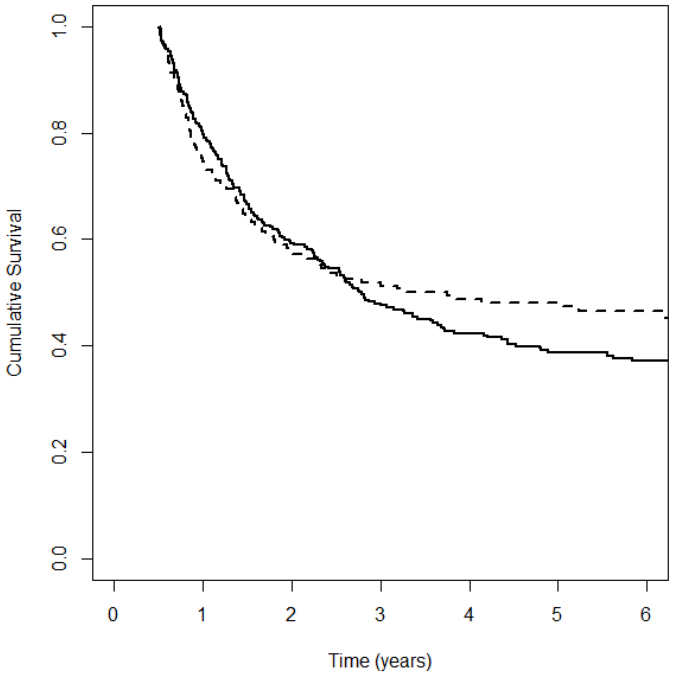


Figure 5 LM-based curves of DFS, with LM set at 6 months from achievement of remission. Transplant = dashed line (193 subjects), Chemotherapy = solid line (300 subjects). A total of 49 subjects do not contribute to the analysis since they developed failure before LM (no censoring observed up to LM). Among the 493 patients failure free at LM, 193 were previously switched to transplant, and the remaining 300 were still under chemotherapy at LM. The size of the sub-sample of 193 patients corresponds to the proportion 35.6% of patients in state 1 at LM time (figure 4, dashed - dotted line = 0.356 –0.020 = 0.376), being 193=0.356*542. Similarly, the size of the sub-sample of 300 patients corresponds to the proportion 55.2% of patients in state 0 at LM time (figure 4, 1 - solid line = 1 – 0.446 = 0.554), being 300=0.554*542. This sub-sample includes 121 patients who have been switched to transplant after LM time.

3.2 Simon and Makuch approach

This method relies on the dynamic update of the risk sets in the two groups depending on the BNRTD variable behavior. This means that patients contribute to the hazard estimate of the chemotherapy group up to their waiting time when they start to contribute to the hazard estimate of the transplant group.

3.2.1 Transplant

The application of the KM formula with this dynamic update leads for the transplant group to the estimator

$$\hat{S}_{tr}^{SM}(t) = \prod_{t_{(j)} \le t} \left(1 - \frac{\sum_{i=1}^{n} \varepsilon_i \cdot I(\varepsilon_i w_i \le t_{(j)}) \cdot I(t_i = t_{(j)}) \cdot \delta_i}{\sum_{i=1}^{n} \varepsilon_i \cdot I(\varepsilon_i w_i \le t_{(j)}) \cdot I(t_i \ge t_{(j)})} \right) \quad (10)$$

where $\varepsilon_i \cdot I(\varepsilon_i w_i \le t_{(j)})$ specifies the dynamic selection of patients transplanted up to $t_{(j)}$. Pragmatically, this is carried out by imposing left truncation to the observed times of transplanted patients.

In the application example, however, there are no patients transplanted at the beginning of the follow-up, thus the estimate of the hazard involved in (10) cannot be calculated before the minimum observed waiting time (48 days). This problem is generally addressed by setting up an early landmark (ELM) for initial estimate of the hazard which guarantees an adequate initial number of subjects in the transplant group who are failure free (50 in the example, leading to an ELM=91 days). The modification of the KM formula (10) including the ELM is

$$\hat{S}_{tr}^{SM-ELM}(t) = \prod_{ELM \le t_{(j)} \le t} \left(1 - \frac{\sum_{i=1}^{n} \varepsilon_i \cdot I(\varepsilon_i w_i \le t_{(j)}) \cdot I(t_i = t_{(j)}) \cdot \delta_i}{\sum_{i=1}^{n} \varepsilon_i \cdot I(\varepsilon_i w_i \le t_{(j)}) \cdot I(t_i \ge t_{(j)})} \right) \qquad t \ge ELM \qquad (11)$$

Of note the cardinality of the risk set for the hazard calculation $\sum_{i=1}^{n} \varepsilon_i \cdot I(\varepsilon_i w_i \le t_{(j)}) \cdot I(t_i \ge t_{(j)})$ is eroded from both the failure development and censoring, but also increased from the entry of newly transplanted patients. For instance, in the application example the initial cardinality of the risk set at ELM is 50, while at $t_{(1)}$ =ELM+54 days is 146. This is due to the entry of 96 patients transplanted between ELM and $t_{(1)}$. The corresponding curve is displayed in figure 6 (dashed line).

The KM estimator reviewed in formula (10) is calculated on patients who switched to transplant and it relies only on the hazard of failure under transplant. Let us consider the case of no censoring. The sampling data used for estimation of the hazard in (10) relies on patients who were observed to be transplanted up to *t* and failure free at *t*, i.e. satisfying the condition

$$(W \le t) \cap (W \le T_{ch}) \cap (T_{tr} > t) \qquad (12)$$

Given the independence between T_{ch} and T_{tr}, the sampling data satisfying (12) can be thought as a random sub-sample of the (unobservable) sample of subjects satisfying the condition $(W \le t) \cap (T_{tr} > t)$. In the presence of random censoring a further random selection is applied on (14), i.e. $(C > t)$. The hazard term in (10) at time $t = t_{(j)}$ corresponds theoretically to

$$\lim_{\Delta t \to 0} P[t < T_{tr} < t + \Delta t \mid (W \le t) \cap (T_{tr} > t) \cap (C > t)] / \Delta t$$

which from the independence between T_{tr} and *C* becomes

$$\lim_{\Delta t->0} P[(t < T_{tr} < t + \Delta t) \mid (W \le t) \cap (T_{tr} > t)] / \Delta t \qquad (13)$$

This is indeed a mixture of hazards since it involves the history of patients who switched to transplant up to *t*. If the Markov assumption holds true, the value of the hazard at time *t* is the same, regardless of the time of treatment switching (*W*). Thus, under this assumption, (13) would also equal the hazard of failure of a patient who was transplanted at the beginning of the follow-up (*W=0*)

$$\lim_{\Delta t->0} P_{W=0}(t < T_{tr} < t + \Delta t \mid T_{tr} > t) \qquad (14)$$

The only one possibility of interpreting the theoretical survival function corresponding to (13) is to think about a patient who is submitted to transplant since the beginning of follow-up, because only the hazard under transplant is involved, as in formula (10). This leads to the quantity

$$P_{W=0}(T_{tr} > t) \qquad (15)$$

Thus, only under the Markov assumption, (10) is a suitable estimator of (15) which answers the following counterfactual question: <<Let us consider a patient for whom transplant can only be administered at the beginning of the follow-up, i.e. at remission time. What is his/her expectation of surviving failure free in time?>>.

Intuitively, at level of estimation, the use of the left truncation in (10) assumes that patients are transplanted since the beginning of the follow-up and yet not observed until the real treatment switch. In the absence of markovianity, the survival function corresponding to the hazard (13), and estimated by (10), does not corresponds to (15) and does not identify the survival function of a single patient. In addition, it cannot be interpreted as an averaged survival over the distribution of the observed waiting times since it does not involve any hazard of failure under chemotherapy which should be accounted for if the waiting time is greater than zero.

3.2.2 Chemotherapy

For the chemotherapy group, the corresponding estimator is

$$\hat{S}_{ch}^{SM}(t) = \prod_{t_{(j)} \le t} \left(1 - \frac{\sum_{i=1}^{n} ((1-\varepsilon_i) + I(\varepsilon_i w_i \ge t_{(j)})) \cdot I(t_i = t_{(j)}) \cdot \delta_i}{\sum_{i=1}^{n} ((1-\varepsilon_i) + I(\varepsilon_i w_i \ge t_{(j)})) \cdot I(t_i \ge t_{(j)})} \right) \qquad (16)$$

where the terms $(1-\varepsilon_i)$ and $I(\varepsilon_i w_i \geq t_{(j)})$ specifies the dynamic selection of patients not transplanted up to $t_{(j)}$ which includes patients who are never observed to switch treatment and patients who are observed to switch after $t_{(j)}$.

For a fair contrast to the transplanted group when the ELM set is needed, the KM formula (16) needs also to be modified. This leads to

$$\hat{S}_{ch}^{SM-ELM}(t) = \prod_{ILM \leq t_{(j)} \leq t} \left(1 - \frac{\sum_{i=1}^{n} (1-\varepsilon_i) \cdot I(\varepsilon_i w_i > t_{(j)}) \cdot I(t_i = t_{(j)}) \cdot \delta_i}{\sum_{i=1}^{n} (1-\varepsilon_i) \cdot I(\varepsilon_i w_i > t_{(j)}) \cdot I(t_i \geq t_{(j)})} \right) \qquad t \geq ELM \qquad (17)$$

Of note the cardinality of the risk set for the hazard calculation $\sum_{i=1}^{n} (1-\varepsilon_i) \cdot I(\varepsilon_i w_i > t_{(j)}) \cdot I(t_i \geq t_{(j)})$ is eroded from the failure development, the censoring process and also from the artificial censoring due to switch to transplant. For instance, in the application example, the initial cardinality of the risk set at ELM is 473, while at $t_{(1)}$ =ELM+8 days is 459. This is due to 1 patient who was truly censored and 13 patients who were censored because switched to transplant between ELM and $t_{(1)}$. The corresponding curve is displayed in figure 6 (solid line).

The KM estimator reviewed in formula (16) is calculated on patients under chemotherapy considering the treatment switching as an additional cause of censoring. Let us consider the case of no censoring. The sampling data used for estimation of the hazards in (16) relies on patients who were not observed to be transplanted up to *t* and who are failure free at *t*, i.e. satisfying the condition

$$(W > t) \cap (T_{ch} > t) \qquad (18)$$

Given the independence between T_{ch} and *W*, the sampling data satisfying (18) can be thought as a random sub-sample of the (unobservable) sample of subjects satisfying the condition $(T_{ch} > t)$. In the presence of random censoring a further random selection is applied on (18), i.e. $(C > t)$. The hazard term in (16) at time $t = t_{(j)}$ hazard corresponds theoretically to

$$\lim_{\Delta t \to 0} P[t < T_{ch} < t + \Delta t \mid (W > t) \cap (T_{ch} > t) \cap (C > t)] / \Delta t$$

which, from the independence between T_{ch} and *W*, T_{ch} and *C*, becomes

$$\lim_{\Delta t \to 0} P(t < T_{ch} < t + \Delta t \mid T_{ch} > t) \qquad (19)$$

The theoretical survival function corresponding to this hazard is the survival probability under chemotherapy

$$P(T_{ch} > t) \qquad (20)$$

answering the following counterfactual question: <<Let us consider a patient whose only treatment option is chemotherapy. What is his/her expectation of surviving failure free in time?>>

3.2.3 Graphical comparison of transplant vs chemotherapy

The contrast between the SM-based theoretical probabilities is fair, since both are built on survival probabilities according to the counterfactual treatment allocation at the beginning of the follow-up by forcing the BNRTD variable to be equal either to *0* or to *1* over the entire follow-up.

In the application example, we need to set an ELM for the SM estimator in both groups, as shown in figure 6. Observing the graph we can state that a counterfactual patient who survived relapse-free until ELM (91 days) under chemotherapy and without the possibility to switch treatment has a better prognosis in the early time (within 2.5 years after ELM) than a counterfactual patient transplanted at ELM; while, in the long term, the patient transplanted has an advantage in DFS. Of note, these considerations would be reliable only under the Markov assumption. However, as noted in chapter 2, the data generating process is semi-markovian.

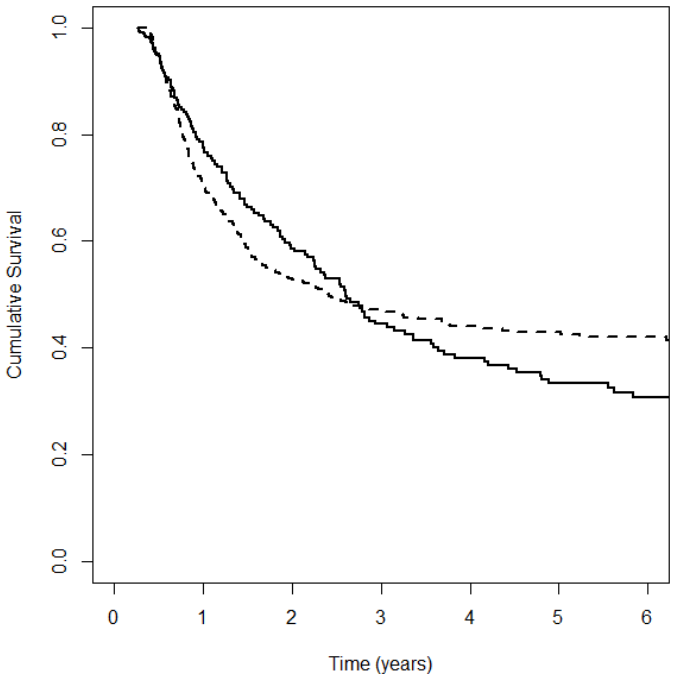


Figure 6. SM-based curves of DFS with ELM set at 91 days. Transplant = dashed line, Chemotherapy = solid line. To ensure an initial number of 50 subjects, an ELM equal to 91 days was chosen. A total of 19 subjects do not contribute to the analysis since they developed failure before ELM (no censoring observed up to ELM), while 473 were still under chemotherapy at ELM.

4 A novel proposal for survival estimation under semi-markovianity

The assumption of markovianity, needed for a suitable estimate of (15) by (10) or (11), is too restrictive in practical applications, such as in the example, where it needs to be relaxed to semi-markovianity. However, the issue of estimating (15) under semi-markovianity can be addressed by noticing that the hazard (13) can be written as

$$\lim_{\Delta t->0} P(t < T_{tr} - W < t + \Delta t \mid T_{tr} - W > t) / \Delta t \qquad (21)$$

which is the hazard of failure under transplant considering only the sojourn time in the transplanted state. In practice, this hazard relies on a different time scale: the time since transplant, kwon as "clock reset" or "clock back" scale (Putter et al. 2007, Iacobelli and Carstensen 2013). The hazard (21) when considering a sojourn time in the transplanted state of $t = t^s_{(j)}$ time units, can be estimated by

$$\frac{\sum_{i=1}^{n} \varepsilon_i \cdot I(t_i - w_i = t^s_{(j)}) \cdot \delta_i}{\sum_{i=1}^{n} \varepsilon_i \cdot I(t_i - w_i \geq t^s_{(j)})}.$$

This means in practice that the survival function (15) can be directly estimated by the "clock back" (CB) estimator

$$\hat{S}^{CB}_{tr}(t) = \prod_{t^s_{(j)} \leq t} \left(1 - \frac{\sum_{i=1}^{n} \varepsilon_i \cdot I(t_i - w_i = t^s_{(j)}) \cdot \delta_i}{\sum_{i=1}^{n} \varepsilon_i \cdot I(t_i - w_i \geq t^s_{(j)})} \right) \qquad (22)$$

This estimator corresponds to the standard KM formula applied on the sub-sample of patients who were observed to be transplanted, provided that the failure time is measured in the "clock back" scale from transplant. This means in practice that, owing to the semi-markov assumption, the hazard under transplant on the CB scale is projected in (22) on a counterfactual subject considered under transplant since the beginning of the follow-up. Although this may appear counterintuitive, it is actually what is done also by the SM estimator (10) through left truncation, improperly using the hazard on the original scale (i.e. time form remission).

The proper graphical comparison of the treatments can now be done under semi-markovianity using the CB estimator (22) for transplant and the SM estimator (16) for chemotherapy, as shown in

figure 7. Of note, the use of an arbitrary ELM is not required in this approach. Apart from this, the chemotherapy curve corresponds to that in figure 6, while the transplant curve now truly represents the survival of a single counterfactual patient transplanted at the beginning of the follow-up.

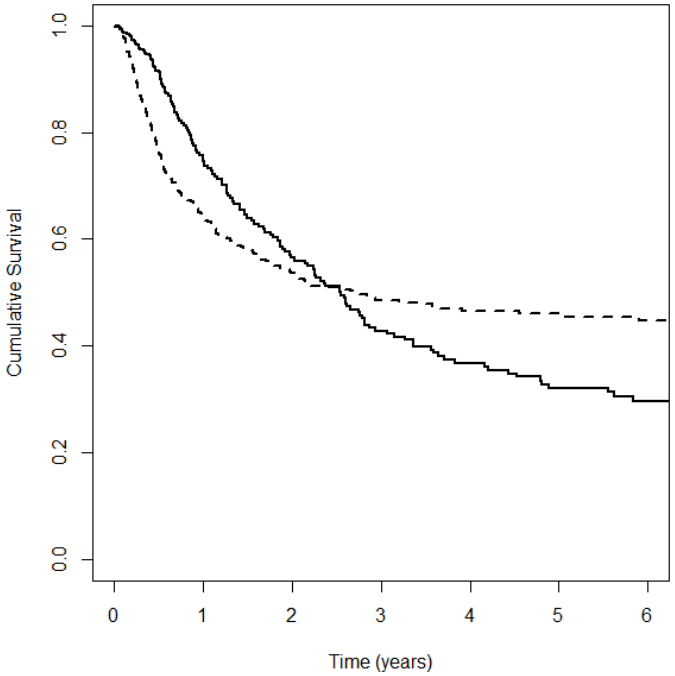


Figure 7. Curves of DFS according to the proposed approach. Transplant = dashed line, Chemotherapy = solid line. The transplant curve is based on the sample of transplanted patients (325) while the chemotherapy curve is based on the whole sample of patients (542).

5 Counterfactual dynamic quantities: definition and estimation under semi-markovianity

The survival experience forcing the treatment to be fixed since the beginning of the follow-up in SM and CB approaches is a pragmatic way to contrast the cumulative effect of treatments in their velocity of failure development. However, in practical applications where the treatment switch right after the beginning of the follow-up is rarely observed or the waiting times are very heterogeneous, this could seem too artificial. In this case, a more realistic approach is the graphical comparison of the survival experience mimicking a counterfactual treatment administration at some time greater than zero. To address this issue, the theoretical survival probability under transplant $P_{W=0}(T_{tr} > t)$ can be modified in the context of counterfactual dynamic prediction arguing as follows. First, a clinically relevant LM time point is set, in the flavour of the reviewed LM approach. Second, the projection of the transplant effect is induced by administering the treatment exactly at LM. This leads to the dynamic counterfactual quantity

$$P_{W=LM}(T_{tr} > t \mid T_{ch} > LM) \qquad (23)$$

which answers the question: <<Let us consider a patient for whom the treatment option is chemotherapy up to LM. If still failure free at LM, what is his/her expectation of surviving failure free forward in time if transplant is administered at LM?>>

A suitable estimator for (23) is

$$\hat{S}_{tr}^{CB-LM}(t) = \prod_{t^s{}_{(j)} \le t-LM} \left(1 - \frac{\sum_{i=1}^{n} \varepsilon_i \cdot I(t_i - w_i = t^s{}_{(j)}) \cdot \delta_i}{\sum_{i=1}^{n} \varepsilon_i \cdot I(t_i - w_i \ge t^s{}_{(j)})} \right) \qquad t \ge LM \qquad (24)$$

which involves the risk of failure under transplant up to $t - LM$ time units of sojourn times. Of note, owing to the semi-markov assumption, the hazard under transplant estimated according to the time units spent in the transplant status is projected on a counterfactual subject considered under transplant since LM.

The corresponding counterfactual quantity for chemotherapy is

$$P(T_{ch} > t \mid T_{ch} > LM) \qquad (25)$$

which answers the question: <<Let us consider a patient for whom the treatment option is chemotherapy. If still failure free at LM, what is his/her expectation of surviving failure free forward in time?>>

Of note, (25) relies only on the velocity of failure development under chemotherapy and does not have the averaging flavor of the dynamic quantity (9). The latter in fact, includes also the impact of treatment switching after the LM time.

A suitable estimator for (25) is

$$\hat{S}_{dh}^{SM-LM}(t) = \prod_{LM \le t_{(j)} \le t} \left(1 - \frac{\sum_{i=1}^{n} (1-\varepsilon_i) \cdot I(\varepsilon_i w_i > t_{(j)}) \cdot I(t_i = t_{(j)}) \cdot \delta_i}{\sum_{i=1}^{n} (1-\varepsilon_i) \cdot I(\varepsilon_i w_i > t_{(j)}) \cdot I(t_i \ge t_{(j)})} \right) \qquad t \ge LM \qquad (26)$$

In figure 8 the counterfactual dynamic survival curves for transplant and chemotherapy based on the novel LM approach are shown for two LM times, i.e. 6 months and 2 years. Of note: i) this representation is reliable for a semi-Markov data generating process and ii) the information provided by the transplanted patients, even from those failed or censored before LM, is used.

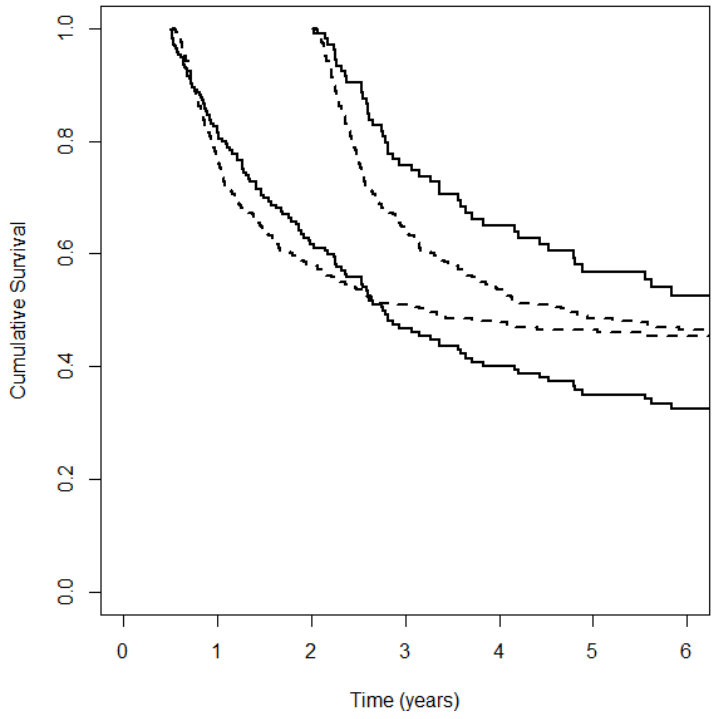


Figure 8. Dynamic curves of DFS at two different LM (6 months and 2 years), according to the novel approach. Transplant = dashed line, Chemotherapy = solid line. Both transplant curves are based on the whole sample of transplanted patients (325) while the chemotherapy curves are based on patients failure free at LM (493 for LM=6 months and 285 for LM=2 years).

6 Simulation Protocol (1)

To investigate the performance of the estimators discussed in chapters 3, 4 and 5, I conducted a simulation protocol using a data generating process compatible with the application example (figure 1). I generated 1000 dataset of 500 observations each. The latent failure times under transplant T_{ptr}, under chemotherapy T_{ch}, and the latent waiting time W (and thus $T_{tr} = W + T_{ptr}$) were simulated using the inversion method of Bender et al. (2005), while the censoring time *C* was directly simulated using a uniform distribution: $C \sim Uniform\ (a=1, b=15)$.

The latent failure time after transplant switching T_{ptr} was generated in three different ways to represent several conditions on process memory:

1. constant hazard under transplant and no effect of the waiting time (homogeneous Markov), by $T_{ptr} \sim Exp(\lambda = 0.08)$;
2. non-monotone hazard under transplant and no effect of the waiting time (semi-Markov), by $T_{ptr} \sim Burr(k = 0.3, c = 1.1, s = 1)$;
3. non-monotone hazard under transplant and proportional hazard effect of the waiting time (extended semi-Markov), by $T_{ptr} \sim Burr\left(k_W = 0.3\left(\frac{W}{50}+1\right), c = 1.1, s = 1\right)$;

T_{ch} was generated in the same way in all the scenarios, by $T_{ch} \sim LogLogistic\ (\alpha = 0.4, \beta = 1)$ as well as the latent waiting time *W*, by $W \sim Exp(\theta = 0.4)$. The details of the simulation protocol are summarized in table I, while the graphical representation of the hazards under transplant and under chemotherapy in the three scenarios are shown in figure 9.

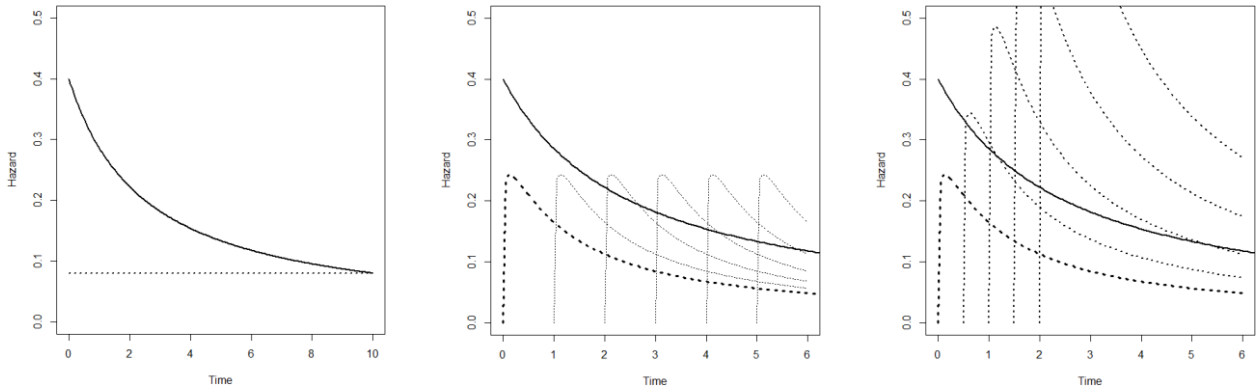
The results on the performance of the SM estimator $\hat{S}_{tr}^{SM}(t)$ (10) and CB estimator $\hat{S}_{tr}^{CB}(t)$ (22) for the theoretical survival of a patient transplanted at *W*=0 $P_{W=0}(T_{tr} > t)$ (15) and the SM estimator $\hat{S}_{ch}^{SM}(t)$ (16) for the theoretical survival of a patient under chemotherapy with no possibility to switch treatment $P(T_{ch} > t)$ (20), are summarized in table II. I chose 4 time points *t*={1,2,5,10} to compare the estimates with the theoretical values of the survival functions. The observed average percentage of censoring at each time point was also reported in table II.

Concerning transplant, we observe that $\hat{S}_{tr}^{SM}(t)$ (10) is unbiased only when data are generated under the (unrealistic) homogenous Markov process (scenario 1), while $\hat{S}_{tr}^{CB}(t)$ (22) is unbiased even

under the semi-Markov assumption, but not under the extended semi-Markov assumption, i.e. when an effect of W on the distribution of T_{ptr} is present (Putter et al. 2007). In the latter case, however, the bias of $\hat{S}_{tr}^{CB}(t)$ (22) is much smaller compared to $\hat{S}_{tr}^{SM}(t)$ (10). Concerning chemotherapy, we see that $\hat{S}_{dh}^{SM}(t)$ (16) is unbiased in all scenarios. The variances of the estimators are consistent with each-other.

The results on the performance of the LM estimator $\hat{S}_{tr}^{LM}(t)$ (2) and CB-LM estimator $\hat{S}_{tr}^{CB-LM}(t)$ (24) for the theoretical survival of a patient failure-free at LM transplanted at W=LM $P_{W=LM}(T_{tr} > t)$ (23) and the old LM estimator $\hat{S}_{dh}^{LM}(t)$ (6) and new LM estimator $\hat{S}_{dh}^{SM-LM}(t)$ (26) for the theoretical survival of a patient failure-free at LM under chemotherapy and with no possibility to switch treatment $P_{LM}(T_{dh} > t)$ (25), are summarized in table III. I chose LM=0.5 and, again, evaluate the estimates at t={1,2,5,10}. The average percentage of censoring at each time point was also reported in table III.

Concerning transplant, we observe that $\hat{S}_{tr}^{LM}(t)$ (2) is biased in all scenarios. This was expected since it estimates a different quantity. $\hat{S}_{tr}^{CB-LM}(t)$ (24) is unbiased in scenarios 1 and 2 but not in scenario 3. In the latter case, however, the bias is very small. Concerning chemotherapy, we see that $\hat{S}_{dh}^{LM}(t)$ (6) is biased in all scenarios while $\hat{S}_{dh}^{SM-LM}(t)$ (26) is always unbiased. The variances of the estimators are consistent with each-other.



1. Homogeneous Markov 2. Semi-Markov 1. Extended semi-Markov

Figure 9. Hazard of failure under transplant (dashed lines) and under chemotherapy (solid lines) in the three simulation scenarios, plotted in the original scale. Hazards under transplant according to different waiting times are shown and it is worth to note that: in scenario 1 they completely overlap; in scenario 2 they are just translated but keeping the same shape; in scenario 3, due to the effect of waiting time, the shape is also modified leading to an increasing peak. The hazard under chemotherapy is the same in all scenarios.

Table I. Details on the simulation protocol.

Scenario	T_{ptr}	T_{ch}	W	C
1. Homogeneous Markov	$h(t)=\lambda$ $S(t)=\exp(-\lambda t)$ $\lambda=0.08$	$h(t)=\frac{(\beta\alpha)(t\alpha)^{\beta-1}}{1+(t\alpha)^{\beta}}$ $S(t)=\frac{1}{1+(t\alpha)^{\beta}}$ $\alpha=0.4$ $\beta=1$	$h(t)=\theta$ $S(t)=\exp(-\theta t)$ $\theta=0.2$	$F(t)=\begin{cases}0 & t<a\\ \frac{t-a}{b-a} & a\le t<b\\ 1 & t\ge b\end{cases}$ $a=1$ $b=15$
2. Semi-Markov	$h(t)=\frac{\frac{kc}{s}\left(\frac{t}{s}\right)^{c-1}}{1+\left(\frac{t}{s}\right)^{c}}$ $S(t)=\exp\left(-k\ln\left(1+\left(\frac{t}{s}\right)^{c}\right)\right)$ $k=0.3$ $c=1.1$ $s=1$	$h(t)=\frac{(\beta\alpha)(t\alpha)^{\beta-1}}{1+(t\alpha)^{\beta}}$ $S(t)=\frac{1}{1+(t\alpha)^{\beta}}$ $\alpha=0.4$ $\beta=1$	$h(t)=\theta$ $S(t)=\exp(-\theta t)$ $\theta=0.2$	$F(t)=\begin{cases}0 & t<a\\ \frac{t-a}{b-a} & a\le t<b\\ 1 & t\ge b\end{cases}$ $a=1$ $b=15$
3. Extended semi-Markov	$h(t)=\frac{\frac{k_W c}{s}\left(\frac{t}{s}\right)^{c-1}}{1+\left(\frac{t}{s}\right)^{c}}$ $S(t)=\exp\left(-k_W\ln\left(1+\left(\frac{t}{s}\right)^{c}\right)\right)$ $k_W=0.3\left(\frac{W}{50}+1\right)$ $c=1.1$ $s=1$	$h(t)=\frac{(\beta\alpha)(t\alpha)^{\beta-1}}{1+(t\alpha)^{\beta}}$ $S(t)=\frac{1}{1+(t\alpha)^{\beta}}$ $\alpha=0.4$ $\beta=1$	$h(t)=\theta$ $S(t)=\exp(-\theta t)$ $\theta=0.2$	$F(t)=\begin{cases}0 & t<a\\ \frac{t-a}{b-a} & a\le t<b\\ 1 & t\ge b\end{cases}$ $a=1$ $b=15$

Table II. Simulation results (1000 dataset of 500 observations) on counterfactual quantities. Bias is calculated as average on the 1000 samples of the difference between the estimated and the true value. Variance (Var) is calculated as variance of the estimate of the 1000 samples.

Scenario	T	Censored (%)	Transplant					Chemotherapy		
			True $P_{W=0}(T_{tr} > t)$	Bias $*10^2$ $\hat{S}_{tr}^{SM}(t)$	Bias$*10^2$ $\hat{S}_{tr}^{CB}(t)$	Var $*10^3$ $\hat{S}_{tr}^{SM}(t)$	Var $*10^3$ $\hat{S}_{tr}^{CB}(t)$	True $P(T_{ch} > t)$	Bias $*10^2$ $\hat{S}_{ch}^{SM}(t)$	Var $*10^3$ $\hat{S}_{ch}^{SM}(t)$
1. Homogeneous Markov	1	0	0.923	0.038	-0.013	3.887	0.363	0.714	0.062	0.459
	2	10.6	0.852	-0.030	-0.051	3.826	0.707	0.556	0.065	0.671
	5	21.7	0.670	-0.063	-0.130	3.277	1.518	0.333	-0.161	0.979
	10	30.1	0.449	-0.002	-0.045	2.793	2.734	0.200	-0.218	1.989
2. Semi-Markov	1	0	0.812	-1.018	0.003	8.037	0.760	0.714	-0.083	0.447
	2	3.0	0.709	-4.239	-0.071	6.788	1.076	0.556	-0.058	0.601
	5	6.2	0.561	-10.218	-0.118	4.178	1.218	0.333	-0.055	0.800
	10	9.0	0.457	-13.135	0.072	2.604	1.280	0.200	-0.132	1.215
3. Extended semi-Markov	1	0	0.812	-1.104	-0.911	8.057	0.813	0.714	-0.083	0.447
	2	3.0	0.709	-4.491	-1.358	6.782	1.087	0.556	-0.058	0.601
	5	6.1	0.561	-10.969	-1.730	4.098	1.234	0.333	-0.055	0.800
	10	8.9	0.457	-14.333	-1.734	2.493	1.243	0.200	-0.132	1.215

Table III. Simulation results (1000 dataset of 500 observations) on dynamic counterfactual quantities (LM=0.5). Bias is calculated as average on the 1000 samples of the difference between the estimated and the true value. Variance (Var) is calculated as variance of the estimate of the 1000 samples.

Scenario	t	Censored (%)	Transplant					Chemotherapy				
			True $P_{W=LM}(T_{tr} > t)$	Bias*10^2 $\hat{S}_{tr}^{LM}(t)$	Bias*10^2 $\hat{S}_{tr}^{CB-LM}(t)$	Var*10^3 $\hat{S}_{tr}^{LM}(t)$	Var*10^3 $\hat{S}_{tr}^{CB-LM}(t)$	True $P_{LM}(T_{ch} > t)$	Bias*10^2 $\hat{S}_{ch}^{LM}(t)$	Bias*10^2 $\hat{S}_{ch}^{SM-LM}(t)$	Var*10^3 $\hat{S}_{ch}^{LM}(t)$	Var*10^3 $\hat{S}_{ch}^{SM-LM}(t)$
1. Homogeneous Markov	1	0	0.961	2.759	-0.001	0.062	0.220	0.857	0.461	-0.026	0.301	0.329
	2	16.6	0.887	6.280	0.045	0.259	0.597	0.667	2.672	0.097	0.596	0.740
	5	28.1	0.698	9.735	-0.009	0.962	1.478	0.400	7.127	0.045	0.807	1.298
	10	36.8	0.468	8.417	-0.083	2.077	3.179	0.240	6.783	0.077	0.946	2.910
2. Semi-Markov	1	0	0.911	5.805	0.098	0.161	0.492	0.857	0.116	-0.026	0.311	0.329
	2	14.9	0.796	9.502	-0.036	0.493	1.023	0.667	0.763	0.097	0.612	0.740
	5	24.8	0.629	5.632	0.026	1.203	1.642	0.400	2.228	0.045	0.790	1.298
	10	33.5	0.513	-0.705	-0.046	1.830	2.602	0.240	4.234	0.077	0.859	2.910
3. Extended semi-Markov	1	0	0.911	5.778	-0.240	0.162	0.505	0.857	0.107	-0.026	0.312	0.329
	2	14.9	0.796	9.373	-0.769	0.497	1.053	0.667	0.706	0.097	0.616	0.740
	5	24.7	0.629	4.968	-1.066	1.224	1.682	0.400	1.874	0.045	0.786	1.298
	10	33.1	0.513	-2.127	-1.309	1.847	2.610	0.240	3.419	0.077	0.852	2.910

Part II
Regression Models

The aim of the second part of this thesis is to discuss the use of regression models in survival analysis in the presence of a BNRTD variable (such as the chemotherapy versus transplant treatment) with possible time-varying effects. I underline the limits of the methods commonly used in practice, pointing out two aspects:

1. a regression model with time measured in the original scale (i.e. from remission), where treatment switch is accounted through left truncation, does not allow to properly assess the impact of waiting time to transplant on the hazard of failure;
2. in the presence of time-varying effect of the BNRTD variable, profile-specific (dynamic) predictions cannot be derived from the extended Cox model.

Concerning issue 1, I show that fitting a regression model only on transplanted patients, measuring time on the clock back scale (i.e. since transplant), leads to a coefficient of waiting time which provides an accurate estimate of the real impact of the waiting on the hazard of failure.

To address issue 2, I review and discuss two recently proposed flexible regression models, namely the “Hanley-Miettinen model” (HM) and the “Landmark regression model” (LMR), that can overcome the limits of the Cox model when estimating survival curves in complex situations.
I show that the HM model, measuring time in the clock back scale for transplanted patients, is a good option both for assessing the impact of waiting time on the hazard of failure and for developing dynamic prediction curves. This is confirmed by comparing the proposed methods through an application on real data and a simulation protocol.

7 Assessment of the impact of waiting time

When the focus is to evaluate the effect of a BNRTD treatment, such as transplant, on the survival experience, one could be interested first in assessing the impact of waiting time to transplant on the disease process through the hazard function. I already investigated the impact of the waiting time to transplant on the hazard of failure, using a graphical method that compares the hazards under transplant estimated on groups of patients with homogeneous waiting times in the clock back scale (see section 2.3). This method is useful to get insights into the (possible) presence of an effect of the waiting time on the hazard of failure after transplant.

To address this issue, however, one could think about using a hazard based regression model, such as the Cox model, where the variable transplant is accounted with (delayed) entry of patients in the transplanted state by left truncation, and the waiting time to transplant is included as an interaction term with the transplant indicator. Doing so one may expect to obtain a reliable estimate of the waiting time to transplant effect on the hazard of failure after transplant.

This approach, however, has two strong limitations concerning:

i) the general validity of the model estimates,
ii) the interpretation of the waiting time coefficient.

On point i) one should be aware that such a model is valid only under the Markov assumption, i.e. the hazard of failure at time t (from remission) for a patient transplanted at $w<t$ is equal to the hazard of failure at time t for a patient transplanted at a different $w^*<t$ (see section 2.3). The reason for this assumption is analogous to the consideration done in section 3.2 for the Simon and Makuch method: using left truncation means assuming that patients are considered to be transplanted since the beginning of the follow-up, but yet not observed until the real treatment switch. Thus, only markovianity can guarantee that a real patient transplanted at $w<t$ shares the same hazard of failure at time t with his counterfactual representation which is transplanted since the beginning.

On point ii), it must be noticed that, in a general setting, what we observe is not the real effect of waiting time, but a mixture of the real one and that of the time past transplant (sojourn time). In fact, measuring time on the original scale, the waiting time w and the sojourn time s in the transplanted state are fully dependent: fixing time since remission t, one can be obtained from the other (e.g. $s=t-w$). This means that at time t from remission, patient at risk with higher waiting time will have shorter sojourn time (and vice versa) and both times will influence the hazard of failure through the same parameter, which becomes not interpretable.

When the process is semi-markovian, i.e. the hazard of failure at time t of a patient with a waiting time $w<t$ and a sojourn time s in the transplant state equals the hazard at t^* of a patient with a different waiting time $w^*< t^*$ and the same sojourn time s, one may obtain a coefficient of the waiting time variable sensibly different from 0. This is due only to a heterogeneity of sojourn times of patients at risk at time t from remission, being in the absence of a true effect of the waiting time.

A reasonable alternative is to fit a model only on transplanted patients (Iacobelli and Carstensen 2013). This naturally leads to measure time since transplant, avoiding all the previously described issues affecting the standard approach. In fact, this model is automatically reliable even in non-Markov settings since we exclude the first state of the process and transitions from that state. Secondly, the coefficient of the waiting time is not influenced by the effect of the sojourn time which is fully captured by the baseline hazard. This means that the obtained parameter for the waiting time truly represents the effect of the hazard of failure after transplant of a one-time-unit increase in the waiting time. When the process is standard semi-markovian, we should observe an estimated coefficient very close to 0.

8 Dynamic prediction

Standard regression models for failure time data, such as the Cox model, can be useful to estimate the treatment effect on the development of a disease, in terms of hazard ratio. However, clinicians would most benefit from the knowledge of the treatment-specific predicted survival to target treatment options (Julian and Hanley 2008). For this reason, one may want to display the profile specific survival or incidence function in time and/or to obtain estimates of these functions at some clinically relevant time points (e.g. 3 years, 5 years).

However, from a patient-oriented point of view, this picture may be unsatisfactory because it gives only information about the prognosis conditional on being at risk at the beginning of the follow-up. In other words, we could not directly predict which would be the probability of a patient that has already survived until $t > 0$ to survive beyond $s > t$, in formulas: $S(s \mid t) = P(T > s \mid T \geq t)$. This quantity represents the probability of being event-free between t and s conditional on being event-free just before t ,In section 5 I showed how to estimate it non-parametrically when a BNRTD variable is involved, fixing s=LM. However, it can be of interest to plot the function shifting t over time. In order to do so and to avoid handling a function with two arguments, we can assume $s = t + k$, where k is a fixed quantity and represents the width of a prediction window starting from t. Thus, switching the point of view from survival to incidence quantities, we can define the following *fixed width conditional failure function*:

$$F_k(t) = P(T < t + k \mid T \geq t) = 1 - S(t + k \mid t) \qquad (27)$$

It represents the probability of failure within a fixed window of width k, given that the patient is still event-free just before t. This function provides a "dynamic" description of the prognosis because it refers to a moving prediction time t on a fixed length time window. Clearly, the choice of the width k of the time window is arbitrary and a proper value should be selected in relation to the length of follow-up and/or other features of the disease of interest. Through the conditional probability law, the following expression can be obtained to compute the estimates of (27) over time:

$$\hat{F}_k(t) = 1 - \frac{\hat{S}(t + k)}{\hat{S}(t)} \qquad (28)$$

where $\hat{S}$ represents an estimates of the (profile specific) survival function.

By displaying the estimate (28) over time one can directly answer the question: << Let us consider a patient failure free at time t, which is his/her expectation of developing failure in the next period

of length *k*? >>. Such a plot is also informative about the time needed to possible "cure" the disease, where the conditional fixed window failure function should drop down to zero.

Finally, from a theoretical point of view, it is worth to note that, taking one single prediction window as large as the whole follow-up, leads to a conditional incidence equal to the standard cumulative incidence. By contrast, reducing the width of the prediction window to an instant, the conditional incidence boils down to the instantaneous hazard. Some examples on the similarity between the shape of the hazard function and the fixed window conditional incidence function are shown in figure 10.

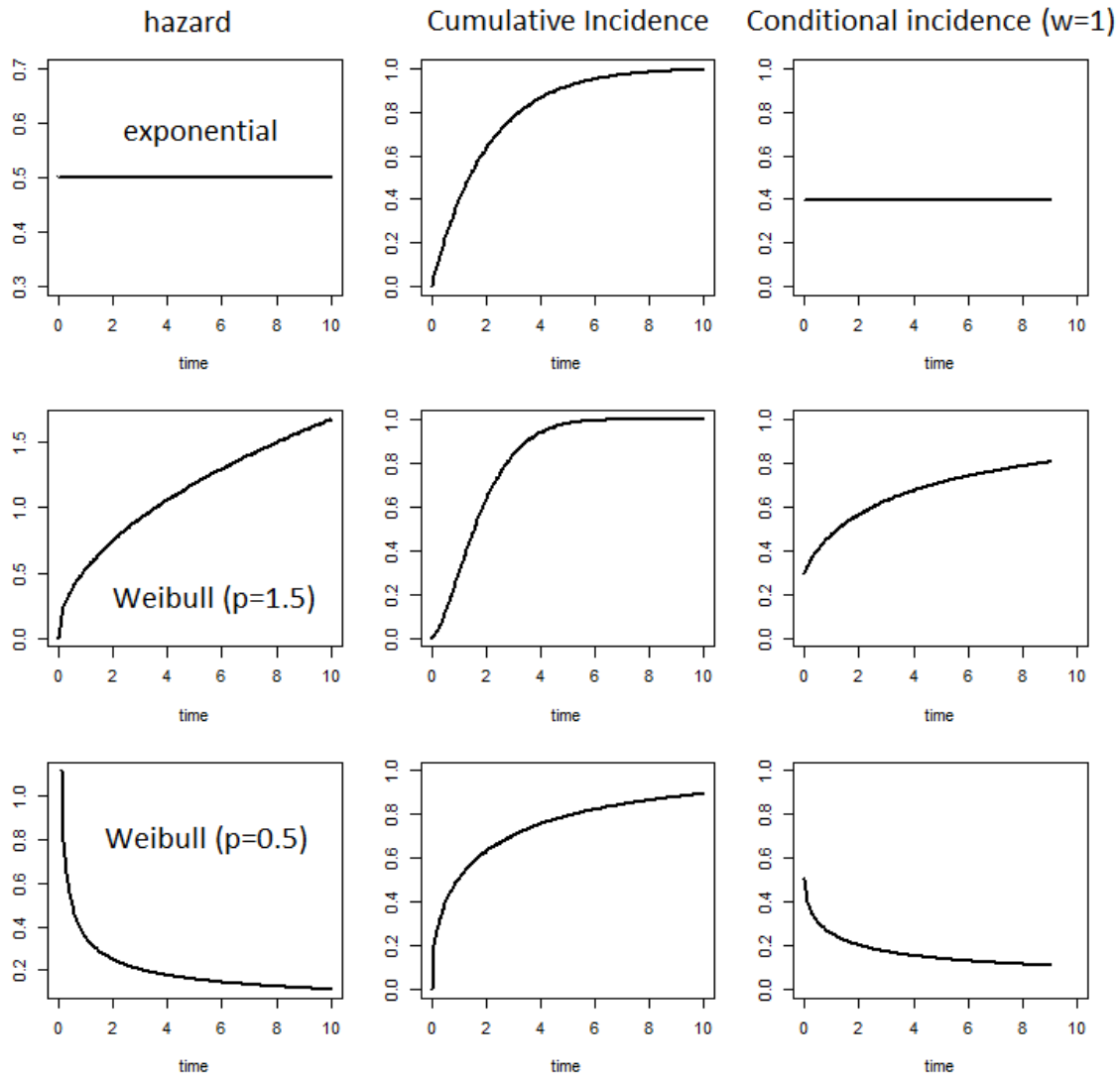


Figure 10. Three examples of hazards shapes (constant, increasing and decreasing) with the corresponding cumulative incidence and fixed window conditional incidence functions. Of note, while the cumulative incidence is always monotonically increasing, the fixed window conditional incidence preserves the trend of the hazard.

9 Review of two recently proposed flexible regression models for prediction with BRNTD treatment

9.1 Hanley-Miettinen model

The specific intent of the work of Hanley and Miettinen (2009), was to find a full-parametric regression method for the hazard function with two main characteristics: first, the possibility to model flexibly the baseline hazard and the interactions between time and covariates (i.e. time-varying effects), and second, the opportunity to compute the estimates in a straightforward way, even in the presence of time-varying covariates, using functions already implemented in standard statistical software.

The basic idea, arising from some methodological papers published by Efron (1977, 1988 and 2002), consisted in fitting a logistic regression model to estimate the hazard function over time on a expanded dataset created by dividing the observed survival time of each subject into a number of time-units and to treat the number of events in every single interval as a Binomial random variable.

To overcome the computational issues arising from this kind of approach due to the potentially huge size of the expanded dataset, Hanley and Miettinen reviewed and extended a method developed by Mantel (1973). They used a specific type of random sampling of the individual time-slices which is supposed not to affect the predicted coefficients of the logistic regression, except for an additive constant that can be accounted by an offset term in the model. The method proposed can be described more formally as follows.

Firstly, one must split the observed survival time t_i of each patient into single time units, generating a new dataset consisting in $B = \sum_{i=1}^{N} t_i$ rows. Subsequently, the final dataset is obtained by combining: i) one subset which includes all the $c = \sum_{i=1}^{N} \delta_i$ rows relative to the time-units where an event as occurred (case-series), and ii) another subset created by selecting further $b = 100 \cdot c$ rows (base series). The latter subset is created through a systematic random sample of all the *B* time-units, thus it includes $b_i = (b/B) \cdot t_i$ time units for each patient $i=1,\ldots,N$. In practice, every row of the base series identifies the sampled event-free time units (denoted by a new event indicator $Y = 0$), while each row of the case series is associated with a time unit where an event as occurred (denoted by $Y = 1$). The sampling procedure is displayed in figure 11.

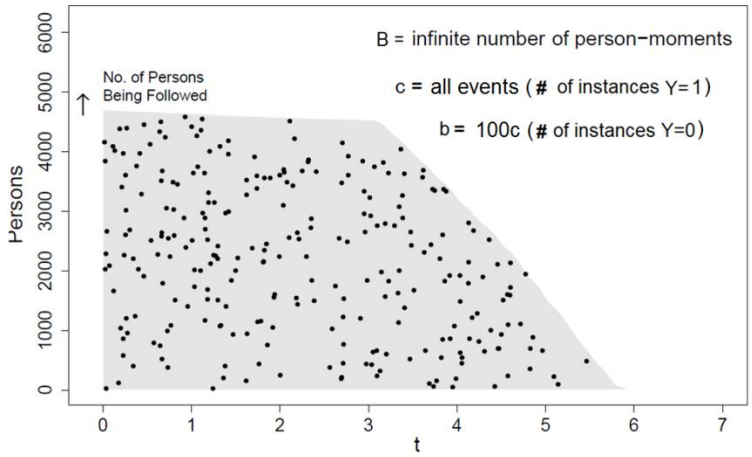


Figure 11. Visual description of the sampling procedure relative to the Hanley-Miettinen model.

Let us observe that at each time *t*, the total number of events $c(t)=\sum_{i=1}^{N} I(t_i=t)\cdot\delta_i$ is known, while the total number of time unit observations $B(t)=\sum_{i=1}^{N} I(t_i=t)$ is not. The instantaneous hazard at *t* would be estimated by $\hat{h}(t)=\hat{P}(T=t\,|\,T\geq t)=\hat{P}(Y=1\,|\,t)=\frac{c(t)}{B(t)}=\frac{\sum_{i=1}^{N} I(t_i=t)\cdot\delta_i}{\sum_{i=1}^{N} I(t_i=t)}$. However, in our dataset, not all the $B(t)$ time units are available, but only a sample of size $b(t)=\frac{b}{B}\cdot B(t)$. Thus, due to the sampling, it is possible to calculate only a biased estimate of the instantaneous failure rate function over time: $\hat{h}(t)_{biased}=\hat{P}(T=t\,|\,T\geq t)_{biased}=\hat{P}(Y=1\,|\,t)=\frac{c(t)}{b(t)}=\frac{\hat{h}(t)\cdot B(t)}{\frac{b}{B}\cdot B(t)}=\hat{h}(t)\cdot\frac{B}{b}$.

By applying the same logic used in epidemiology for case-control studies, since the biased conditional survival is close to one, $\hat{P}(T>t\,|\,T\geq t)_{biased}=\hat{P}(Y=0\,|\,t)=1-\hat{h}(t)_{biased}\cong 1$, we could substitute $\hat{h}(t)_{biased}$ by the odd: $\frac{\hat{h}(t)_{biased}}{1-\hat{h}(t)_{biased}}=\frac{\hat{P}(T=t\,|\,T\geq t)_{biased}}{1-\hat{P}(T=t\,|\,T\geq t)_{biased}}=\frac{\hat{P}(Y=1\,|\,t)}{\hat{P}(Y=0\,|\,t)}=\hat{h}(t)\cdot\frac{B}{b}$. As a consequence, it is possible to model the odd of event as a function of time *t* by a logistic regression approach, specifying log(*B*/*b*) as an offset (i.e. a term whose coefficient is forced to be equal to 1) to obtain valid estimates for the hazard: $\hat{h}(t)=\frac{b}{B}\cdot\exp\left(\alpha+\gamma\cdot g(t)\right)$. By this formulation, one can

set the assumed form of the baseline hazard by specifying an appropriate function of time $g(t)$; for example $g(t)=\log(t)$ implies a Weibull distribution for baseline hazard, while $g(t)=t$ leads to a Gompertz distribution. Spline functions represents an attractive alternative to model more complex hazard shapes and can also be included in the Hanley-Miettinen model.

In this context, accounting for fixed or time-dependent covariates becomes straightforward, as well as to relax the proportional hazard assumption by including time dependent effects (i.e. interactions between covariates and functions of time) in the model. Using the previous notation one can write: $\hat{h}(z,t) = \exp\left(\log\frac{B}{b} + \alpha + \gamma \cdot g(t) + \beta(t) \cdot Z(t) \right)$, where $Z(t)$ is the matrix of the (time-dependent) covariates for every time unit included in the dataset, $g(t)$ the chosen function of time, α, γ and β the estimated coefficients (β is a coefficients vector) from the logistic regression fitting and $\log\frac{B}{b}$ is the offset. Using the well-known relationship between the hazard and the survival functions, $S(\mathrm{z},t) = \exp[-\int_0^t h(\mathrm{z},u)du]$, the method can ultimately be used to generate and draw smooth-in-time survival functions specific for patient profile and treatment.

If the focus is on dynamic prediction, it is very straightforward to derive an estimate of the profile-specific fixed window conditional incidence function $\hat{F}_{HM,k}(t \mid x)$ through the Hanley-Miettinen model. The following formulation can be postulated:

$$\hat{F}_{HM,k}(t \mid x) = 1 - \exp(-\int_t^{t+k} \hat{h}_{HM}(u \mid x)du)$$

where $\hat{h}_{HM}(t \mid x)$ is the estimated profile-specific hazard function provided by the Hanley-Miettinen model.

Some considerations about the choice of the b/c ratio are finally reported: the size of the sample affects the variance of the regression coefficients and, thus, should be chosen as large as possible according to computing software capacity. However, as Mantel already noted in his 1973 paper, "there is little to be gained by letting the size of one series, b, become arbitrarily large if the size of the other series, c, must remain fixed". By choosing $b/c = 100$, we would expect only a 1% increase in the variance of the coefficients computed in the case where b approaches infinity and thus is considered a reliable preference.

9.2 Landmark Cox model

The "landmark" model was designed specifically for dynamic predictions. The genesis of the model has a starting point in the following consideration that applies to the issue of time-varying effects in standard survival analysis. Different ways of modeling the violation of the proportional-hazards assumption lead to very different estimates for hazard functions while the differences between the survival functions are much smaller. This can be intuitively explained by considering that, since the survival function is directly related to the cumulative hazard $H(t \mid x)$, the time dependent effect of the covariates on the survival goes through a weighted average of the $\beta(t)$ over the time interval $[0,t]$, with weights proportional to the baseline hazard. In fact, the following approximation holds:

$$H(t \mid x) \approx H_0(t) \exp(x^T \overline{\beta}(t)), \qquad (29)$$

where

$$\overline{\beta}(t) = \frac{\int_0^t h_0(s)\beta(s)ds}{\int_0^t h_0(s)ds}.$$

Van Houwelingen (2007), starting from this consideration, attempts to study the use of the Cox model for predictive purposes even if it is known that $\beta(t)$ varies over time. His findings illustrate that, if the data are used with administrative censoring at t_{hor} and random censoring before t_{hor}, the Cox model gives approximately correct predictions for the cumulative hazard (and hence the survival) function at t_{hor} even if the true effect of the covariates is time-dependent. In formulas:

$$H_{true}(t_{hor} \mid x) \approx H_{Cox}(t_{hor} \mid x).$$

This approximation holds better as long as t_{hor} becomes smaller and $\beta(t)$ is small or does not vary too much. It is worth nothing that no insights are given about the goodness of prediction in the whole interval $[0,t_{hor}]$.

Given these premises, one can argue that imposing administrative censoring to the data at each time point where a standard prediction is of interest and fitting a Cox model at each prediction time can be a reliable method to obtain pointwise survival estimates.

Using this approach, one can also address the issue of survival prediction from a dynamic point of view, i.e. focusing on the *fixed width conditional failure function*. Within this framework the previous considerations about the PH model are still valid but a crucial aspect must be kept in mind: if the interest is to make prediction at $t + k$, given that the patient has survived up to t, the approximation (29) still holds but in this case the average of $\beta(s)$ over the interval $[t,t + k]$ is needed instead of the average over the whole follow-up period $[0,t_{hor}]$.

As a consequence, dynamic prediction in the interval $[t,t + k]$ should be based on data regarding individuals still at risk at a "landmark" time $t=t_{LM}$. In the context of dynamic predictions, the landmark time t_{LM} is the lower limit of the prediction window of width k. To obtain a prediction within that window, we can fit, on data with administrative censoring at $t_{LM} + k$ and left truncation at t_{LM}, the following *landmark model*:

$$h(t \mid x, t_{LM}, k) = h_0(t \mid t_{LM}, k)\exp(x^T \beta_{LM}), \qquad t_{LM} \le t \le t_{LM} + k \tag{30}$$

It is worth noting that the model: i) applies only for the patients still at risk at t_{LM} and ignores any event after t_{LM} +k, ii) it gives estimates of both $h_0(t \mid t_{LM}, k)$ and β_{LM}.

In practice, a reliable estimate of $\hat{F}_k(t_{LM} \mid x) = 1 - \exp(-H(t_{LM} + k \mid t_{LM}, x))$ is obtained through

$$H(t_{LM} + k \mid t_{LM}, x) = \exp(x^T \beta_{LM}) H_0(t_{LM} + k \mid t_{LM}).$$

From the approach sketched so far, it follows that, potentially, we need to fit a different *landmark model* on each *landmark dataset* related to the time points where the prediction of the conditional incidence within window is needed. However, a more elegant solution to the problem of fitting such dynamic prediction model has been suggested by Van Houwelingen (2007). The procedure requires some data-manipulation steps described in the following:

1. fix the prediction window width w;
2. select a set of prediction time points $\{s_1, \ldots, s_L\}$;
3. create a prediction data set for each $t_{LM}=s_l$ by truncation and administrative censoring;
4. stack all those data sets into a single "super prediction data set".

The choice of w should be based on clinical criteria according to the type of disease. For what concerns the selection of the prediction time points, one should be aware of the fact that, in the final combined data set, the $t_{LM}=s_l$ represent the strata. As the window slides along the follow-up time, it moves from one stratum to another, possibly producing a variation in the risk set: patients are lost if s_l passes an event time and gained if $s_l + k$ passes an event time. A simple but useful way to choose the prediction time points $\{s_1, \ldots, s_L\}$ could be to take an equidistant grid of points within the overall prediction interval.

As already pointed out, the landmark model described by (30) is designed to obtain a prediction in a specific window or, in other words, for a single stratum. To build a model that, working on the "super data set", allows for prediction in each sliding window defined by $t_{LM}=s_l$, we should account for the following two aspects: i) the regression coefficients can be thought as stratum-varying as we can let them depend on $t_{LM}=s_l$; ii) we must assume a different baseline hazard for each stratum too.

The first condition can be translate conveniently into practice by modeling the dependence of the coefficients form proper functions of the strata (e.g. considering polynomial functions or splines) in

a linear way. Regarding ii), we can either run a stratified model to obtain separate estimated baseline hazards for each stratum or, alternatively, we could model this dependence smoothly as was done for the coefficients: in the application example and in the simulation study only the second approach was considered.

An example of a "super" prediction model can be defined as:

$$h(t \mid x, t_{LM} = s, k) = h_0(t \mid s, k)\exp(x^T \beta_{LM}(s)), \qquad s \le t \le s + k \qquad (31)$$

where $\beta_{LM}(s) = \gamma \cdot f\left(\frac{s}{s_L}\right)$ and $h_0(t \mid s, k) = h_0(t)\exp\left(\theta \cdot f\left(\frac{s}{s_L}\right)\right)$. The function of time *f* can be chosen as a quadratic or cubic function or can be modeled more flexibly using splines.

The fitting of the model is done by applying a Cox model on the "super" prediction dataset: this corresponds to maximize an *integrated* (over *s*) *partial log-likelihood* (details can be found in Van Houwelingen, 2007). As a form of *pseudo partial log-likelihood* it yields consistent estimates but the standard errors cannot be obtained through the second derivative and some robust procedure is required.

Extensions of the Cox model allowing for external time-dependent covariates $x(t)$ and time-dependent effects $\beta(t)$ are well known. However, these models cannot be used for prognostic purposes unless the distribution of $x(t)$ is known.

Nevertheless, within the landmark model, we can update the information at each stratum *s*. Thus, for a single landmark *s* we can postulate the model:

$$h_s(t \mid x(s)) = h_{s,0}(t)\exp(x(s)^T \beta_s(t)), \qquad t \ge s \qquad (32).$$

When $x(t)$ is measured on a continuous scale, such as a biomarker, one must be aware of the fact that the landmark value $x(s)$ might rapidly loose its relevance as a proxy of the true $x(t)$. In contrast, when $x(t)$ is a binary time-dependent indicator, the landmark model represents a flexible way to account for this variation and does not require further efforts (such as the need to build complex joint models) to obtain predictions. In addition, with the opportunity of updating the available information dynamically in time to adjust the prediction, this method is able to handle directly time-dependent external covariates too.

10 Application of flexible models to ALL data

In this chapter I present some results about the motivating example on pediatric ALL described in section 2.1. Specifically, I wish to achieve the following three goals:

1. estimating the true effect of waiting time on the hazard of failure after transplant. I address this issue using Cox, HM and LMR models to compare the two approaches to handle time described in chapter 7;
2. developing profile-specific dynamic prediction curves using HM and LMR regression models based on both the clock-back and the original scale;
3. comparing the predictive performance of the two flexible models within and between the two time scales by using the pseudo-value based time-dependent Brier score to compute the prediction error.

10.1 Impact of waiting time

To assess the effect of waiting time I considered the two regression-based approaches described in chapter 7:

i) fitting the models only on transplanted patients measuring time on the clock-back scale;
ii) fitting the models on all subjects measuring time on the original scale.

To compare the methods I applied Cox, HM and LMR regression models. Some issue about the parameter specification of the last two models is described in the following.

The HM approach was applied considering "days" as the single time-units: the dataset used for the fitting of the model was made combining all the c=333 event time-units and a random sample of b=33300 time units. The baseline hazard was modeled using a restricted cubic spline with 3 knots corresponding to the quartiles of the sampled time-points. An interaction term between transplant and time was included to accommodate for the time-varying effect of transplant.

For the LMR, I selected the values of t_{LM} on an equidistant grid ranging from 0 to 5 years with distance 1/12 (i.e. approximately every month). The dependence of the effect of transplant and of the baseline hazard on the landmarks was modeled using a restricted cubic spline with 3 knots corresponding to the quartiles of the landmarks. I considered a width of the prediction window of 1 year.

The estimates of the effect of waiting time on the hazard of failure after transplant estimated by the three considered regression models using the traditional approach based on the original time scale (since remission) are displayed in the right part of table V. The estimates of the three models are very similar and suggest the presence of a possibly harmful effect of waiting time (all p values are close to the nominal 0.05).

These results are in contrast with the findings of section 2.3, where we conclude that no influence of waiting time was present. As explained in chapter 7, this is due to the (wrong) choice of the time scale. In fact, in the estimates displayed in the right part of table V, the real effect of waiting time is mixed with the effect of the sojourn time: comparing the hazards at t of two patients transplanted at different waiting times (before t), the patient with a lower waiting time has consequently a longer sojourn time and thus, given the decreasing behavior of the hazard under transplant in time (section 2.3), he/she is experiencing a lower hazard than the other patient.

A possible solution is to consider a model fitted only on transplanted patients measuring time on the clock-back scale (since transplant). With this approach, in fact, the effect of the sojourn time is naturally accounted in the time-scale and the true effect of waiting time can be quantified by the estimated coefficient. On the application example data, the estimated effect of waiting time in the clock-back scale is very close to 0 for all the three models (left part of table V). This is in line with the conclusions derived from figure 3 in section 2.3, which is a further indication that this analysis is reliable to assess the impact of the waiting time. A further advantage of the clock back scale is that the estimated effect of the waiting time can also be accounted for in the (standard or dynamic) profile-specific prediction curves where the semi-Markov property is not fulfilled.

Table IV. Assessment of the effect of waiting time to transplant on the hazard of the transplanted patients. Estimates by the Cox, HM and LMR model in the original and clock back scale are shown.

Model	Clock back scale			Original scale		
	log(HR)	Standard error	p-value	log(HR)	Standard error	p-value
Cox	0.00007	0.00072	0.922	0.0016	0.00072	0.030
HM	0.00004	0.00072	0.953	0.0014	0.00074	0.054
LMR (k=1)	0.00084	0.00091	0.360	0.0018	0.00078	0.023

10.2 Dynamic predictions

In this section I present results about the profile-specific dynamic prediction curves estimated using the HM and LMR flexible models. I considered two different approaches of handling time:

i) original scale for patients treated with chemotherapy-only and for the transplanted patients during the period before transplant, "clock-back" scale for the transplanted patients during the period after transplant,
ii) original scale for all patients, using left truncation to account for treatment switch.

Some further aspects about the parameter specification of the two flexible models is described in the following. The HM approach was applied considering "days" as the single time-units: the dataset used for the fitting of the model was made combining all the c=333 event time-units and a random sample of b=33300 time units. The baseline hazard was modeled using a restricted cubic spline with 3 knots corresponding to the quartiles of the sampled time-points. An interaction term between transplant and time was included to accommodate for the time-varying effect of transplant.
For the LMR, I selected the values of t_{LM} on an equidistant grid ranging from 0 to 5 years with distance 1/12 (i.e. approximately every month). The dependence of the effect of transplant and of the baseline hazard on the landmarks was modeled using a restricted cubic spline with 3 knots corresponding to the quartiles of the landmarks. I considered a width of the prediction window of 1 year.
In figure 12 the logarithm of the hazard ratio (HR) for transplant vs chemotherapy obtained by the HM model (A) and by the LMR model (B) using the clock-back scale and by the two models using the original scale (C and D, respectively) are shown. Despite the different scale used, the shape of the coefficient of treatment is similar between A and C (as well as between B and D) due to the adjusting effect of the coefficient of waiting time which acts in the models based on the original scale.
It is worth noting that this time-varying coefficient of transplant has a different interpretation between the two models: for the HM model it represents the standard time-varying log(HR) over time t while for the LMR model it is a function of t_{LM} and represents the estimated β_{LM} over the window $[t_{LM}\ ,\ t_{LM}\ +k]$ assuming proportional hazard within the window. Thus, to obtain the prediction of the fixed window conditional incidence at $t + k$, in the case of the HM model the whole interval of values between t and $t + k$ of the log(HR) are involved, whereas in the case of the LMR model the value of the log(HR) at each $t = t_{LM}$ is used.

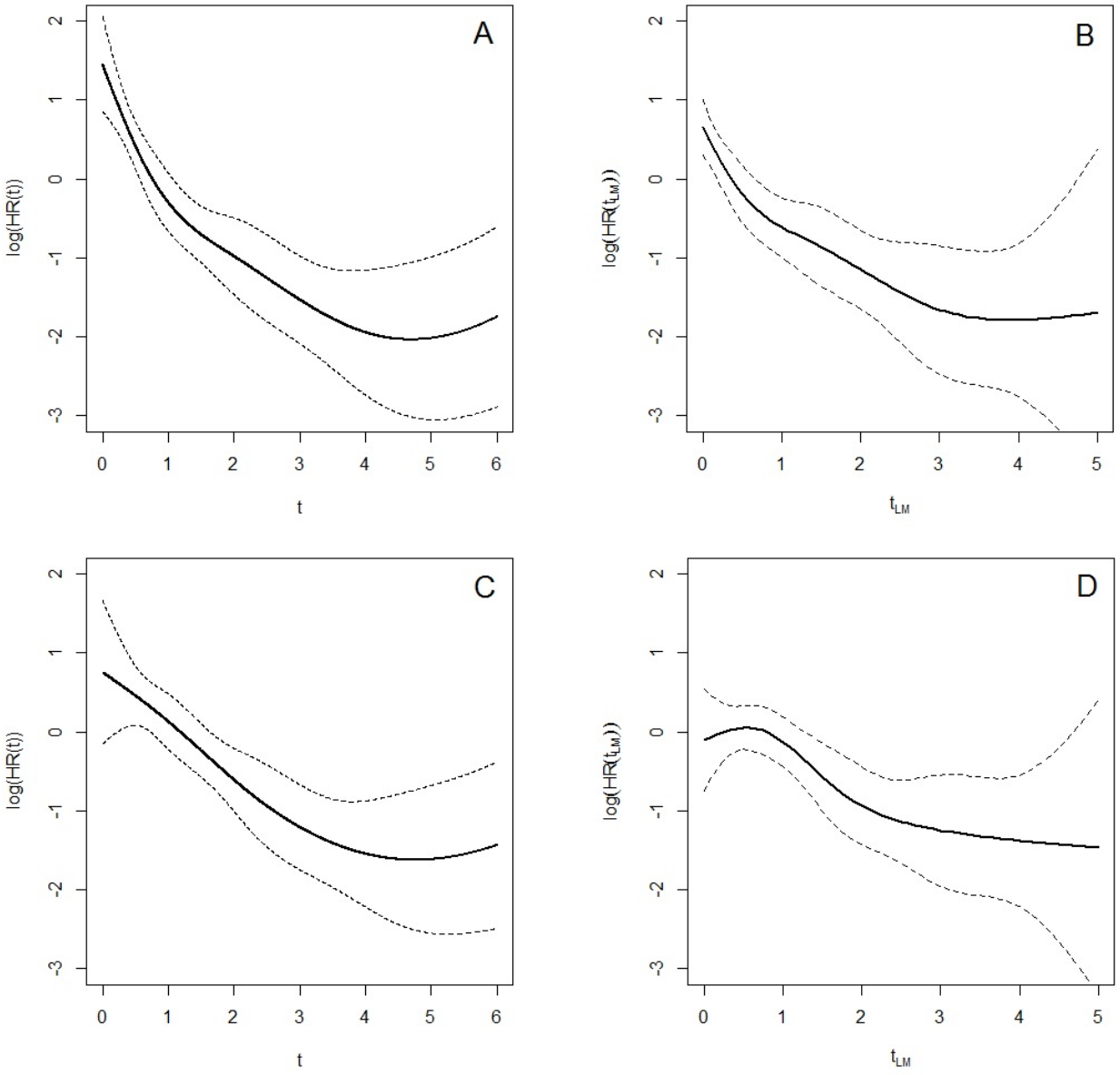


Figure 12. Estimated log(HR) for transplant vs chemotherapy by the HM model (left panels) and by the LMR model (right panels) measuring time on the clock-back scale for transplanted patients (upper panels) and measuring time in the original scale for all patients (lower panels). Dashed lines represent the point-wise 95% confidence interval.

Dynamic predictions obtained by the two models measuring time on the clock-back scale and on the original scale are displayed in figure 13 (panels A and B, respectively). For both models, the chosen width of the prediction window *k* was 1 year. The curves show dynamic predictions of patients with a median risk profile (black and gray lines), defined as the median of the prognostic score $x^T\beta$ which is based on the effects, estimated by the two models, of all the risk factors considered other than treatment (none of these factors showed a time-varying effect), reported in table V. Also predictions of patients with a risk profile equal to the first (blue and light blue curves) and third (red and orange) quartile are displayed. Predictions are shown for transplanted and non-transplanted

patients separately. The solid curves at each time point *t* describe the probability of occurrence of relapse or death within the next year for a patient event-free and still not transplanted at *t*. The dashed curves at each time point *t* describe the probability of occurrence of relapse or death within the next year for a patient event-free at *t* and transplanted before *t*.

It is worth to note that, despite of the different machinery working behind the two models, the results in terms of prediction are very similar. In fact, even if the model parameters estimated by LMR model are not reliable relatively to the instantaneous hazard, the estimated incidence based on these parameter is good due to averaging effects of integrated quantities (van Houwelingen 2008).

Dynamic prediction curves based on the two different time-handling approaches are also similar. The only evident difference regards, in the initial times, the curves of transplanted patients. This can be due to the fact that the estimates based on the original scale are less precise due to the occurrence of few failures in the initial times. It also should be noted that, even if the parameter estimates of the original scale based models are valid only under markovianity, the dynamic predictions are consistent even for a semi-Markov process. This is consistent with the validity of the estimates on the state probabilities evaluated in section 2.3 (Glidden 2002).

Table V. Estimated coefficients and standard errors of the baseline risk factors in the ALL data.

Baseline risk factors	Clock-Back scale				Original scale			
	HM		LMR		HM		LMR	
	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Male (ref. female)	0,104	0,116	0,082	0,122	0,091	0,116	0,108	0,127
Age at diagnosis, years (ref. >15)								
0-3	-0,792	0,271	-1,012	0,311	-0,818	0,270	-1,201	0,321
3-6	-0,430	0,231	-0,479	0,262	-0,437	0,231	-0,580	0,271
6-10	-0,313	0,225	-0,325	0,255	-0,304	0,225	-0,446	0,266
10-15	-0,247	0,222	-0,250	0,251	-0,250	0,222	-0,342	0,260
Leukocyte count, per mL (ref. > 100)								
0-10	-0,762	0,164	-0,653	0,173	-0,765	0,164	-0,674	0,179
10-25	-0,607	0,167	-0,481	0,181	-0,614	0,167	-0,513	0,189
25-50	-0,471	0,183	-0,391	0,198	-0,487	0,183	-0,449	0,205
50-100	-0,477	0,175	-0,393	0,193	-0,470	0,175	-0,391	0,199
Poor response (ref. good response)								
Not known	0,231	0,161	0,198	0,179	0,215	0,161	0,153	0,190
in bone marrow	0,179	0,153	-0,039	0,180	0,201	0,153	-0,064	0,185
in peripheral blood	0,686	0,240	0,714	0,226	0,708	0,240	0,781	0,218

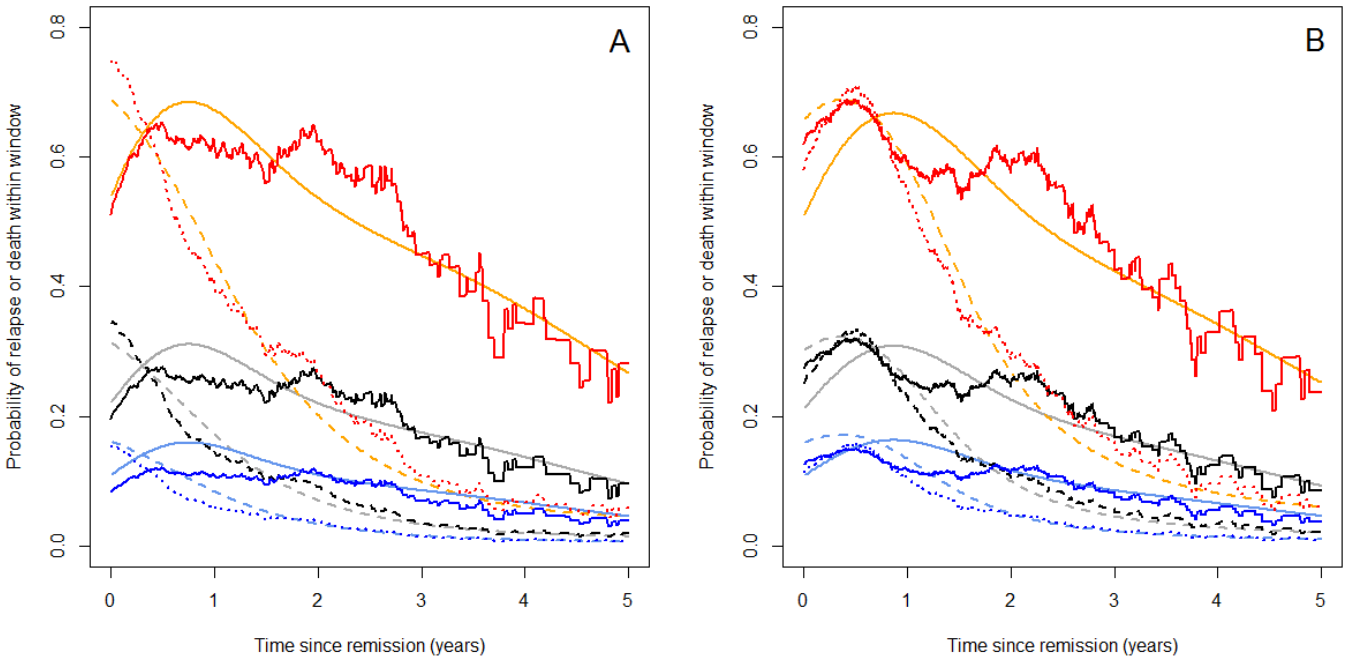


Figure 13. Dynamic predictions of relapse or death within 1 year obtained by the HM (subtle colors) and the LMR (loud colors) models based on the clock back scale (A) and on the original scale (B). Transplant=dashed lines, Chemotherapy=solid lines. The predictions are shown for a patient with median risk profile (black and gray lines) and with a risk profile equal to the first quartile (blue and light blue lines) and third quartile (red and orange lines).

10.3 Prediction error: comparison of flexible models

In this section I wish to estimate the prediction error of the dynamic incidence $F_k(t)$ at different prediction times $t \in \{0;0.5;1;2;4\}$ considering a prediction window of width k=1 estimated by the two flexible models using the two time-handling approaches on the ALL data. The model specifications and the time-scales considered are the same used in section 10.2 to develop the profile specific dynamic prediction curves. To this purpose I used the procedure described by Cortese et al. (2013), which is based on the application of the time-dependent Brier Score (BS) (Schoop et al. 2008) using the Pseudo-Values method (Andersen and Pohar Perme 2010) to solve the problem of censoring. Since the models have generally a worse predictive performance when applied to data different than those used for the estimation, a cross validation technique was used, consisting in an iterative procedure that can be described as follows.

The models were fitted on a training data set, obtained by randomly selecting 2/3 of the subject included in the original ALL dataset. Secondly, the pseudo-value based time-dependent BS at each

prediction time was estimated on a testing dataset formed by the remaining 1/3 of subject excluded for the training data set. This procedure was repeated 100 times and the averaged BS over the total iterations was computed.

In table VI, the results of this analysis are displayed. It can be noticed that the prediction error is very close among all methods, i.e. using either the HM or the LMR model and measuring time either in the clock-back or original scale, although the HM model on the clock-back scale generally shows a better performance.

Table VI. Averaged Pseudo-Value-based Brier score $BS(t, t+k)$ on the 100 testing datasets.

t	$t+k$	Average Brier score			
		Clock Back scale		Original scale	
		HM	LMR	HM	LRM
0	1	0.172	0.172	0.177	0.173
0.5	1.5	0.217	0.220	0.218	0.218
1	2	0.185	0.186	0.188	0.185
2	3	0.130	0.130	0.130	0.129
4	5	0.055	0.055	0.055	0.056

11 Simulation protocol (2)

11.1 Simulation settings

To support the conclusions on both the issue regarding the estimation of the impact of waiting time and the dynamic prediction, I conducted another simulation protocol, similar to that described in chapter 6. I generated 1000 dataset of both 400 and 500 observations each. The latent failure times under transplant T_{ptr}, under chemotherapy T_{ch}, and the latent waiting time W (and thus $T_{tr} = W + T_{ptr}$) were simulated using the inversion method of Bender et al. (2005), while the censoring time *C* was directly simulated using a uniform distribution: $C \sim Uniform(a = 1, b = 15)$.

The latent failure time after transplant switching T_{ptr} was generated in three different ways to represent several conditions on process memory:

1. non-monotone hazard under transplant and no effect of the waiting time (semi-Markov), by $T_{ptr} \sim Burr(k = 0.3, c = 1.1, s = 1)$;
2. non-monotone hazard under transplant and proportional hazard effect of the waiting time (extended semi-Markov), by $T_{ptr} \sim Burr\left(k = 0.3\exp(\beta W), c = 1.1, s = 1\right)$;
3. decreasing hazard under transplant and non-proportional hazard effect of the waiting time (extended semi-Markov), by modelling the baseline hazard as a $Weibull\left(\lambda = 0.2\right)$, leading to a non-definable function of T_{ptr}.

T_{ch} was generated in the same way in all the scenarios, by $T_{ch} \sim LogLogistic(\alpha = 0.4, \beta = 1)$ as well as the latent waiting time *W*, by $W \sim Exp(\theta = 0.4)$. The details of the simulation protocol are summarized in table VII, while the graphical representation of the hazards under transplant and under chemotherapy in the three scenarios are shown in figure 14.

11.2 Impact of waiting time

To compare the two different approaches of handling time when using a regression model to assess the impact of waiting time to transplant on the hazard of failure after transplant (discussed in chapter 7), I considered the Cox model as well as the HM and LMR models (using the same parameter settings described in section 10.1).

I focused on the estimated coefficient of waiting time. In the first scenario (semi-Markov) this was set equal to 0, indicating that the waiting time has no influence on the hazard of failure after

transplant. In the second scenario (extended semi-Markov with PH effect) the coefficient set equal to 0.05, meaning that a longer waiting time causes a worse prognosis. In the third scenario (extended semi-Markov with non-PH effect), I set a time-varying effect (considering a logarithm transformation of time): the waiting time coefficient was correctly specified in all models and was evaluated at t=1. The result of the simulations are summarized in table VIII.

Considering the results about the estimated coefficient using the clock-back scale approach, we note that the average estimates of the coefficient of all models are very close to the real value.

In contrast, when adopting the original time scale, the estimated effect of waiting time is strongly biased by all the three models, due to the wrong interpretation of these coefficients as the impact of waiting time on the hazard after transplant. The variances of the estimators are consistent with each-other.

Concerning the third scenario we also note that, even if the model is not correctly specified, but a flexible specification (i.e. using splines) is adopted, the estimated time-varying coefficient of the waiting time using regression models based on the clock-back scale is approximately similar to the true one (figure 15).

11.3 Dynamic predictions

Using the same simulation protocol I compare the performance of the HM and LMR flexible models with both time-handling approaches to develop profile specific dynamic predictions. The same parameter settings used in section 10.2 are considered. In the third scenario a flexible specification (using splines) of the time-varying nature of the waiting time effect was adopted.

For each scenario, I compared the true and the estimated values of the fixed window conditional failure function $F_k(t)$ at different prediction times $t \in \{0;0.5;1;2;5\}$ choosing a prediction window of width k=1. Three different treatment profiles at time t were considered: chemotherapy only, transplanted at time t and transplanted at t-1 (only for t=2 and t=5). For each comparison the mean square error (MSE) was computed.

Observing the results, reported in table IX, we note that both models show a better performance in terms of MSE when the clock-back approach to measure time is adopted, in all three scenarios. This is probably due to the fact that using different time scales to model the baseline hazard of patients under chemo and of patients already transplanted is more efficient than assuming that the effect of both waiting and sojourn time is captured by the coefficients of the treatment and waiting time covariates.

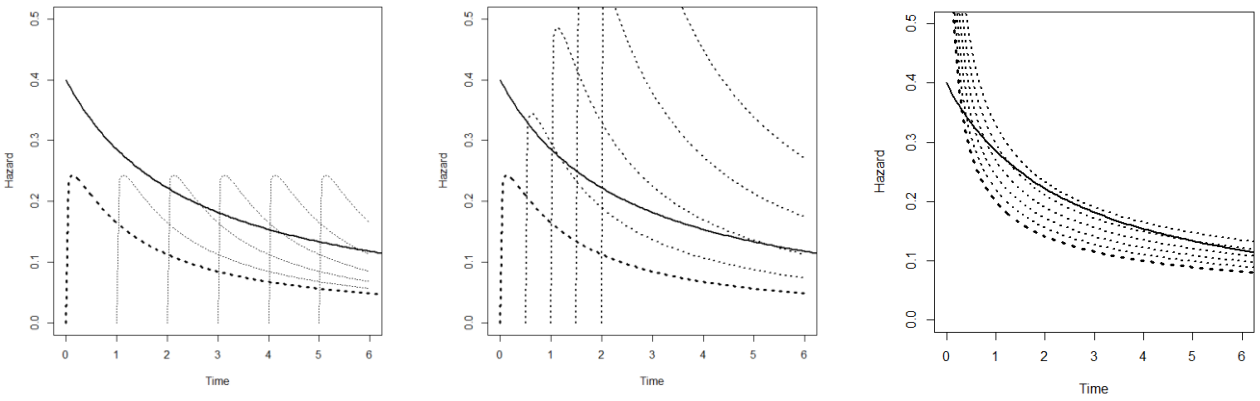
Concerning the comparison of the two models, we observe that the HM model shows systematically lower MSEs than the LMR model. This could lead to the conclusion that a flexible full parametric

model (such as the HM) allowing for an explicit definition of time as a covariate is a better option the a Cox-based model (such as the LMR) that, although the possibility to model flexibly the time-dependent nature of the covariates effects, still relies on a semi-parametric specification of the baseline hazard.

As a final consideration, it must be noted that the predictions at time t=0 for patients transplanted at t (i.e. at remission) are generally worse than on other time-points (except for the clock-back scale in the first scenario). The error increases dramatically when the model is not correctly specified: predictions for transplanted at t=0 and t=0.5 in the third scenario (results not shown) are strongly erroneous due to the very extreme settings of the simulation scenario. This can be explained by observing that the model is used to estimate a counterfactual quantity, since there are not patients receiving transplant at t=0. As a consequence, very few events occur among transplanted at the beginning of the follow-up causing the model to underestimate the true fraction of patients that would have failed if they had received transplant since the beginning.

Table VII. Details on the simulation protocol.

Scenario	T_{ptr}	T_{ch}	W	C
1. Semi-Markov	$h(t)=\dfrac{\frac{kc}{s}\left(\frac{t}{s}\right)^{c-1}}{1+\left(\frac{t}{s}\right)^{c}}$ $S(t)=\exp\left(-k\ln\left(1+\left(\frac{t}{s}\right)^{c}\right)\right)$ $k=0.3$ $c=1.1$ $s=1$	$h(t)=\dfrac{(\beta\alpha)(t\alpha)^{\beta-1}}{1+(t\alpha)^{\beta}}$ $S(t)=\dfrac{1}{1+(t\alpha)^{\beta}}$ $\alpha=0.4$ $\beta=1$	$h(t)=\theta$ $S(t)=\exp(-\theta t)$ $\theta=0.2$	$F(t)=\begin{cases}0 & t<a\\ \dfrac{t-a}{b-a} & a\le t<b\\ 1 & t\ge b\end{cases}$ $a=1$ $b=15$
2. Extended semi-Markov with PH effect of W	$h(t)=\dfrac{\frac{kc}{s}\left(\frac{t}{s}\right)^{c-1}}{1+\left(\frac{t}{s}\right)^{c}}\exp(\eta W)$ $S(t)=\exp\left(-k\exp(\eta W)\ln\left(1+\left(\frac{t}{s}\right)^{c}\right)\right)$ $k=0.3$ $c=1.1$ $s=1$ $\eta=0.05$	$h(t)=\dfrac{(\beta\alpha)(t\alpha)^{\beta-1}}{1+(t\alpha)^{\beta}}$ $S(t)=\dfrac{1}{1+(t\alpha)^{\beta}}$ $\alpha=0.4$ $\beta=1$	$h(t)=\theta$ $S(t)=\exp(-\theta t)$ $\theta=0.2$	$F(t)=\begin{cases}0 & t<a\\ \dfrac{t-a}{b-a} & a\le t<b\\ 1 & t\ge b\end{cases}$ $a=1$ $b=15$
3. Extended semi-Markov with non-PH effect of W	$h(t)=\lambda t^{\lambda-1}\exp(\eta W+\gamma W\log(t))$ $S(t)=\exp\left(-\dfrac{\lambda\exp(\eta W)t^{\gamma W+\lambda}}{\gamma W+\lambda}\right)$ $\lambda=0.2$ $\eta=0.1$ $\gamma=0.3$	$h(t)=\dfrac{(\beta\alpha)(t\alpha)^{\beta-1}}{1+(t\alpha)^{\beta}}$ $S(t)=\dfrac{1}{1+(t\alpha)^{\beta}}$ $\alpha=0.4$ $\beta=1$	$h(t)=\theta$ $S(t)=\exp(-\theta t)$ $\theta=0.2$	$F(t)=\begin{cases}0 & t<a\\ \dfrac{t-a}{b-a} & a\le t<b\\ 1 & t\ge b\end{cases}$ $a=1$ $b=15$



1. Semi-Markov | 2.Extended semi-Markov, PH effect of W | 3. Extended semi-Markov, non PH effect of W

Figure 14. Hazard of failure under transplant (dashed lines) and under chemotherapy (solid lines) in the three simulation scenarios, plotted in the original scale. Hazards under transplant according to different waiting times are shown (bold dashed lines correspond to hazard under transplant with W=0).

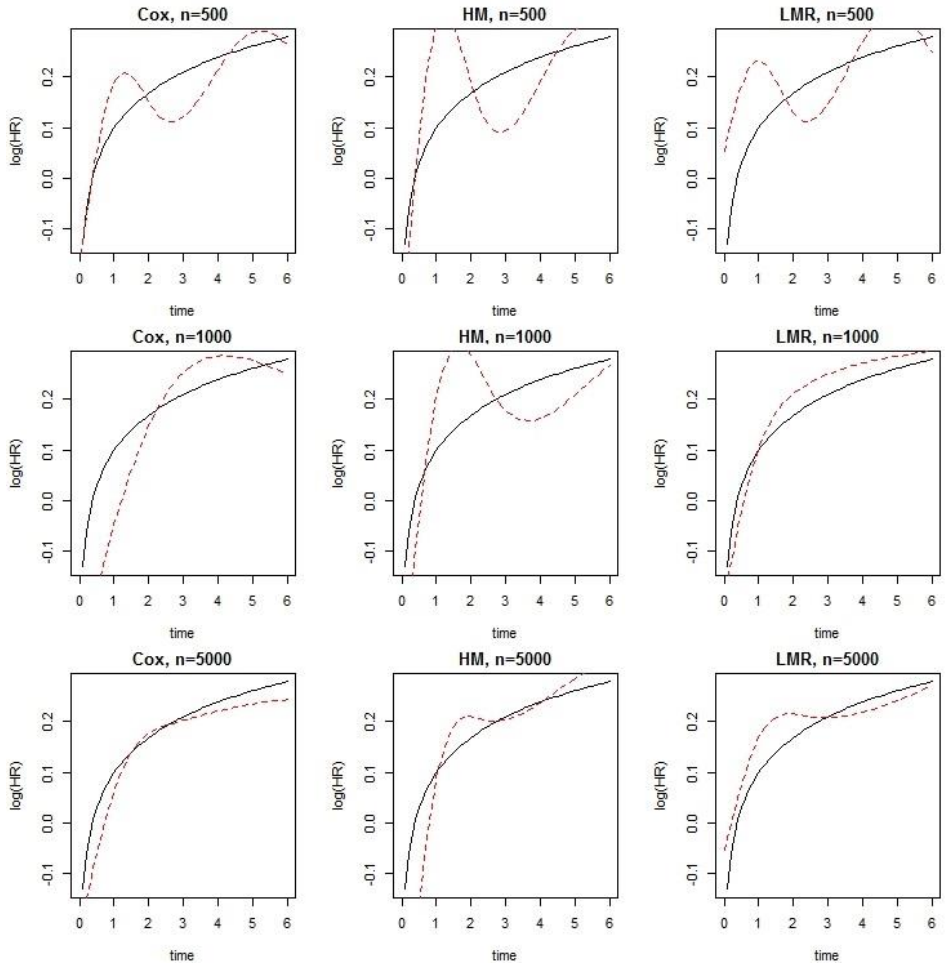


Figure 15. Approximation of the estimated time-varying coefficient of waiting-time modeled with spline functions (dashed line) to the true one (solid line) in the 3rd simulation scenario, using the Cox, HM and LMR models in the clock-back scale. For each plot, a single dataset of 500, 1000 and 5000 observations was set up for the estimation.

Table VIII. Simulation results (1000 datasets of 400 and 1000 of 500 observations) about the performance of Cox, HM and LMR models with the two time-scales on the effect of waiting time estimation. Bias is calculated as average on the 1000 samples of the difference between the estimated and the true value. Variance (Var) is calculated as variance of the estimate of the 1000 samples.

Scenario	η	Cens. (%)	Model	Clock Back scale						Original scale					
				n=400			n=500			n=400			n=500		
				Mean ($\hat{\beta}$)	Bias $*10^2$	Var $*10^2$	Mean ($\hat{\beta}$)	Bias $*10^2$	Var $*10^2$	Mean ($\hat{\beta}$)	Bias $*10^2$	Var $*10^2$	Mean ($\hat{\beta}$)	Bias $*10^2$	Var $*10^2$
1. Semi-Markov	0	33.0	Cox	-0.002	-0.178	0.432	-0.003	-0.272	0.328	0.213	21.307	0.590	0.206	20.600	0.503
			HM	-0.002	-0.182	0.436	-0.003	-0.270	0.330	0.214	21.428	0.604	0.207	20.654	0.504
			LMR	-0.003	-0.263	0.519	-0.003	-0.321	0.395	0.209	20.855	0.932	0.200	20.036	0.686
2. Extended semi-M. with PH effect of W	0.05	31.7	Cox	0.050	0.033	0.352	0.049	-0.076	0.267	0.261	21.113	0.561	0.263	21.307	0.435
			HM	0.051	0.061	0.354	0.049	-0.062	0.269	0.262	21.228	0.566	0.263	21.337	0.428
			LMR	0.050	0.014	0.426	0.049	-0.137	0.317	0.262	21.246	0.754	0.259	20.926	0.554
3. Extended semi-M. with non-PH effect of W (log(t))	0.1	22.9	Cox	0.100	-0.005	0.247	0.101	0.057	0.243	0.181	8.117	0.287	0.181	8.128	0.189
			HM	0.099	-0.063	0.207	0.100	0.036	0.207	0.178	7.772	0.745	0.180	7.965	0.537
			LMR	0.100	-0.022	0.396	0.100	-0.043	0.298	0.255	15.540	0.417	0.274	17.440	0.370

Table IX. Simulation results (1000 datasets of 400 and 1000 of 500 observations) about the performance (measured in terms of MSE) of the flexible models with the two time-scales in dynamic predictions.

Scenario	Profile	T	$t+k$	$F_k(t)$	Mean square error * 1000							
					Clock-back scale				Original scale			
					N=400		N=500		N=400		N=500	
					HM	LMR	HM	LMR	HM	LMR	HM	LMR
1. Semi-Markov	Chemo	0	1.0	0.286	0.516	0.674	0.378	0.478	0.490	1.010	0.349	1.037
	Tr. at *t*	0	1.0	0.188	0.901	1.414	0.688	1.116	5.300	12.278	3.909	13.433
	Chemo	0.5	1.5	0.250	0.576	0.697	0.461	0.572	0.534	0.721	0.414	0.538
	Tr. at *t*	0.5	1.5	0.188	0.901	1.414	0.688	1.116	1.922	3.575	1.351	2.480
	Chemo	1	2.0	0.222	0.484	0.796	0.395	0.666	0.514	0.939	0.423	0.624
	Tr. at *t*	1	2.0	0.188	0.901	1.414	0.688	1.116	1.740	2.929	1.275	1.989
	Chemo	2	3.0	0.182	0.665	1.135	0.591	0.921	0.620	1.023	0.548	1.005
	Tr. at *t*	2	3.0	0.188	0.901	1.414	0.688	1.116	1.399	3.100	1.216	2.245
	Tr. at *t-1*	2	3.0	0.127	0.479	0.720	0.396	0.586	0.587	1.486	0.475	0.950
	Chemo	5	6.0	0.118	1.006	2.854	0.850	2.163	1.245	3.004	1.033	1.863
	Tr. at *t*	5	6.0	0.188	0.901	1.414	0.688	1.116	2.782	6.849	2.361	5.267
	Tr. at *t-1*	5	6.0	0.127	0.479	0.720	0.396	0.586	0.925	2.825	0.673	2.087
2. Extended semi-Markov with PH effect of W	Chemo	0	1.0	0.286	0.507	0.644	0.379	0.481	0.485	1.154	0.356	1.090
	Tr. at *t*	0	1.0	0.188	1.318	1.962	1.026	1.519	5.340	14.720	3.906	10.837
	Chemo	0.5	1.5	0.250	0.580	0.702	0.473	0.556	0.554	0.613	0.453	0.555
	Tr. at *t*	0.5	1.5	0.192	1.162	1.737	0.898	1.352	1.938	3.367	1.373	2.752
	Chemo	1	2.0	0.222	0.493	0.810	0.401	0.705	0.521	0.825	0.441	0.703
	Tr. at *t*	1	2.0	0.196	1.050	1.580	0.807	1.235	1.758	3.250	1.272	2.039
	Chemo	2	3.0	0.182	0.666	1.072	0.567	0.957	0.645	1.195	0.528	0.882
	Tr. at *t*	2	3.0	0.205	0.976	1.480	0.743	1.162	1.557	3.007	1.357	2.790
	Tr. at *t-1*	2	3.0	0.196	0.559	0.820	0.464	0.710	0.642	1.145	0.524	1.062
	Chemo	5	6.0	0.118	1.017	2.745	0.757	2.133	1.269	1.956	0.944	1.982
	Tr. at *t*	5	6.0	0.234	2.362	3.339	1.814	2.618	3.657	8.470	3.245	6.422
	Tr. at *t-1*	5	6.0	0.224	0.982	1.327	0.769	1.115	1.084	3.268	0.874	2.687
3. Extended semi-Markov with non-PH effect of W	Chemo	0	1.0	0.286	0.525	0.691	0.394	0.530	0.493	1.264	0.362	1.127
	Chemo	0.5	1.5	0.250	0.561	0.719	0.453	0.572	0.578	0.627	0.473	0.526
	Chemo	1	2.0	0.222	0.535	0.827	0.432	0.691	0.496	0.833	0.412	0.704
	Tr. at *t*	1	2.0	0.357	1.658	3.735	1.323	2.854	6.772	13.134	5.566	9.862
	Chemo	2	3.0	0.182	0.599	0.933	0.511	0.882	0.694	1.195	0.582	0.991
	Tr. at *t*	2	3.0	0.263	1.777	4.892	1.507	4.234	3.421	9.562	2.725	5.685
	Tr. at *t-1*	2	3.0	0.167	1.508	1.350	1.279	1.149	1.505	2.425	1.160	1.721
	Chemo	5	6.0	0.118	0.890	2.660	0.670	2.137	1.139	2.212	0.847	1.767
	Tr. at *t*	5	6.0	0.176	2.706	4.695	2.238	3.750	12.750	45.453	11.399	39.809
	Tr. at *t-1*	5	6.0	0.295	2.165	2.574	1.600	2.171	6.352	9.476	5.914	6.122

12 Discussion

The graphical representation of the survival experience according to a non-reversible time-dependent treatment is still not well established in the statistical literature. It is well known in fact that the naive application of the Kaplan-Meier estimator according to the treatment given is affected by the “immortal time bias”.

The most popular methods specifically developed to address the issue of a time-dependent treatment were proposed focusing on the practical estimation, but lacking a proper analysis of the corresponding theoretical quantities. This might led to the general belief that the Simon and Makuch approach, which allows to dynamically updating the risk sets, provides a more reliable analysis than the pragmatic Landmark approach, which is based on the fixed classification according to the treatment status at the landmark time. In addition, the interpretability of the curves estimated using the Simon and Makuch approach has been heavily criticized (Beyersmann 2006). I very carefully went through all approaches used with time dependent treatment indicators, reviewing the estimators and identifying the corresponding theoretical quantities.

Concerning the landmark method, I reviewed as the estimated quantities are conditional on being failure free at the landmark and conditional on treatment classification at landmark. They are useful for a pragmatic representation of the “average” survival experience of transplanted patients up to landmark accounting for the natural heterogeneity of the waiting times, and the “average” survival experience of patients not transplanted up to landmark accounting for the natural heterogeneity of the times to treatment switch occurring after landmark. Additionally, the landmark curves refer to "real" populations and summarize the expected survival considering the possibility that patients who have not yet switched to the alternative treatment by time LM might do the switch afterwards, as in real life. This approach is suitable to be used in the context of dynamic prediction since the landmark could be moved forward in time.

Concerning the Simon and Makuch method, I proved that it provides estimates of quantities defined on the ground of a counterfactual treatment administration set at the beginning of the follow-up. The effect of transplant is represented by the survival experience of a counterfactual patient as if he/she would receive transplant since the beginning of the follow-up. The effect of chemotherapy is represented estimating the survival experience of a counterfactual patient as if he/she would be treated with chemotherapy only during all follow-up. The early landmark setting is needed only to enable the initial estimate of the hazard in the transplanted group and does not share interpretative

aspects with the landmark method. The validity of the estimates for the transplanted group is guaranteed only under the restrictive Markov assumption. I proposed a simple modification of the Simon and Makuch estimator related to the transplant group to address the case of semi-markovianity. The hazard under transplant was estimated on the clock back scale to benefit from the semi-Markov assumption, but then it was projected on a counterfactual subject considered in the original scale as if he/she was under transplant since the beginning of the follow-up. Although this may appear counterintuitive, this is actually what is done by the SM estimator (10) directly in the original scale through an improper use of left truncation. Left truncation, in fact, would imply that patients are transplanted since the beginning of the follow-up and yet not observed until the real treatment switch. The counterfactual survival experience obtained forcing the treatment to be fixed since the beginning of the follow-up in SM and CB approaches is a pragmatic way to contrast the cumulative effect of treatments in their velocity of failure development. It may happen, however, that early treatment switch is rarely observed and/or the waiting times are very heterogeneous. Thus, the quantities estimated by the Simon and Makuch and the proposed approach for semi-Markov data could seem too artificial. As an alternative, I proposed to focus on dynamic counterfactual survival probabilities defined on a patient under chemotherapy and failure-free at a chosen landmark. The contrast between the survival functions considers either a treatment switch exactly at landmark or continuing chemotherapy. This would also provide an approach to get insights into the impact of different waiting times to transplant on survival when compared to that under chemotherapy only.

The methods reviewed and the novel approaches proposed answer in a simple way the issue of estimating survival curves conditional on a BNRTD treatment administration. Multi-state models (Putter et al. 2007, Andersen and Keiding 2002) can also be used to analyze the type of data presented. This will enable to describe the evolution of the process through transition probabilities. In particular, multi-state models could provide landmark curves on the future state occupation conditional on the status occupied at the landmark time (Putter et al. 2007, Iacobelli et al. 2008). These curves, however, in the application example used through the paper, would be based on the joint distribution between the survival time and the treatment, which would naturally change in time. This would lead to a different analysis than that based on the counterfactual treatment administration here presented. In addition, the large potential of multi-state models (including the possibility of applying regression models to adjust the estimates for relevant covariates) is highly counterbalanced by the complexity of the approach, also in practical terms. The estimation is rather feasible under the Markov assumption (Putter et al. 2007, de Wreede et al. 2010), and more difficult in non-Markov settings (Datta et al. 2000, Meira-Machado et al. 2006, Andersen and Perme 2008).

I proposed a simple modification of the Simon and Makuch estimator related to the transplant group to address the case of semi-markovianity. The dynamic flavor of the landmark method and the counterfactual interpretation of the Simon and Makuch method motivated the proposal of a novel approach based on counterfactual dynamic quantities and suitable estimators.
These quantities enlighten the impact on survival of a treatment switch administered at given landmark times with respect to the expected survival without the treatment switch. This provides a novel tool for dynamically assessing the impact of the time of treatment administration.

The results presented in the first part of the thesis were useful to make some considerations on a further aspect of survival clinical studies: the need of regression models to study the impact of covariates (other than treatment) on the hazard of failure and to account for these effects in developing survival/incidence prediction curves.
In fact, especially from the point of view of a newly diagnosed patient, the choice of the best treatment option, according to the patient's individual characteristics, can be investigated focusing on the profile-specific prediction. The Cox model, which is the most popular regression model in applied survival studies, presents two strong limitations to address the issue of profile-specific survival curves: the fact that in the presence of time-varying effects and/or time-dependent covariates it does not allow to derive survival/incidence predictions; the non-parametric nature of the baseline hazard estimator that causes a non-attractive steps-in-time shape of the curves.
Besides standard parametric models, as a reasonable alternative to handle non-standard situations, I reviewed two flexible modeling solutions which do not require assuming a well-defined density distribution of time, and enable to account for time-varying effects and time-dependent covariates: the Hanley-Miettinen model and the Landmark regression model.
The first is a fully-parametric model that uses a sampling method of the time-units of each patient to obtain estimates of the hazard function through logistic regression while the second is a semi-parametric model fitted through the maximization of a pseudo-partial likelihood and is based upon the idea of combining several Cox models fitted to data including patients at risk at the beginning of a sliding fixed-width time-window and with administrative censoring at the end. A comparative summary of the characteristics of the two models is given in table X.
Both models, in addition to providing a solution to overcome the limits of the Cox model when estimating the survival/incidence function, can be implemented within any standard statistical software and require only small computational efforts for data preparation.
Another regression method based on Pseudo-Values was considered at the beginning of my research (Andersen and Pohar Perme 2010, Ambrogi et al. 2008). Although it represents a flexible

regression approach of direct estimation of survival/incidence probabilities, it is not reliable in the presence of time-dependent covariates, since in this situation the Pseudo-Values are affected by immortal time bias. An extension of this method allowing to overcome this issue is currently under investigation: this was the topic of an oral presentation (Potschger et al. 2014) at the 35th annual conference of the International Society for Clinical Biostatistics, held in Vienna on August 2014.

Table X. A comparative summary of the characteristics of the HM and LMR models.

Characteristics	**Hanley-Miettinen model**	**Landmark regression model**
Shape of curves	smooth function	step function
Baseline hazard	full parametric specification	piecewise semi-parametric specification
Time-varying effects	full parametric specification	piecewise parametric specification
Time-dependent covariates	value updated in time through reshape of the data	value updated in time through reshape of the data
Fitting procedure	generalized linear model	pseudo-partial likelihood

I discussed and applied the use of the Cox model and the two flexible regression models specifically to cover two aspect. First, to study the role of the waiting time to transplant on the development of the disease. The traditional approach to this problem is to use a regression model measuring time on the original scale (since remission) for all patients and setting the waiting as the (delayed) entry for transplanted patients. As a first limit, I point out that the estimates of this model are valid only under the Markov assumption, i.e. when the hazard of failure under transplant in time does not depend of the waiting time to transplant. Secondly, this would induce to observe a spurious effect of the waiting time which is indeed a mixture including the true effect and that of the time since transplant.

As a possible alternative solution I propose a simple approach that consist in fitting a model only on the transplanted patients. This will allow to measure time on the clock-back scale since the disease process of every patients starts from the entry in the transplant state. As a consequence, the estimates are automatically valid even in non Markov contexts. Moreover, the coefficient of waiting time truly represents the impact of the waiting time to transplant on the hazard of failure after transplant, since the effect of the sojourn time is fully captured by the baseline hazard.

The second aspect, which can be addressed only by the flexible regression models, is about developing profile-specific dynamic predictions. This means to study the probability of failure of patients within a time window, of fixed width *k*, that moves forward in time. In other words, the focus is on predictions of failure within *k* time-units, conditional on being failure-free at the beginning of the time window.

The choice of the width of the time-window should depend on clinical considerations. It is interesting to note, however, that choosing a short window will lead the curves to approximately represent the instantaneous hazard function. For this reason, this approach allows to figure out more clearly the evolution of the disease than standard survival/incidence curves, but keeping the clear interpretation of the estimates as simple probabilities.

I discussed the issue regarding the choice of the time-scale when developing dynamic prediction. Models based on the original time scale should provide good estimates of the survival/incidence quantities both in non Markov settings, due to the averaging effects of integrated quantities. However, a clock-back scale approach, i.e. measuring time since remission for patients under chemotherapy and resetting the clock to zero as patients become transplanted, is again advisable. Firstly, because the estimated effect of the covariates on the hazard are clearly interpretable. Secondly, because the predicted probabilities are even more reliable due to the more efficient model parametrization.

Despite the conceptual differences between the HM and LMR regression techniques, the models showed (similar) good performances in term of prediction error both on applied data and on simulations. This is true even under the third scenario, i.e. extended semi-Markov with time-varying effect of waiting time, when the models were not correctly specified but some additional flexibility is introduced. If the estimates of the time-varying parameters are of interest, however, the HM model should be preferred since it allows to explicitly model time as a covariate leading to a clearer interpretation of the impact of covariates; in contrast, the LMR model should be used only for predictive purposes.

As a closing remark, when the interest is on studying the effect of a binary non-reversible time-dependent treatment, such as transplant, on the disease process of a group of patients, I recommend the use a clock-back time scale to model the baseline hazard after transplant. This can be applied within the adoption of flexible regression methods based on generalized linear models, such as the HM model (other examples are the method of Efron based on the Poisson model), both to study the impact of waiting time on the covariates and especially to derive profile-specific dynamic predictions.

Bibliography

Aalen O O, Borgan O, Gjessing HK. Survival and Event History Analysis: A Process Point of View, Statistics for Biology and Health, Springer, New York (2008).

Ambrogi F, Biganzoli E, Boracchi P. Estimates of clinically useful measures in competing risks survival analysis. *Statistics in Medicine* (2008).

Andersen PK, Keiding N. Multi-state; models for event history analysis. *Statistical Methods Medical Research*; 11:91–115. (2002).

Andersen PK, Perme MP. Pseudo-observations in survival analysis. *Statistical methods in medical research*, (2010).

Anderson JR, Cain KC, Gelber RD, Analysis of survival by tumor response. *J Clin Oncol*; 1:710–719. (1983),

Arico M, Schrappe M, Hunger SP, et al. Clinical outcome of children with newly diagnosed Philadelphia chromosome-positive acute lymphoblastic leukemia treated between 1995 and 2005. *J Clin Oncol*; 28(31):4755-4761. (2010).

Arnold J, Yusuf S, Young J, et al. Prevention of heart failure in patients in the Heart Outcomes Prevention Evaluation (HOPE) study. *Circulation*;107:1284-1290. (2003).

Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*; 24:1713–1723. (2005).

Beyersmann J, Gastmeier P, Wolkewitz M, Schumacher M, An easy mathematical proof showed that time-dependent bias inevitably leads to biased effect estimation, *Journal of Clinical Epidemiology*; 61(12): 1216-1221. (2008).

Beyersmann, J. Gerds, T. and Schumacher. M. Letter to the editor for the paper of Snapinn et al, 2005, *The American Statistician*; 60:296–296. (2006).

Brand R, Putter H, van Biezen A, Niederwieser D, Martino R, et al. Comparison of Allogeneic Stem Cell Transplantation and Non-Transplant Approaches in Elderly Patients with Advanced Myelodysplastic Syndrome: Optimal Statistical Approaches and a Critical Appraisal of Clinical Results Using Non-Randomized Data. *PLoS ONE*; 8(10): e74368. (2013).

Cortese G, Gerds TA, Andersen PK. Comparing predictions among competing risks model with time-dependent covariates. *Statistics in Medicine*; 32:3089-3101. (2013).

Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*; 34 (2): 187–220, (1972).

Dafni U, Landmark Analysis at the 25-Year Landmark Point, *Circulation: Cardiovascular Quality and Outcomes*; 4: 363-371. (2011).

Datta S, Satten GA, Datta S. Nonparametric estimation for the three-stage irreversible illness-death model. *Biometrics*; 56:841-847. (2000)

Datta S, Satten G.A. Validity of the Aalen Johansen estimators of stage occupation probabilities and Nelson Aalen integrated transition hazards for Non-Markov models. *Statistics and Probability Letters*; 55:403-411. (2001).

Dubin JA, Müller HG, Wang JL. Event history graphs for censored survival data. *Statistics in Medicine*; 20: 2951–2964. (2001).

Efron B. Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve. *Journal of the American Statistical Association*; 83, 414-425, (1988).

Efron B. The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association*; 72, 557-565, (1977).

Efron B. The two-way proportional hazards model. *Journal of the Royal Statistical Society. Series B 64*; 899-909, (2002).

Gail MH. Does cardiac transplantation prolong life? A reassessment. *Ann Intern Med*; 76: 815–7. (1972).

Giobbie-Hurder A, Gelber RD, Regan MM. Challenges of guarantee-time bias. *J Clin Oncol*; 31(23): 2963–2969. (2013).

Glesby MJ, Hoover DR. Survivor treatment selection bias in observational studies: examples from the AIDS literature. *Ann Intern Med*; 124: 999–1005. (1996).

Glidden, DV. Robust inference for event probabilities with non-Markov event data. *Biometrics*; 58:361–8. (2002).

Hanley JA, Carrieri MP, Serraino D. Statistical fallibility and the longevity of popes: William Farr meets Wilhelm Lexis. *Int J Epidemiol*; 35: 802–5. (2006).

Hanley JA, Miettinen OS. Fitting smooth-in-time prognostic risk functions via logistic regression. *International Journal of Biostatistics*, (2009).

Iacobelli S, Carstensen B. Multiple time scales in multi-state models, *Statistics in Medicine*; 32:5315–5327. (2013).

Julien M, Hanley JA, Profile-specific survival estimates: Making reports of clinical trials more patient-relevant. *Clinical Trials*; 5: 107–115, (2008).

Kaplan E L, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*; 53:457–481, (1958).

Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*; 78: 13–22, (1986).

Mantel N, Byar DP. Evaluation of Response-Time Data Involving Transient States: An Illustration Using Heart-Transplant Data. *Journal of the American Statistical Association*; 69, 345. (1974).

Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*; 50 (3): 163–70, (1966).

Mantel N. Synthetic retrospective studies and related topics. *Biometrics*; 29, 479-486, (1973).

Marubini E, Valsecchi MG. Analysing Survival Data from Clinical Trials and Observational Studies. London, United Kingdom, John Wiley & Sons, (2004).

Mathew J, Sleight P, Lonn E, Johnstone D et al., for the Heart Outcomes Prevention Eval uation (HOPE) Investigators. Reduction of Cardiovascular Risk by Regression of Electrocardiographic Markers of Left Ventricular Hypertrophy by the Angiotensin-Converting Enzyme Inibitor Ramipril. *Circulation*; 104:1615-1621. (2001).

Meira-Machado L, Una-Alvarez J, Cadarso-Suarez C. Nonparametric estimation of transition probabilities in a non-Markov illness-death model. *Lifetime Data Analysis*; 12(3):325-44. (2006).

Meira-Machado L, Una-Alvarez J, Cadarso-Suarez C, Andersen PK. Multi-state models for the analysis of time-to-event data. *Stat Methods Med Res*; 18(2):195-222. (2009).

Potschger U, Heinzl H, Valsecchi MG, Mittlbock M. Using pseudo-values for comparing long-term survival after stem-cell transplantation (SCT) with long-term survival after chemotherapy. Oral presentation at the 35th conference of the International Society for Clinical Biostatistics. (2014).

Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine*; 26:2389 – 2430. (2007).

Rebora P, Salim A and Reilly M. bshazard: Nonparametric Smoothing of the Hazard Function. R package version 1.0. http://CRAN.R- project.org/package=bshazard. (2014).

Rodríguez-Girondo M, de Uña-Álvarez J. A nonparametric test for Markovianity in the illness-death model. *Statistics in Medicine*; 31(30):4416-27. (2012).

Rothman KJ. Longevity of jazz musicians: flawed analysis. *Am J Public Health*; 82: 761. (1982).

Schoop R, Graf E, Schumacher M. Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates. *Biometrics*; 64:603-610. (2008).

Simon R, Makuch RW. A non-parametric graphical representation of the relationship between survival and the occurrence of an event: application to responder versus non-responder bias. *Statistics in Medicine*; 3(1): 35-44. (1984).

Snapinn S, Jiang Q, Iglewicz B. Illustrating the impact of a time-varying covariate with an extended Kaplan–Meier estimator. *The American Statistician*; 59: 301–307. (2005).

Suissa S. Effectiveness of inhaled corticosteroids in chronic obstructive pulmonary disease: immortal time bias in observational studies. *Am J Respir Crit Care Med*; 168: 49–53. (2003).

Sylvestre M, Huszti E, Hanley JA. Do Oscar Winners Live Longer than Less Successful Peers? A Reanalysis of the Evidence. *Ann Intern Med.*; 145:361-363. (2006).

Therneau TM, Grambsch PM, Modeling Survival Data: Extending the Cox Model, Springer Science & Business Media, (2000).

Van Houwelingen HC, Putter H. Dynamic prediction by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data. *Lifetime Data Analysis*; 14: 447–436, (2008).

Van Houwelingen HC, Putter H. Dynamic prediction in clinical survival analysis. Chapman & Hall, (2012).

Van Houwelingen HC. Dynamic prediction by landmarking in event history analysis. *Scand. J. Stat*; 34: 70–85, (2007).

Van Walraven C, Davis D, Forster AJ, Wells GA. Time-dependent bias was common in survival analyses published in leading clinical journals. *J Clin Epidemiol*; 57: 672–682. (2004).

Webster NR. Professors live longer than doctors: immortality bias in survival analysis. *British Journal of Anaesthesia*; 106(2): 161–3. (2011).