MSc Mathematics

Track: Stochastics

Master thesis

Relapse prediction for acute myeloid leukemia patients

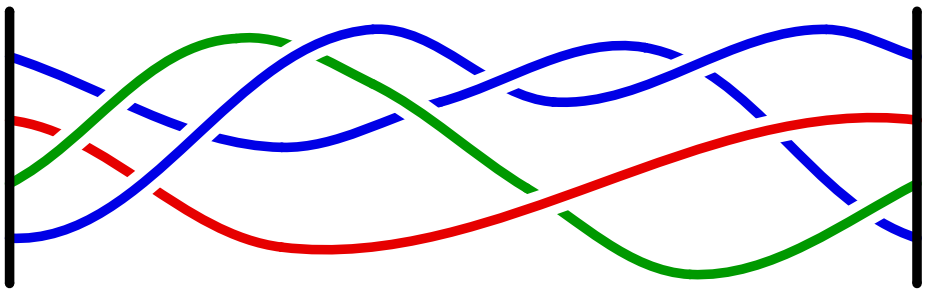
by

Alexandra Vegelien

Oktober 2020

Supervisor: dr. Dennis Dobler

Second examiner: dr. Joost Berkhout



Department of Mathematics

Faculty of Sciences

VU UNIVERSITY AMSTERDAM

Contents

Abstract

Acute myeloid leukemia is a malignant bone marrow disease with poor survival probability. Most patients reach complete remission, but many relapse, which confers a strong prognostic factor regarding their overall survival. Thus, relapse should be prevented. To prevent relapse, patients undergo a consolidation treatment based on a population-based risk classification system: low risk patients receive an autologous stem cell transplantation whereas high risk patients undergo an allogeneic stem cell transplantation that reduces the relapse risk but has severe side effects. This population based stratification is unsatisfactory because it misclassifies on an individual level. The goal of this project is to develop a clinical decision support tool to enable clinicians to make decisions regarding consolidation therapy, while considering an individual patient's relapse risk. To achieve this, we developed a relapse prediction model based on patient characteristics and leukemia-specific parameters using the Fine-Gray survival analysis model for competing risks. We identified several (both previously described as uncovered) factors as highly predictive for relapse. This model serves as the basis for an individualized relapse prediction model and warrants validation in future studies.

Title: Relapse prediction for acute myeloid leukemia patients

Author: Alexandra Vegelien, a.g.j.vegelien@vu.nl, 2552901
Supervisor: dr. Dennis Dobler
Second examiner: dr. Joost Berkhout

Date: Oktober 2020

Department of Mathematics
VU University Amsterdam
de Boelelaan 1081, 1081 HV Amsterdam
`http://www.math.vu.nl/`

1. Popular Summary

Acute myeloid leukemia is a cancer of the bone marrow with poor survival. Most patients reach complete eradication of the cancer, but in many patients the leukemia recurs, conferring a strong prognostic factor for overall survival. Thus, relapse should be prevented. To prevent relapse, patients undergo a consolidation treatment based on a population-based risk classification system: low risk patients receive a stem cell transplantation with their own (healthy) bone marrow cells, whereas high risk patients undergo an allogeneic stem cell transplantation from a donor. This reduces the relapse risk but has severe side effects.

In this research, using patient characteristics, we leverage data and mathematics to improve the prediction of relapse for patients. If we can predict the risk of relapse, the clinicians and patients can use this to determine the best treatment strategy, which ultimately leads to fewer relapses and decreased treatment related side effects. A good prediction model can improve the quality of life and life expectancy of patients with acute myeloid leukemia.

We developed a specific regression model for the survival time of relapse using patient and disease characteristics (such as age and mutations) as predictors. We identified factors as the most predictive for relapse which were previously described in other medical research, and pinpointed additional, variables with high predictive capacity. The model and its results warrant validation in future studies.

Survival time, which is the time between an initiating event and the event of interest (in our case relapse), is not a typical outcome variable. It is not a variable that can be observed or measured at a given time (e.g. the white blood cell count of a patient). Instead, we observe an individual from the initiating event until the event of interest takes place. However, this event will almost certainly not take place for every individual during the study. This is not the same as a missing value because we do know that, for a certain length of time, the event has not happened; thus, we have a lower bound for the survival time. Survival analysis is a branch of statistics that provides the theory to handle this type of data and to build regression models using all the information we have.

The other main theme of the research is how to handle missing data, which is an important problem in (medical) data. Only a fraction of each variable is missing, but if we only include patients with complete information, we would only use a fraction of the data. For the analysis we wanted to perform, the data became almost useless. We investigated and implemented imputation methods that "fill in" missing values while aiming to avoid distorting the dataset. This allowed us to optimize the use of costly collected data.

2. Introduction

2.1. Problem description

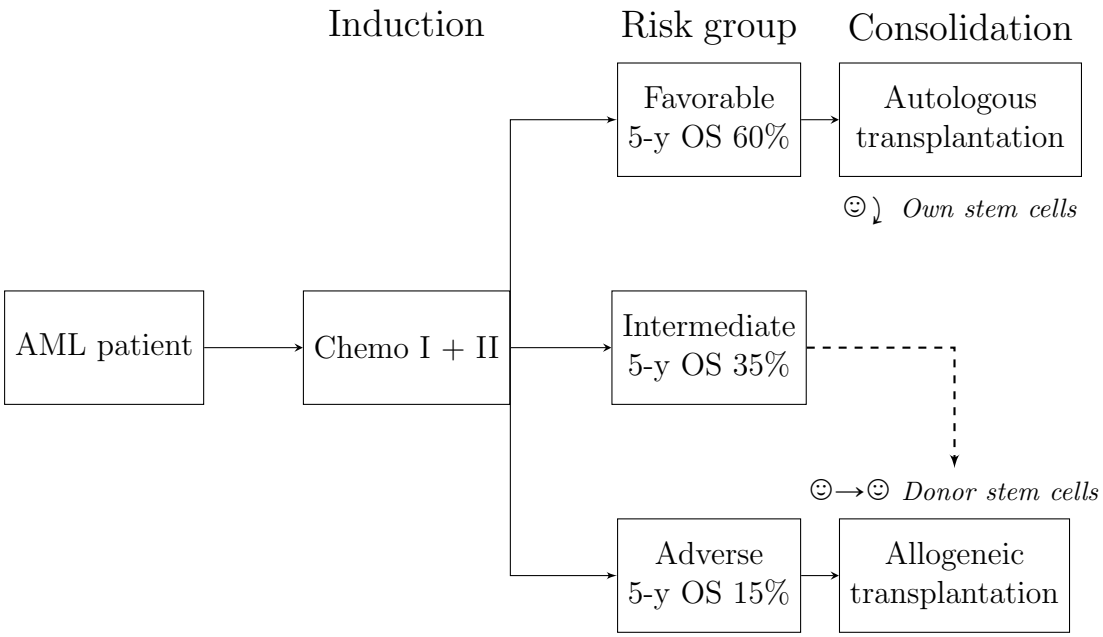


Figure 2.1.: Treatment protocol

Acute myeloid leukemia (AML) is a malignant bone marrow disease; patients with the disease have poor survival probabilities. Patients undergo two rounds of chemotherapy (called induction therapy) to eliminate the leukemic cells. In the majority of cases, patients experience complete remission after this treatment (Jongen-Lavrencic et al., 2018). Consolidation therapy is given to prevent relapse, but unfortunately relapses are common (Jongen-Lavrencic et al., 2018). Relapses are a strong adverse prognostic factor for overall survival and should thus be prevented. The choice of the type of consolidation therapy depends on the risk classification, which is based on the ELN 2017 recommendations (Döhner et al., 2017). Patients are classified within one of three risk groups based on genetic abnormalities, as assessed at time of diagnosis.

When a patient has a "favorable" risk (i.e., a 5-year overall survival probability [5-y OS] of 60%), the patient undergoes autologous stem-cell transplantation. Healthy stem cells are harvested from the bone marrow of the patient. Then, the bone marrow is emptied with aggressive chemotherapy and the healthy stem cells are returned. If the harvesting is unsuccessful, a third cycle of chemotherapy is given.

A patient with an "adverse" risk and thus a very low survival probability (i.e. a 5-year overall survival probability of 15%), undergoes allogeneic stem-cell transplantation. First, a donor's stem

cells are harvested from his or her bone marrow, and the patient's bone marrow is emptied. Then, the donor stem cells are administered to the patient, restoring the bone marrow. The advantage of this treatment is that the donor cells recognize the patient's cells as foreign cells; thus, residual leukemic cells will be targeted and suppressed by the donor's immune cells. Unfortunately, the donor cells also recognize healthy tissue as foreign, which causes severe side-effects and the need for life-long administration of immunosuppressive medication.

Although allogeneic stem cell transplantation has stronger anti-leukemic effects, it is also associated with increased treatment-related toxicity and death. Therefore, it is essential to give patients who are at risk of relapse the necessary treatment and to not over treat patients who are at low risk of relapsing.

Several characteristics are associated with relapse risk. The ELN 2017 risk stratification is a population-based tool used to predict the outcome of therapy in AML patients and to inform treatment selection. However, on an individual patient level, it cannot predict prognosis and/or relapse accurately enough. Favorable-risk patients sometimes relapse, whereas adverse-risk patients sometimes do not, indicating that this classification fails to predict relapse on an individual level. Other characteristics may have additional prognostic value. For example, in patients with measurable residual disease (MRD) after the second chemotherapy cycle – indicating surviving leukemic cells – the risk of relapse is higher than in patients who are MRD-negative (Jongen-Lavrencic et al., 2018). Furthermore, the presence of so-called leukemic stem cells (LSCs) at diagnosis, which are highly immature blood cells with high resistance to chemotherapy, is another adverse risk factor. Although MRD and LSC measurements have prognostic value, they poorly predict relapse as a single variable in individual patients. Moreover, current risk stratification does not always consider the interaction effect of genetic aberrations on relapse probability. Thus, to improve treatment selection and thus the survival of AML patients, there is an urgent need to develop a patient-specific relapse prediction model that includes several risk factors that are not incorporated into current prognostic models.

2.2. Objectives

The goal of this project is to build a clinical decision support tool to enable clinicians to make decisions for consolidation therapy based on the risk classifications of individual patients. To achieve this, we use a data-focussed approach to estimate the probability of relapse for leukemia patients based their personal characteristics (e.g. age and sex) and leukemia-specific parameters (e.g. WBC, MRD and molecular aberrations). For practical reasons, it is most useful time if risk is classified after induction because the treatment decision is made at that time. The aim of this paper is to estimate the probability of relapse within five years of induction; after five years, the risk of relapse is so low that we assume the relapse afterwards is unrelated. The success criterion is that the new relapse prediction model should be superior to the current ELN 2017 model.

In the data from the clinical studies that are used in the analysis, there are approximately 60 parameters for each patient. The main goals of the analysis are (1) to identify the most predictive variables and (2) to develop a relapse prediction model using these variables. First, we will exclude MRD and LSC measurements because these measurements are not performed in every hospital; thus a model without these variables is practical. Moreover, we investigate if the most predictive variables in our analysis are also included in the ELN classification, which is more biologically motivated. Second, we develop a prediction model with MRD and LSC measurements added. Therefore, we

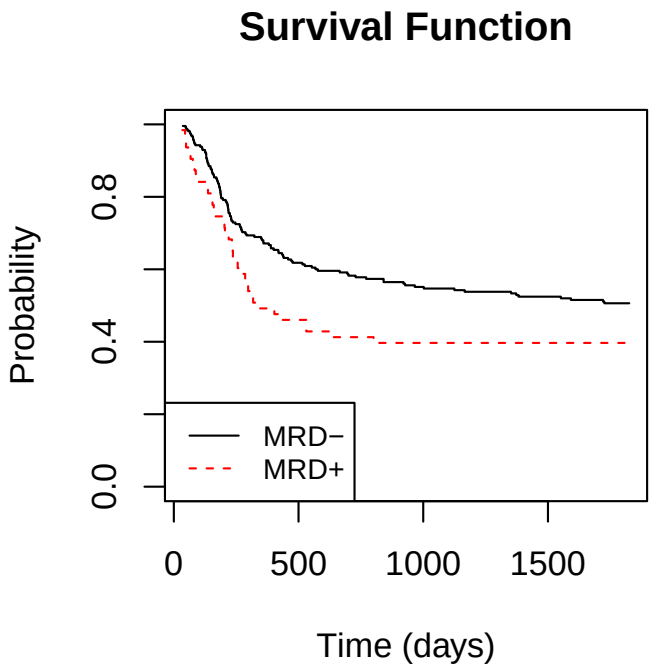


Figure 2.2.: Kaplan Meier curves for event free survival time (the time between induction and the first event: relapse or death) stratified on MRD positive (line below) and MRD negative (above) patients.

analyze the additional predictive value of MRD and LSCs.

2.3. Survival analysis

The goal of this analysis is to predict the relapse probability of leukemia patients, but this is not a standard Y variable or outcome variable. We could eliminate the time element of the relapse by picking a fixed time point, such as the five-year time frame mentioned above. But the problem that then arises is that the outcome variable will still not binary. The patient could, for example, leave the study prematurely or die from treatment-related risks. In these cases, we would not be able to whether the patient has relapsed within five years. If we omitted these cases from the dataset, we would lose valuable data and introduce selection bias. Thus, eliminating time does not solve all problems, and instead, this oversimplification leads to a loss of information; there is a significant difference between having an aggressive tumor and dying within weeks and having five more years of life.

Therefore, in this paper, we model the relapse survival time of patients. Survival time is a peculiar type of variable and is the origin of a branch of statistics called survival analysis. The survival time is the time between an initiating event (in our case, the completion of induction therapy) and the event of interest (in our case, relapse), but the issue is that it is impossible to observe the event of interest for all individuals. If a patient leaves the study healthy in the last year, we do not

know whether the patient would have relapsed. However, we would know it has not happened in the first four years. The Kaplan-Meier curve is an estimation of survival time using all available information regarding a group of patients. It can be used to create an estimate of the percentage of people who have survived (have not relapsed yet) as a function of time. It can also be interpreted as the probability of survival as a function of time. An example of a Kaplan-Meier plot is given in Figure 2.2.

Our main interest is to predict relapses. However, other events may occur, such as treatment-related mortality (e.g., caused by Graft-versus-host disease after allogeneic stem cell transplantation or toxicity after chemotherapy). Relapse and treatment-related mortality are called competing risks; they compete as we can only observe the event that happens first (we cannot know when a patient would have relapsed after they have died). The Fine-Gray model is designed specifically for modelling competing risks in survival analysis, so that the effect of the regression coefficient on the probability of a specific event is interpretable. Therefore, we will build a Fine-Gray model to make a predictive model. Chapter 3 provides a more in-depth explanation of survival analysis, competing risks and the Fine-Gray model.

2.4. Data description

The data we use in this research is from a phase III clinical trial called HOVON 102. The aim of clinical trials is to prove superiority of a medication over the standard treatment for a disease. In the case of HOVON 102, the combination of Clofarabine – a chemotherapeutic used for cancers other than leukemia – and standard chemotherapy was compared to standard chemotherapy alone. The HOVON 102 is a collaboration of 45 hospitals in Europe with a total of 862 patients aged 18 to 65. The study ran between 2009 and 2013 and their progress was followed for up to 10 years.

In a clinical trial, each patient is assigned to either the control group (who receive the standard treatment) or the study group (who receive the new medicine, in this case Clofarabine) upon inclusion. Patients are randomly assigned to the groups[1] to prevent bias by minimizing the population differences that can impact the effect of the medication. The patients are then observed during their treatment and followed for a sufficient amount of time afterwards to study their long-term response to the treatment.

During the study, disease-specific and response data are collected. Examples of general data are patient age and gender. Examples of disease-specific data are white blood cell (leukocyte) count at diagnosis, genomic mutations, and chromosomal abnormalities. Examples of response data are date of complete remission, date of relapse, and overall survival time. The aims of the data collection are 1) to investigate possible relationships between patient characteristics and treatment responses and 2) to answer additional clinical questions (for example, questions regarding the relationship between a mutation and survival).

The HOVON-102 study is special because for the first time, both MRD and 'leukemic stem cells' (LSCs) were measured in AML patients at multiple points in time. LSCs are chemo-resistant cells that can cause relapse (Zeijlemaker and Schuurhuis, 2013). In a preliminary analysis, including these variables in a prognostic model improved the predictions of relapse compared with current methods (unpublished data). These variables provide a unique opportunity to improve current models.

[1]Note that this is only for the treatment under study. The consolidation treatment is not randomized.

2.5. Missing data challenge

One of the main challenges of the data that has hindered previous survival analysis is missing data. We focus our research on patients who finished induction treatments and reached complete remission, which results in 663 patients in the dataset. We aim to investigate the relevance of all variables that may be medically relevant (about 60). Statistical methods, such as the survival analysis of the Fine-Gray model, are designed for complete records, and partially incomplete records are disregarded. Removing all cases with incomplete information yields only 123 cases ($\approx 20\%$), which has previously proven to be insufficient for the intended analysis. However, including patients with nearly complete information ($\leqslant 3$ missing values) yields 468 cases (almost 80%)!

2.6. Results

We implemented two imputation models, both specifically designed for survival analysis data. We designed a multistate model and used the imputed datasets to train the Fine-Gray models for the event of interest: relapse. Using the step-up method, we identified the most predictive variables for relapse: monosomal karyotype, white bloodcell count, whether a patient received allogeneic or another consolidation treatment, a mutated npm1 gene, translocation t(8;23) (chromosomal abnormality), age, the chemotherapy medication and chromosomal abnormality abn11q23. These variables have also been proven to be predictive in previous medical research. In addition, the effect of the variables (whether it is a good or bad prognosis) is in line with previous research. Therefore, the model seems promising and warrants validation in future research.

2.7. Outline

This thesis starts with the necessary theoretical background for our intended analysis. In Chapter 3 we examine the survival analysis theory and in Chapter 4 we examine the theory of missing data and imputation. We then move on to our application. We start by discussing how the dataset differs from the protocol and how we cleaned it in Chapter 5. Then, we review how we imputed the data in Chapter 6. Finaly, we present how we designed the model. We finish with the results in Chapter 8, the discussion in Chapter 9, and the conclusions in Chapter 10.

3. Survival analysis

In this chapter, we provide an introduction to survival analysis. Afterward, the focus is the theory needed to build our intended model.

3.1. The basics

Survival analysis is centered around a particular type of variable: **survival time**. Survival time is the time T between an initializing event and the event of interest. Examples are the time between hospital admittance and death, between complete remission and relapse, between marriage and divorce or between infection and recovery.

Survival time is special because it is, in many cases, impossible to observe the event of interest for all individuals. Common reasons are that the individual may never experience the event (luckily, not all marriages end in divorce and not all cancer patients die), the patient may leave the study before the event occurs (for example, due to relocation, side effects of the study, or unrelated death), or the study may end before the event occurs (the end date of the study can be fixed, so patients included later in the study are observed for shorter amount of time). The left portion of Figure 3.1 illustrates the case where the study ends (the dotted vertical line) before the event happens for the top and middle individuals.

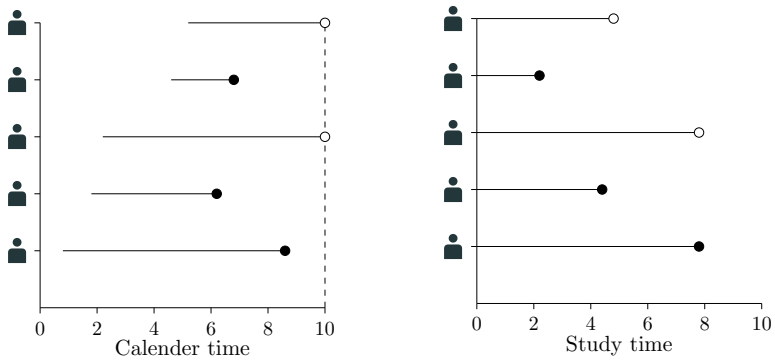


Figure 3.1.: Illustrative example of survival time for five individuals on calender time (left) and study time (right). On the left, the x-axis is the calander time with at time 0 the start of the study and time 10 the end. On the right, the x-axis is the study time with time 0 the initiating event. The horizontal lines starts left at the initializing event and ends right with a black dot, at the event of interest, or with an open dot, when the patient left the study.

When the survival time is not observed for an individual, we say the survival time is **censored**, and the censoring time is the time the individual leaves the study. Although the survival time is

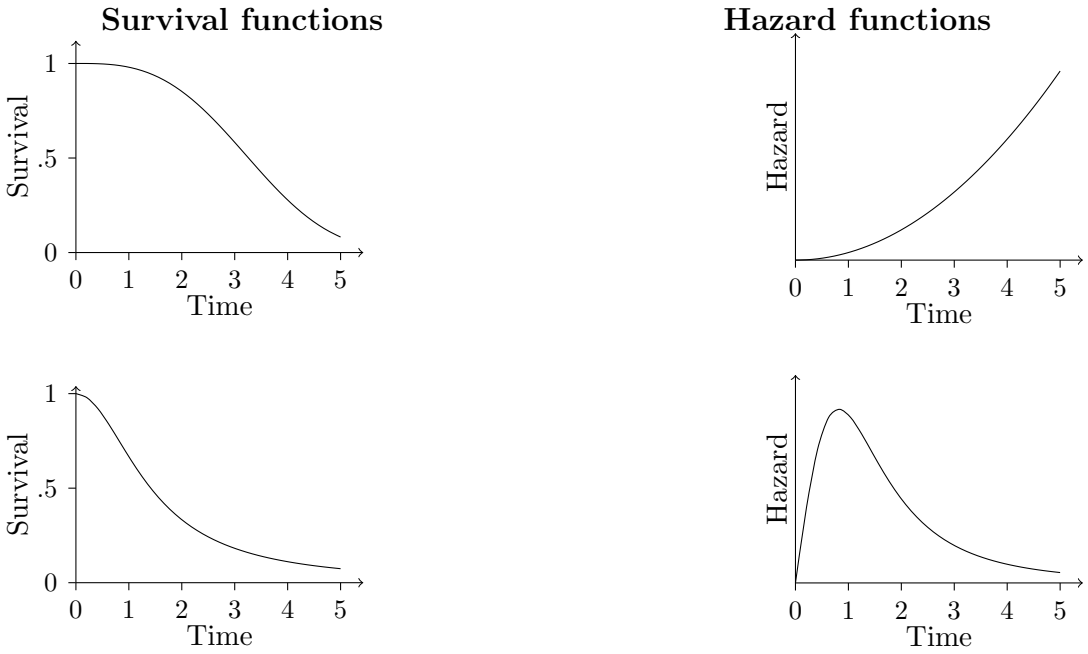


Figure 3.2.: On the left, two typical survival functions are shown with their corresponding hazard function on the right.

not observed for censored individuals, we do know the event has not happened before the censoring time and must have occurred after the patient left the study. Discarding censored survival times would bias the distribution; cases with large or infinite survival times (marriages without divorce are an example of the latter) would be more likely to be excluded. Due to this problem, there is a need for a special statistical approach to survival time.

In Figure 3.1, the two dimensions of survival time are shown: calendar time and study time, where in terms of study time, the initiating event is time 0. The **risk set** at time t is defined in terms of study time and refers to the number of individuals who are still at risk, meaning the event has not occurred and the individuals are still under study. In Figure 3.1, at time 0, all individuals are in the risk set, whereas only the bottom and middle individual are in the risk set at time 7.

The distribution of the (random variable) survival time T is characterized by a **survival function**,

$$S(t) = P(T > t) = 1 - F(t) \tag{3.1}$$

with F being the cumulative distribution function of T. This survival function at t, $S(t)$, can be interpreted as the proportion of people for whom the event has not happened by time t; the survival function is 1 at time 0 and monotonically decreases over time. In Figure 3.2 two typical survival functions are shown on the left. The two graphs have similar endpoints. Both graphs are 1 for $t = 0$ and are about 0.08 for $t = 5$. The middle is, however, different; at $t = 2$, the graph above is still above 0.85, whereas the graph below is already at 0.33.

It is often individual's survival function that people are interested in: what is the patient's probability of surviving until time t given certain characteristics (e.g., probability of not having lung cancer at age t after 30 years of smoking)? This can be obtained by modeling the underlying

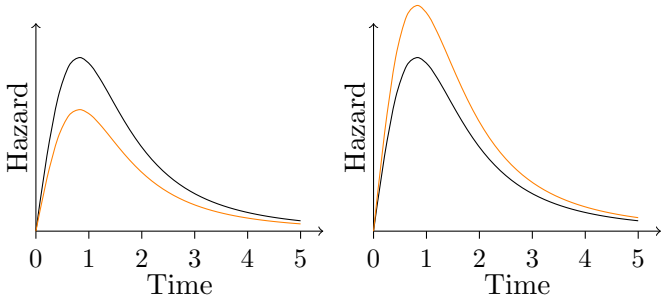


Figure 3.3.: Illustration of the hazard ratio in a Cox model with one covariate: $\alpha(t|x) = \alpha_0(t)\exp(\beta x)$. The black graphs represent the baseline hazard $\alpha_0(t)$ and the orange graphs the hazard $\alpha(t|x)$ for $x = 1$. The left figure is for $\beta < 0$ and the right for $\beta > 0$

.

process, the **hazard rate**,

$$\alpha(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leqslant T < t + \Delta t | T \geqslant t), \tag{3.2}$$

which can be interpreted as the "force of mortality" at time t. In other words, the hazard rate is the instantaneous risk at time t of experiencing the event given that it has not happened yet, so it can be seen as the pull or the force of the event at time t.

In Figure 3.2, the hazard rates are to the right of their corresponding survival functions. For the survival function above, we noted that it decays slowly at first and fast later; the corresponding hazard rate is thus small at first and increases over time. An example is the hazard rate for lung cancer among smokers: the risk is small at first and increases over time. On the other hand, the survival function below drops rapidly and stabilizes afterward; the corresponding hazard thus peaks early on and steeply decreases afterward. This could be the hazard rate for an infection (the event of interest) after an operation (the initializing event): the pull is large right after surgery and steeply decreases after recovery.

Usually, the relationship between covariates and the survival is of interest, and the hazard rate is used to model this. The most used model is that of Cox (Cox, 1972), where the hazard rate is supposed to be of the form

$$\alpha(t|x_1, ..., x_p) = \alpha_0(t)\exp(\beta_1 x_1 + ... + \beta_p x_p), \tag{3.3}$$

where $\alpha_0(t)$ is called the baseline hazard and $\exp(\beta_1 x_1 + ... + \beta_p x_p)$ is called the hazard ratio. The baseline hazard is the only factor that is dependent on t and thus determines the shape of the function, whereas the hazard ratio determines the relative risk based on the covariates. This is visualized in Figure 3.3, with the baseline hazard represented by the black graphs and the corresponding hazards by orange graphs for negative and positive $\beta_1 x_1 + ... + \beta_p x_p$ on the left and right, respectively.

The survival function $S(t)$ can be expressed in terms of the hazard rate $\alpha(t)$ as $S(t) = \exp(-\int_0^t \alpha(s)ds) = \exp(-A(t))$, where $A(t) = \int_0^t \alpha(s)ds$ is the cumulative hazard rate. In particular, for the hazard of the form of Equation (3.3), the survival function is $S(t) = \exp[-\exp(\beta_1 x_1 + ... + \beta_p x_p)\int_0^t \alpha_0(s)ds]$; this implies that for $\beta_i > 0$ the survival is smaller for an individual with $x_i = 1$ compared to an

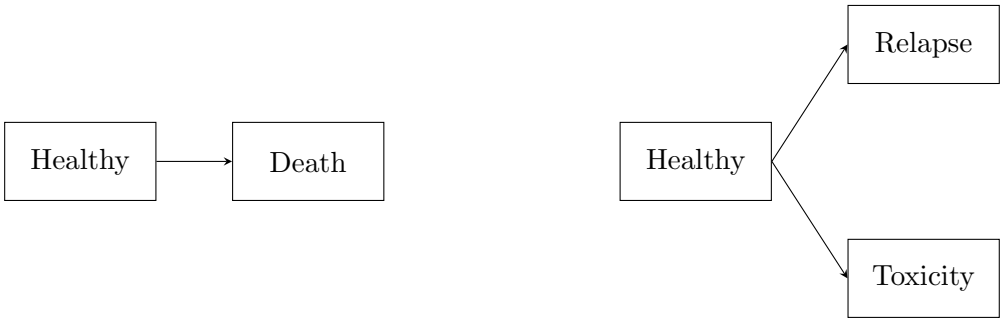


Figure 3.4.: Single risk on the left and competing risks on the right.

individual with the same covariates but with $x_i = 0$. The effect of a covariate on the survival function is thus easily interpreted with the Cox model.

3.2. Multistate and competing risks

Until now, we have only considered a simple model: we observe an individual starting at the initial event until a specific type of event occurs or the individual leaves the study. We can extend this model and our terminology to a more general setting, called a **multistate model**, in which we allow for more than one type of event. A popular example is the illness-death model which is used to study diseases and survival in which an individual begins healthy and then dies or begins healthy, becomes diseased, and then dies.

The jump process of the individual can be given by a continuous stochastic process $X = (X_t)_{t\in[0,\infty)}$ taking values in the state space $\mathcal{I} = \{0, 1, 2, ..., k\}$, which denote all states in our multistate model. The jump process denotes where the individual is at a certain time, so if the individual is in state g at time t, $X(t) = g$. In the standard survival model discussed above we had $\mathcal{I} = \{0, 1\}$ and $T = \min\{t|X(t) = 1\}$.

3.2.1. Competing risks

A simple, yet important, multistate model is the model of **competing risks** in which there is not one cause of failure but multiple causes that compete with each other. A cancer patient, for example, can die because of the disease at time T_1 or because of consequences of the treatment at time T_2 but we would only observe the first event at time $T = \min(T_1, T_2)$ – we assume here that the two survival times are independent. In this sense, the events are competing. In this model, the initial state 0 is the only transient state, and the individual can from here jump to any of the k absorbing states, so $\mathcal{I} = \{0, 1, 2, ..., k\}$. As in the example, all we observe is the event time $T = \min\{t|X_t \neq 0\}$ and the type of event $\delta = X_T$. The difference between a single risk and competing risks is visualized in Figure 3.4 for the competing risks relapse and toxicity.

The aim of a competing risk model is usually to study or model the survival for a specific type of event. We can extend our terminology to that of the competing risk model. Before, we looked at the survival function $S(t)$ which gives the probability of not having an event before time t. Now, we want to investigate the probability $P_{0h}(t) = P(T \leqslant t, X_T = h)$ of having a specific type of event h before time t, which is called the **cumulative incidence** function. In the standard survival model

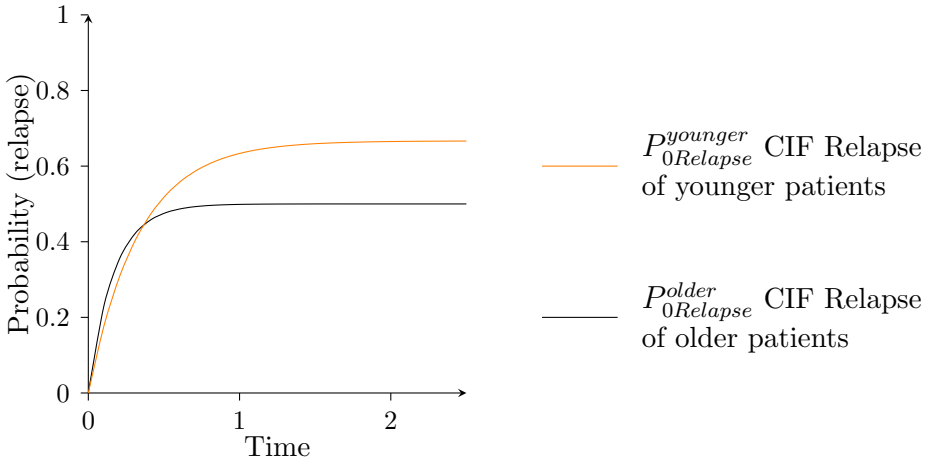


Figure 3.5.: Example cumulative incidence functions of relapse in case of competing risk for two risk groups, patients younger or older than 30 years.

with a single type of event ($k = 1$), the cumulative incidence function reduces to $P_{01}(t) = 1 - S(t)$. Similarly, we can define a cause-specific hazard $\alpha_{0h}(t) = \lim_{\delta t \to 0} \frac{1}{\delta t} P(X(t+\delta t) = h | X(t-) = 0)$ with $X(t-) = \lim_{s \uparrow t} X(s)$ the state just before time t, which is the instantaneous risk of having event h at time t given that no event occurred before time t.

Unfortunately, the cumulative incidence function is not a function of the cause-specific hazard. Instead,

$$P_{0h}(t) = P(T \leqslant t, X_T = h) = \int_0^t P(T > u-)\alpha_{0h}(u)du \tag{3.4}$$

where the all-cause survival function $P(T > u)$ depends on all hazards. The all-cause hazard rate corresponding to the survival function is the sum of all cause specific hazards, $\alpha(t) = \sum_{h=1}^{k} \alpha_{0h}(t)$ and thus, as before, $P(T > u) = \exp(-\int_0^t \alpha(s)ds) = \exp(-\int_0^t \sum_{h=1}^{k} \alpha_h(s)ds)$.

A simple example of a Cox model with a single variable and constant cause-specific hazards with two competing risks shows how a lower cause-specific hazard for one patient group compared to another can even result in a higher cumulative incidence function (Gray, 1988). Suppose we look at cancer patients after they have finished chemotherapy; there are two competing risks, relapse and toxicity, and the single variable x_y in the model is if the patient is younger than 30. Let the cause-specific hazards be a Cox model of the form Equation (3.3), where $\alpha_{0Relapse} = \alpha_0^R(t) \exp(\beta_{young}^R * x_y)$ and $\alpha_{0Toxicity} = \alpha_0^T(t) \exp(\beta_{young}^T * x_y)$ with $\alpha_0^R(t) = \alpha_0^R(t) = 3$ and $\exp(\beta_{young}^R) = \frac{2}{3}$ and $\exp(\beta_{young}^T) = \frac{1}{3}$. Now, using the relations from above, we get for patients above the age of 30 ($x_y = 0$)

$$\begin{aligned}
P_{0Relapse}^{\text{older}}(t) &= \int_0^t P(T > u - | x_y = 0)\alpha_{0Relapse}(u | x_y = 0)du \\
&= \int_0^t \exp\left(-\int_0^u \alpha_{0Relapse}(s|x_y = 0) + \alpha_{0Toxicity}(s|x_y = 0)ds\right) \alpha_0^R(u) \exp(\beta_{young}^R * 0)du \\
&= \int_0^t \exp\left(-\int_0^u 3 + 3ds\right) 3du \\
&= \int_0^t e^{-6u} 3du = \frac{1}{2}(1 - e^{-6t}).
\end{aligned}$$

Similarly for patients who are under the age of 30 ($x_y = 1$), we get $P^{\text{younger}}_{0Relapse}(t) = \frac{2}{3}(1-e^{-3t})$; both cumulative incidence functions are graphically presented in Figure 3.5. The cause-specific hazard for relapse is lower for younger patients than for older patients but, as shown in Figure 3.5, the cumulative incidence function is higher after a certain time. The issue is that the variable effect of age reduces the hazard of competing risk toxicity even more than it reduces the hazard of relapse.

3.2.2. Fine-Gray model

In general cases of competing risks, the analysis of the covariate effect on the cumulative incidence function can become complicated and requires all cause-specific hazards to be modeled. Fine and Gray developed a model where the cumulative incidence function is modeled directly by assuming it is of the form

$$P_{0h}(t) = P(T \leqslant t, X_T = h|x_1, ..., x_p) \tag{3.5}$$

$$= 1 - \exp(-\int_0^t \lambda_h(s|x_1, ..., x_p)ds) \tag{3.6}$$

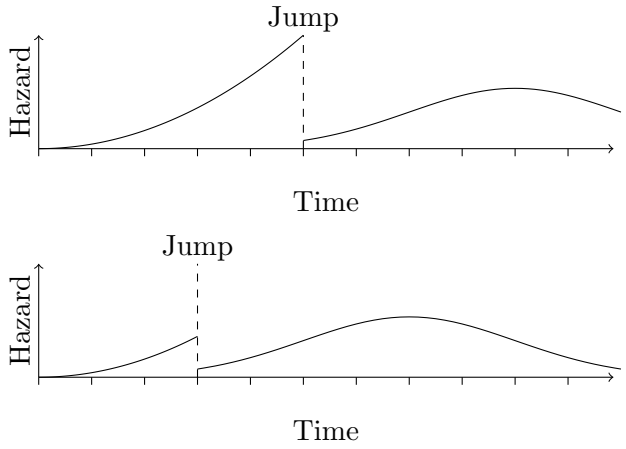
$$= 1 - \exp(-\int_0^t \lambda_{h0}(s)ds \exp(\beta_1 x_1 + ... + \beta_p x_p)), \tag{3.7}$$

where $\lambda_h(t|x_1, ..., x_p) = \lambda_{h0}(t)\exp(\beta_1 x_1 + ... + \beta_p x_p)$ is the corresponding hazard, $\lambda_{h0}(t)$ is the baseline hazard (dependent on time) and $\exp(\beta_1 x_1 + ... + \beta_p x_p)$ is the hazard ratio (not dependent on time). The form of the cumulative incidence function is the same as that of the distriubtion function $P(T \leqslant t) = 1 - S(t)$ of the Cox model in case of a single risk. Furthermore, note that the cumulative incidence function no longer depends on other hazards, so we can focus our analysis on the risk of interest, which we will call 1.

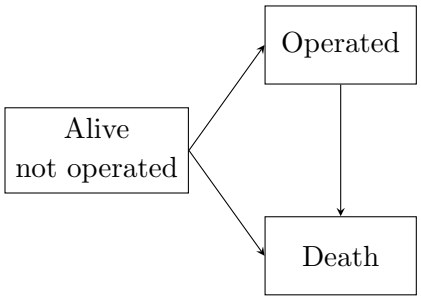
The hazard λ_h in the Fine-Gray model is called a **subdistribution hazard** because it is the hazard function of a subdistribution of an improper random variable. Because of other hazards, the function $P_{01}(t) = P(T \leqslant t, X_T = 1)$ does not converge to one as t goes to infinity, which means it is not a distribution but a subdistribution (as it is non-decreasing, right continuous, and $\lim_{t\downarrow 0} P_{01}(t) = 0$). In Figure 3.5 we can see that the cumulative incidence functions do not converge to 1.

The cumulative incidence function is thus not the distribution of a random variable but instead that of the improper random variable $T^* = 1_{\delta=1} \times T + 1_{\delta\neq 1} \times \infty$. The random variable equals the event time if the event of interest has taken place and infinity otherwise, which can be seen as the event of interest never taking place. The (sub)distribution of T^*, which equals P_{01}, thus has a point mass at $t = \infty$equal to $P(\delta \neq 1) = P(T \leqslant \infty, \delta \neq 1) = 1 - P_{01}(\infty)$, which in the case of the orange hazard in Figure 3.5 is equal to 1/3.

The risk set of T^* is difficult to interpret. If a competing risk has occurred we get $T^* = \infty$, meaning that the individual will forever be in the risk set even though we can be sure they are no longer at risk of the event of interest. For example, in cases, of multiple causes of death, after patients have died, they can no longer die from another cause of death. It is more realistic in other cases, for instance in the case of a cure model where a patient can be cured and the event of interest will never take place. In this case we do not observe when the patient is cured and he will thus always stay in the risk set.



(a) Visualization of two possible hazards in a semi-Markov multistate model for different jump times.



(b) Multistate model with three states and no loops.

Figure 3.6.: Multistate model with cause specific hazard function for death

3.2.3. Homogeneous semi-Markov Models

The competing risk model is a specific multistate model with the only transitions possible between the original state and absorbing states; thus we only consider the time spent in the original state. In general, this can become arbitrarily complicated with many possible transitions and we might also be interested in modeling the survival time in other states than the origin state. The hazard in these states can depend on the past in a complicated way (for example on the time spent in previous states and the states visited), which complicates analysis. Therefore, assumptions are usually made on the relation of the hazard and the past.

A common example is the Markov assumption in which the hazard and transition probabilities are independent of the past given the present state, which means that knowing the past, such as all past states visited, provides no information concerning the future if we know the present, i.e.:

$$P(X(t) = g|X(s) = h, X(r) \text{ for } r < s) = P(X(t) = g|X(s) = h).$$

We can now define the transition probability $P_{hg}(s,t)$ and transition intensity $\alpha_{hg}(t)$ similar to how we defined the cumulative incidence function and the cause-specific hazard function respectively:

$$P_{hg}(s,t) = P(X(t) = g|X(s) = h),$$
$$\alpha_{hg}(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(X(t + \Delta t) = g|X(t-) = h).$$

A special class of multistate models, which has nice analytical properties, is that of **homogeneous semi-Markov models**. The assumption is that the transition probabilities only depend on the current state and the time spent in the current state, often called the "duration" or "sojourn time". This differs from the Markov assumption: we now allow the future to depend on the time spent in a state, whereas in the Markov assumption this has no influence.

In Figure 3.6, we present a multistate model with three states (see Figure 3.6b): alive and not operated, alive and operated and death. For example, for a diseased patient (e.g., leaking heart valves) who requires an operation. Now we made an imaginative hazard of death for a patient who

at time zero is alive and not operated and after a certain amount of time jumps to operated. The hazard of death before the operation will increase over time. In this case, we can imagine that when the patient is operated, the risk of death is no longer dependent on how long the patient needed to wait for the operation; the risk of death of the patient now only depends on the time since the operation. This is visualized in Figure 3.6a for different jump times.

We furthermore assume that it is not possible to return to a state that has been visited in the past. Because the hazard and transition probabilities only depend on the time in the current state, we can introduce a new time variable. First, let us introduce notation (that we borrowed from (Gill, 1980)): again let $X = (X_t)_{t\in[0,\infty)}$ denote the jump process, let $J_0, J_1, J_2, ...$ denote the consecutive states and let $T_1, T_2, T_3, ...$ be the corresponding sojourn times in these states – note that X gives us $J_0, J_1, ...$ and $T_1, T_2, ...$ and conversely. Now, suppose $J_i = h, J_{i+1} = g$, then we can introduce a new time scale $d_i = t - \sum_{j=1}^{i} T_j$ which is zero right after the i-th transition. The cause specific hazard in state h, or transition intensity, is a function of the transformed time:$\alpha_{hg}(d_i) = \lim_{\Delta t\to 0} \frac{1}{\Delta t} P(X(\sum_{j=1}^{i} T_j + d_i + \Delta t) = j | X(\sum_{j=1}^{i} T_j + d_i -) = h)$. This comes down to resetting the time to 0 when a jump is made to another state and this new state can be considered the original state. Analyzing the survival time in this new state and the hazards is reduced to the competing risk setting with this new state as the original state and with all states with non-zero transition probability as the competing risks. We can therefore model the transition using the Fine-Gray model.

Returning to our example in Figure 3.6, we might be interested in the probability of dying before a time t for a patient recently diagnosed. The patient can die either before the operation or afterward. We can model the first, the direct transition to death, with a Fine-Gray model and determine the cumulative incidence function for the patient. The probability of first getting operated and afterward dying before a time t for a patient can be determined by modeling the two transitions separately by Fine-Gray models. We can then combine these results to obtain the two-step probability as follows:

$$\begin{aligned} P_{A,O,D}(t) &:= P(X(0) = \text{Alive and not operated}, X(r) = \text{Operated, for some } r \in (0,t), X(t) = \text{Death}) \\ &= \int_{u=0}^{t} P(T_1 > u-)\alpha_{AO}(u) \left[\int_{v=0}^{t-u} P(T_2 > v-)\alpha_{OD}(v)dv \right] du, \end{aligned}$$

(Cortese and Andersen, 2009) where the states "Alive", "Operated" and "Death" are denoted by A, O and D respectively. Now you can recognize two cumulative incidence functions, one for each jump, in the two step probability, similar to what we had in Equation (3.4). Now, again, we can model the cumulative incidence functions directly with the Fine-Gray model, as in Equation (3.7), where we use subdistribution hazards instead of cause specific hazards. So we get:

$$P_{A,O,D}(t|Z) = \int_{u=0}^{t} P_{OD}(t-u|Z)P_{AO}(du|Z),$$

$$P_{AO}(t|Z) = 1 - \exp(-\int_{v=0}^{t} \lambda_{AO}(v|Z)dv) = 1 - \exp(-\int_{0}^{t} \lambda_{(AO)0}(s)ds \exp(\beta_1^{AO} z_{AO,1} + ... + \beta_n^{AO} z_{AO,n}))$$

and

$$P_{OD}(t|Z) = 1 - \exp(-\int_{u=0}^{t} \lambda_{OD}(u|Z)du) = 1 - \exp(-\int_{0}^{t} \lambda_{(OD)0}(s)ds \exp(\beta_1^{OD} z_{OD,1} + ... + \beta_m^{OD} z_{OD,m})),$$

where Z is the covariate vector – note that the two models can have different covariates.

3.2.4. Time-dependent variables

The models we have discussed can only include baseline covariates: a variable that is known at the initiating event. The models can be extended to include time-dependent variables $Z(t)$. But this is only of prognostic value if the covariate function $Z(t)$ of an individual is known at the initiating event. This is the case for, for example, age. However, in case the covariate is a random variable, such as blood pressure over time, this is highly unrealistic.

A simple, yet important, example of a random variable is that of a binary time-dependent variable. An example is if a cancer patient has reached complete remission (meaning cancer has almost completely disappeared) or if someone has had the intended operation; these variables are initially 0 until an event takes place and afterward remains 1. If we include such a variable in the model and we want to predict the probability of the event of interest taking place, we need to the covariate function beforehand: when the patient will reach complete remission or when they receive their operation. However, this information is not always available beforehand: the date of complete remission is not predetermined and the operation data could depend on the recovery of the patient.

In the case of the binary random variables described above, it is tempting to include the variable as a baseline covariate. For the example above, we could include the covariate if a patient eventually reaches complete remission or eventually gets their operation. This method, however, introduces an "immortal time bias". The individuals who have reached the event (such as complete remission or the operation) must have stayed alive until that happened, so they are "immortal" until the event happens. The other individuals might have reached the event too if they would have stayed alive (Klein and Klein, 2013).

We can include a binary random time-dependent covariate (and similarly a k-level categorical covariate) by identifying different states with the categories of the covariate (Beyersmann and Schumacher, 2008). In other words, we define an underlying multistate model with additional non-absorbing states which denote the current covariate value of the individual. Take again the example of the binary variable if the patient has had their operation and death as the event of interest. The underlying multistate model would in this case look as in Figure 3.6b.

(Beyersmann and Schumacher, 2008) designed a multistate model that allows the effect of the time-dependent covariate to be modeled, similarly to a non-time-dependent covariate. However, to obtain the probability of a certain path, the covariate path needs to be known at the initiation event. Suppose, instead, the interest is not in quantifying the effect of a time-dependent variable but needs to be included because of the effect on the hazard. If it is moreover reasonable to assume the time spent in the current state depends only on the duration in that state, we can use the theory described in Section 3.2.3. This allows us to model the transitions between the covariate values and thus provide a probability estimation without knowing the covariate at baseline.

4. Handling missing data

Complete data is rare in clinical databases, and most statistical methods require complete records. One solution is to discard incomplete cases. However, this leads to small sample sizes and sample bias which hampers the generalization of results. In the current dataset, we face a similar problem: the data consists of 663 patients and 60 parameters that may be medically relevant but removing all cases with incomplete information would yield only 123 cases ($\approx 20\%$). Including patients with near complete information ($\leqslant 3$ missing values), would yield 468 cases (almost 80%). This illustrates that complete case analysis is inefficient.

In the final relapse prediction model, which we will call the substantive model, we will most likely only use a fraction of all parameters included. Thus, we could focus on complete parameters (observed for all patients), which amounts to 32 parameters. Even though the substantive model will not include nearly as many parameters, this approach would mean filtering the parameters based on their completeness instead of predictiveness, which introduces selection bias and leads to the loss of potentially valuable information.

Solely based on the scale of the missingness, complete case analysis is not feasible with the current data. We thus need a method to deal with the missingness. Imputation is such a method, which involves preprocessing the dataset by filling in missing values. The main advantage is that the intended analysis does not have to be adapted to allow for missing data. Instead, after imputation, the intended analysis can be performed on a complete dataset.

4.1. Missingness patterns

An important concept when studying missing data is the mechanism of missingness, which was first introduced by (Rubin, 1976). By understanding how the missingness occurs, we can understand its effect on the data. The resulting terminology is now standard and is used to specify conditions under which methods in missing data work (Donders et al., 2006) (Little and Rubin, 2020) (Bartlett and Taylor, 2016) (White and Royston, 2009).

First, we will introduce notation to aid the definition of the mechanisms. Let $Y = (y_{ij})$ denote the full dataset as a $n \times K$-matrix with n the number of cases and K the number of parameters. A row $y_i = (y_{i1}, ..., y_{iK})$ denotes the values of all variables for case i and a column $(y_{1j}, ..., y_{nj})$ denotes the values of all cases for a specific parameter. This matrix is complete and can be interpreted as capturing the underlying truth. However, in reality, some cases y_{ij} are unobserved. The matrix $M = (m_{ij})$ is the missingness indicator matrix of Y where $m_{ij} = 1$ if y_{ij} is unobserved and 0 otherwise. We assume that the rows (y_i, m_i) are identically and independently distributed. Based on the missingness indicator, we can partition Y_i into the subvector of observed values Y_i^o, where $m_{ij} = 0$, and the subvector of unobserved values Y_i^m, where $m_{ij} = 1$. The mechanisms of missingness characterize the conditional distribution $f_{M|Y}$.

There are several mechanisms of missingness. The first possibility is that data are **missing**

completely at random (MCAR). In this case, missingness does not depend on the values of the data: no values are more or less likely to be missing; thus, for any distinct y_i, y_i^* and all i we have $f_{M|Y}(m_i|Y_i = y_i) = f_{M|Y}(m_i|Y_i = y_i^*)$. This is a valid assumption in case of administrative mistakes or clumsiness in testing. It is less likely when there is for example a sense of shame around a certain answer so that specific outcomes are less likely to be reported, or the other way around: it is more likely that one will forget to note someone has not had cancer than has had cancer. These are two examples when some values are more or less likely to be missing than other values. If this is the case, the subset of the sample with only complete cases is a random sample of the original sample and is thus a random sample as well. This means that we can base the statistical analysis on the complete cases only without getting biased results.

Missing completely at random is a strong assumption and is often too restrictive; **missing at random (MAR)** is a less restrictive assumption and is often sufficient. Data is missing at random when missingness can depend on the observed values but not the unobserved values – that is, if for any distinct vectors $y_i = (y_i^o, y_i^m), y_i^* = (y_i^o, y_i^{m*})$ that are equal on the observed subvectors, it holds that $f_{M|Y}(m_i|Y_i = (y_i^o, y_i^m)) = f_{M|Y}(m_i|Y_i = (y_i^o, y_i^{m*}))$. An example of this is a case where the hospital has the white blood cell count for every cancer patient but not for patients with a fracture. The probability of knowing the white blood cell count only depends on the type of patient (e.g., fractures or cancer), and the probability of missingness remains unchanged if we knew the actual white blood cell count. Complete case analysis with MAR data complicates the analysis not only by reducing the subset but also, more importantly, by distorting the distribution. In the previous example, only the white blood cell count of cancer patients would be included and would thus not represent that of any patient.

Missing not at random (MNAR) is when neither is the case and the missingness depends on the values of missing data – that is, when there is some i and pair of distinct vectors $y_i = (y_i^o, y_i^m), y_i^* = (y_i^o, y_i^{m*})$ equal on the observed subvectors such that $f_{M|Y}(m_i|Y_i = (y_i^o, y_i^m)) \neq f_{M|Y}(m_i|Y_i = (y_i^o, y_i^{m*}))$. An example is a variable indicating whether a patient has had cancer. If this value is missing in the database, the patient has most likely not had cancer. This means that the distribution of the observed and the missing values are different and are based on the unobserved values. In this case, complete case analysis would also produce biased results.

4.2. Quick fix imputation methods

The first known report of a possible solution to missing data dates back to 1926 [handbook MD], which is before the modern computer was invented; the methods developed were thus limited in terms of computational time. The classic imputation methods are intended to be a quick fix and generally lack statistical foundation. The consequences of this is that they (1) distort the distribution, (2) create biased results, (3) weaken the relationships with other variables, or (4) reduce the variance. Therefore, the use of these ad hoc methods is discouraged except for specific cases. In this section, we will introduce well known quick fix methods and discuss their shortcomings.

There are multiple simple imputation methods. The best-known method is mean imputation (Buuren, 2018) (Little and Rubin, 2020): missing values of a variable are imputed by its mean or the mode in case of a categorical variable. This method is straightforward but distorts the distribution of the data by not incorporating variance and the relationship with other variables; thus, it should be avoided (Buuren, 2018). An extension of this method is regression imputation:

missing values are predicted based on other variables. Although relationships are taken into account, the variability is still underestimated. Moreover, this method poorly deals with missing values of the predictive variables.

Another simple method which is popular in public health and epidemiology, is the indicator method (Buuren, 2018) (Donders et al., 2006). Missing values are replaced by 0 for continuous variables or an extra category for categorical variables. The missingness indicator is included as an extra variable. In specific cases, this method can be useful, but in general, it produces severely biased results (Buuren, 2018) (Donders et al., 2006).

A different option would be hot-deck or cold-deck imputation, where missing values are replaced by the value of a similar record from the same dataset or a different dataset, respectively. This is only feasible in datasets with relatively small dimensions. The dataset we use, for example, is too elaborate (600x60) and we have no mean to define similarity well (Little and Rubin, 2020).

4.3. Imputation as resampling

Since the 1950s, methods have been developed that address the issues with the quick fix methods. Sophisticated solutions now exist and are implemented and ready to use for researchers. These methods are discussed in this section.

A more sophisticated method of imputation, with statistical foundation, treats imputation as replacement (Donders et al., 2006); all incomplete records are replaced with new records drawn from the source population. In clinical practice, this would mean new patients should be included and followed for years, which is unfeasible. Instead, we draw this new record from an estimation of the source population.

(Donders et al., 2006) explain this underlying concept as follows:

In the classical (frequentistic) statistical view, conclusions drawn from any study should not depend on the sample that is involved in the study. Should the study be repeated with a different sample, nearly identical results should be obtained. The conclusions do not depend on the given set of subjects in the sample. This implies that every subject in a randomly chosen sample can be replaced by a new subject that is randomly chosen from the same source population as the original subject, without compromising the conclusions. Imputation techniques are also based on this basic principle of replacement. ... Imputation of missing data on a variable is replacing that missing value by a value that is drawn from an estimate of the distribution of this variable.

In practice, this would mean that the source population should be modelled and each missing value is filled in based on a drawing of this estimated population. The methods previously discussed were meant to find the most suitable value for imputation, but they distort the distribution of the population. In this case, we intend to do the exact opposite: we do not necessarily have the true value that we impute, and we are determined to leave the underlying distribution intact.

4.4. Multiple imputation

Until now, we have only considered examples of single imputation: where each missing value is imputed by a single value. If the estimated source population is the same as the real source population, this method would be the same as replacement. Even though this is not realistic, the

imputed dataset can provide unbiased results (Donders et al., 2006). But when the final statistical analysis is performed, it is as if there were never any missing values, which is misleading. The added uncertainty of imputation is thus not taken into account, which results in a variance for the substantive model that is smaller than in reality.

Rubin's approach to incorporating the uncertainty because of imputation is elegant in its simplicity: he proposed that researchers should repeat the imputation in order to create multiple imputed datasets. One can view this approach as taking random samples from the estimated source population and these samples together represent the uncertainty of this estimation. The statistical analysis is performed on each data set as if there were no missing data and the results are then combined.

The imputation and analysis are thus still conveniently separated so that the intended complete-data method of analysis can be used (Rubin, 1976) (Molenberghs, 2015). This makes the method widely applicable and appealing. Another benefit is that it is a statistical method that provides an intuitive interpretation of multiple possible completions of the dataset where the variety expresses the uncertainty. If a variable is infrequently observed and its pattern is difficult to understand, the unobserved values can be filled in in many different ways. This has a large effect on the variability as opposed to a variable with only a few missing values and an easy to understand pattern. This convenience and the intuitive appeal are the main reasons for the popularity of multiple imputation (MI) (Molenberghs, 2015) (Buuren, 2018) (Bartlett and Taylor, 2016) (Donders et al., 2006).

4.4.1. Building MI models

Multiple imputation is just the outline of an imputation algorithm. Mean imputation, for example, fully describes how to perform the imputation whereas MI omits an important step: how to do the imputation. In this section, we discuss an important aspect of building an MI model: the relation between the imputation model and the intended analysis. In this project the intended analysis is survival analysis in the setting of competing risks. In Section 4.5, we discuss two algorithms adapted for a Fine-Gray substantive model.

Compatibility between the imputation and the substantive model is an important concept to take into account when deciding on the imputation model. We mentioned before that the power of MI is that the imputation step is separated from the complete data analysis, but to generate unbiased results, the imputation model chosen should be compatible with the substantive model (Bartlett et al., 2014).

The main idea behind this model is that there is an underlying joint model, "the truth", which captures the structure and relationships of all variables (including the outcome variable), which is the underlying model of the source population. Taking a draw from this distribution would be the same as adding a new record to the dataset. The substantive model is obtained from this joint model by conditioning on all predictive covariates; this conditioning models how the outcome relates to these predictors (a simple example of this is linear regression $Y_i = \beta_0 + \beta_1 * X_i + \epsilon_i$ with $\epsilon_i \sim N(0, \sigma^2)$). Similarly, the imputation model can be obtained from a joint model that specifies how the missing values depend on the observed data.

The imputation and substantive models are compatible when such an overarching joint model exists (for a formal definition, see (Liu et al., 2013) or (Meng, 1994)). To safeguard this condition, it may seem straightforward to start by defining the joint model and then define the imputation and substantive model. For some specific joint distributions, specifically multivariate normal dis-

tribution, this is feasible, and methods exists. However, in general, this approach is too complex, especially with many partially observed variables, some continuous some discrete (Bartlett et al., 2014) (Molenberghs, 2015).

The practical implication of this method is that the two models should be related, and if one model captures a true underlying relationship, then the other should too. If this is not the case, the estimated association will be underestimated, and the results will be biased. Suppose again that the substantive model is a simple linear regression model with $\log(X)$ as independent variable (so $Y_i = \beta_0 + \beta_1 * \log(X_i) + \epsilon_i$ with $\epsilon_i \sim N(0, \sigma^2)$) and this is the true relationship between the variable X and the outcome parameter Y. Now suppose X has missing variables and the imputation model is a simple linear regression of the form $X_i = \beta_{I,0} + \beta_{I,1} * Y_i + \epsilon_{I,i}$ with $\epsilon_i \sim N(0, \sigma_I^2)$. The imputation model does not allow for the true underlying relationship that is captured in the substantive model, and the result will be that this true underlying relationship will be weakened by the imputation.

Importantly, the outcome parameter of the substantive model should be allowed as a predictive variable in the imputation model, as was seen in the previous example (Moons et al., 2006). This may at first seem like a self-fulfilling prophecy and circular reasoning, which would amplify the association. This is, however, untrue: including the outcome parameter as a predictor in the imputation model reduces the bias of the association parameter and omitting it would result in diluted (biased) estimated association parameter.

4.4.2. Multivariate imputation: fully conditional specification

In case of a single variable with missing data, an imputation model is similar to a prediction model. One needs to find a model that best describes the relationship between the outcome variable (in this case the variable with missing data) and the predictive variables (the complete variables and the outcome variable of the substantive model) and train the model with the data. We could then use this model to predict the values that are missing , and we can add noise to the model to allow for variability because the true underlying model is most likely not deterministic (see Section 4.2 for the consequences of not incorperating noise). We could also incorporate parameter uncertainty; we do not know the parameters in our model, so we should incorporate the uncertainty of the estimation. Each missing value can be imputed by sampling from the estimated distribution conditional on the predictive variables in the imputation model. Repeating this procedure provides multiple imputed datasets.

The fully conditional specification (FCS) (Buuren, 2018) is based on similar principles as described above but is suited for more complicated cases with multiple variables with missing data. Instead of defining the joint distribution of the multivariate model (which is difficult in general cases), we implicitly define the joint distribution by defining the conditional univariate models. Suppose we have two covariates X_1, X_2 with missing data and other covariates Z and outcome Y, where X_1 is in the prediction model for X_2 and the other way around. But if both X_{j1} and X_{j2} are missing in some case j , we cannot apply the method as described for the univariate case because to impute one, we need the other.

The FCS method solves the problems due to multivariate missing data by iteratively imputing each variable with missing values based on its defined conditional model. Again, let $Y = (y_{ij})$ denote the full dataset as an $n \times K$-matrix, and for simplicity, let the first p variables have missing values, where Y_i^m will denote the unobserved subvector of Y_i and Y_i^o denotes the observed. First, all missing values $Y_1^m, ..., Y_p^m$ are imputed with an initial guess. Then, at each iteration we redo the

imputation on a variable-by-variable basis, by imputing Y_{i^m} again by sampling from the estimated distribution of Y_{i^m} conditional on the other variables $Y_1, ..., Y_{i-1}, Y_{i+1}, ..., Y_K$, for $i = 1, ..., p$. So, at each iteration, we impute through every variable individually and repeat this multiple times to get an imputed dataset. Applying this method multiple times gives us multiple imputed datasets.

The MICE algorithm is an implementation of the FCS in R and is a Markov chain Monte Carlo method. If the conditional distributions belong to an overarching joint model, the algorithm is a Gibbs sampler. This condition is theoretically important but has in practice not proven to be an important issue.

4.4.3. Combining results

After the data were multiple-imputed and the intended analysis was performed on each of these m datasets, the results must be combined. Usually, the objective is to estimate some parameter β (possibly multivariate), for example, the coefficients in a regression model. These m estimated parameters $\hat{\beta}^{(1)}, \hat{\beta}^{(2)}, ..., \hat{\beta}^{(m)}$ of the different datasets then must be combined into a single parameter $\hat{\beta}$, and the added uncertainty of the imputation must be incorporated.

The most well-known procedure to combine the results of multiple imputation is called Rubin's rules. The final parameter is estimated using the average of the estimated parameters, $\hat{\beta}^{MI} = \frac{1}{m}\sum_{i=1}^{m}\hat{\beta}^{(i)}$. The variance of the parameter V^{MI} (which is a measure of the uncertainty) is estimated by $V^{MI} = \bar{V}_W + (1 + \frac{1}{m})V_B$. The first component $\bar{V} = \frac{1}{m}\sum_{i=1}^{m}\text{Var}(\hat{\beta}^{(i)})$ reflects the "within-imputation" variance, the usual sampling variability, which is the uncertainty we have in the case of no missing data or if we imputed the data using the true underlying distribution (which is impossible in practice). The second component, with $V_B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\beta}^{(i)} - \hat{\beta}^{MI})^2$, reflects the added uncertainty because of the imputation. This is intuitively justifiable. If we have a large proportion of missing data, it is difficult to impute the values and each imputed dataset would be quite different, leading to considerable variance between the m estimated parameters (Molenberghs, 2015).

4.5. Imputation for survival analysis

To obtain the most meaningful imputed dataset, the outcome variable of the final model should be included as a predictive variable in the imputation model (Moons et al., 2006). In survival analysis, including the outcome variable (such as the survival time) requires special care because survival time is a peculiar variable (see Chapter 3). Furthermore, the Fine-Gray model (or Cox model) is non-linear, so the imputation model should allow for the same type of association to prevent the dilution of the relationship between the covariate and the outcome in the substantive model. In this section, we discuss two imputation algorithms adept to include survival data with competing risks.

Unfortunately, currently no imputation model exists that is compatible with a Fine-Gray substantive model (Lau and Lesko, 2018). However, two approaches exist for the competing risk setting when the substantive model are Cox proportional hazards models for each cause-specific hazard: the Resche-Rignon method (Resche-Rignon et al., 2020) (based on (White and Royston, 2009)) and the Bartlett method (Bartlett and Taylor, 2016). These methods might not be compatible with a Fine-Gray substantive model, but no better model exists and so we will focus on these methods.

We will first introduce the notation used (Bartlett et al., 2014). Let $X = (X_1, ..., X_p)$ denote the vector of the p partially observed covariates and let $Z = (Z_1, ..., Z_q)$ denote the q fully observed covariates. In both models we assume that the substantive model outcome is the survival time subject to censoring $Y = \min(T, C)$ and the type of event δ which equals 0 in case of censoring and otherwise the event $k \in \{1, ..., K\}$. For each cause k, we assume the cause specific hazard $\alpha_h(t)$ is a Cox proportional hazard function, thus $\alpha_h(t|X, Z) = \alpha_0(t) \exp(\beta X + \gamma Z)$. Now note that the likelihood of the observed event or censoring $(Y = t, \delta = d)$ for an individual is

$$\begin{aligned} f(Y = t, \delta = d|X, D) \propto & P(Y > t - |X, Z)\alpha_d(t|X, Z) \\ & = \left(\exp(-\int_0^t \sum_{k=1}^{K} \alpha_K(s|X, Z)ds) \right) \prod_{k=1}^{K} \alpha_k(t|X, Z)^{1_{d=k}}. \end{aligned}$$

The aim is now to define an imputation model that is compatible with the substantive model (Cox proportional cause specific hazards). Note that by Bayes theorem we have that the likelihood of a value for an unobserved covariate is

$$\begin{aligned} f(X_j|Y, D, X_{-j}, Z) &= \frac{f(X_j, Y, D, X_{-j}, Z)}{f(Y, D, X_{-j}, Z)} \\ &= \frac{f(Y, D|X_j, X_{-j}, Z)f(X_j, X_{-j}, Z)}{f(Y, D, X_{-j}, Z)} \\ &= f(Y, D|X_j, X_{-j}, Z)f(X_j|X_{-j}, Z)\frac{f(X_{-j}, Z)}{f(Y, D, X_{-j}, Z)} \\ &\propto f(Y, D|X_j, X_{-j}, Z)f(X_j|X_{-j}, Z). \end{aligned}$$

This means that the imputation model should be proportional to substantive model and an imputation model using only the other covariates as predictors, called the exposure model.

(White and Royston, 2009) approximate the posterior distribution and (Resche-Rignon et al., 2020) extends this to the competing risks setting. They show that using Taylor series approximation for different exposure models that the imputation model is approximately compatible if the covariates Z, D and the cumulative baseline hazard function at the event or censoring time $H_{0k}(Y)$. They propose to estimate the cumulative baseline hazard function by the Nelson-Aalen estimates. This imputation method is only approximately compatible but simulations show that there is little bias.

However, if some covariates have a large effect in the substantive model, the bias is expected to be larger. The smcfcs method, developed by Bartlett, samples from the posterior distribution directly and is thus fully compatible with the substantive model. First, the substantive model and exposure model is fitted conditioned on previous the imputed values in the previous step. This then gives us the distributions $f(Y, D|X_j, X_{-j}, Z)$ and $f(X_j|X_{-j}, Z)$. No closed form is in general available for the posterior distribution to sample from the distribution directly, so the method of rejection sampling is used (Bartlett et al., 2014) (Bartlett and Taylor, 2016).

5. Data

We introduced the data in Section 2.4 and described the treatment protocol in Section 2.1. The HOVON put much effort into preserving and managing the data, so the data are already clean and organized. However, one of the problems we have encountered is that the reality is not the same as the protocol as the personal considerations of the patients have also impacted treatment decisions. To provide more insight into the data, we discuss the treatment protocol realization using a visualization in Section 5.1. Moreover, the data was not acquired and organized with our study in mind, so we need to adjust the data for our application. We want to model the survival time between an initiating event (finishing induction treatment) and an outcome (relapse, fraft versus host disease (GvHD) or toxicity). We do this in order to aid consolidation treatment decision making. In Section 5.2, we discuss the decisions made to prepare the data for our analysis.

5.1. Treatment protocol realization

In Section 2.1, we discussed the consolidation treatment protocol: patients with intermediate or adverse risk will receive an allogeneic stem cell transplantation (SCT), or an autologous SCT if this is not possible, whereas good risk patients will receive an autologous SCT, or a third chemo if this is not possible. In reality, however, this is not always true. To gain insight into deviations from the protocol, we provide a visualization of the treatment paths of the patients in Figure 5.1. This flowchart displays the treatments patients had before any event has occured. The number in the box of the state is the number of patients ending their treatment path in this state and the pie chart captures the risk distribution of these patients. Note that this is the HOVON-102 classification, similar to the ELN classification but the adverse risk is split into adverse (orange) and very adverse (red).

The first deviation is the arrow between chemo I and allo, meaning that 23 patients did not receive both rounds of chemotherapy before starting consolidation treatment. A possible explanation is that these patients suffered severe side effects from chemotherapy and therefore the clinician decided to deviate from the protocol.

The second deviation is that some patients may receive two consolidation treatments: five patients received an autologous SCT first and then an allogeneic SCT, and 48 patients received a third chemo and an allogeneic SCT. The patients did not have any event (such as relapse) between the treatments, so this cannot explain the extra treatment. It is likely that (1) there was no stem cell donor available at the time and a third chemo or autologous SCT was meant to bridge the time until there was a donor, or (2) the autologous SCT was unsuccessful (for example because the patient did not recover well) and an allogeneic SCT was given to solve this.

The third deviation is that good-risk patients (green) may receive an allogeneic SCT: in all three boxes for allogeneic SCT, there is a slice of green in the pie chart. Possible explanations are (1) the patients responded poorly to initial therapy even though they were classified as good risk and

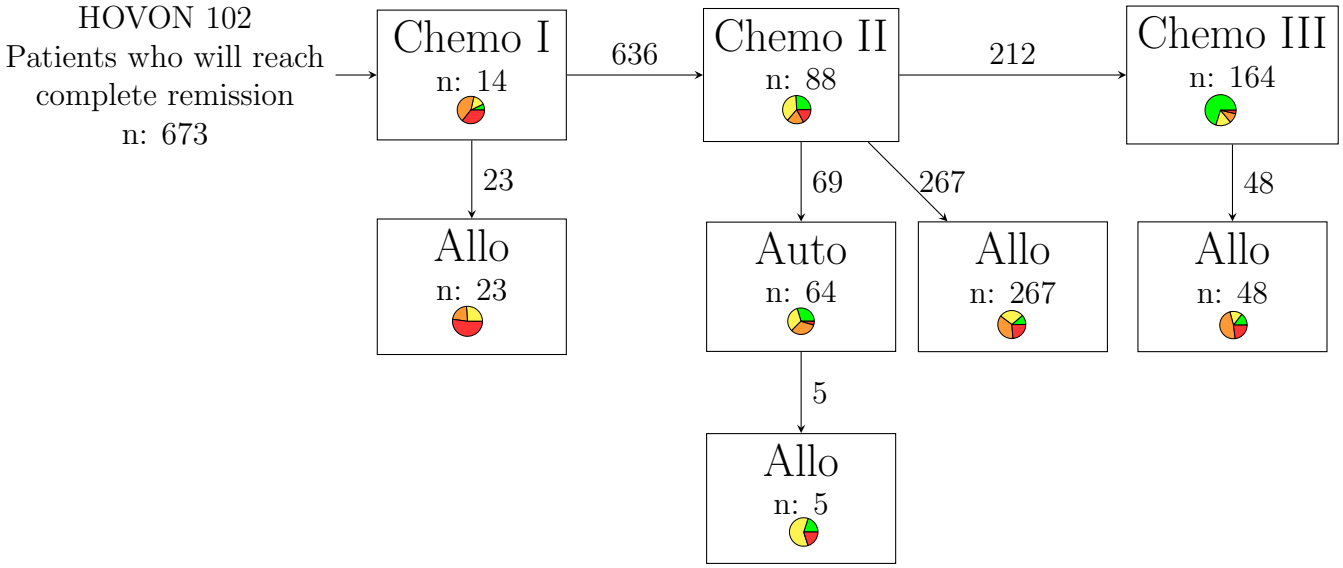


Figure 5.1.: Visualisation of treatment paths taken by patients in the HOVON 102 study who eventually reach complete remission. The numbers in the boxes are the number of patients who end their treatment path at that point (for example: 64 people got two rounds of chemo and an autologous stem cell transplantation) and the number along the arrows represent the number of patients having made that transition (for example: 212 patients got a third chemo after the second chemo). The pie charts in the boxes visualise the distribution of the risk classification of the patients who ended their treatment there, where green represents favorable risk, yellow intermediate risk, orange adverse risk and red very adverse risk (for example: almost three quarter of patients who only had three rounds of chemo were good risk patients).

thus the clinician was afraid of a relapse or (2) as before, the patient did not recover well from the autologous SCT or the third chemo and an allogeneic SCT was given.

The data lacks information on the intended consolidation treatment. Consider, for example, the state of the third chemo. About 75% of the patients who have only had three chemotherapies (and no SCT) have positive risk profiles. According to the treatment protocol, we would expect mostly good risk patients as this is their alternative treatment. The non-good risk patients (yellow, orange or red) would only get a third chemo if both SCTs are not possible, which seems unlikely. For the quarter of non-good risk patients who have only gotten three rounds of chemo, it is impossible to deduce their intended treatment.

Lastly, the number of patients getting an autologous SCT versus a third chemo is strikingly high. According to the protocol, a third chemo is only indicated when an autologous SCT is not possible, suggesting that most patients would get the autologous SCT. Instead, 164 patients received a third chemo as consolidation treatment whereas only 64 patients received an autologous SCT.

5.2. Data curation

The aim of this study is to aid consolidation treatment decision making. Patients first need to reach complete remission before they can undergo consolidation treatment (which is meant to keep the patient in complete remission). If the patient fails to reach complete remission during induction treatment, they will be treated off-protocol. Furthermore, we focus on predicting relapse, and a patient first needs to reach complete remission before they can relapse. Thus, we have decided to only include patients who have reached complete remission during induction treatment.

In the HOVON 102, two dosing regimens of the experimental drug were investigated. In patients treated with the highest clofarabine doses, a unexpectedly high toxicity rate was observed, which led to dose reduction during the rest of the study. Because the treatment related risks of these patients were much higher than in the other arms, including this cohort would lead to bias of treatment related toxicity and mortality analyses. Therefore, we excluded patients who were treated in this experimental arm (32 patients).

Initially, we expected all patients to receive two cycles of chemotherapy before starting their consolidation treatment, and the end of induction therapy would be the initializing event. However, 14 + 23 patients only received a single round of chemotherapy, as displayed in Figure 5.1. For these patients, we have used? the completion of the first chemotherapy treatment as the initializing event.

Some patients received an allogeneic SCT after an autologous SCT. Allogeneic SCT provides a much stronger anti-leukemic activity but also the higher risks of treatment related toxicity, compared to autologous SCT. Thus, these patients will be included as if they only received an allogeneic SCT. Contrarily, this is not necessarily the case for a third chemo, so we include both treatments in case of a third chemo and an allogeneic SCT.

Last, we have imputed missing values based on deduction as much as possible. For example, when the value for the variable indicating whether a patient previously had chemotherapy was missing, we assumed this was not the case for the patient or else it would have been included in the record. When the value of the cause of death variable was missing, we imputed this based on the more extensive description of the cause of death.

6. Imputation

6.1. Implementation

6.1.1. Algorithm 1: White and Royston and Resche-Rignon

The first step of the imputation method described by (White and Royston, 2009) is to estimate the (marginal) cumulative cause-specific hazard functions by the Nelson-Aalen estimates for the causes relapse, graft versus host disease (GvHD) and toxicity (chemotherapy related risk). We can do this in R using the following function:

```
1 mvna(DF, state.names, transitions, cens.name),
```

where DF has columns id (of the patient), from (from state of transition), to (end state of transition) and time (of transition); state.names contains the state names; transitions is the matrix of all possible transitions; and cens.name is the censoring name. For each patient, the estimated cumulative cause-specific hazards at the time of the event (even if the patient is censored) are added to the data as covariates. So, in our case, three columns are added to the dataset (one for each cause).

A practical complication is that only patients who have had allogeneic SCT can suffer from GvHD. This would imply that the cumulative hazard of GvHD is constantly zero, whereas it is at some point larger than zero for patients who have had an allogeneic SCT. If we estimated the GvHD cumulative hazard with the full dataset, we would overestimate the hazard for patient without allogeneic SCT and underestimate it for patients with allogeneic SCT. Furthermore, we know that an allogeneic SCT reduces the relapse hazard more than an autologous SCT or third chemo because of the suppresive effect of graft vs. leukemia (which is why it is given to the worst-risk patients), so the cumulative relapse hazard is also unlikely to be the same.

Because of this complication, we have decided to stratify the dataset based on whether the patients have had an allogeneic treatment and estimate the cumulative hazards separately for these sets. These estimates are depicted in Figure 6.1. Note that the cumulative hazards for patients who have had an allogeneic SCT is zero for some time. This is the consequence of the immortality bias; we have only included patients who have made it to the consolidation therapy, but in reality, these patients did not have zero hazard until the treatment. In contrast, the cumulative toxicity hazard increases right after time zero for the patients who have not had an allogeneic SCT, but we should note that the data subset used to estimate hazard contains all cases of patients suffering from toxicity before their intended consolidation therapy. Note that the cumulative relapse hazard is indeed larger for patients who have not had allogeneic SCT. This could also be, in part, a consequence of the opposite of immortality bias. However, the difference becomes most apparent after 50 days, which is when most patients have had their intended consolidation therapy, so it is

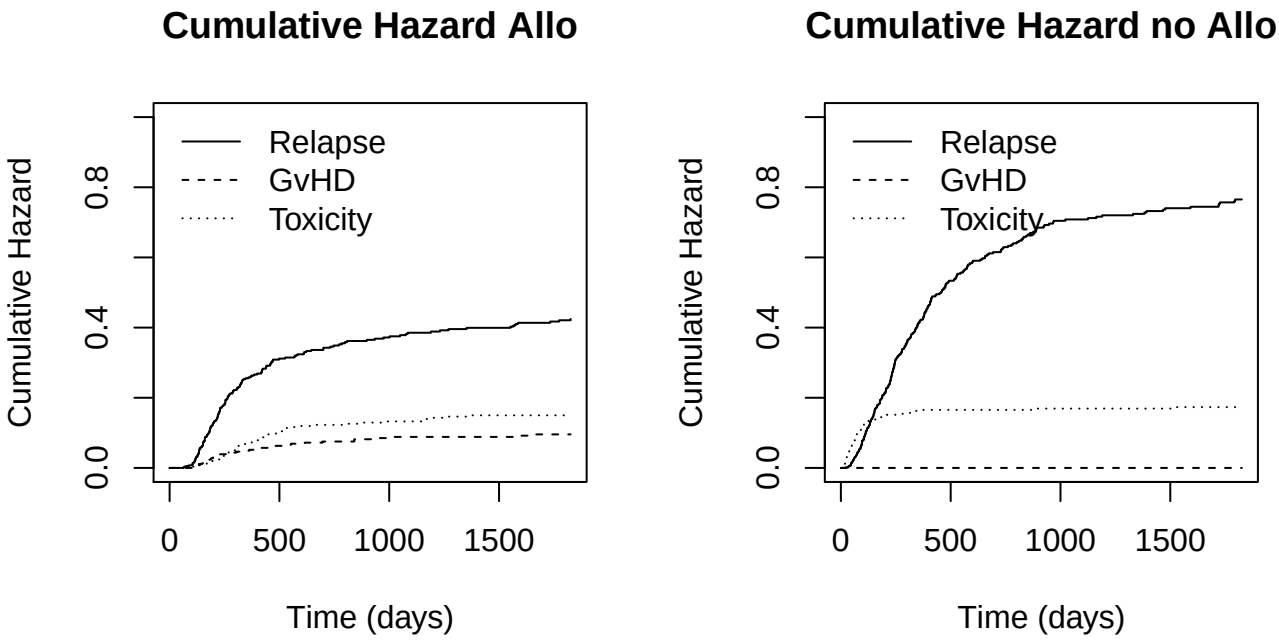


Figure 6.1.: Cumulative baseline hazard functions stratified on allogeneic SCT. Left shows the cumulative hazard functions for patients who have had allogeneic SCT for the three possible transitions Relapse, GvHD and Toxicity. Right showns this for patients who have not had allogeneic SCT. Time is given in days.

more likely a consequence of the type of consolidation.

Before we can use the MICE algorithm for the imputation, the data need to be compatible with the algorithm. First, the discrete variables need to be coded as factors and discrete ordered variables should be included as ordered factors. Next, if any variable with missing values is constant, the algorithm is not able to impute these values. Similarly, the algorithm is not able to impute in the case of perfect correlation between the outcome and a predictor in the imputation model. For example, imputing the gender of the donor (coded as (1) same gender, (2) different gender and male, or (3) different gender and female) with the gender of the patient yields a perfect correlation; different gender female donor are only present in male patients. Lastly, the missing values should be NA, but values that are intentionally missing (for example the gender of the donor when the patient has not had a transplantation) should not be NA but filled in instead.

We have decided to use the quickpred function of MICE to choose the predictors for each imputation model. The output of this function is a binary matrix $P = (p_{ij})$ where a cell p_{ij} is one if variable j is included in the imputation of variable i. This matrix can also be filled in manually; however, the output of quickpred was validated by three acute myeloid leukemia experts, who deemed the matrix justifiable. We manually set columns to zero for variables in the dataset without a predictive value (such as the index).

Finally, the imputation was performed using the function "mice" and the output of quickpred.

6.1.2. Algorithm 2: Bartlett

The Bartlett imputation method, called smcfcs, uses the substantive model (which they assume to be a Cox proportional hazard model for each cause-specific hazard). They assume the true substantive model to be known but this is in reality often not true. We thus now need the substantive model for the imputation and we need the imputed dataset to obtain the substantive model. We think it would be best to iterate and at each iteration improve the imputation model by using the newest substantive model and then use the improved imputation model to improve the substantive model until we reach a form of convergence.

However, because of time restrictions, we have only built the substantive model for the dataset imputed with the Resche-Rignon method. Note that we can not use this substantive model for the smcfcs imputation method because the smcfcs method requires a Cox proportional model for every cause-specific hazard. Instead, we have developed a Fine Gray model for a single cause-specific hazard. To implement the smcfcs imputation, we have modeled the cause-specific hazards on the complete case dataset using a step-up method.

The complete case dataset is small, especially for competing risks for which few events were observed. We have therefore also modeled the cause-specific hazards by including variables that were shown to be predictive in previous research. However, these results did not directly translate to our dataset because variables are defined differently. Moreover, there were hardly any results for the competing risks toxicity and Graft versus Host disease. This resulted in the imputation model not being able to converge.

6.2. Results

It is difficult to summarize the results or asses the quality of the imputation based on the results of the imputation alone. The quality would be best established by the quality of the resulting predictive substantive model, which can be determined by validating the model with a different dataset. We used a diagnostic plots to assess the convergence of the MICE algorithm.

Figure 6.2 shows the convergence plot of the statistics mean and standard deviation of three mutation variables – kitex8, asxl1 and idh1. Not having a mutation is in this case coded as 1 and having the mutation as 2, so the mean is between 1 and 2. Each plot has 20 different colored graphs belonging to the different imputed datasets. If the plots showed a clear trend that does not stabilize, this would be a sign of non-convergence and we could increase the number of iterations. This is not the case in the plot and it looks like healthy convergence. (Buuren, 2018) presents more diagnostic tools to assess the quality of the imputation.

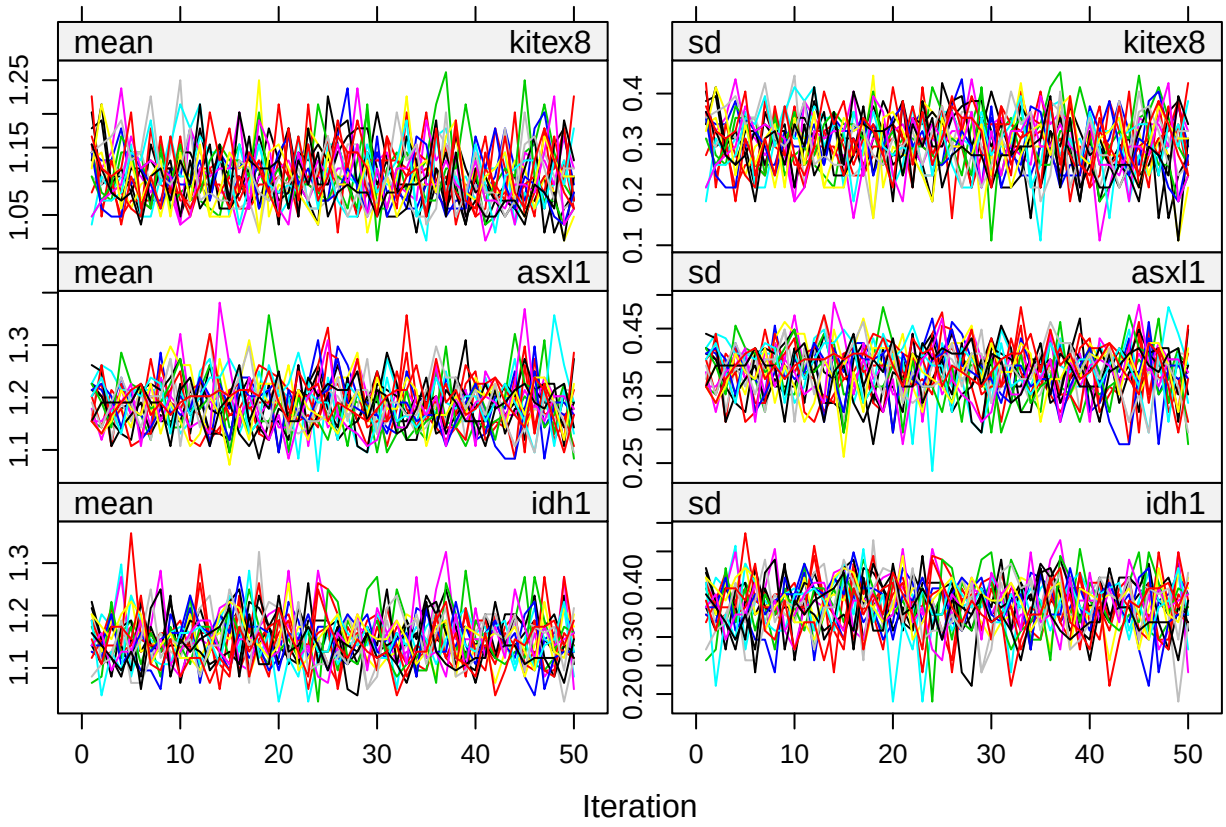


Figure 6.2.: Convergence plot of MICE algorithm for three variables: the statistics mean and standard deviation of the imputed values as a function of the iteration number. The twenty differently colored graphs correspond to the twenty imputed datasets.

7. Modelling methodology

In this section, we discuss our methodology for building a Fine-Gray model to predict relapse in acute myeloid leukemia patients based on the HOVON 102 dataset (see Chapter 5). The aim of the model is to aid clinicians and patients in the consolidation[1] treatment decision making; after induction[2] the clinician and patient need to decide between allogeneic SCT or autologous SCT (Chapter 2). Therefore, the initializing event, time zero, will be at the completion of induction therapy. The focus is on predicting the probability of having a relapse within five years based on the covariates of the patient and to compare the risk of the two different treatment strategies.

7.1. Competing risks

We use a Fine-Gray model because there are competing risks: relapse, graft versus host disease and toxicity (which we define as all treatment-related risks except GvHD). We cannot avoid the competing risks because they cannot be combined into one risk; treatment and disease risks are opposing forces, and we are specifically interested in relapse. An alternative to the Fine-Gray model is modelling a cause-specific hazard for relapse (Section 3.2). However, our aim is to model the probability of relapse – that is, the cumulative-incidence function (CIF). The Fine-Gray model is developed for the covariates that directly affect the CIF; this allows for a better interpretation of the covariate effect on the probability of relapsing.

The consolidation-treatment decision involves a balance: it should eliminate disease-related risks but should not be so severe as to cause treatment-related risks. Relapse is a disease-related risk, which is a strong adverse prognostic factor for overall survival. Graft versus host disease is a treatment risk of allogeneic stem cell transplantation because the donor cells attack healthy cells as well as cancerous cells. This can be suppressed by medication, but it can still lead to death. Another important treatment-related risk is due to chemotherapy, which is meant to kill all cancer cells but simultaneously kills healthy rapidly dividing cells. Consequently, the immune system is temporarily compromised which increases the risk of dying of infections. For this risk, we include all patients who are overtreated by chemo: both therapy-related deaths and patients leaving the protocol because of overtreatment (due to hypoplasia[3] and toxicity).

7.2. Multistate model

In order to deal with the time dependency of the treatment effect and the immortality bias, we use a multistate model that is visualized in Figure 7.1. An individual starts in state complete remission

[1] stem cell transplantation or third chemo given to patients in complete remission to prevent relapse
[2] the first treatment, consisting of two rounds of chemo therapy, to reach complete remission
[3] hypoplasia in leukemia is the inadequate recovery of healthy stem cells, which is a type of toxicity

(CR) after induction and can then transition to either the consolidation treatment or a competing risk, such as relapse or chemo-related risk; this additional step is taken to model whether the patients reach their intended consolidation treatment. When a patient receives the consolidation treatment, the model time is reset to zero because we assume the consolidation treatment effect depends on the time since treatment instead of time since induction[4]. From the consolidation treatment state, the patient can also transition to any of the competing risks. In this subsection, we discuss why we think this is the best solution, and we explain our decision to use this specific approach.

The consolidation-therapy-related variables, such as the therapy type or the donor type, are inherently time dependent and require special care, as explained in Section 3.2.4. The treatment decisions are made right after completion of the induction chemotherapy cycles, which is our initial event, so most variables are determined at time zero – except for possible alterations afterwards (for example switching treatments because no donor was found or because of a failed stem cell harvest). However, the intended consolidation treatment is not available and cannot be reconstructed because the reality is not the same as the protocol Section 5.1; all we know is the treatment that was given. If we were to use this as a baseline variable, we would introduce an immortality bias (Section 3.2.4).

If we had the intended consolidation treatment data, it would be unreasonable to assume that the effect of the treatment is constant over time and does not depend on the time since the treatment. First, the effect of the treatment will evidently be different after the treatment was given compared to before the treatment. Second, the time between completion of induction and consolidation treatment varies between individuals.

A plausible assumption is that the consolidation treatment effect only depends on time since consolidation and not on the time between induction and the consolidation. A possible approach would thus be to take the consolidation treatment (instead of the induction therapy completion) as the initializing event. This would solve the problem of the time dependent effect of the treatment. This, however, does not solve the immortality bias; to obtain the probability of relapsing within five years of induction, we need to model whether each patient receives their intended treatment, as this is not certain.

7.2.1. The states

We can structure our multistate model in different ways based on how we choose the states. We chose the structure depicted in Figure 7.1. We have chosen to group all consolidation treatments into a single state. For the transitions from consolidation treatment to the competing riks, we have chosen not to stratify based on consolidation treatment. Instead, the type of consolidation is included as a covariate. In this subsection, we introduce the alternatives and justify our choice.

We still have to explain how the treatment-related states are incorporated as we have already decided in the previous section that the three competing risks (relapse, GvHD and toxicity) will be separate states in the model. As an alternative to our decision to group all treatments into a single state, every treatment could be a separate state, as depicted in Figure 7.2. A second alternative is to have consolidation as a single state and stratify the patients reaching the "consolidation" state on consolidation treatment.

The reason to group the consolidation treatments into a single state is because otherwise the different treatments (autologous SCT, allogeneic SCT, or a third chemo) would interact as competing

[4]This assumption was deemed reasonable by the acute myeloid leukemia experts at AUMC

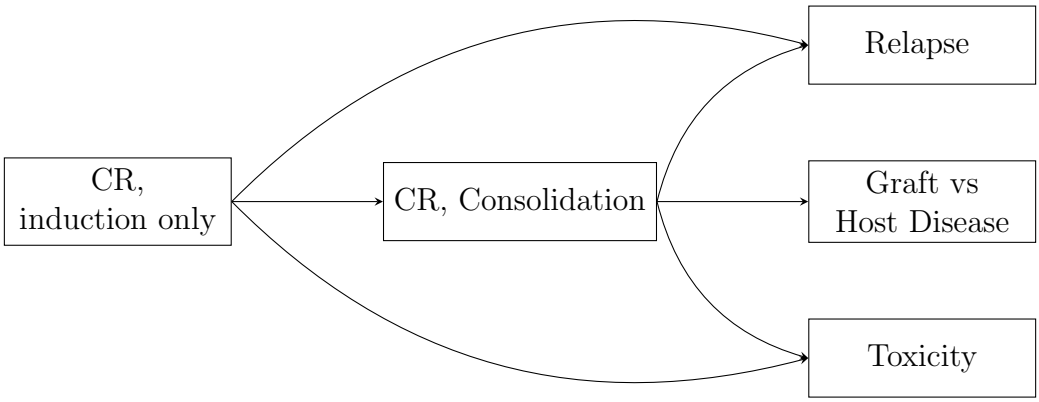


Figure 7.1.: Multistate model.

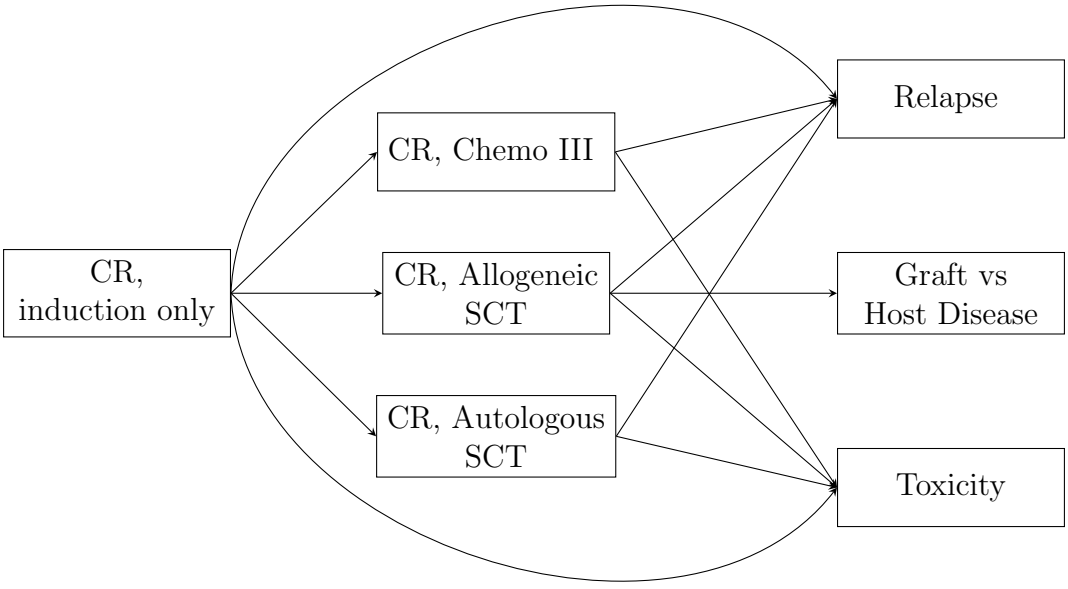


Figure 7.2.: Multistate model.

risks. The probability of having an allogeneic SCT would, in the Fine-Gray model, thus be reduced because the patient might also get an autologous SCT. First, this would not correspond to reality because the type of consolidation treatment is set after completion of induction chemotherapy (our time zero), so alternative treatments cannot thwart the intended treatment. Second, the probability of a transition to a specific treatment would be more difficult to interpret: what proportion of patients not receiving the treatment do not receive it because something negative has occured and what proportion simply receive a different treatment? This is a particularly serious issue when we consider the probability of two steps in the multistate model: first receiving the treatment and then relapsing.

The reason not to stratify on type of consolidation treatment is because only 64 patients received an autologous SCT, whereas, by comparison, over 300 patients received an allogeneic SCT. The complexity of the model from autologous SCT to relapse would thus be limited by size of the dataset (a model with many variables included needs sufficient data to train). We assume the effect of the covariates is the same for patients who have had different consolidation treatments. Suppose the assumption holds; then combining the patients with different consolidation treatments would add power to the model as there are more cases to estimate the coefficients of the model. This would allow for an extensive model for all treatment groups.

However, it is uncertain if the assumption holds. Stratifying based on the consolidation treatment group would allow for an interaction effect between variables and the consolidation treatment and would allow for a different baseline hazard for different treatments. If sufficient data is present, we would opt for stratification because the assumption might be too strong.

7.2.2. Two-step probability

Ultimately, we want to model relapse survival probability based on patient characteristics and compare the probability for patients who have received an allogeneic SCT to that for those who have received an autologous SCT. In other words, we are interested in the following two step probabilies in our multistate model:

$$P(X(0) = CR, X(r) = \text{autologous SCT for some r between 0 and t}, X(t) = \text{relapsed}|Z), \quad (7.1)$$
$$P(X(0) = CR, X(r) = \text{allogeneic SCT for some r between 0 and t}, X(t) = \text{relapsed}|Z), \quad (7.2)$$

where $X : [0, \infty) \to \mathcal{I}$ is the stochastic process of the jumps in the statespace $\mathcal{I} = \{0, \text{consolidation treatment, relapse, GvHD, toxicity}\}$ with $X(t) = g$ if the individual is in state g at time t. These probabilities can then be used to decide on the consolidation treatment. Note that the probability of not receiving consolidation treatment also reduces this two step probability. When the probabilities in 7.2 are used in practice, the CIF from CR to relapsed should also be taken into account.

7.3. Variable selection

We will use the step-up method for variable selection for each model, which means that one variable is added at each step. We do this until adding a new variable does not significantly improve the model anymore. Specifically, starting with an empty model, we will add the variable with the highest pseudo log likelihood score, which we will use as a measure of improvement of the model.

The coefficient for the variable needs to be significant at the 0.05 level. Furthermore, we require the minimum category (by which we mean the category that occurs least often) of discrete variables to be at least 5% of the total number of cases the model is trained on (e.g., if only 3 out of 300 patients have a mutation, this variable will be excluded). We do this to prevent overfitting and so that the model is for the majority of patients. Furthermore, we will not use the ELN-risk classification as a variable in the model because this is a combination of other variables.

7.4. Multiple imputation adaptions

As discussed in Chapter 4, multiple imputation results in $m = 20$ complete datasets and combining the results requires special care, which is discussed in Section 4.4.3. Instead of training the model once, the model needs to be trained on all m datasets and the results are then combined to correct for the added uncertainty due to imputation. We have adjusted the p-value by correcting the variance by Rubin's rules (Section 4.4.3). The new pseudo-loglikelihood is the median of the twenty datasets obtained by the different datasets, and similarly, the coefficient is the mean.

7.5. MRD and LSC

Amsterdam UMC is especially interested in the added value of the MRD and LSC variables. Unfortunately, more than half of the cases miss these values, so imputation is likely not reliable. For this reason, we have included both the imputed and the non-imputed MRD and LSC variables; furthermore, we have also included a discretized version where we used a cut-off value of 0.1 for MRD and 0 for LSC.

For every step in the step-up plan, we will compare the newly added variable to the MRD and LSC variables. For a fair comparison of pseudo log likelihoods of different models, we need the models to be trained on the same dataset. Therefore, the pseudo log likelihood of the model with the newly added variable will also be trained on the datasets for which the MRD is known and similarly for the LSC. However, we will not include the non-imputed MRD and LSC variables even if their log likelihoods are higher because we want the variable selection to be on the complete dataset. Instead, the MRD and LSC variables will be added after the variable selection is completed.

7.6. Implementation

The models will be estimated with the cmprsk package in R. To implement the multistate model, the data need to be adapted; instead of one row in the dataframe per patient, we need one row per transition in the multistate model and when a patient is censored instead of transitioning. Then, the transitions of interest are modelled separately with the dataset including all rows with the same origin state.

8. Results

In this section, we will first present an extensive description of the results of the step-up procedure, described in Section 7.3, of the Fine-Gray models for the transitions CR and induction only to Consolidation treatment. The model started with a single variable and the top results for all possible variables are summarized in the tables in Appendix A. At each step, the best variable is added, of which the results are summarized in the tables, until no other suitable variable exists. Similarly, the model for the transition from consolidation to relapse was built.

Because of special interest in the MRD and LSC-related variables (both continuous and discrete and both imputed and not imputed), the top four variables based on the log-likelihood scores are also presented in the table. In order to compare the variable that were added to the model with the MRD and LSC variables, a footnote is added with the log-likelihood test scores of the potentially new model on the reduced subsets for which the MRD and LSC are known.

The final results of both models are shown and discussed. An illustration of the cumulative incidence function is given for different covariate combinations to illustrate the effect of the variable coefficients on the probability. Lastly, the two models are combined using Equation (7.2) into the two-step probability given the covariates in both models combined. An illustration of the two-step probability is given for different sets of covariate values.

8.1. Step up variable selection: Transition from CR and induction only to consolidation

First, the treatment arm, a variable that is 1 if clofarabine[1] was given and 0 otherwise, was the first variable we added to the Fine-Gray model of the transition from CR and induction only to consolidation. Single variable models were trained for every variable, and the top 6 models, based on the log-likelihood test score and a significance level of 0.1, are summarized in Table A.1. Based on our step up protocol given in Section 7.3, the treatment arm was added to our model; this variable has the highest pseudo log-likelihood test score (36.706) among the variables that were significant at the 0.05 level (p-value of 1.7×10^{-9}) and had a minimal category size of $0.05 \cdot 663 = 33$ (314). The coefficient of the variable in the final model is -0.528 with $\exp(-0.528) = 0.590$.

In the lower half and the footnote of Table A.1, note that our current model with the variable treatment arm has a log-likelihood test score of 7.613 on the subset for which the LSC is known whereas the model with the continuous LSC on the same subset has a higher log-likelihood of 8.161. The coefficient of continuous LSC is -7.453 (with $\exp(-7.453) = 0.001$), which is further away from zero than the other coefficients are. However, there are only 13 cases where the continuous LSC is larger than 0.001 and $exp(0.001 \cdot (-7.453)) = 0.993$ which is almost equal to one, and thus, hazards

[1]The clinical study from which the dataset originates was aimed to evaluate the effect of the chemotherapeutic medicine Clofarabine.

Table 8.1.: Top six variables and top four MRD/LSC related variables as variable 1 in the model for transition CR to consolidation

Variable	LL test	p-value adj	p-value	Coef	exp(Coef)	Minimal Cat.
Treatment_arm	36.706	0.000	0.000	-0.518	0.596	314 / 663
age	29.956	0.000	0.000	-0.019	0.981	1 / 663
Clofa&intRisk	16.046	0.000	0.000	-0.486	0.615	92 / 663
risk102_1	14.908	0.000	0.000	0.365	1.440	194 / 663
risk102_2	13.924	0.000	0.000	-0.361	0.697	170 / 663
asxl1	11.027	0.003	0.001	-0.454	0.635	75 / 663
LSC_cont	8.161	0.016	0.016	-7.453	0.001	1 / 290
MRD_cont	3.029	0.325	0.325	-0.029	0.971	1 / 286
LSC_0	1.156	0.283	0.283	-0.143	0.867	96 / 290
MRD_.1	0.172	0.699	0.699	-0.064	0.938	61 / 286

[a] A Fine Gray model is trained with a single variable.

[b] LL score of new model on the MRD subset: 4.598352, on the LSC subset: 7.612600

of at most 13 cases are affected, which is less than $5\% \cdot 290 = 14.5$ of the total number of cases. The binary LSC variable is not significant at the 0.05 level (p-value of 0.283) and has a log-likelihood test score of 1.156, which is significantly lower than that of our model on the same subset.

On the subset for which the MRD is known, the log-likelihood score of the current model is 4.598, which is higher than the model with a continuous MRD variable which has log-likelihood score 3.029. The variable is also not significant at the 0.05 level (p-value of 0.325). The binary MRD performs worse based on both p-value and log likelihood test score.

Second, the ages of the patients was added to the model with the treatment arm variable. Table A.2 summarizes the top six two-variable models from CR and induction only to consolidation with treatment arm and a potential new variable. The variable has a log-likelihood test statistic of 61.612 and an adjusted p-value of 2.3×10^{-6}. The minimal category is one, but as this is a continuous variable, this statistic is irrelevant. The coefficient of the variable in the final model is -0.015 with $\exp(-0.015) = 0.985$. The ages of the patients in the study are between 18 and 66 with a mean (median) age of 50 (53).

Now, consider the lower half of Table A.2. The continuous LSC again has again the highest log-likelihood for the MRD and LSC variables that are significant at the 0.05 level. Moreover, the log-likelihood of the model with the treatment arm and the continuous LSC of 17.473 is higher than that of our current model with the treatment arm and age of 16.516. However, the effect for more than 95% of the cases on the hazard is $\exp(0.001 \cdot (-8.470)) = 0.992$ or even closer to one. The binary LSC has a lower log-likelihood test score (10.454) than our current model and is not significant at the 0.05 level (p-value of 0.098).

The log-likelihood of the continuous MRD (8.874) is slightly less than that of our current model but is not significant at the 0.05 level (p-value of 0.299). The coefficient of the continuous MRD is -0.036, which is close to zero, but note that the 95th percentile of MRD is 1.828, at which $\exp(1.828 \cdot (-0.036)) = 0.936$. The binary MRD performs worse in terms of log-likelihood (4.996)

and p-value (0.563).

Third, the genetic mutation asxl1 is added to the model with treatment arm and age. Table A.3 summarizes the top six two-variable models from CR to consolidation with a potential new variable. The variable with the highest log-likelihood test score is risk102_1, which represents the good-risk patients from the ELN classification (score from 1 to 4, all added as dummy variables). However, in Section 7.3, we determined this variable should not be included. The variable with the second highest log-likelihood test score, 72.417, is the mutation asxl1 with an adjusted p-value of 0.006. The coefficient of the variable in the final model is -0.459 with $\exp(-0.459) = 0.632$.

The same observations hold for the unimputed LSC variables in the lower half of Table A.3 as that of Table A.2. Note that now the log likelihood of the model with continuous MRD (13.616) is higher than that of our current model with asxl1, but the p-value is not significant (0.297). The log-likelihood of the binary MRD is lower and has a higher p-value (0.688).

Fourth, the mutation nras is added to the model with treatment arm, age and asxl1. Table A.3 summarizes the top six three-variable models from CR and induction only to consolidation with a potential new variable. The top-two log-likelihood scores are of ELN risk classification variables, which will not be added to the model. Note that the adjusted p-value of 0.027 is almost twice as high as the mean p-value of 0.014. The coefficient of the variable in the final model is 0.253 with $\exp(0.253) = 1.288$.

The observations in the lower half of Table A.3 also hold for Table A.4: the model with continuous LSC and MRD added outperform the current model with nras, but the LSC has hardly any effect for more than 95% of the cases, and MRD is not significant at the 0.05 level (0.298). The binary cases perform worse than the current model based on the log-likelihood test score and are not significant.

Fifth, the continuous variable bmblast (percentage of blasts in bone marrow) is added to the model because it has the highest log likelihood score (85.009) and is significant at the 0.05 level (0.027), which can be found in Table A.5. The corresponding coefficient in the final model is 0.005 with $\exp(0.005) = 1.005$. Because the variable is a percentage, values are between 0 and 100 (but because of imputation the values are between -3.8 and 100) with a mean of 52 and median of 54. Similar observations hold for the lower half of Table A.5 as before.

Sixth, the mutation FLT3-ITD is added to the model, which now has a log-likelihood test score of 92.111, and the variable has a p-value of 0.014 (see Table A.6). The coefficient of the variable in the final model is -0.224 with $\exp(-0.224) = 0.799$. Note that the follow-up variable is npm1&flt3; npm1&flt3 is one if the patient has both mutation FLT3-ITD and NPM1. In the lower half of the table, note that after adding continuous LSC and MRD, our model now no longer outperforms the model with the chosen variable. Otherwise, similar observations hold as before.

At the seventh step, we have reached the end of the step-up procedure, and the results which can be found in Table A.7. The variable with the highest log likelihood is npm1, now on its own instead of as the interaction variable in the previous table but it is not significant at the 0.05 level. The only variable that is significant at the 0.05 level is chromosomal abnormality abn17_p, but the smallest category of this variable has size 22 which is less than 5% of the total number of cases. Because we have concluded our step-up procedure, the lower half of the table will be the same as in the summary of all MRD and LSC variables that can be added to our final model, which can be found in Table 8.4.

Table 8.2.: Final model estimations of transition CR to consolidation

	coef	exp(coef)	p-value	2.5%	97.5%
Treatment_arm	-0.533	0.587	0.000	0.493	0.698
age	-0.016	0.984	0.000	0.977	0.992
asxl1	-0.459	0.632	0.005	0.458	0.871
nras	0.262	1.300	0.033	1.021	1.654
bmblast	0.005	1.005	0.004	1.002	1.008
flt3itd	-0.274	0.760	0.014	0.612	0.946

[a] p-values are adjusted for added incertaity due to imputation
[b] LL score of final model: 92.111444

8.2. Transition CR to consolidation

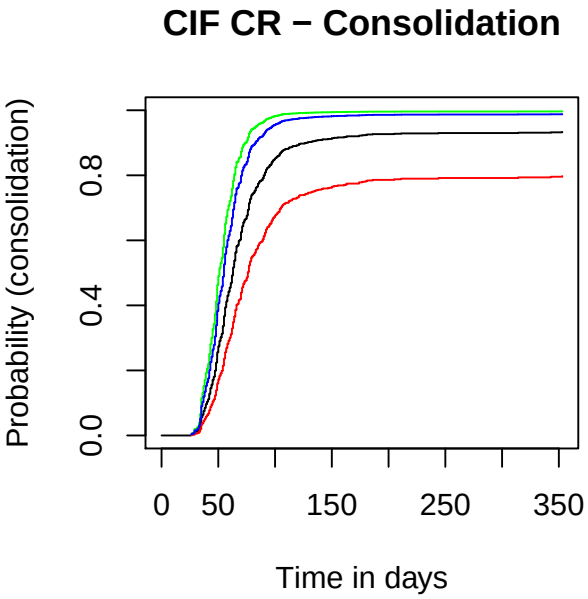


Figure 8.1.: Cumulative incidence function for different covariates of the final model for transition CR to consolidation.

	Black	Red	Green	Blue
Treatment_arm	0	1	0	0
age	50	50	18	18
asxl1	0	0	0	0
nras	0	0	1	0
bmblast	54	54	54	54
flt3itd	0	0	0	0

Table 8.3.: Legenda of Figure 8.1

The final model and the variable estimations are summarized in Table 8.2. The variables included in the model are treatment arm, age, genetic mutations asxl1, FLT-3ITD and nras and the percentage bone marrow blasts. Note that all variables in the final model are significant at the 0.05 level – this is not a consequence of the added variables being significant because adding a new variable can make previously added variables insignificant. The log likelihood score of the final model is 89.380.

Table 8.4.: Summary of the MRD/LSC related variables added to the model for transition CR to consolidation

Variable	LL test	p-value adj	p-value	Coef	exp(Coef)	Minimal Cat.
MRDimp_cont	97.782	0.289	0.077	-0.017	0.983	1 / 663
MRDimp_.1	96.784	0.259	0.096	-0.152	0.859	267 / 663
LSCimp_cont	93.573	0.357	0.093	-0.659	0.518	1 / 663
LSCimp_0	92.289	0.626	0.500	-0.059	0.942	285 / 663
LSC_cont	45.825	0.016	0.015	-7.601	0.001	1 / 290
LSC_0	39.083	0.430	0.425	-0.115	0.891	96 / 290
MRD_cont	29.566	0.344	0.343	-0.033	0.968	1 / 286
MRD_.1	26.093	0.812	0.811	-0.042	0.959	61 / 286

[a] Additional variables in the models: Treatment_arm, age, asxl1, nras, bmblast

[b] LL score of final model on the MRD subset: 24.873125, on the LSC subset: 37.811238

The corresponding cumulative incidence function is illustrated in Figure 8.1 for four different covariate profiles presented in Table 8.3. The black line corresponds to a patient who has an average age, median bone marrow blast percentage, none of the mutations in the model, and is in the control group. We use this graph as a baseline for reference of other covariate profiles. The red graph has the same profile but has had the experimental drug Clofarabine, which is the covariate with the lowest coefficient in the final model (-0.528). The red graph lies below the black graph; the red graph is 0.79 at 350 days whereas the black graph is 0.93. The blue graph represents a patient similar to that of the black graph but is not of average age but instead is the youngest patient in the study. The blue graph lies above the black graph and is 0.99 at 350 days. The green graph is of a patient of similar age as that of the blue line but additionally has mutation nras which has coefficient 0.253. The green graph lies above the blue and black line and is 1 at 350 days.

Table 8.4 summarizes the models adding an MRD or LSC variable to our final model of the step-up procedure. The imputed MRD and LSC variables have higher log likelihood test scores than the non-imputed variables, but these are trained on the full dataset, and their adjusted p-values are far from significant at the 0.05 level even though some mean p-values are. Furthermore, the log likelihood scores are lower than our current model. Note the large difference between their adjusted p-values, which take into account the added uncertainty of imputation, and the mean p-values. Moreover, note that the continuous variant in each case outperforms the binary variant.

Furthermore, we see that adding the continuous LSC to the final model increases the log-likelihood from 37 to 45 and is significant at the 0.05 level (0.016). But as noted before, the effect of the variable on the hazard is close to 1 for most of the cases. The other variables are not significant at the 0.05 level of significance.

Table 8.5.: Final model estimations of transition consolidation to relapse

	coef	exp(coef)	p-value	2.5%	97.5%
cyt6_6	0.939	2.557	0.000	1.658	3.943
wbc	0.007	1.007	0.000	1.005	1.009
Allo	-0.689	0.502	0.000	0.373	0.674
npm1	-0.699	0.497	0.000	0.353	0.699
cyt6_1	-1.206	0.299	0.011	0.119	0.754
age	0.019	1.019	0.003	1.006	1.032
Treatment_arm	-0.414	0.661	0.004	0.499	0.875
abn11q23	0.655	1.926	0.007	1.194	3.107

[a] p-values are adjusted for added incertaity due to imputation

[b] LL score of final model: 89.191372

8.3. Transition from consolidation to relapse

The final model and the variable estimations are summarized in Table 8.5. The variables included in the model are cyt6_6 (monosomal karyotype), wbc (white blood cell count), allogeneic SCT, mutation npm1, cyt6_1 (chromosomal abnormality t(8;21)), age, treatment arm (clofarabine or control group), and chromosomal abnormality abn11q23. Note that, again, all variables in the final model are significant at the 0.05 level of significance. The log-likelihood score of the final model is 89.191.

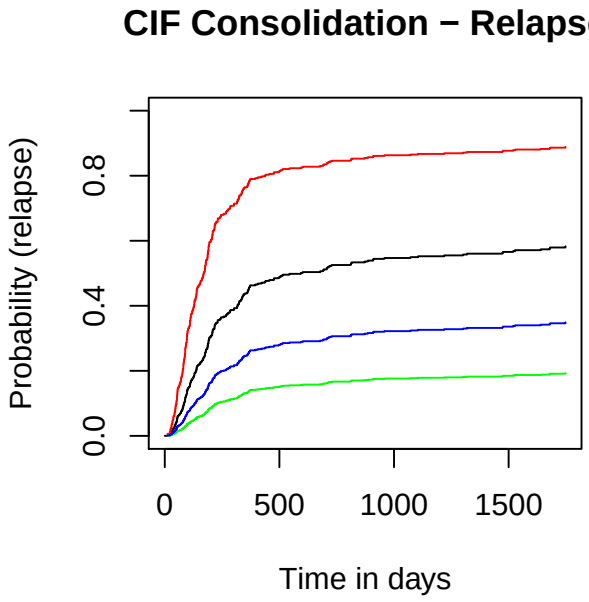


Figure 8.2.: Cumulative incidence function for different covariates of the final model for transition consolidation to relapse.

	Black	Red	Green	Blue
cyt6_6	0	1	0	0
wbc	8	8	8	8
Allo	0	0	0	1
npm1	0	0	0	0
cyt6_1	0	0	1	0
age	50	50	50	50
Treatment_arm	0	0	0	0
abn11q23	0	0	0	0

Table 8.6.: Legenda of Figure 8.2

The effect of some of the variables are shown in Figure 8.2 by the resulting cumulative incidence functions for four different covariate combinations presented in Table 8.6. The black line corresponds to a patient who has none of the chromosomal abnormalities or mutations in the model, has median white blood cell count and age, and is in the control group. We use this graph as a baseline for reference for other covariate profiles. The red graph has the same profile but has a monosomal karyotype, which is the covariate with the highest coefficient in the final model (0.939). The red graph lies above the black graph; the red graph is 0.89 after 5 years whereas the black graph is 0.58. The patient represented by the blue graph is similar to that of the black graph but has had an allogeneic SCT instead of another consolidation treatment (corresponding coefficient of -0.689). The blue graph lies below the black graph and is 0.35 after five years. The green graph represents a patient similar to that of the black graph but has chromosomal translocation t(8;23) (cyt6_1 in the model) which has the lowest corresponding coefficient (-1.206). The green graph lies below the blue and black line and is 0.19 after five years.

Table 8.7 summarizes the results of adding an MRD or LSC variable to our final model. We see again that the imputed variables have a larger log-likelihood test statistic but are far from significant at the 0.05 level. The binary variants (cutoff values: MRD 0.1 procent and LSC 0)) of the MRD and LSC variables outperform the continuous variants based on the-log likelihood, which is the opposite of what we saw for the model of the transition from CR and induction only to consolidation in Table 8.4. Furthermore, all non-imputed MRD and LSC variables are significant at the 0.05 level.

As the non-imputed, binary MRD variable and LSC variable to our final model are significant at the 0.05 level of significance and have minimal category sizes that are greater than 5%, we

Table 8.7.: Summary of MRD/LSC related variables as variable 9 in the model for transition consolidation to relapse

Variable	LL test	p-value adj	p-value	Coef	exp(Coef)	Minimal Cat.
LSCimp_0	92.628	0.180	0.087	0.236	1.266000e+00	264 / 606
LSCimp_cont	91.429	0.660	0.473	0.542	1.719000e+00	1 / 606
MRDimp_.1	90.419	0.664	0.561	0.084	1.088000e+00	229 / 606
MRDimp_cont	90.120	0.455	0.353	0.012	1.012000e+00	1 / 606
MRD_.1	55.282	0.030	0.030	0.527	1.694000e+00	59 / 280
MRD_cont	51.907	0.006	0.006	0.022	1.023000e+00	1 / 280
LSC_0	49.125	0.008	0.008	0.523	1.688000e+00	92 / 279
LSC_cont	45.494	0.000	0.000	45.611	6.434061e+19	1 / 279

[a] Additional variables in the models: cyt6_6, wbc, Allo, npm1, age, cyt6_1, Treatment_arm, abn11q23

[b] LL score of final model on the MRD subset: 50.555904, on the LSC subset: 42.706825

investigated the added value of these variables to the prediction model in more detail. The summary of the results of including them separately to the model are given in Table A.17 and Table A.18, and including both in Table A.19. Note that the variables npm1, cyt6_1, and abn11q23 are no longer significant at the 0.05 level when LSC_0 is added to the model, whereas all variables remain significant when MRD_.0 is added. The coefficients of LSC (0.523) and MRD (0.527) are similar and positive. When both are added to the model, they remainl significant at the 0.05 level and the coefficients of both are slightly higher (0.623 for MRD and 0.740 for LSC).

In Figure 8.3 the cumulative incidence functions of our new models with LSC and MRD individually and together are presented for different values of LSC and MRD and all other variables equal to that of the black graph in Figure 8.2 presented in Table 8.6. In the left figure, the CIF at 5 years for the LSC- case (black graph) is 0.52 and 0.70 for the LSC+ case (red graph). Similarly, in the middle figure, the CIF at five years for the MRD- case (solid graph) is 0.54 and for the MRD+ case (dotted graph), it is 0.73. In the right figure, the CIF at five years for the model with both MRD and LSC is 0.46 in the case of both MRD- and LSC- (black solid graph); in case of LSC- and MRD+ (red solid graph), it is 0.69; in case of LSC+ and MRD- (black dotted graph), it is 0.73; and in case both MRD+ and LSC+ (red dotted graph), it is 0.91.

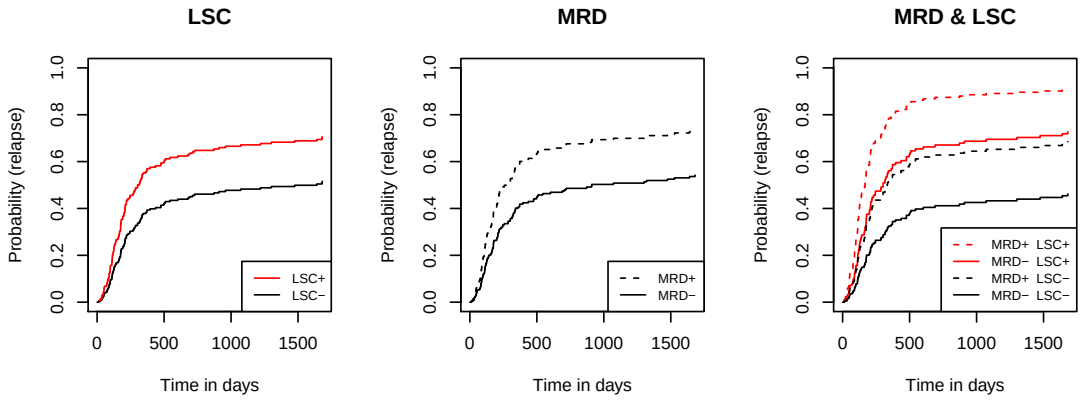


Figure 8.3.: Cumulative incidence functions for final model of transition Consolidation to Treatment with additional variable LSC (left), MRD (middle) and with both variables (right). The MRD and LSC variables are included as binary variables in the models and the different graphs in the figures represent different values for the LSC and MRD variables with all other variables remaining the same (with the same values as the black graph in Figure 8.2 presented in Table 8.6).

8.4. Models combined: two step probability

The two models can be combined to obtain the two step probability

$$P_{0,C,R}(t|Z) = P(X(0) = 0, X(r) = \text{consolidation for some r between 0 and t}, X(t) = \text{relapsed}|Z), \tag{8.1}$$

of first transitioning to the consolidation treatment and then to relapse for a given covariate profile Z. The covariates in both final models are included in Z.

Figure 8.4 illustrates cumulative incidence functions of the two models and their combination in the two-step probability for different covariate combinations. The left figure shows the cumulative incidence function for the transition from CR and induction only to consolidation; the red graph represents a patient who has mutation asxl1 (which has coefficient -0.456 in the first model) but not mutation npm1 (which has coefficient 0.253 in the first model) and the black graph is the other way around. The middle figure shows the cumulative incidence function of the transition from consolidation to relapse where the striped graph represents a patient who has a monosomal karyotype (called cyt6_6 in the tables and has coefficient 0.920 in the second model), which is not the case of the solid graph. Note that the x-axes are different for the left and middle figures.

The right figure in Figure 8.4 combines the two cumulative incidence function into the two-step probability. The black graphs are very similar to the graphs in the middle figure, except that they are constant at zero at the start. The red graphs are similar to this but seem to be a factor smaller.

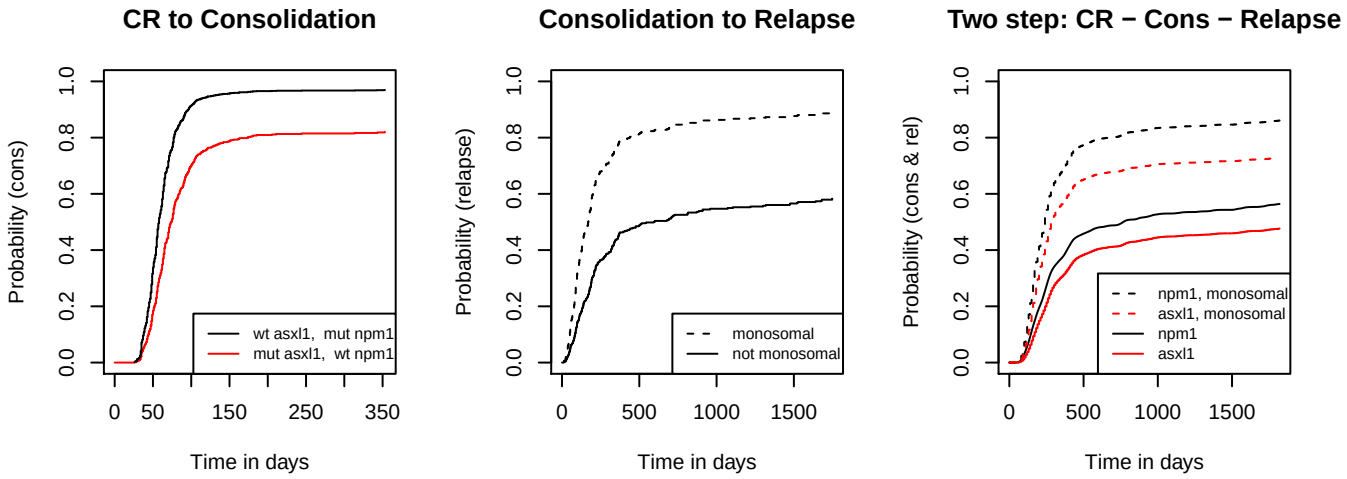


Figure 8.4.: Illustration of the two cumulative incidence functions combined to the two step probability from CR to consolidation to relapse for four different covariate combinations. The patients represented by the graphs were all in the control group (so no Clofarabine), have a mean age of 50, have a median bone marrow blast percentage of 54, do not have mutation flt3itd, a white bloodcell count of 8, got either a autologous SCT or third chemo as consolidation, do not have mutation npm1 or chromosomal abnormalities t(8;21) (named cyt6_1 in the tables) or abn11q23. The red graph represents a patient who has mutation asxl1 but not mutation npm1. This is the other way around for the black graph. The striped graph represents a patient who has a monosomal karyotype (called cyt6_6 in the tables and has coefficient 0.920 in the second model) which is not the case of the solid graph.

9. Discussion

In this section, we first discuss the results of the modelling and its implications. Then, we discuss the limitations of the model and prospects for further research.

9.1. Results

In this section, we discuss the results of the step-up variable selection of the models of transition from CR and induction only to consolidation and from consolidation to relapse.

9.1.1. Transition CR and induction only to consolidation

First, the treatment arm, a variable that is 1 if Clofarabine[1] was given and 0 otherwise, was the first variable we added to the Fine-Gray model of the transition from CR and induction only to consolidation. The coefficient of the variable in the final model is -0.528 with $\exp(-0.528) = 0.590$. The value does not have a medical interpretation but means that the subdistribution hazard is reduced by a factor 0.590.

The coefficient of the treatment arm is negative, which means that patients who were treated with clofarabine had a reduced probability of transitioning to their consolidation treatment compared to patients with the same variable values but who did not have clofarabine. Based on the transition numbers in Figure 9.1, relatively fewer patients who had clofarabine reach consolidation because relatively more patients suffered from toxicity (13% vs. 3% in control group). In research conducted on the effect of clofarabine, the authors state "The latter results are indicative of a greater antileukemic effect of the clofarabine schedule and of a concurrent greater toxicity profile, resulting in an enhanced death rate in CR" (**?**). This is in accordance with what we observe in the model: we have included complete remission patients only, so the coefficient of the treatment arm does not capture the antileukemic effect, which would be refleted in more patients reaching complete remission, but only the toxicity effect of the treatment (as this is modelled as a competing risk).

Second, the patient's age was added as a continuous variable to the model. The coefficient of the variable in the final model is -0.015 with $\exp(-0.015) = 0.985$ but this value needs to be interpreted cautiously as the variable is continuous. The patients are between age 18 and 66 and have a mean age of 50. Suppose we have three similar patients of different ages with covariate vectors Z_1, Z_2 and Z_3, which are identical except $Z_{1age} = 18$, $Z_{2age} = 50$ and $Z_{3age} = 66$. The relative estimated sub-distribution hazard ratio of the youngest patient compared to the average aged patient is $\frac{\lambda_{0Treatment}(t|Z_1)}{\lambda_{0Treatment}(t|Z_2)} = \exp(\beta Z_1 - \beta Z_2) = \exp(-0.015 * (18 - 50)) = 1.616$, meaning the subdistribution hazard of the youngest patient is a factor of 1.616 larger than that of the average-aged patient and

[1]The clinical study from which the dataset originates was aimed to evaluate the effect of the chemotherapeutic medicine Clofarabine.

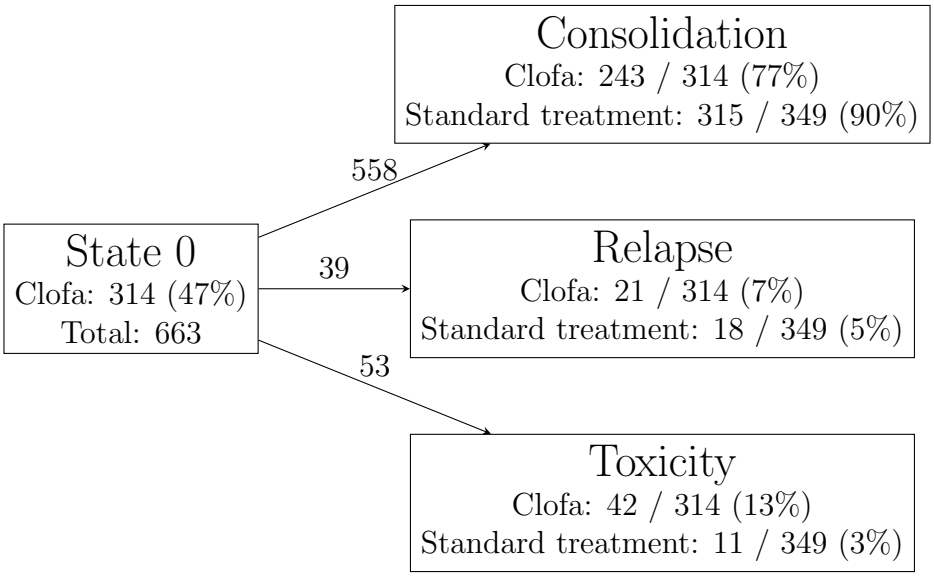


Figure 9.1.: Transition numbers from state 0 stratified on the treatment arm (Clofarabine versus standard treatment).

thus the youngest patient has a higher probability of reaching the consolidation. In Figure 8.1, the effect of age on the cumulative incidence function is shown for an average aged patient (black) compared to the youngest (blue). Similarly, the relative sub-distribution hazard ratio of the oldest patient compared to the average aged patient is $\frac{\lambda_{0Treatment}(t|Z_3)}{\lambda_{0Treatment}(t|Z_2)} = \exp(-0.015 * (66 - 50)) = 0.787$, meaning the oldest patient has a decreased probability of reaching consolidation treatment.

Older age is associated with higher treatment mortality as they have a lower tolerance for chemotherapy, with most deaths occurring within a month after chemo (Döhner et al., 2017) (Walter et al., 2011). From the data we see that less that 30% of the patients are over 60 years of age whereas this is 58% of patients suffering from toxicity before reaching consolidation. In current treatment protocols, age is taken into account to determine if the patient is fit for chemotherapy and to determine a suitable dose. Therefore, an explanation of our results could be that older patients are less likely to reach consolidation – the median number of days between induction and consolidation is 63 and the minimum is 26 – because of an increased risk of chemotherapy-related mortality.

The other variables are more difficult to compare to previous research because research is usually aimed at the association with the survival outcome instead of reaching consolidation. The variable axsl1 has a negative coefficient, indicating that it reduces the probability of reaching consolidation treatment, which is in line with previous research showing poor survival probabilities (Döhner et al., 2017).

9.1.2. Transition Consolidation to Relapse

The variable cyt6_6, which encodes whether a patient has a monosomal karyotype, was the first to be added to the model and has the highest corresponding coefficient of 0.939. The increasing effect on the hazard is in line with previous research showing that patients with a monosomal karyotype have poor survival probabilities (a four-year survival of 9%) (Kayser et al., 2012). Furthermore, the

risk classification of a patient with a monosomal karyotype is adverse because of its "independent association with adverse risk" (Döhner et al., 2017).

The variable Allo, which is 1 if the patient received an allogeneic SCT instead of a third round of chemotherapy or an autologous SCT, has a negative corresponding coefficient. This is in line with the intended result. Allogeneic SCT is given to patients who have the worst risk of relapse because this consolidation treatment most reduces the relapse hazard. This means that a selection bias exists (patients with high relapse risk receive allogeneic SCT), which results in a positive coefficient, and a treatment effect exists, which results in a negative coefficient. The model can capture the effect of the treatment, which is good, but the magnitude of the selection bias remains unclear.

The treatment arm (whether a patient received clofarabine or was in the control group) was also included in this model and has a negative corresponding coefficient, indicating a protective effect. We previously noted that clofarabine is associated with higher toxicity rates. However, clofarabine has also shown a greater antileukemic effect compared to standard treatment. Thus, patients who have reached consolidation treatment recovered from induction treatment with clofarabine and thus have not experienced severe toxicity. These patients benefit from the antileukemic effect, which explains the opposite sign of the variable's coefficient in this model compared to the previous model (**?**).

The variables cyt6_1 (encoding translocation t(8;21), a chromosomal abnormality) and npm1 have a negative corresponding coefficient. Both are included in the favorable risk category of the ELN risk classification. The chromosomal abnormality abnq23 has a positive coefficient and is also included in the ELN risk classification in the adverse risk category.

The variable age now has a positive coefficient, so it is again an adverse prognostic factor for the patient. However, we identify age as a prognostic factor, especially for the competing risk relapse. An explanation is that age is associated with specific genetic abnormalities that increase the likelihood of chemotherapy resistance. Older patients might not reach a deep state of complete remission and are more likely to relapse (Döhner et al., 2017).

Last, we found the binary MRD and LSC variables to be significant, both separately and combined. The corresponding coefficients were positive in all models. This is in line with previous research showing the predictive power of both variables (Zeijlemaker and Schuurhuis, 2013) (Jongen-Lavrencic et al., 2018). Some variables in the final model were no longer significant after the LSC variable was added to the model. The most likely reason for the npm1 variable is that the method to measure LSC depends on the cell marker CD38-, which is not present in the case of a mutated npm1 gene. The other variables abn11q23 and cyt6_1 are most likely no longer significant because the minimal category is too small on the reduced datase.

9.2. Limitations

The Fine-Gray model assumes that the effect of a covariate on the subdistribution hazard is constant over time (the proportionality assumption), but it has yet to be tested regarding whether this is the case for the variables in our model. It could be unreasonable to assume the effect of consolidation type is constant over time. Perhaps the effect of autologous SCT and a third round of chemotherapy treatment is mostly directly after the treatment and reduced afterwards, whereas this is not the case for the allogeneic SCT.

The second possible limitation of the model is that we have not incorporated interaction ef-

fects of the variables in the model. The subdistribution hazard is of the form $\lambda_h(t|x_1, ..., x_p) = \lambda_{h0}(t) \exp(\beta_1 x_1 + ... + \beta_p x_p)$; thus, the combined effect of $x_i = 1$ and $x_j = 1$ on the hazard is simply $\exp(\beta_i + \beta_j)$. However, previous research has demonstrated that an interaction effect can take place, which would make this assumption unrealistic. This could be dealt with by including the interaction term $\beta_{ij} x_i x_j$. However, for many combinations of variables we only have few patients who have both, which makes the interaction coefficient β_{ij} difficult to estimate.

The third main limitation is the possible distorting effect of the imputation on the final model. First, the significance of the variables in the model also depends on the percentage of missing data for that variable and the quality of the imputation. Consequently, which variables are chosen by the step-up method are partly determined by the quality of the data and the imputation of the variable. Second, the imputation can only recreate associations if these are allowed in the imputation model. Thus, misspecification can distort the distribution of the data. We assumed a different imputation model (competing risks, stratified on whether patient received an allogeneic SCT) and a substantive model (general multistate model with a separate state for consolidation treatment, not stratified). This could have led to a significant difference between allogeneic and non-allogeneic consolidation treatment but an insignificant difference between a third chemotherapy round and autologous SCT.

9.3. Future perspectives

The proposed model is of valuable practical use to clinicians and patients if the relapse prediction is superior to the current ELN 2017 model. The prediction accuracy should be established with a separate dataset. A new clinical study, the HOVON 132, has recently been completed and the corresponding dataset will be available shortly. The model will be validated using this dataset as a test set. (Zhang et al., 2018) present methods for validating Fine-Gray models for a fixed point in time (this would, in our case, be five years after induction therapy) or as a function of time.

Furthermore, the current model should be optimized, for example by including interaction terms for the purpose of correcting for the additivity of two variables in the model or allowing for variables that are stronger in combination than alone. One example of the latter is age, which is a better predictor in combination with other variables (Döhner et al., 2017). Furthermore, optimization should determine whether a variable should be included as continuous or discrete variable; age could be included as a categorical variable instead of a continuous variable and the binary LSC and MRD variables might be more predictive for a different cut off variable.

The model could also be improved by evaluating modelling decisions on a validation set and tuning the model based on the results. First, the effect of the different imputation methods (White and Royston versus Bartlett) should be determined. Second, the inclusion of the binary MRD and LSC variables should be verified. Third, the effect of the chosen states in the multistate model (as opposed to the other possibilities discussed in Section 7.2.1) should be evaluated.

Finally, we present some possible improvements in the imputation model. Because of time limitations, we were not able to adjust the imputation model based on the substantive model with the extra state for consolidation therapy. The data could be imputed according to a multistate model, meaning each transition in the multistate model is a separate line. The dataset could then be split based on the starting state, and each dataset can be imputed separately. This may improve the imputation. However, the interpretation of the imputation is more difficult because a patient could have differently imputed values at different transitions even though, in reality, the variables remain

the same throughout. Even if we did this, no imputation model exists yet that is compatible with the Fine-Gray model. Perhaps solutions to this problem will be developed in the future.

10. Conclusion

The primary goal of this thesis was to model the relapse of patients with leukemia based on the patient characteristics and leukemia-specific parameters to better able to predict relapse on an individual level. Improved models support clinicians and patients in their decisions regarding the type of consolidation treatment (after finishing induction therapy). Furthermore, we aimed to identify the most predictive variables and analyze the predictive value of MRD and LSC.

To achieve this, we cleaned the data and implemented two multiple imputation algorithms to maximize our data utility. Next, we designed a multistate model with competing risks and an intermediate consolidation state to handle the time varying property of the treatment. Finally, we used our imputed datasets to identify the most predictive variables using a step-up approach and trained two Fine-Gray models with these variables. These results were then combined into the two-step probability.

This final two-step model can be used to estimate the probability of relapse before time t after consolidation treatment as a function of time, conditional on the patient's characteristics. One of the characteristics is the type of consolidation therapy; therefore, filling in both allows us to compare the relapse risks of the two treatments.

Based on the data alone, we pinpointed variables known to be important based on previous medical research and combined these in relevant models that predict relapse in individual patients. Moreover, we successfully imputed the dataset and thus were able to use data from all patients. Imputed variables were still found to be significant in the final model, and more importantly, were clinically relevant and in accordance with the previous research. The imputation methods enabled us to explore other incomplete clinical datasets, which were previously of less use due to missing data.

Besides identifying variables with known prognostic value, we demonstrated the added value of MRD and LSC for relapse prediction. We are the first to successfully combine these variables in an individual relapse prediction model. With these variables being relatively new in the diagnosis of patients with AML, we contributed to the establishment of added value for risk stratification of MRD and LSC. We anticipate that these variables can be included in novel risk stratification models in the future.

For the model to be of practical use for clinicians and patients, the model requires independent validation. Using a validation method, the effect of the modeling decisions and imputation method can be determined and optimized. This research can serve as the foundation for building a relapse prediction model to support treatment decisions. Ultimately, this will lead to improved patient outcomes with fewer relapses, decreased treatment-related side effects, and an improved quality of life.

11. Acknowledgements

I am very grateful for my supervisors David Cucchi, Dennis Dobler, and John Jacobs, who has been there at every step and whom without this project would not be possible. I would like to thank my supervisor at ORTEC John Jacobs for his infectious enthusiasm and guidance (both with the project and careerwise). I am grateful that he introduced me to the project and all the people who have helped me along the way. I would like to thank David Cucchi for helping me build the bridge between acute myeloid leukemia and mathematics, challenging me and supporting me. I want to thank Dennis Dobler for introducing me to survival analysis and helping me mathematically (and for the Eureka moments).

Next, I would like to thank Jacqueline Cloos for allowing me to do this project and introducing me to acute myeloid leukemia. I want to thank everyone from the weekly AML meeting: Lok Lam Ngai, Costa Bachas, and Jesse Klaver. Furthermore, I would like to thank Jeroen Janssen and Gert Ossenkoppele for their insights and helping with the interpretation of the data and the results. I also want to thank ORTEC and Menno Brandjes for this wonderful opportunity and the healthcare development team for letting me in their team and the many games at the virtual watercooler.

I want to thank everyone else who has helped me along the way, such as Jonathan Bartlett, Jan Beyersmann, Ed Bonneville, and Ronald Buitenhek. Finally, I thank Boele Camps, my brother George Vegelien and my mother Monique Vegelien for their support and encouragement.

A. Results of step-up process

Table A.1.: Top six variables and top four MRD/LSC related variables as variable 1 in the model for transition CR to consolidation

Variable	LL test	p-value adj	p-value	Coef	exp(Coef)	Minimal Cat.
Treatment_arm	36.706	0.000	0.000	-0.518	0.596	314 / 663
age	29.956	0.000	0.000	-0.019	0.981	1 / 663
Clofa&intRisk	16.046	0.000	0.000	-0.486	0.615	92 / 663
risk102_1	14.908	0.000	0.000	0.365	1.440	194 / 663
risk102_2	13.924	0.000	0.000	-0.361	0.697	170 / 663
asxl1	11.027	0.003	0.001	-0.454	0.635	75 / 663
LSC_cont	8.161	0.016	0.016	-7.453	0.001	1 / 290
MRD_cont	3.029	0.325	0.325	-0.029	0.971	1 / 286
LSC_0	1.156	0.283	0.283	-0.143	0.867	96 / 290
MRD_.1	0.172	0.699	0.699	-0.064	0.938	61 / 286

[a] A Fine Gray model is trained with a single variable.

[b] LL score of new model on the MRD subset: 4.598352, on the LSC subset: 7.612600

Table A.2.: Top six variables and top four MRD/LSC related variables as variable 2 in the model for transition CR to consolidation

Variable	LL test	p-value adj	p-value	Coef	exp(Coef)	Minimal Cat.
age	61.612	0.000	0.000	-0.017	0.983	1 / 663
risk102_1	50.149	0.000	0.000	0.346	1.414	194 / 663
asxl1	49.077	0.003	0.001	-0.471	0.624	75 / 663
risk102_2	48.720	0.001	0.001	-0.336	0.715	170 / 663
bmblast	47.634	0.001	0.001	0.005	1.005	1 / 663
nras	45.787	0.016	0.005	0.313	1.368	140 / 663
LSC_cont	17.473	0.023	0.023	-8.470	0.000	1 / 290
LSC_0	10.454	0.098	0.098	-0.228	0.796	96 / 290
MRD_cont	8.874	0.299	0.299	-0.036	0.965	1 / 286
MRD_.1	4.996	0.563	0.563	-0.098	0.907	61 / 286

[a] Additional variables in the models: Treatment_arm

[b] LL score of new model on the MRD subset: 9.671593, on the LSC subset: 16.516137

Table A.3.: Top six variables and top four MRD/LSC related variables as variable 3 in the model for transition CR to consolidation

Variable	LL test	p-value adj	p-value	Coef	exp(Coef)	Minimal Cat.
risk102_1	73.890	0.001	0.001	0.331	1.392	194 / 663
asxl1	72.417	0.006	0.002	-0.446	0.640	75 / 663
risk102_2	71.113	0.002	0.002	-0.301	0.740	170 / 663
nras	69.108	0.022	0.009	0.285	1.330	140 / 663
bmblast	68.246	0.013	0.011	0.004	1.004	1 / 663
npm1	68.193	0.012	0.011	0.239	1.270	214 / 663
LSC_cont	25.249	0.015	0.015	-8.318	0.000	1 / 290
LSC_0	18.493	0.177	0.177	-0.192	0.825	96 / 290
MRD_cont	13.616	0.297	0.297	-0.034	0.967	1 / 286
MRD_.1	9.861	0.688	0.688	-0.068	0.934	61 / 286

[a] Additional variables in the models: Treatment_arm, age

[b] LL score of new model on the MRD subset: 11.883900, on the LSC subset: 23.855882

Table A.4.: Top six variables and top four MRD/LSC related variables as variable 4 in the model for transition CR to consolidation

Variable	LL test	p-value adj	p-value	Coef	exp(Coef)	Minimal Cat.
risk102_1	81.941	0.003	0.003	0.296	1.345	194 / 663
risk102_2	80.940	0.004	0.004	-0.285	0.752	170 / 663
nras	79.423	0.027	0.014	0.274	1.315	140 / 663
bmblast	77.453	0.028	0.026	0.004	1.004	1 / 663
idh1	77.435	0.051	0.030	0.296	1.344	77 / 663
npm1	76.204	0.066	0.060	0.180	1.198	214 / 663
LSC_cont	32.385	0.012	0.011	-8.382	0.000	1 / 290
LSC_0	24.186	0.280	0.275	-0.153	0.858	96 / 290
MRD_cont	16.549	0.298	0.297	-0.034	0.967	1 / 286
MRD_.1	12.827	0.675	0.674	-0.073	0.930	61 / 286

[a] Additional variables in the models: Treatment_arm, age, asxl1

[b] LL score of new model on the MRD subset: 13.001255, on the LSC subset: 24.465530

Table A.5.: Top six variables and top four MRD/LSC related variables as variable 5 in the model for transition CR to consolidation

Variable	LL test	p-value adj	p-value	Coef	exp(Coef)	Minimal Cat.
risk102_1	87.393	0.009	0.008	0.268	1.308	194 / 663
risk102_2	87.029	0.006	0.006	-0.277	0.758	170 / 663
bmblast	85.009	0.027	0.025	0.003	1.003	1 / 663
idh1	84.748	0.056	0.031	0.295	1.342	77 / 663
flt3itd	83.664	0.093	0.082	-0.178	0.837	146 / 663
abn17_p	82.812	0.040	0.030	-0.455	0.635	22 / 663
LSC_cont	33.620	0.011	0.010	-8.221	0.000	1 / 290
LSC_0	25.387	0.353	0.348	-0.131	0.877	96 / 290
MRD_cont	18.118	0.307	0.306	-0.033	0.967	1 / 286
MRD_.1	14.460	0.694	0.694	-0.068	0.934	61 / 286

[a] Additional variables in the models: Treatment_arm, age, asxl1, nras

[b] LL score of new model on the MRD subset: 15.174108, on the LSC subset: 27.452654

Table A.6.: Top six variables and top four MRD/LSC related variables as variable 6 in the model for transition CR to consolidation

Variable	LL test	p-value adj	p-value	Coef	exp(Coef)	Minimal Cat.
risk102_2	92.979	0.005	0.004	-0.288	0.750	170 / 663
risk102_1	92.673	0.011	0.010	0.262	1.300	194 / 663
flt3itd	92.111	0.014	0.011	-0.274	0.760	146 / 663
npm1&flt3	91.235	0.019	0.017	-0.322	0.725	83 / 663
idh1	89.376	0.075	0.048	0.272	1.312	77 / 663
abn07_min	88.336	0.064	0.050	0.306	1.358	54 / 663
LSC_cont	36.100	0.011	0.011	-8.068	0.000	1 / 290
LSC_0	27.682	0.382	0.378	-0.125	0.883	96 / 290
MRD_cont	19.970	0.293	0.292	-0.032	0.968	1 / 286
MRD_.1	16.257	0.665	0.664	-0.075	0.927	61 / 286

[a] Additional variables in the models: Treatment_arm, age, asxl1, nras, bmblast

[b] LL score of new model on the MRD subset: 24.873125, on the LSC subset: 37.811238

Table A.7.: Top six variables and top five MRD/LSC related variables as variable 7 in the model for transition CR to consolidation

Variable	LL test	p-value adj	p-value	Coef	exp(Coef)	Minimal Cat.
risk102_2	95.962	0.053	0.050	-0.221	0.802	170 / 663
risk102_1	95.690	0.067	0.064	0.202	1.224	194 / 663
npm1	95.526	0.079	0.071	0.194	1.215	214 / 663
idh1	95.276	0.098	0.065	0.259	1.295	77 / 663
abn17_p	94.766	0.043	0.033	-0.459	0.632	22 / 663
abn11q23	93.964	0.078	0.075	-0.277	0.758	37 / 663
LSC_cont	45.825	0.016	0.015	-7.601	0.001	1 / 290
LSC_0	39.083	0.430	0.425	-0.115	0.891	96 / 290
MRD_cont	29.566	0.344	0.343	-0.033	0.968	1 / 286
MRD_.1	26.093	0.812	0.811	-0.042	0.959	61 / 286

[a] Additional variables in the models: Treatment_arm, age, asxl1, nras, bmblast, flt3itd

[b] LL score of new model on the MRD subset: 29.072927, on the LSC subset: 40.157914

Table A.8.: Top six variables and unimputed MRD/LSC related variables as variable 1 in the model for transition consolidation to relapse

Variable	LL test	p-value adj	p-value	Coef	exp(Coef)	Minimal Cat.
cyt6_6	24.527	0.000	0.000	1.033	2.808000e+00	59 / 606
risk102_4	21.270	0.000	0.000	0.780	2.182000e+00	106 / 606
risk102_1	17.218	0.000	0.000	-0.658	5.180000e-01	177 / 606
abn05_min	17.218	0.000	0.000	1.368	3.927000e+00	18 / 606
abn07_min	17.177	0.000	0.000	0.950	2.585000e+00	50 / 606
complex	15.357	0.000	0.000	0.766	2.150000e+00	71 / 606
LSC_0	10.601	0.001	0.001	0.655	1.925000e+00	92 / 279
MRD_.1	3.477	0.056	0.056	0.432	1.540000e+00	59 / 280
MRD_cont	2.497	0.000	0.000	0.032	1.033000e+00	1 / 280
LSC_cont	1.711	0.000	0.000	32.304	1.070604e+14	1 / 279

[a] A Fine Gray model is trained with a single variable.

[b] LL score of new model on the MRD subset: 14.994160, on the LSC subset: 16.510625

Table A.9.: Top six variables and unimputed MRD/LSC related variables as variable 2 in the model for transition consolidation to relapse

Variable	LL test	p-value adj	p-value	Coef	exp(Coef)	Minimal Cat.
wbc	37.638	0.000	0.000	0.005	1.005000e+00	1 / 606
risk102_1	34.747	0.002	0.002	-0.530	5.890000e-01	177 / 606
Allo	33.438	0.003	0.003	-0.427	6.520000e-01	277 / 606
AutoORcIII	33.438	0.003	0.003	0.427	1.533000e+00	277 / 606
Treatment_arm	32.069	0.007	0.007	-0.386	6.800000e-01	260 / 606
risk102_3	30.956	0.009	0.009	0.397	1.488000e+00	183 / 606
LSC_0	25.668	0.002	0.002	0.610	1.841000e+00	92 / 279
LSC_cont	18.484	0.000	0.000	35.312	2.167145e+15	1 / 279
MRD_.1	17.459	0.132	0.132	0.363	1.438000e+00	59 / 280
MRD_cont	15.560	0.054	0.054	0.013	1.014000e+00	1 / 280

[a] Additional variables in the models: cyt6_6

[b] LL score of new model on the MRD subset: 22.464827, on the LSC subset: 21.231777

Table A.10.: Top six variables and unimputed MRD/LSC related variables as variable 3 in the model for transition consolidation to relapse

Variable	LL test	p-value adj	p-value	Coef	exp(Coef)	Minimal Cat.
risk102_1	48.158	0.001	0.001	-0.538	5.840000e-01	177 / 606
Allo	47.041	0.002	0.002	-0.439	6.440000e-01	277 / 606
AutoORcIII	47.041	0.002	0.002	0.439	1.552000e+00	277 / 606
Treatment_arm	44.884	0.008	0.008	-0.379	6.850000e-01	260 / 606
abn11q23	43.708	0.004	0.004	0.610	1.840000e+00	40 / 606
III	43.530	0.014	0.014	0.347	1.415000e+00	260 / 606
LSC_0	30.401	0.002	0.002	0.611	1.842000e+00	92 / 279
MRD_.1	25.295	0.100	0.100	0.392	1.480000e+00	59 / 280
LSC_cont	23.498	0.000	0.000	38.523	5.376914e+16	1 / 279
MRD_cont	23.061	0.050	0.050	0.014	1.014000e+00	1 / 280

[a] Additional variables in the models: cyt6_6, wbc

[b] LL score of new model on the MRD subset: 28.332720, on the LSC subset: 25.221228

Table A.11.: Top six variables and unimputed MRD/LSC related variables as variable 4 in the model for transition consolidation to relapse

Variable	LL test	p-value adj	p-value	Coef	exp(Coef)	Minimal Cat.
risk102_1	76.650	0.000	0.000	-0.992	3.710000e-01	177 / 606
risk102_3	58.627	0.001	0.001	0.575	1.777000e+00	183 / 606
npm1	56.183	0.003	0.002	-0.502	6.050000e-01	202 / 606
abn11q23	54.135	0.003	0.003	0.670	1.954000e+00	40 / 606
Treatment_arm	52.914	0.017	0.017	-0.343	7.100000e-01	260 / 606
cyt6_5	52.475	0.016	0.016	0.410	1.507000e+00	168 / 606
LSC_0	35.223	0.001	0.001	0.641	1.898000e+00	92 / 279
MRD_.1	32.105	0.052	0.052	0.457	1.579000e+00	59 / 280
MRD_cont	29.089	0.028	0.028	0.016	1.016000e+00	1 / 280
LSC_cont	28.046	0.000	0.000	44.193	1.558288e+19	1 / 279

[a] Additional variables in the models: cyt6_6, wbc, Allo

[b] LL score of new model on the MRD subset: 29.427684, on the LSC subset: 25.567888

Table A.12.: Top six variables and unimputed MRD/LSC related variables as variable 5 in the model for transition consolidation to relapse

Variable	LL test	p-value adj	p-value	Coef	exp(Coef)	Minimal Cat.
risk102_1	80.175	0.000	0.000	-0.914	4.010000e-01	177 / 606
cyt6_1	66.431	0.014	0.011	-1.160	3.140000e-01	34 / 606
risk102_3	65.276	0.002	0.002	0.508	1.663000e+00	183 / 606
age	63.806	0.012	0.012	0.017	1.017000e+00	1 / 606
Treatment_arm	63.345	0.009	0.009	-0.377	6.860000e-01	260 / 606
abn3q26_5	62.551	0.061	0.058	-1.381	2.510000e-01	11 / 606
LSC_0	35.640	0.001	0.001	0.643	1.903000e+00	92 / 279
MRD_.1	33.042	0.058	0.058	0.446	1.563000e+00	59 / 280
MRD_cont	30.199	0.027	0.027	0.016	1.016000e+00	1 / 280
LSC_cont	28.325	0.000	0.000	43.567	8.338232e+18	1 / 279

[a] Additional variables in the models: cyt6_6, wbc, Allo, npm1

[b] LL score of new model on the MRD subset: 35.147025, on the LSC subset: 29.732573

Table A.13.: Top six variables and unimputed MRD/LSC related variables as variable 6 in the model for transition consolidation to relapse

Variable	LL test	p-value adj	p-value	Coef	exp(Coef)	Minimal Cat.
risk102_1	83.323	0.000	0.000	-0.808	4.460000e-01	177 / 606
age	74.581	0.008	0.008	0.017	1.017000e+00	1 / 606
Treatment_arm	74.436	0.005	0.005	-0.405	6.670000e-01	260 / 606
abn3q26_5	72.940	0.055	0.053	-1.429	2.400000e-01	11 / 606
risk102_3	72.707	0.009	0.009	0.424	1.528000e+00	183 / 606
cyt6_2	72.672	0.026	0.026	-0.864	4.210000e-01	25 / 606
LSC_0	39.193	0.001	0.001	0.625	1.869000e+00	92 / 279
MRD_.1	38.902	0.051	0.051	0.457	1.579000e+00	59 / 280
MRD_cont	35.948	0.023	0.023	0.016	1.017000e+00	1 / 280
LSC_cont	32.302	0.000	0.000	41.900	1.573585e+18	1 / 279

[a] Additional variables in the models: cyt6_6, wbc, Allo, npm1, cyt6_1

[b] LL score of new model on the MRD subset: 40.690269, on the LSC subset: 34.060398

Table A.14.: Top six variables and unimputed MRD/LSC related variables as variable 7 in the model for transition consolidation to relapse

Variable	LL test	p-value adj	p-value	Coef	exp(Coef)	Minimal Cat.
risk102_1	89.956	0.000	0.000	-0.777	4.600000e-01	177 / 606
Treatment_arm	82.536	0.005	0.005	-0.406	6.670000e-01	260 / 606
risk102_3	81.323	0.007	0.007	0.441	1.554000e+00	183 / 606
abn11q23	80.899	0.006	0.006	0.637	1.890000e+00	40 / 606
abn3q26_5	80.832	0.052	0.050	-1.420	2.420000e-01	11 / 606
cyt6_2	79.486	0.044	0.044	-0.779	4.590000e-01	25 / 606
MRD_.1	43.912	0.073	0.073	0.425	1.529000e+00	59 / 280
LSC_0	42.750	0.002	0.002	0.601	1.825000e+00	92 / 279
MRD_cont	41.203	0.086	0.086	0.013	1.013000e+00	1 / 280
LSC_cont	37.226	0.000	0.000	48.558	1.225882e+21	1 / 279

[a] Additional variables in the models: cyt6_6, wbc, Allo, npm1, cyt6_1, age

[b] LL score of new model on the MRD subset: 45.942661, on the LSC subset: 39.563035

Table A.15.: Top six variables and unimputed MRD/LSC related variables as variable 8 in the model for transition consolidation to relapse

Variable	LL test	p-value adj	p-value	Coef	exp(Coef)	Minimal Cat.
risk102_1	99.631	0.000	0.000	-0.817	4.420000e-01	177 / 606
abn3q26_5	90.505	0.037	0.035	-1.556	2.110000e-01	11 / 606
abn11q23	89.191	0.007	0.007	0.655	1.926000e+00	40 / 606
risk102_3	88.622	0.012	0.012	0.413	1.511000e+00	183 / 606
cyt6_2	87.935	0.035	0.035	-0.811	4.440000e-01	25 / 606
npm1&flt3	87.826	0.022	0.017	0.685	1.984000e+00	74 / 606
MRD_.1	48.897	0.088	0.088	0.406	1.501000e+00	59 / 280
LSC_0	46.946	0.004	0.004	0.559	1.748000e+00	92 / 279
MRD_cont	46.820	0.020	0.020	0.018	1.018000e+00	1 / 280
LSC_cont	42.100	0.000	0.000	42.956	4.524565e+18	1 / 279

[a] Additional variables in the models: cyt6_6, wbc, Allo, npm1, age, cyt6_1, Treatment_arm

[b] LL score of new model on the MRD subset: 50.555904, on the LSC subset: 42.706825

Table A.16.: Top six variables and unimputed MRD/LSC related variables as variable 9 in the model for transition consolidation to relapse

Variable	LL test	p-value adj	p-value	Coef	exp(Coef)	Minimal Cat.
risk102_1	104.598	0.000	0.000	-0.788	4.550000e-01	177 / 606
abn3q26_5	97.472	0.039	0.038	-1.595	2.030000e-01	11 / 606
abn05_min	94.887	0.033	0.031	0.881	2.413000e+00	18 / 606
npm1&flt3	94.652	0.020	0.016	0.694	2.002000e+00	74 / 606
flt3itd	94.498	0.029	0.022	0.423	1.526000e+00	138 / 606
risk102_3	94.298	0.021	0.020	0.380	1.462000e+00	183 / 606
MRD_.1	55.282	0.030	0.030	0.527	1.694000e+00	59 / 280
MRD_cont	51.907	0.006	0.006	0.022	1.023000e+00	1 / 280
LSC_0	49.125	0.008	0.008	0.523	1.688000e+00	92 / 279
LSC_cont	45.494	0.000	0.000	45.611	6.434061e+19	1 / 279

[a] Additional variables in the models: cyt6_6, wbc, Allo, npm1, age, cyt6_1, Treatment_arm

[b] LL score of new model on the MRD subset: 50.555904, on the LSC subset: 42.706825

Table A.17.: Final model with LSC estimations of transition consolidation to relapse

	coef	exp(coef)	p-value	2.5%	97.5%
cyt6_6	1.197	3.312	0.000	1.764	6.215
wbc	0.005	1.005	0.009	1.001	1.010
Allo	-0.653	0.520	0.005	0.328	0.825
npm1	-0.428	0.652	0.100	0.388	1.093
cyt6_1	-1.063	0.345	0.081	0.104	1.142
age	0.018	1.019	0.045	1.000	1.037
Treatment_arm	-0.469	0.626	0.031	0.408	0.959
abn11q23	0.573	1.774	0.140	0.836	3.764
LSC_0	0.523	1.688	0.008	1.146	2.486

[a] p-values are adjusted for added incertaity due to imputation

[b] LL score of final model: 49.125347

Table A.18.: Final model with MRD estimations of transition ronsolidation to relapse

	coef	exp(coef)	p-value	2.5%	97.5%
cyt6_6	1.141	3.131	0.001	1.579	6.211
wbc	0.008	1.008	0.000	1.004	1.012
Allo	-0.684	0.505	0.001	0.333	0.764
npm1	-0.491	0.612	0.037	0.386	0.972
cyt6_1	-1.282	0.278	0.031	0.086	0.892
age	0.025	1.025	0.007	1.007	1.043
Treatment_arm	-0.495	0.610	0.017	0.406	0.916
abn11q23	0.964	2.622	0.006	1.318	5.217
MRD_.1	0.527	1.694	0.030	1.051	2.730

[a] p-values are adjusted for added incertaity due to imputation
[b] LL score of final model: 55.282193

Table A.19.: Final model with MRD and LSC estimations of transition consolidation to relapse

	coef	exp(coef)	p-value	2.5%	97.5%
cyt6_6	1.493	4.451	0.000	2.263	8.754
wbc	0.008	1.008	0.000	1.004	1.012
Allo	-0.748	0.473	0.003	0.291	0.771
npm1	-0.379	0.685	0.170	0.398	1.178
cyt6_1	-0.913	0.401	0.130	0.122	1.325
age	0.023	1.023	0.017	1.004	1.042
Treatment_arm	-0.503	0.605	0.030	0.383	0.954
abn11q23	0.598	1.818	0.150	0.800	4.133
MRD_.1	0.623	1.865	0.013	1.140	3.049
LSC_0	0.740	2.095	0.000	1.383	3.173

[a] p-values are adjusted for added incertaity due to imputation
[b] LL score of final model: 62.550402

Bibliography

Hovon clinical picture: Aml (acute myeloide leukemia), *http://www.hovon.nl/studies/studies-per-ziektebeeld/aml.html?action=showstudiestudie$_i$d = 72categorie$_i$d = 4*.

Aalen, O., Borgan, and G. Hakon (2011). *Survival and event history analysis: a process point of view*. Springer.

Andersen, P. K. (1993). *Statistical models based on counting processes*. Springer.

Austin, P. C., A. Latouche, and J. P. Fine (2019). A review of the use of time-varying covariates in the fine-gray subdistribution hazard competing risk regression model. *Statistics in Medicine 39*(2), 103–113.

Bartlett, J. W., S. R. Seaman, I. R. White, and J. R. Carpenter (2014). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research 24*(4), 462–487.

Bartlett, J. W. and J. M. G. Taylor (2016). Missing covariates in competing risks analysis. *Biostatistics 17*(4), 751–763.

Beyersmann, J. and M. Schumacher (2008). Time-dependent covariates in the proportional subdistribution hazards model for competing risks. *Biostatistics 9*(4), 765–776.

Buuren, S. v. (2018). *Flexible imputation of missing data*. Taylor Francis Group CRC Press.

Cortese, G. and P. K. Andersen (2009). Competing risks and time-dependent covariates. *Biometrical Journal*.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological) 34*(2), 187–202.

Donders, R., G. van der Heijden, T. Stijnen, and K. Moons (2006, 11). Review: A gentle introduction to imputation of missing values. *Journal of clinical epidemiology 59*, 1087–91.

Döhner, H., E. Estey, D. Grimwade, S. Amadori, F. R. Appelbaum, T. Büchner, H. Dombret, B. L. Ebert, P. Fenaux, R. A. Larson, R. L. Levine, F. Lo-Coco, T. Naoe, D. Niederwieser, G. J. Ossenkoppele, M. Sanz, J. Sierra, M. S. Tallman, H.-F. Tien, A. H. Wei, B. Löwenberg, and C. D. Bloomfield (2017, 01). Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood 129*(4), 424–447.

Fine, J. P. and R. J. Gray (1999a). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association 94*(446), 496–509.

Fine, J. P. and R. J. Gray (1999b). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association 94*(446), 496–509.

Gill, R. D. (1980). Nonparametric estimation based on censored observations of a markov renewal process. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 53*(1), 97–116.

Gray, R. J. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics 16*(3), 1141–1154.

Jongen-Lavrencic, M., T. Grob, D. Hanekamp, F. G. Kavelaars, A. al Hinai, A. Zeilemaker, C. A. Erpelinck-Verschueren, P. L. Gradowska, R. Meijer, J. Cloos, B. J. Biemond, C. Graux, M. van Marwijk Kooy, M. G. Manz, T. Pabst, J. R. Passweg, V. Havelange, G. J. Ossenkoppele, M. A. Sanders, G. J. Schuurhuis, B. Löwenberg, and P. J. Valk (2018). Molecular minimal residual disease in acute myeloid leukemia. *New England Journal of Medicine 378*(13), 1189–1199. PMID: 29601269.

Kayser, S., M. Zucknick, K. Döhner, J. Krauter, C.-H. Köhne, H. A. Horst, G. Held, M. V. Lilienfeld-Toal, S. Wilhelm, M. Rummel, and et al. (2012). Monosomal karyotype in adult acute myeloid leukemia: prognostic impact and outcome after different treatment strategies. *Blood 119*(2), 551–558.

Klein, J. P. and Klein (2013). *Handbook of Survival Analysis*. CRC Press.

Klein, J. P. and M. L. Moeschberger (2011). *Survival analysis: techniques for censored and truncated data*. Springer.

Lau, B. and C. Lesko (2018). Missingness in the setting of competing risks: from missing values to missing potential outcomes. *Current Epidemiology Reports 5*(2), 153–159.

Little, R. J. A. and D. B. Rubin (2020). *Statistical analysis with missing data*. Wiley.

Liu, J., A. Gelman, J. Hill, Y.-S. Su, and J. Kropko (2013). On the stationary distribution of iterative imputations. *Biometrika 101*(1), 155–173.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science 9*(4), 538–558.

Molenberghs, G. (2015). *Handbook of missing data methodology*. Chapman Hall/CRC.

Moons, K. G., R. A. Donders, T. Stijnen, and F. E. Harrell (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology 59*(10), 1092–1101.

Resche-Rignon, M., I. White, and S. Chevret (2020). Imputing missing covariate values in presence of competing risk.

Rubin, D. B. (1976, 12). Inference and missing data. *Biometrika 63*(3), 581–592.

Walter, R. B., M. Othus, G. Borthakur, F. Ravandi, J. E. Cortes, S. A. Pierce, F. R. Appelbaum, H. A. Kantarjian, and E. H. Estey (2011). Prediction of early death after induction therapy for newly diagnosed acute myeloid leukemia with pretreatment risk scores: A novel paradigm for treatment assignment. *Journal of Clinical Oncology 29*(33), 4417–4424.

White, I. R. and P. Royston (2009). Imputing missing covariate values for the cox model. *Statistics in Medicine 28*(15), 1982–1998.

White, I. R. and S. G. Thompson (2005). Adjusting for partially missing baseline measurements in randomized trials. *Statistics in Medicine 24*(7), 993–1007.

Zeijlemaker, W. and G. Schuurhuis (2013). Minimal residual disease and leukemic stem cells in acute myeloid leukemia. *Leukemia*.

Zhang, Z., G. Cortese, C. Combescure, R. Marshall, M. Lee, H. Lim, and B. Haller (2018). Overview of model validation for survival regression model with competing risks using melanoma study data. *Annals of Translational Medicine 6*(16), 325–325.