TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

DIPLOMARBEIT

A Predictive Microsimulation Approach for Modelling the Burden of Mental Disorders in Austria

Ausgeführt am Institut für

Analysis & Scientific Computing

der Technischen Universität Wien
unter der Anleitung von

Ao.Univ.Prof. Dipl.–Ing. Dr.techn. Felix Breitenecker

durch

Andreas Bauer

Schaumannstraße 8
A–2100 Korneuburg

Wien, im Dezember 2015

Abstract

This work aims to analyze the pathways of patients with mental diseases through the health care system. The prediction of the burden of disease is necessary to provide sufficient capacities in the hospitals and to adjust for changes in the structure of the population. In this work emphasis is put on the analysis of reimbursement data of affected people and on the parametrization of a microsimulation model. Also, regional differences of Lower Austria compared to the entire Austrian population are analyzed.
The available data sets contain information about patient attributes as well as times of admissions to hospitals, ambulant contacts to psychiatrists and deaths.
Selected methods from survival analysis and model selection are compared and used to analyze the given data in terms of the readmission times depending on patient parameters. The Cox regression and the model selection methods are used to determine significant parameters for the simulation model. The hazard rates of the particular events in the simulation are estimated using an extension of the Cox model for multiple events.
Using the previous results, a microsimulation model is built to simulate the pathways of mentally ill patients. The considered events are readmissions to hospital, contacts to an ambulant psychiatrist and death. Every patient is classified according to a particular set of parameters and can be in one of several predefined, exclusive states. The events are implemented as state changes.
Three scenarios of simulations are defined to test the consequences of using differently detailed patient-level data on result quality. The first one only takes data of the first readmissions of a patient into account, the second scenario takes all readmissions into account but without any order and the third takes all readmissions into account with order.
All simulations and analyses of the results are performed in R. The simulations are based on a fixed cohort and the duration is a fixed time span.
The overall numbers and times of patients events are analyzed as well as the number of events per patient. Typical pathways of patients are defined to make a more detailed analysis possible.
The differences between the results for the different scenarios and for the various subpopulations regarding patient parameters are pointed out. Further analyses regarding the connection between ambulant contacts and readmissions to the hospital are performed. Finally, an intervention strategy with compulsory ambulant contacts is examined.

Kurzfassung

Diese Arbeit beschäftigt sich mit der Analyse der Wege von psychisch erkrankten Patienten durch das Gesundheitssystem. Die dabei gewonnenen Informationen sollen dazu dienen, entsprechende Kapazitäten in den Spitälern bereitzustellen und Anpassungen bezüglich Veränderungen der Bevölkerungsstruktur vorzunehmen. Das Hauptaugenmerk dieser Arbeit liegt auf der Analyse der Verrechnungsdaten der Patienten und auf der Parametrisierung eines Mikrosimulationsmodells. Weiters werden regionale Unterschiede zwischen der niederösterreichischen und der gesamtösterreichischen Bevölkerung untersucht.

Die verwendeten Datensätze enthalten Informationen über Eigenschaften der Patienten sowie die Zeiten der Wiederaufnahmen, der ambulanten Psychiaterkontakte und Todeszeitpunkte. Ausgewählte Methoden der Ereigniszeitanalyse (engl. survival analysis) und zur Modellauswahl (engl. model selection) werden verglichen und im Folgenden herangezogen, um die Wiederaufnahmezeiten in Abhängigkeit der Patientenparameter zu bestimmen. Die Cox-Regression und die Modellauswahlmethoden werden zusätzlich verwendet, um die signifikanten Parameter für das Simulationsmodell zu eruieren. Die Hazardraten der einzelnen Ereignisse in der Simulation werden mit einer Erweiterung der Cox-Regression geschätzt.

Unter Verwendung dieser Resultate wird ein Mikrosimulationsmodell zur Simulation der Wege von psychisch erkrankten Patienten erstellt. Die untersuchten Ereignisse sind Wiederaufnahmen, ambulante Kontakte zu Psychiatern und der Tod. Jeder Patient besitzt bestimmte Eigenschaften und befindet sich zu jedem Zeitpunkt der Simulation in einem von mehreren vordefinierten und voneinander abgegrenzten Zuständen. Die Ereignisse sind als Zustandswechsel implementiert.

Um die Auswirkungen der Verfügbarkeit der Daten auf die Qualität der Ergebnisse zu untersuchen, werden drei Szenarien definiert, die die Daten zu unterschiedlichen Detailgraden verwenden. Im ersten Szenario werden ausschließlich Daten der ersten Wiederaufnahmen verwendet, im zweiten Daten aller Wiederaufnahmen ohne Reihenfolge, während im dritten Daten aller Wiederaufnahmen mit Reihenfolge herangezogen werden.

Alle Simulationen und Auswertungen werden mit R durchgeführt. Den Simulationen liegt eine fixe Kohorte zugrunde.

Die Gesamtanzahl und die Zeitpunkte der Ereignisse der Patienten sowie die Anzahl der Ereignisse pro Patient werden analysiert. Um eine detailliertere Analyse zu ermöglichen, werden typische Patientenpfade definiert.

Die Unterschiede in den Resultaten der einzelnen Szenarien und Subpopulationen werden herausgearbeitet. Weitere Analysen zum Zusammenhang zwischen ambulanten Psychiaterkontakten und Wiederaufnahmen werden durchgeführt. Abschließend wird eine Interventionsstrategie mit verpflichtenden ambulanten Psychiaterkontakten getestet.

Contents

1 Introduction

1.1 Motivation

In the field of chronic, mental diseases it is important to understand the conditions under which patients are likely to be readmitted to hospital in order to provide sufficient care. The ambulant treatment between the admissions is also involved in this process. With this information the burden of mental disease can be predicted more accurately and the health care needs can be adjusted. Besides, the increasing number of elderly people will probably lead to an increasing number of patients and additional demand for treatment. The health care management has to deal with this challenge.

1.2 Goals

The first goal is to determine the factors which influence the readmission times of patients. This information is also used for the determination of the pathways of patients with mental diseases through the health care system. The overall number of readmissions to hospital is estimated as well as the numbers of events for subpopulations defined by certain patient characteristics to determine the required capacity and its changes over time. The effect of an aging population is examined.
Also, the influence of ambulant contacts to the psychiatrist on time and number of readmissions is examined. The question is if the contacts help patients to stay away from hospitals or if they are an indicator for a worsening patient condition and an intermediate step on the way to a readmission to hospital.
The consideration of regional aspects is important for the accuracy of the simulation results. So, the situation of patients with mental diseases for Lower Austria is investigated in detail and compared with the situation for entire Austria to find out possible specific characteristics of the population of Lower Austria.
The availability of data is often a problem especially when using sensitive patient data. In these cases often privacy protection only allows usage of k-anonymized data. So, the question arises if it is possible to get meaningful simulation results with the given data. Differently detailed patient-level data of the same set is used to analyze the effect on the result quality.
Overall, the goal of this work is to improve simulation models for planning of the re-

sources and capacities in hospitals by putting emphasis on the analysis of the data and the parametrization of the model.

1.3 Tasks

For the achieving the above described goal a number of tasks is defined in this section. At the beginning, a descriptive statistical analysis of the available data sets containing information about patient attributes is performed as well as times of admissions to hospitals, ambulant contacts to psychiatrists and death.
Then, selected methods from the fields of survival analysis and model selection are introduced and compared. These are used for the determination of significant parameters for the readmission times and moreover for the parametrization of the model.
The next step is to build the actual microsimulation model. The considered events are readmissions to hospital, contacts to an ambulant psychiatrist and death. Every patient is classified according to a particular set of parameters and can be in one of several predefined, exclusive states. The events are implemented as state changes.
Three scenarios of simulations are defined and compared. The first one takes only the first readmissions of a patient into account, the second scenario takes fixed number of several readmissions without any order into account and the third one takes the same fixed number of several readmissions into account chronologically. All scenarios are executed with and without contacts to a psychiatrist.
Finally, simulations for two populations, one from Lower Austria and one from entire Austria, are performed and a detailed analysis of the numbers and times of the events is carried out and all results for the different scenarios and subpopulations are presented and compared.

1.4 Overview

In Chapter 2, the given data samples are presented and described. An introduction to survival analysis is given in Chapter 3 and the methods which are applied in this work are presented. Also, an overview of selected model selection methods is given. In Chapter 4, the data is analyzed with the methods from Chapter 3. Microsimulation models are introduced in Chapter 5. After that, the model and its implementation are described. The model also is validated. For the simulations, scenarios are defined in Chapter 6. The model is examined for stability and the results of the simulations of the various scenarios are presented and compared. Finally, the results of the work are summarized and conclusions are given in Chapter 7.

2 Data

Three data samples of patients data are available through the CEPHOS-LINK project (Comparative Effectiveness Research on Psychiatric Hospitalisation by Record Linkage of Large Administrative Data Sets, number = 603264). The data samples were recorded in the years 2006 and 2007 in hospitals in Austria and are used for the parametrization and the sampling of the population of the simulation model.

2.1 Sample Austria

The patient sample *dataaut* consists of over 240000 records of about 19000 patients from Austria. Information about birth year, sex, date of death, dates of admissions and releases, type of admission and release, postal code, length of stay in the hospital, diagnosis in ICD-10 code [1], main or additional diagnosis, department in which the patient stayed and length of stay in the psychiatric department of the hospital are contained.
Since the simulation starts with the first release from hospital, only patients with first admissions are considered. A first admission is defined that there was no admission in the previous half year, because, according to experts, patients with psychiatric diseases (F20-69) normally are readmitted to hospital at least every half of a year. The given data is already adjusted to this criteria, so the first admission in the dataset is the actual first admission according to this definition.
The data is filtered for various criteria. All incomplete data sets are eliminated and only stays with main diagnosis psychiatric disease (F20-69 in ICD-10) and stays in the psychiatric department of the hospital are considered. Thereafter the sample consists of almost 30000 records of 18638 patients.
A second data set *datapsy* with times of ambulant contacts to a psychiatrist (outpatient contacts = OPC) is used. This data set is merged with dataset *dataaut* by assigning the contacts to assign the contacts to patients already contained in the first set.

Patient parameters

Table 2.1 shows an overview of the four patient parameters, sex, age, length of stay in the psychiatric department of the hospital and diagnosis, that are included in the model and their values. The diagnoses are split into six diagnosis groups for the simulations.

Short	Name	Values
S	Sex	categorical: male, female
A	Age	ordinal: 18-97
L	Length of Stay	ordinal: 1-430
D	Diagnosis	categorical: F2x, F30+F31, F32-F39, F4x, F5x, F6x

Table 2.1: Parameters of the full model

The distribution of these four patient parameters in the patient sample *dataaut* is illustrated below.
Table 2.2 shows that almost 60 percent of the patients are female.

Sex	Number	Percentage
female	10911	58.72
male	7727	41.28

Table 2.2: Distribution of sexes in data sample *dataaut*

The histogram in Figure 2.1 shows the distribution of age in the data sample *dataaut*. The median is at the age 43 and the range goes from 18 to 97 years. The group between 40 and 45 years is the biggest one. Almost two third of the patients are aged between 30 and 60 years.

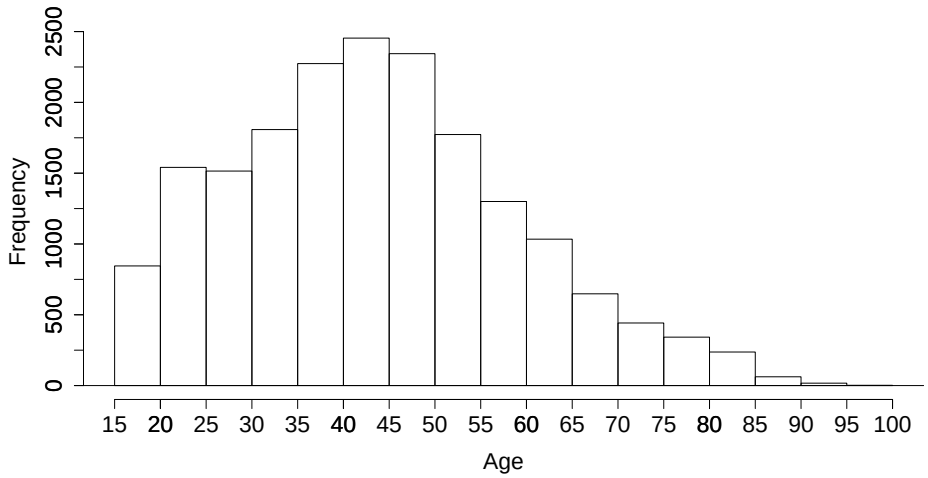


Figure 2.1: Histogram of age distribution in patient sample *dataaut*

In Figure 2.2, a histogram for the length of stay in the psychiatric department of the hospital is shown. The distribution is nearly exponential with about 6000 stays shorter

than 10 days and 5000 between 10 and 19 days. Only 231 stays are longer than 100 days. This is slightly over one percent of all stays.

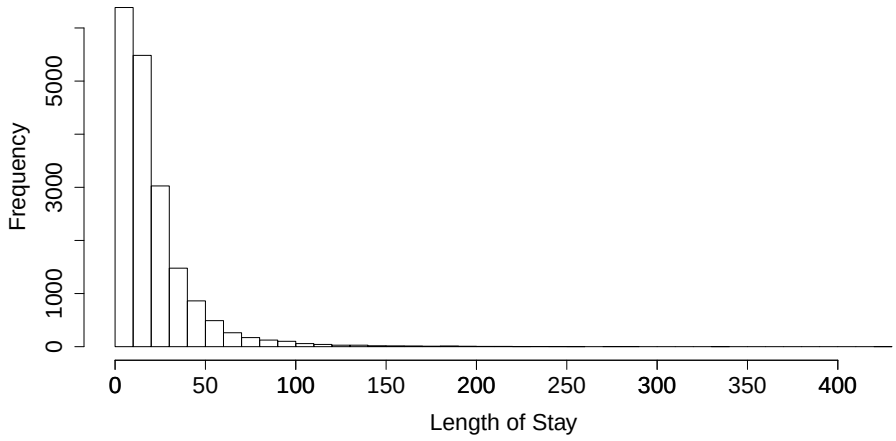


Figure 2.2: Histogram of the distribution of the lengths of stay in patient sample *dataaut*

The distribution of the diagnosis groups is presented in Figure 2.3. The most common diagnosis group is F32-F39 containing about 7000 patients. The group F5x is the most uncommon with only 146 patients.

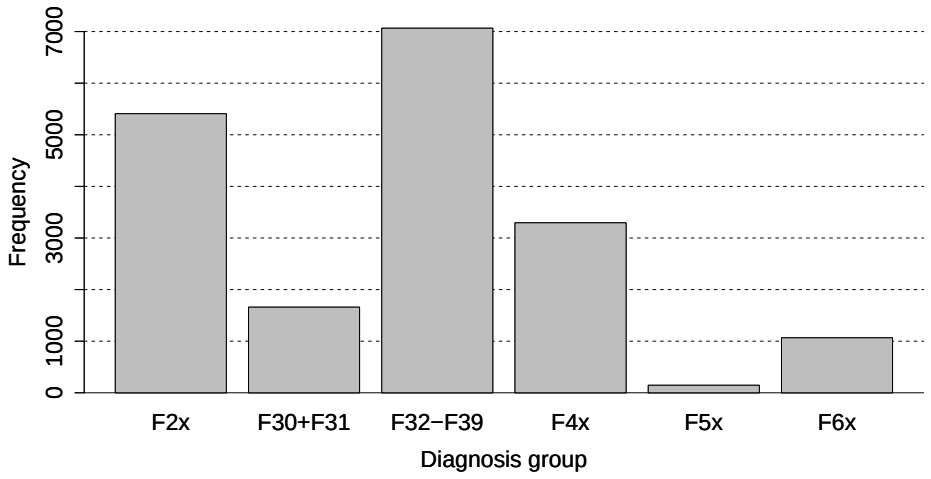


Figure 2.3: Barplot of the distribution of the diagnosis groups

Distribution of events and event times

In the following the times and numbers of the readmissions, psychiatrist contacts and death are analyzed.

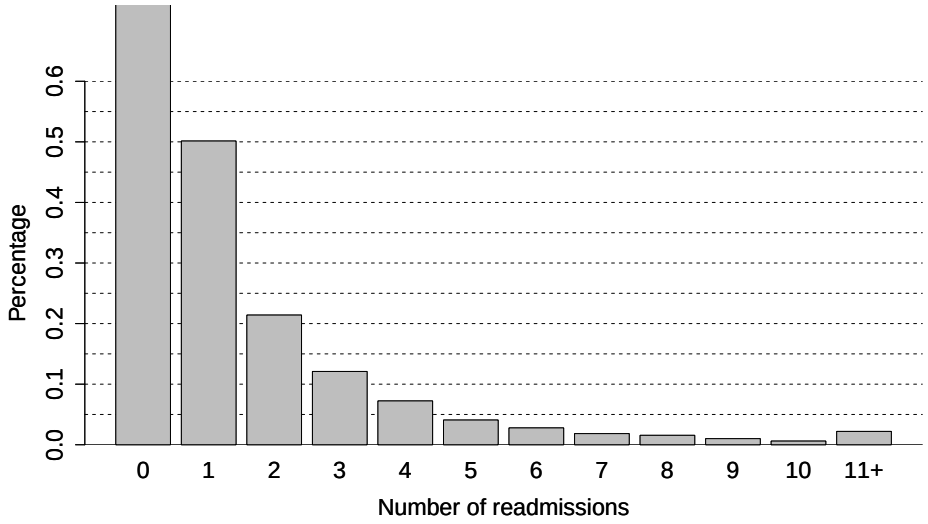


Figure 2.4: Numbers of readmissions per person

In Figure 2.4, an overview of the distribution of the numbers of readmissions is presented. The biggest share with about 58% are patients with no readmission. The number of patients decreases with an increase of the number of readmissions. Patients with more than ten readmissions comprise less than 1%.

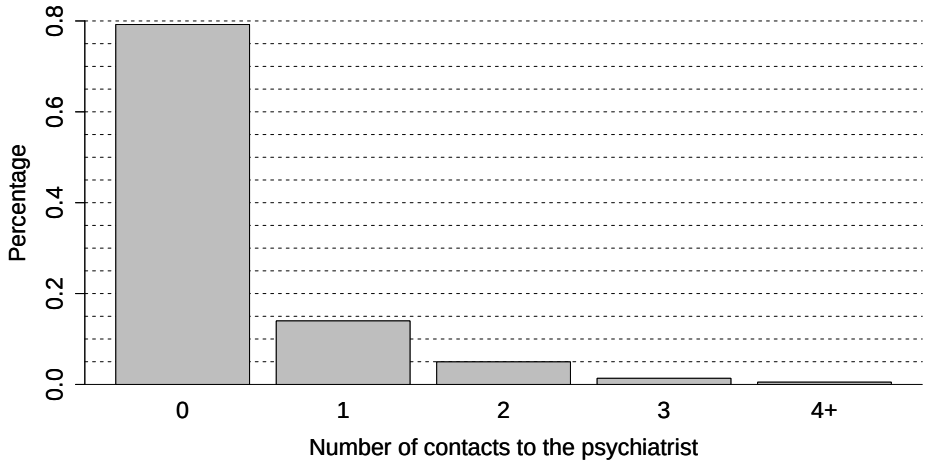


Figure 2.5: Numbers of recorded psychiatrist contacts per person

Figure 2.5 shows the distribution of the numbers of contacts to a psychiatrist. Almost 80% of the patients have no contact to the psychiatrist, about 14% have one contact and about 5% have two contacts. The patients with more than three contacts comprise less than one percent of all patients.

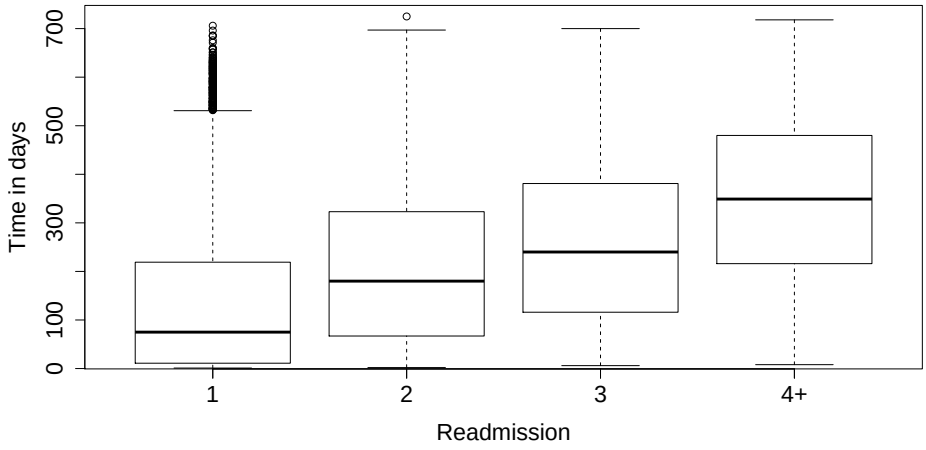


Figure 2.6: Boxplot for distribution of the times of the readmissions

The distributions of the times of the first, second, third and all higher readmissions are displayed in Figure 2.6. The median gets higher with every readmission. Half of the first readmissions occur within 100 days after the initial release.

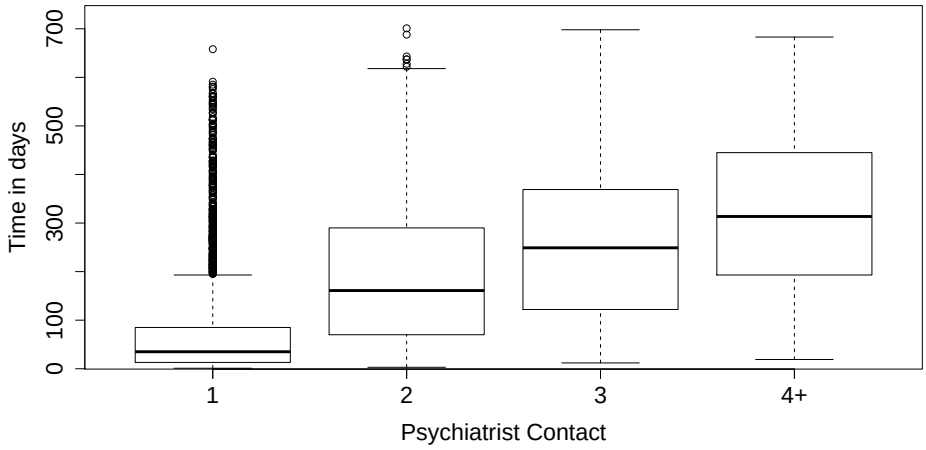


Figure 2.7: Boxplot for distribution of the times of the contacts to the psychiatrist

The distributions of the times of the first, second, third and all higher ambulant contacts to a psychiatrist are displayed in Figure 2.7. It can be seen that the first contact to the

psychiatrist is often very shortly after the initial release from hospital. Half of them are within 40 days after the release and more than three fourths within 100 days. Almost all contacts occur within 500 days after the release.

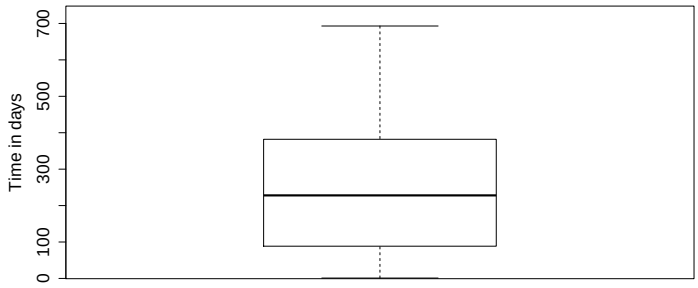


Figure 2.8: Boxplot of the death times

A boxplot of the death times is presented in 2.8. The median is at about 230 days and almost 75 percent of the deaths occur within the first year after the release. About 3 percent of the patients died during the recording of the data.

2.2 Sample Lower Austria

The sample *datanoe* contains information about sex, age, diagnosis group and length of stay in the hospital from 6822 patients from Lower Austria. This data set is used as population sample for simulations.
In the following paragraph, the distribution of the four parameters in the sample *datanoe* is illustrated.
Table 2.3 shows that about 58 percent of the patients are female.

Sex	Number	Percentage
female	3982	58.37
male	2840	41.63

Table 2.3: Distribution of sexes in data sample *datanoe*

The histogram in Figure 2.9 shows the distribution of age in the data sample *datanoe*. The median is at the age 43 and the range goes from 18 to 93 years. The group between 35 and 40 years is the biggest one. Almost two third of the patients are aged between 30 and 60 years.

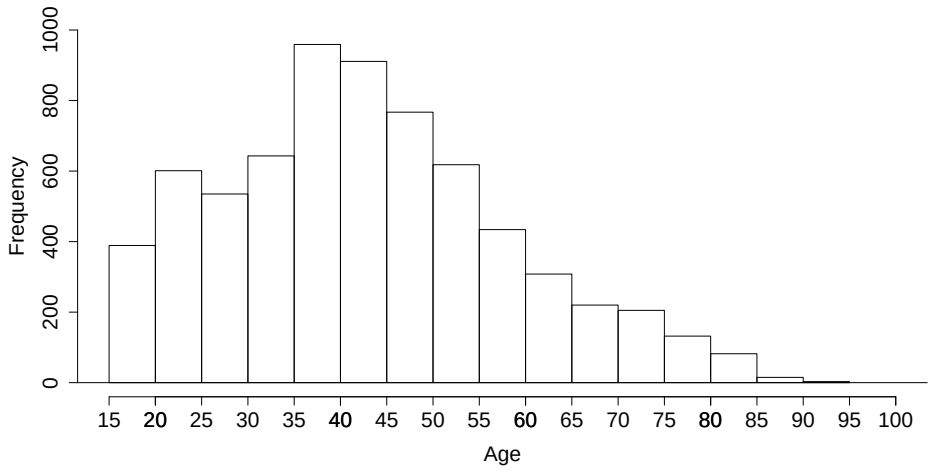


Figure 2.9: Histogram of the age distribution in patient sample *datanoe*

In Figure 2.10, a histogram for the length of stay in the psychiatric department of the hospital is shown. The distribution is nearly exponential with about 2500 stays shorter than 10 days and 2000 between 10 and 19 days. Only 104 stays are longer than 100 days. This is about 1.5 percent of all stays.

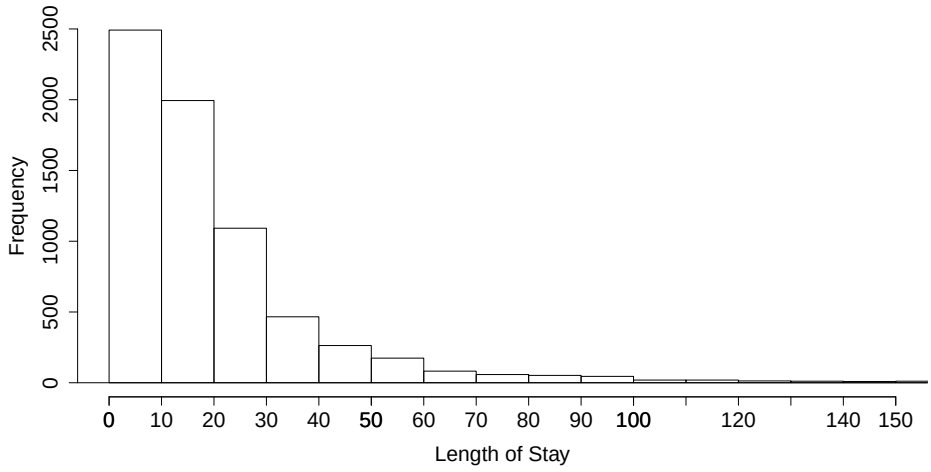


Figure 2.10: Histogram of the distribution of the lengths of stay in patient sample *datanoe*

The distribution of the diagnosis groups is presented in Figure 2.11. The most common diagnosis group is $F1x$ with about 2300 patients. The group $F5x$ is the most uncommon with only 64 patients.

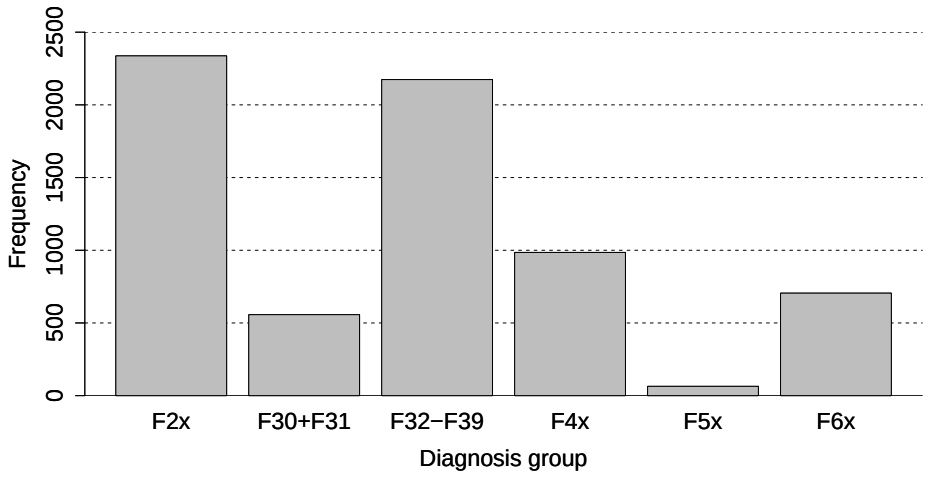


Figure 2.11: Barplot of the distribution of the diagnosis groups in patient sample *datanoe*

3 Methods

In this chapter, methods to build the statistical models including methods from survival analysis and model selection as well as methods used for the parametrization of the simulation model are presented.

3.1 Introduction to survival analysis

Survival analysis deals with the analysis of data of the time until the occurrence of a particular event. This kind of data is frequently encountered in medical research and also in other areas of application and referred as survival data. However, the event of interest is not always death. Also, other types of events such as hospitalization and a change of diagnosis are possible. A critical issue in this context that makes standard statistical methods inapplicable is censoring. This is the case when the data collection ends before the event of interest occurs.
In this chapter, an overview of methods of survival analysis used in this thesis is given; for example: the Kaplan-Meier estimate, the Nelson-Aalen estimate and the Cox model.

Definitions

Let T be a non-negative continuous random variable representing the time until the occurrence of an event. The probability density function (p.d.f.) of T is denoted by f and the cumulative distribution function (c.d.f.) by F. In survival analysis often the complement of the c.d.f. is used. The survival function S is defined as

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

which gives the probability of being alive at time t.
By derivation follows: $S'(t) = -F'(t)$.
Another useful function for survival analysis is the hazard function λ defined as

$$\lambda(t) := \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t}. \tag{3.2}$$

The hazard function gives the instantaneous rate of occurrence of the event and can be written in terms of the survival function and the p.d.f. as seen in Equation (3.3).

$$\begin{aligned}\lambda(t) &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} = \\ &= \lim_{\Delta t \to 0} \frac{P(T > t | t < T \leq t + \Delta t) P(t < T \leq t + \Delta t)}{P(T > t) \Delta t} = \\ &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{P(T > t) \Delta t} = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{S(t) \Delta t} = \frac{f(t)}{S(t)} \end{aligned} \tag{3.3}$$

In the second step, starting from the definition of the hazard function Bayes' Theorem is applied. The third expression holds due to $P(T > t | t < T \leq t + \Delta t) = 1$. Then, the definition of the survival function from Equation (3.1) is applied. Since $-f$ is the derivative of S Equation (3.3) can be rewritten as

$$\lambda(t) = -\frac{d}{dt} \log S(t). \tag{3.4}$$

By integrating Equation (3.4) and introducing the condition $S(0) = 1$ that describes the fact that it is sure that the event has not occurred by time 0, the survival function S can be expressed in terms of the hazard function λ.

$$S(t) = \exp\left(- \int_0^t \lambda(x) dx \right) \tag{3.5}$$

The integral in Equation (3.5) is called the cumulative hazard function Λ and is denoted by

$$\Lambda(t) = \int_0^t \lambda(x) dx. \tag{3.6}$$

So, the survival function and the cumulative hazard function are connected by

$$S(t) = \exp\left(-\Lambda(t) \right). \tag{3.7}$$

Representation of the data

In [2], two ways to represent survival data of individuals are presented. A short overview of both ways is given in the following.
Firstly, the data of individual i can be represented as a pair of variables (t_i, δ_i). Let t_i^* be the survival time of the individual and c_i^* be the censoring time. So, only one of the two times is known exactly depending on whether the event of interest or censoring happens earlier. t_i is the minimum of these two times: $t_i = \min(t_i^*, c_i^*)$. $\delta_i = I_{(t_i^* \leq c_i^*)}$ is an event indicator which is 1, if t_i^* is observed and 0, if the observation is censored.
The other, more general, formulation is the counting process formulation which replaces the pair of variables (t_i, δ_i) with a pair of functions $(N_i(t), Y_i(t))$:

- $N_i(t) = I_{(t_i \leq t, \delta_i = 1)}$
- $Y_i(t) = I_{(t_i \geq t)}$

That means $N_i(t)$ is 1, if an event already happened until time t and 0 if not and $Y_i(t)$ indicates, if individual i is at risk at time t.
The counting process formulation can easily be extended to represent multiple events. In this case $N_i(t)$ counts the number of observed events up to time t.

Censoring

Another distinctive aspect regarding survival data is censoring. This issue renders standard methods inappropriate for the analysis of survival data [3].
In medical studies data collection often ends before the event of interest occurred for all individuals. For those individuals with no event until the end of the data collection, the survival time cannot be determined exactly. It is only sure that the survival time exceeds the time span of the observation of the particular patient.
The first distinction is made between informative and non-informative censoring. Non-informative censoring means that the probability of being censored is unrelated to the probability that the event of interest occurs. In this work, non-informative censoring is assumed [3]. For informative censoring special methods have to be applied.
There are several types of non-informative censoring depending on, if the starting point, the end point of the observation or both are censored. In the data sets at the base of this work, only right-censoring occurs. That means that for some individuals the observation ended or another competing event occurred before the event of interest occurred. More details on competing events are given in Section 3.1.3.

3.1.1 Parametric estimators

This section lists the Weibull and the Exponential distribution which are the two most common distributions used for parametric estimators in the context of survival analysis and is based on [3].
In special settings a probability distribution is assumed for the survival times. In these cases, parametric models can be used. Parametric estimators of the hazard function can also be used for the baseline function of the Cox model introduced in Section 3.1.3.

Exponential distribution

The simplest model for the hazard function is to assume that it is constant over time. Under this model, the hazard function may be written as $\lambda(t) = \alpha$. From Equation (3.5) follows that the corresponding survival function S is $S(t) = \exp(-\alpha t)$. Since

$-f(t) = S'(t)$, the implied probability density function f is $f(t) = \alpha \exp(-\alpha t)$. This is the p.d.f. of a random variable which has an Exponential distribution with mean α^{-1}.

Weibull distribution

The assumption of a constant hazard function is rarely tenable. A more general form is $\lambda(t) = \alpha\gamma t^{\gamma-1}$ with $\alpha, \gamma > 0$. The special case of $\gamma = 1$ leads back to the exponential distribution. Following the procedure from above, the implied p.d.f. is $f(t) = \alpha\gamma t^{\gamma-1} \exp(-\alpha t^{\gamma})$. This is the p.d.f. of a random variable which has a Weibull distribution with shape parameter α and scale parameter γ.

3.1.2 Nonparametric estimators

This section defines the two most common nonparametric estimators, the Kaplan-Meier estimate and the Nelson-Aalen estimate, in the context of survival analysis and is based on [4].

Kaplan-Meier estimate

The Kaplan-Meier estimate is an estimate for the survival function S. It uses the information of the exact time of the occurrence of the event. The estimated survival probability s_t at time t is:

$$s_t = \frac{n_t - d_t}{n_t} \tag{3.8}$$

n_t is the number of people at risk at time t and d_t is the number of people that experience the event at time t. So, s_t is the ratio of the number of people who have no event at time t to the number of people that are at risk at time t. Let t_i denote the event times. Thus, the probability of surviving up to time t_j is calculated with the so-called product-limit formula:

$$S(t_j) = \prod_{i:t_i \leq t_j} s_{t_i} \tag{3.9}$$

Without censoring, the Kaplan-Meier estimate is equivalent to the complement of the cumulative density function.

Nelson-Aalen estimate

The Nelson-Aalen estimate is an estimate for the cumulative hazard function Λ. Let d_t and n_t again denote the number of people that experience the event at time t respectively are at risk at time t. Let t_i denote the event times again. Then Λ can be estimated by

$$\hat{\Lambda}(t_j) = \sum_{i:t_i \leq t_j} \frac{d_{t_i}}{n_{t_i}}. \tag{3.10}$$

Based on the connection of S with Λ in Equation (3.7) the Nelson-Aalen estimate can also be used as an estimate for the survival function:

$$\hat{S}(t) = e^{\sum_{i:t_i \leq t} \frac{d_{t_i}}{n_{t_i}}} = \prod_{i:t_i \leq t} e^{\frac{d_{t_i}}{n_{t_i}}}. \tag{3.11}$$

$\hat{S}(t)$ is the so-called Breslow estimate.

3.1.3 Cox model

This section gives an overview of the Cox model and is based on [2] and [4] as well as [5].

Definition and properties

The Cox model focuses on modeling the hazard function. It was initially defined by David Cox in [6]. The hazard at time t for individual i with covariate vector X_i is assumed to be

$$\lambda_i(t) = \lambda_0(t) \exp(X_i \beta) \tag{3.12}$$

where λ_0 is an unspecified nonnegative function called baseline hazard function and β is an n-dimensional vector of regression coefficients. The baseline hazard function describes the hazard function for an individual with reference values of $X_i = 0$. The hazard function as expressed in Equation (3.12) is the product of two functions. The function λ_0 characterizes how the hazard function changes as a function of survival time. The second factor characterizes how the hazard function changes as a function of subject covariates X_i. So, the baseline hazard function is multiplied with a factor depending on the covariates of the particular individual.
The proportional hazard assumption states that the ratio of the hazards of different groups remains constant over time.

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t) \exp(X_i \beta)}{\lambda_0(t) \exp(X_j \beta)} = \frac{\exp(X_i \beta)}{\exp(X_j \beta)} = C, \quad \forall i, j \tag{3.13}$$

Because of this property, the Cox model is also called proportional hazards model.

It it also possible to include interaction terms in the Cox model. That means that additionally to the single variables interaction variables, implemented as product of the involved single variables, enter the model. This product can consist of several variables. For example, for two parameters S and A with values s and a, the Cox model with interaction is $\lambda(t) = \lambda_0(t) \cdot \exp(\beta_S \cdot s + \beta_A \cdot a + \beta_{SA} \cdot s \cdot a)$.

The Cox model is very convenient for estimating hazard ratios. When the hazard function or survival function is explicitly needed, the baseline hazard also has to be estimated.

This can be done by using either non-parametric estimates such as the Nelson-Aalen estimate or the Kaplan-Meier estimate or parametric estimates, if the data follows a particular probability distribution.

Likelihood

In order to calculate the regression coefficient vector β suppose a subject is observed at time t_i. If the subject died at t_i, its contribution to the likelihood function L_i is the p.d.f. at that time indicating $L_i = f(t_i)$. If the subject is alive, the survival time is greater than t_i. So, the probability of this censored observation is $L_i = S(t_i)$. With the assumption of the independence of the observations the overall likelihood function L is the product of the single likelihood functions L_i. The identity $f(t) = S(t)\lambda(t)$ leads to the last expression

$$L = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} = \prod_{i=1}^{n} \lambda(t_i)^{\delta_i} S(t_i). \tag{3.14}$$

This expression can be used for parametric models where the survival and the hazard function can be stated explicitly. For the Cox regression, the partial likelihood function proposed by Cox has to be used which is described below.

Suppose that data is available for n independent observations of individuals of whom m have distinct and $n - m$ right-censored survival times. The m distinct ordered event times are denoted by $t_1 < t_2 < \cdots < t_m$. It is assumed that there are no tied survival times. The so-called risk set $R(t_j)$ consists of the uncensored individuals at risk just prior to t_j. The partial likelihood proposed by Cox for β is

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(X_i\beta)}{\sum_{j \in R(t_i)} \exp(X_j\beta)} \right]^{\delta_i} = \prod_{i=1}^{m} \frac{\exp(X_i\beta)}{\sum_{j \in R(t_i)} \exp(X_j\beta)}. \tag{3.15}$$

The second product excludes the factors when $\delta_i = 0$. So, the product is only over the m distinct event times and X_i is the covariate vector for the subject with survival time t_i.
For the actual calculation, the log partial likelihood function $\log L(\beta) := l(\beta)$ is used.

$$l(\beta) = \sum_{i=1}^{m} \left[(X_i\beta) - \log \left(\sum_{j \in R(t_i)} \exp(X_j\beta) \right) \right] \tag{3.16}$$

The maximum likelihood estimate of β can be found by maximizing this log likelihood function. This maximization is generally accomplished using the Newton-Raphson algorithm.

Let $u(\beta) = \nabla l(\beta)$ be the so-called score vector of the first derivatives of $l(\beta)$ with respect to β and $\mathcal{I}(\beta)$, the negative $p \times p$ matrix of the second derivatives of l. So, $\mathcal{I}(\beta) = -H_l(\beta)$ is the negative Hessian of l with respect to β.
The Newton-Raphson algorithm is an iterative algorithm to solve the partial likelihood equation $u(\hat{\beta}) = 0$. It starts with an initial guess $\hat{\beta}^{(0)}$ and the iteration is

$$\hat{\beta}^{(n+1)} = \hat{\beta}^{(n)} + \mathcal{I}^{-1}(\hat{\beta}^{(n)})u(\hat{\beta}^{(n)}). \tag{3.17}$$

The iteration stops, when $\hat{\beta}^{(n+1)} \approx \hat{\beta}^{(n)}$. The algorithm is very robust. The default initial value is $\hat{\beta}^{(0)} = 0$.

Analysis of the Cox model

The most common analysis tools for the Cox model are hazard ratios and p-values for each variable. An overview is given in the following paragraphs. More details can be found in [4].
The hazard ratio (HR) for a variable x_k with coefficient β_k is HR $= \exp(\beta_k)$. This results from the proportional hazards assumption, because the baseline hazard and the summands regarding the fixed variables cancel out as shown in Equation (3.13).
For continuous variables, the value of $\exp(\beta_k)$ provides the factor the hazard function is multiplied with when the particular variable is increased by one unit given that all other variables in the model are fixed. For dichotomous variables, the value of $\exp(\beta_k)$ provides the factor the hazard function is multiplied with when the particular variable is one in contrast to being zero given that all other variables in the model are fixed. Thus, a hazard ratio greater than one means that the hazard increases when the particular parameter is increased. By contrast, when the hazard ratio is less than one the hazard decreases when the particular parameter is increased.
The p-values for single variables of the Cox model are calculated by the z-test. The z-test statistic for regression coefficient β is $z = \frac{\beta}{s.e.}$. The null hypothesis is that the parameter β is zero. $s.e.$ denotes the standard error. Then, the p-value is calculated using the normal distribution $p = P(x > |z|)$ for $x \in \mathcal{N}(0, 1)$. The p-value indicates if the null hypothesis holds. If the p-value is smaller than the significance level, the null hypothesis is rejected and the parameter is supposed to be significant for the model.

Tests on coefficients

The standard *Wald*, *score* and *likelihood ratio* tests can be used to test hypotheses about the true parameter vector β. The global null hypothesis is H_0: $\beta = \beta^{(0)}$. $\hat{\beta}$ denotes the final estimate by the Newton-Raphson algorithm. Below the test statistics for the three tests are listed. l denotes the log partial likelihood function.

- *Likelihood ratio* test statistics: $2(l(\hat{\beta}) - l(\beta^{(0)}))$

- *Wald* test statistics: $(\hat{\beta} - \beta^{(0)})'\hat{\mathcal{I}}(\hat{\beta} - \beta^{(0)})$ with $\hat{\mathcal{I}} = \mathcal{I}(\hat{\beta})$
- *Score* test statistics: $u'(\beta^{(0)})\mathcal{I}(\beta^{(0)})^{(-1)}u(\beta^{(0)})$

The null hypothesis distribution of each of these tests is a χ^2 with p degrees of freedom. As usual, they are asymptotically equivalent, but in finite samples they may differ. If so, the *likelihood ratio* test is generally considered the most reliable.
For a single variable, the *Wald* test reduces to the usual z-statistics.
These tests can also be used to test if all parameters of a certain subset are zero. So, the parameter vector β is partitioned into two sets β_1 and β_2 with p_1 elements in β_1 and p_2 elements in β_2. The null hypothesis is $H_0 : \beta_2 = 0$. This can be inserted into the test statistics and the null hypothesis distribution of each of these tests is now a χ^2 with p_2 degrees of freedom. This method can be used in the process of model selection, for example, in the stepwise selection of the covariates of the model as described in Section 3.2.3.
Confidence intervals for the coefficients are usually created based on *Wald* statistics. The lower and upper 95% confidence interval values are $\exp(\hat{\beta} \pm 1.96 s.e.(\hat{\beta}))$.

Preparation of data for R

Basically, an individual is represented by a time-to-event along with censoring status, stratum and covariate variables. The stratified Cox model is explained in the next section in detail.
In order to extend the possibilities of the Cox model the data is cast into a counting process form [2]. The only difference is that individuals have not only a time-to-event but a interval of risk (start, stop]. The interval of risk is always open on the left and closed on the right. Now, a subject can be represented by a set of observations, containing an interval of risk (start, stop] along with status, strata and covariate variables. This can be useful for time-dependent covariates, time-dependent strata and multiple events per subject. In Table 3.1, an example for time-dependent strata is given. Subject 1 with age 67 is in stratum 1 until day 157 and afterwards in stratum 2 until the end of follow-up after 205 days.

Id	Interval	Status	Age	Stratum
1	(0,157]	0	67	1
1	(157,205]	0	67	2

Table 3.1: Example for a dataset in the counting process form

Stratified Cox model

This extension of the Cox model allows multiple strata. The strata divide the subjects into disjoint groups and each subject is member of exactly one stratum. Each of which has a distinct baseline hazard function but common values for the coefficient vector β. The hazard for individual i belonging to stratum k is

$$\lambda_k(t)e^{X_i\beta}. \tag{3.18}$$

The overall log likelihood is the sum of the log likelihoods of each stratum.

$$l(\beta) = \sum_{k=1}^{K} l_k(\beta) \tag{3.19}$$

The major application of the stratified Cox model is to adjust for a confounding variable whose effect does not have to be taken into account in the model.

Extension of the Cox model to multiple events

In order to apply survival analysis to data sets with multiple events per subject, the Cox model can be extended. This section about the Cox model for multiple events is based on [2].
A common approach is the marginal approach. The marginal approach is carried out in three steps:

- Decide on a model and structure the data set accordingly
- Fit the data as an ordinary Cox model, ignoring possible correlations
- Replace the standard variance estimate with one which is corrected for the possible correlations

The ordinary Cox model estimate of the variance for $\hat{\beta}$ treats each of the observations as independent. This assumption does not hold, when a given subject may contribute multiple events. A possible correction is the use of the so-called jackknife estimate for the variance. For data where the correlation is restricted to disjoint groups (subjects) the obvious choice is a grouped jackknife estimate that leaves out one subject at a time rather than one observation at a time. More details on the jackknife estimate can be found in [2].

Unordered multiple events

For unordered, but correlated events the data set contains one stratum for each type of event. So, it consists of (number of patients) $\cdot$ (number of types of events) observations.

The time-to-event is censored for all types except for the type of event that occurred at first. The censoring time is the time to the first event. Commonly, the data is stratified by the type of event. An analysis without stratification would lead to a time-to-first-event analysis.
In Table 3.2, an example with two patients and three types of unordered events is presented. The types of events are called A, B, C. For reasons of simplicity, only one covariate, age, is considered. The first patient at the age of 55 experiences event A after 80 days and the second patient at the age of 67 event C after 157 days. The status is one for the type of event that occurred and zero for all the others which are censored.

Id	Time	Status	Type of Event	Age
1	80	1	A	55
1	80	0	B	55
1	80	0	C	55
2	157	0	A	67
2	157	0	B	67
2	157	1	C	67

Table 3.2: Example of a dataset with unordered multiple events

Ordered multiple events

Three different approaches for ordered multiple events are presented. All of them belong to the group of marginal models: Anderson-Gill model (AG), Wei-Lin-Weißfeld model (WLW), Conditional model (Cond). All three approaches use the counting process style of data input.
In the AG model, each subject is represented as a series of observations with time intervals (0, first event], (first event, second event],...,(mth event, last follow-up]. This model is similar to the original Cox model with only one difference that in the AG model the subject is still at risk after the first event. The AG model is suited for situations of mutual independence of the observations within a subject.
The hazard function for the ith subject is:

$$Y_i(t)\lambda_0(t)\exp(X_i(t)\beta) \tag{3.20}$$

The WLW model treats the ordered outcome data as an unordered competing risk set. First of all, the maximum number of events n per subjects is determined. Then, the data set is split into n strata, so there are n rows for each subject in the analysis, one for each stratum, even if the subject has less than n events. In that case, the baseline hazard can be different for each stratum.

The hazard function for the jth event for the ith subject is:

$$Y_{ij}(t)\lambda_{0j}(t)\exp(X_i(t)\beta_j) \tag{3.21}$$

The Cond model assumes that a subject is not at risk for event m until event m-1 occurs. The input style is similar to the AG model but each event is assigned to a separate stratum. Again, the baseline function can vary for different events. The hazard is formally identical to the hazard of the WLW model with only one difference that Y_{ij} is zero until the (j-1)th event and only then becomes one.
As example, a person with events at times 14 and 35 and follow-up until time 47 is displayed for all three models in Table 3.3.

	Interval	Stratum
AG	(0,14]	1
	(14,35]	1
	(35,47]	1
WLW	(0,14]	1
	(0,35]	2
	(0,47]	3
Cond	(0,14]	1
	(14,35]	2
	(35,47]	3

Table 3.3: Representation of a subject for the three marginal models: AG, WLW and Cond

The example in Table 3.3 shows that both, the AG and the Cond method, treat the data as time-ordered outcome, differing only in their use of stratification. In the AG model, there is only one stratum, so the subjects always stay in stratum 1, whereas in the Cond model, the subjects move to the next stratum after each event. In contrast to that, the WLW model has a row for each possible event and all intervals start at time zero. Thus, the WLW style data set is usually larger than the sets of the other styles, because not every possible event occurs for each person.

Multi-state/Combination models

The three presented approaches can be combined to a multi-state model. In this framework, every transition between states is possible. Thus, all of the three approaches must be combined. The WLW model for competing risks, the AG model for subjects that can reenter states and the conditional model for series of disjoint states in particular order. In the given situation, after a subject is released from hospital, it is at risk for either a visit to a psychiatrist or for readmission to hospital. So, this is a case of competing

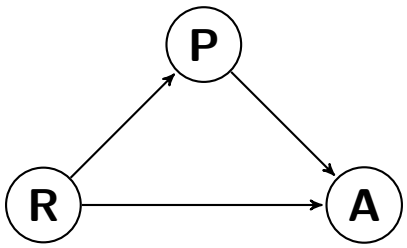


Figure 3.1: Graph of possible transitions between release (R), readmission (A) and visit to psychiatrist (P)

events (WLW). After the visit to the psychiatrist, the patient can still be readmitted to hospital, but after a readmission no visit to the psychiatrist is possible. So, we also have an order of events (Cond). In Figure 3.1, a directed graph with all possible transitions is shown. There are three possible transitions.
After the release, there can be a transition to the psychiatrist or to hospital and after the visit to the psychiatrist can be a readmission to hospital. The visit to the psychiatrist acts as a censoring event for the time from release directly to readmission to hospital. Therefore, the data set is extended that every patient has three observations, one for each transition. It consists of the given covariates of the particular patient, the time in the according state and the censoring status, if the transition actually happened or was censored. Then, the data is stratified into three strata.
For example, consider a patient with a visit to the psychiatrist on day 80 after the release from hospital and a readmission on day 120. In this framework, there are three strata. So, the patient is encoded as three observations. The first one in stratum "release-psych" with time 80 and status one, the second one in stratum "release-readmission" with time 80 and status zero and the third one in stratum "psych-readmission" with time 40 and status one. In Table 3.4, the according data sample for this patient is shown.

Time	Status	Stratum
80	1	release-psych
80	0	release-readmission
40	1	psych-readmission

Table 3.4: Example for representation of a subject in the combination model

Transition probabilities

The transition probabilities can be calculated from the cumulative transition hazards [7]. Let $(X_t)_{t\geq 0}$ be a time-inhomogeneous Markov process with state space $\{0, 1, ..., J\}$. It is assumed that $(X_t)_{t\geq 0}$ has right-continuous sample paths, which are constant between

the transition times. On any finite interval are only finitely many transitions.
The matrix of the transition probabilities is defined as

$$P(s,t) := (P_{jk}(s,t))_{j,k}, \quad j,k \in \{0,1,...,J\} \tag{3.22}$$

with transition probabilities

$$P_{jk}(s,t) := P(X_t = k | X_s = j), \quad s \leq t. \tag{3.23}$$

Let $a(t) := (P(X_t = i))_{i=0,...,J}$ be the state occupation probabilities vector. It can be calculated by multiplying the initial state occupation vector with $P(0,t)$:

$$a(t) = P(0,t) \cdot a(0) \tag{3.24}$$

The cumulative transition hazards $A_{jk}(s)$, $j \neq k$, can be estimated through the multi-state extension of the Cox model and $A_{jj}(t) := -\sum_{k=0,k\neq j}^{J} A_{jk}(t)$.
Consider times $s < v < t$. Using the Markov property:

$$P_{jk}(s,t) = \sum_{z=0}^{J} P(X_v = z | X_s = j) \cdot P(X_t = k | X_v = z) \tag{3.25}$$

If v is close to t, the usual interpretation of the transition hazards is $P(X_t = k | X_v = z) \approx \Delta A_{zk}(t), z \neq k$ and consequently $P(X_t = z | X_v = z) \approx 1 + \Delta A_{zz}(t)$ with $\Delta A_{zk}(t) = A_{zk}(t) - A_{zk}(v)$. Putting this into Equation (3.25) leads to

$$P_{jk}(s,t) \approx \sum_{z=0}^{J} P(X_v = z | X_s = j) \cdot (1_k(z) + \Delta A_{zk}(t)). \tag{3.26}$$

The matrix version of Equation (3.26) is

$$P(s,t) \approx P(s,v)(I + \Delta A(t)) \tag{3.27}$$

I denotes the $(J+1) \times (J+1)$ identity matrix and $\Delta A(t)$ the $(J+1) \times (J+1)$ matrix with $(\Delta A(t))_{j,k} = \Delta A_{jk}(t)$.
A fine partition $(t_i)_{i=0,...,L}$ with $s = t_0 < t_1 < ...t_{L-1} < t_L = t$ of the interval $[s,t]$ is considered. Then, an approximation for $P(s,t)$ is

$$P(s,t) \approx \prod_{l=1}^{L} (I + \Delta A(t_l)) \tag{3.28}$$

3.2 Selected methods for model selection

The goal of model selection is to find the model that fits the data "best" from a predefined set of models at hand. In the course of this work, often Cox models that only differ by the set and usage of the covariates are compared. So, this could also be called covariate selection.
The presented methods are the Lasso method, "Akaike's Information Criterion" (AIC) and a stepwise selection procedure by statistical tests.

3.2.1 Lasso method

The Lasso (Least Average and Shrinkage Operator) method is used in the process of model selection to eliminate coefficients in regression analysis. Particularly, this is applied to the Cox model. This method was proposed by Robert Tibshirani in [8].
The regression coefficients are calculated as usual by minimizing the partial log likelihood, but additionally the sum of the absolute values of the regression coefficients is bounded by a constant s that can either be chosen arbitrarily or automatically based on the data. Two methods for automatic choice are described in the next paragraph. The regression coefficients $\hat{\beta}$ are estimated via the following criterion (l denotes the partial log likelihood):

$$\hat{\beta} = \min_{\beta} l(\beta), \quad \text{subject to } \|\beta\|_1 = \sum_{i=1}^{n} |\beta_i| \leq s \tag{3.29}$$

An alternative formulation of the problem is:

$$\hat{\beta} = \min_{\beta}(l(\beta) + \lambda \cdot \|\beta\|_1) \tag{3.30}$$

The tuning parameter λ controls the strength of the penalty. The greater λ is, the more the norm of β is penalized.
An advantage of this method is that often some coefficients are exactly zero, so the related variables are removed from the model and the model becomes smaller and better interpretable.
One way to calculate the constraint s automatically is to use an approximate generalized cross-validation (GCV) statistic. Let l_s be the log partial likelihood for the constrained fit with constraint s and $p(s)$ the effective number of parameters, the GCV-style statistic is constructed as follows:

$$GCV(s) = \frac{1}{N} \cdot \frac{-l_s}{N[1 - p(s)/N]^2} \tag{3.31}$$

Also, an AIC-style criterion can be used. The AIC is described in the following section.

3.2.2 Akaike's Information Criterion (AIC)

This section introduces Akaike's Information Criterion and is based on [9].
Kullback-Leibler (K-L) Information I between the truth f and the model g that approximates the truth is defined as

$$I(f, g) := \int f(x) \ln \left(\frac{f(x)}{g(x|\theta)} \right) dx. \tag{3.32}$$

I denotes the information loss, when g is used to approximate the truth f. It can also be interpreted as distance from g to f. Now we want the model g that loses the least

information regarding to f. Thus, we have to minimize I with fixed f, and g varying over a space of models denoted by θ.
$E_y E_x[\ln(g(x|\hat{\theta}(y)))]$ is the target of all model selection approaches based on K-L information. This results into "Akaike's Information Criterion" (AIC):

$$AIC = -2\log(L(\hat{\theta}|y)) + 2K \tag{3.33}$$

K denotes the number of parameters in the considered model, y denotes the given data, $\hat{\theta}$ denotes the maximum-likelihood estimator (MLE), the expression $\log(L(\hat{\theta}|y))$ is the numerical value of the log likelihood at its maximum point.
In order to compare different models, AIC differences are computed, because the relative values of the AIC are more important than the absolute values. An AIC difference below two is an indicator for substantial support of the according model. The AIC differences are calculated by subtracting the AIC value of the model with the least AIC from the AIC values of each model. The AIC difference for model i is calculated by

$$\Delta_i = AIC_i - AIC_{min}. \tag{3.34}$$

While the AIC differences are used to rank the models, it is also possible to quantify the plausibility of each model as being the actual K-L best model. In order to do this, the concept of likelihood is expanded to the concept of the likelihood of the model given the data, hence $L(g_i|x)$. The likelihood $L(g_i|x)$ of the model g_i is proportional to $\exp(-\frac{1}{2}\Delta_i)$:

$$L(g_i|x) \propto \exp\left(-\frac{1}{2}\Delta_i\right) \tag{3.35}$$

This statement can be used to calculate the so-called Akaike weights w_i given the data and a set of models:

$$w_i = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{r=1}^{R}\exp(-\frac{1}{2}\Delta_r)} \tag{3.36}$$

The Akaike weight w_i is considered the weight of evidence in favor of model i being the actual K-L best model for the situation at hand given a set of R models.
An important aspect regarding the AIC is that models can only be compared when they have been fit to exactly the same set of data.

3.2.3 Tests

Statistical tests can be used for a stepwise selection of the model. It is tested if the extension of a model provides additional information. Two models are nested when the parameter set of one is a subset of the parameter set of the other model. Various test statistics are used. The most commonly used tests are the *likelihoodratio* test, the *score* test and the $Wald$ test which have been presented in Section 3.1.3.

This instructions roughly follows the procedure presented in [10].

- Step 0: The p-values for all covariates in univariable models are determined.
- Step 1: A model containing all variables with a p-value in the univariable model under 0.25 and all variables that are considered important for other reasons is fit.
- Step 2: Covariates with higher p-values from the $Wald$ test might be deleted from the model.
- Step 3: The reduced model is fit and checked, if there is an "important" change in the coefficients. If the excluded variable is an important confounder, it should be added back into the model. This process is continued until no covariate can be deleted.
- Step 4: One at a time, all variables excluded from the initial multivariable model are added to the model to confirm that they are neither statistically significant nor an important confounder.
- Step 5: The scale of the continuous covariates is examined.
- Step 6: The final step in the variable selection process is to determine whether interactions are needed in the model.
- Step 7: It does not become the final model until it is throughly evaluated. Model evaluation should include: checking for adherence to key model assumptions using case wise diagnostic statistics to check for influential observations and testing for overall goodness-of-fit.

4 Data analysis and variables of interest

In this chapter, the results of the analysis of readmission times of data set *dataaut* are presented and compared using various methods of survival analysis and model selection introduced in Chapter 3. The determination of significant parameters for readmissions is the aim. Also, a comparison between different groupings of the diagnoses is performed.

4.1 First readmission

In this section, the goal is to estimate the time span between the initial release and the second admission to hospital. So, the event in the context of this section is the first readmission to the hospital. Thus, the survival function gives the probability of not being readmitted to hospital up to a certain point of time.
In this work, the classical terminology from survival analysis is used and has to be interpreted in the given context.

In the following, the four categories sex, age, length of stay and diagnosis are used as parameters for the models. The parameters and the according abbreviations and types are listed in Table 4.1.

Short	Name	Type
S	Sex	nominal: binary
A	Age	ordinal: integer
L	Length of Stay	ordinal: integer
D	Diagnosis	nominal: 5 groups

Table 4.1: Set of parameters of the full model

Sex is an ordinal variable with two values, male and female. The parameter age at first admission is given in years. The length of stay is the number of days the patient stays at the psychiatric department of the hospital during the initial stay.

The diagnoses are given in ICD-10 code [1]. For the model, they are split in five groups of similar diagnoses:

- D_1: F20-F29
- D_2: F30-F39
- D_3: F40-F48
- D_4: F50-F59
- D_5: F60-F69

4.1.1 Parametric estimators

A Weibull distribution is used to model the given times between the initial release and the first readmission from the dataset *dataaut*. The cumulative hazard function is $H(t) = \left(\frac{t}{b}\right)^a$. The result of the fit for the parameters is $a = 1.02$ and $b = 217.89$. In Figure 4.1, the cumulative hazard function of the fitted Weibull distribution is displayed.

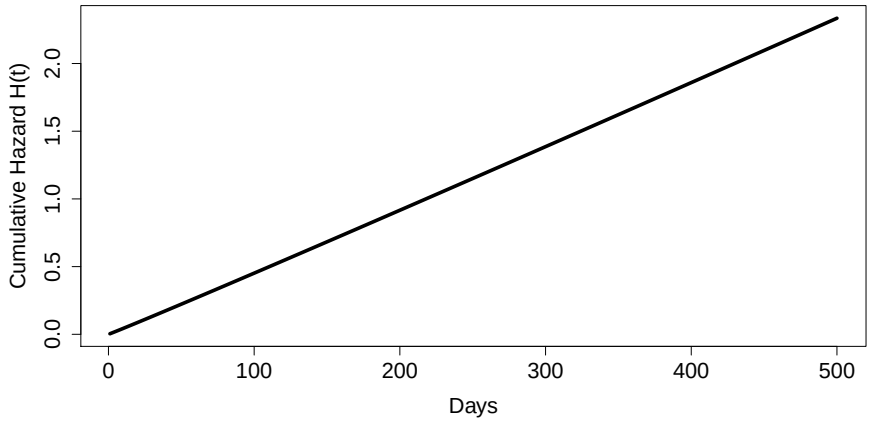


Figure 4.1: Cumulative hazard function for the fitted Weibull distribution

Since a is almost equal to one, the cumulative hazard function is close to a straight line with slope $\frac{1}{b}$, thus the hazard is almost constant over time.

4.1.2 Nonparametric estimators

The Kaplan-Meier method estimates the survival function. It is calculated for the whole population, for each sex separately and for each diagnosis group separately.
In Figure 4.2, the Kaplan-Meier estimate with the two-sided 95% confidence interval for

the whole population is shown. The figure represents an almost exponential decrease of the survival function over time until at about 550 days a value of 0.7 is reached.

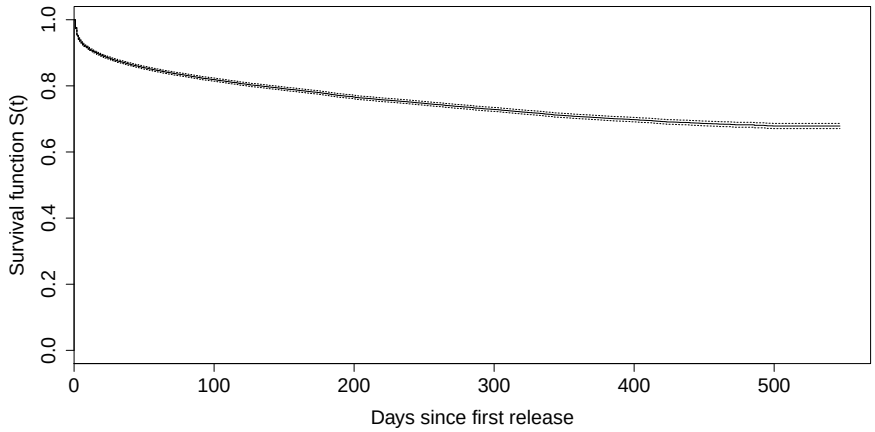


Figure 4.2: Survival curve for the whole population with confidence interval

Figure 4.3 shows a comparison between the Kaplan-Meier estimates for male and female patients from a model with sex as only parameter. The survival curve for females is always lower than the curve for males. That is an evidence that the proportional hazards assumption for the parameter sex holds. According to the log-rank test with a p-value of 0.07 the difference between the curves is almost significant.

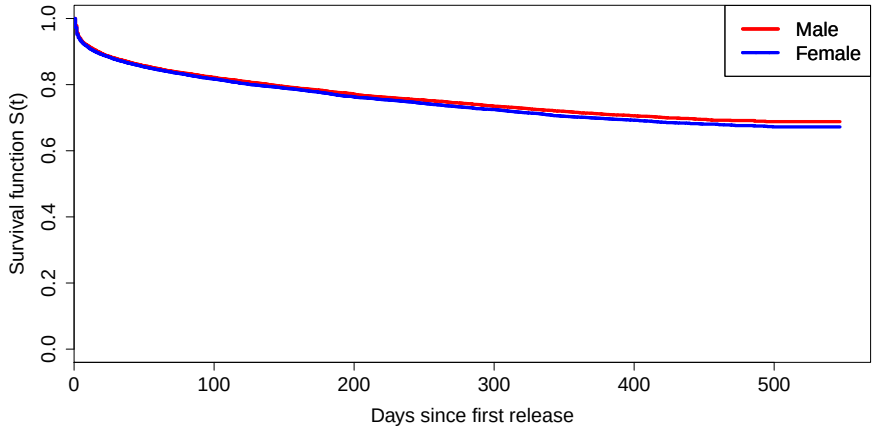


Figure 4.3: Survival curves for each sex

Figure 4.4 shows a comparison between the Kaplan-Meier estimates for the five groups of diagnosis of a model with diagnosis as only parameter.
According to the log-rank test with p-value smaller than 0.01 the differences between

the curves are significant. Diagnosis group F40-48, represented by the green dotted line, always has the highest survival probability. Group F50-59, represented by the blue dash-dotted line, has the lowest survival probability until about 250 days. Afterwards, group F20-29 has the lowest survival probability.
The blue line of group F50-59 crosses two of the other lines which can be an evidence of a violation of the proportional hazards assumption. In the following it is assumed that the possible violation is small enough to be neglected. Another possibility would be to stratify the data in group F50-59 and the remaining other groups of diagnosis.

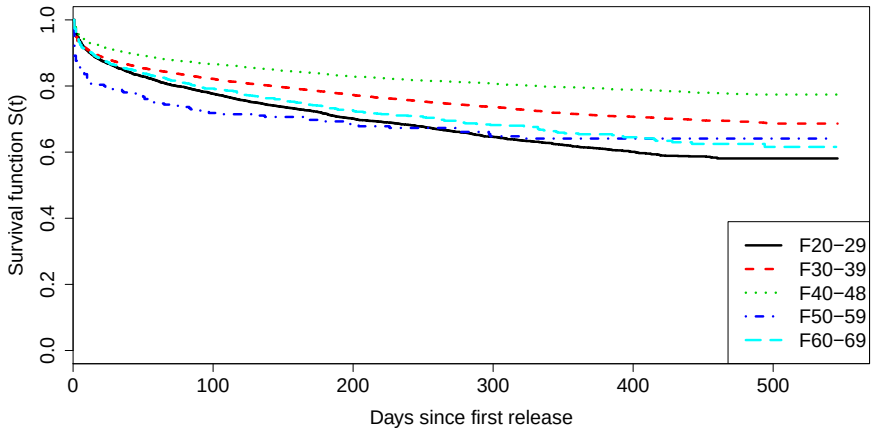


Figure 4.4: Survival curves for each group of diagnosis

In Figure 4.5, the Breslow estimate, which is derived from the Nelson-Aalen estimate for the cumulative hazard function, and the Kaplan-Meier estimate are compared.

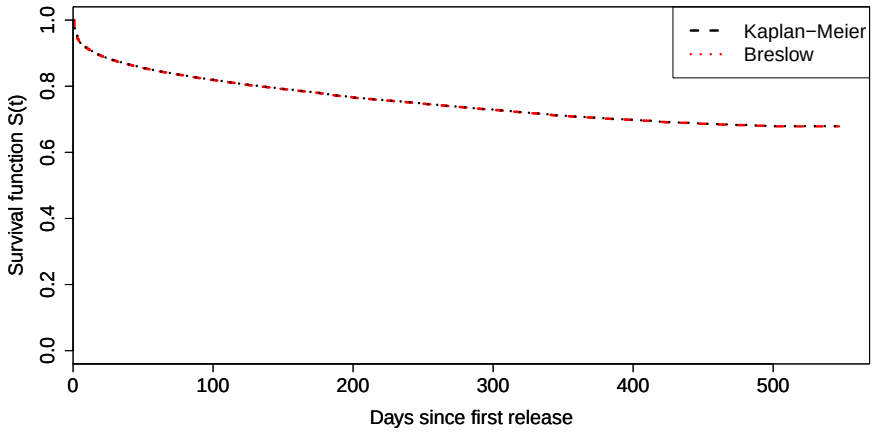


Figure 4.5: Breslow and Kaplan-Meier estimates for the whole population

There is barely a visible difference between the two curves. This observation is consistent with the arguments found in the literature [2]. It is argued that there is no reason not to use the well-known Kaplan-Meier estimate since the estimates are asymptotically equivalent and the differences are very small except for small data samples and a large number of tied survival times. Both limitations are not relevant in this situation.

4.1.3 Cox model

Linear Cox model

The Cox model estimates the hazard function λ. It is performed for all 16 possible subsets of the parameter set of the full model $\{S, A, L, D\}$. In this scenario, only single variables and no interaction terms are considered. For example, for two parameters S and A with values s and a, the Cox model is $\lambda(t) = \lambda_0(t) \cdot \exp(b_S \cdot s + b_A \cdot a)$.
Patients without a readmission enter the model. They are marked as censored with censor time from the first release until the end of the follow-up.
For each parameter, a z-test is performed comparing the given model with the model without the particular parameter. The null hypothesis means that the parameter is zero. So, it can be eliminated from the model without a loss of information. The significance level is set to 0.05. Parameters with p-values between 0.05 and 0.07 are also listed in the tables and marked with the sign *.
In the implementation, the full model has in fact 7 parameters since the categorical parameters have a dummy parameter for each group but one. The coefficient of the dummy parameter for the first group can be set to zero because of the proportional hazards assumption. This means that there is a dummy parameter for females and four for the remaining diagnosis groups.

Model	Significant variables (HR)
Null	
S	$S_2(1.048)^*$
A	$A(0.992)$
L	$L(1.002)$
D	$D_2(0.730)$, $D_3(0.512)$; $D_5(0.895)^*$
SA	$S_2(1.087)$, $A(0.991)$
SL	$L(1.002)$
SD	$S_2(1.069)$, $D_2(0.725)$, $D_3(0.510)$; $D_5(0.898)^*$
AL	$A(0.991)$, $L(1.002)$
AD	$A(0.990)$, $D_2(0.767)$, $D_3(0.492)$, $D_5(0.792)$
LD	$D_2(0.730)$, $D_3(0.515)$; $D_5(0.898)^*$
SAL	$S_2(1.087)$, $A(0.991)$, $L(1.002)$
SAD	$S_2(1.117)$, $A(0.990)$, $D_2(0.761)$, $D_3(0.487)$, $D_5(0.790)$
SLD	$S_2(1.069)$, $D_2(0.727)$, $D_3(0.513)$; $D_5(0.900)^*$
ALD	$A(0.990)$, $D_2(0.769)$, $D_3(0.496)$, $D_5(0.795)$
SALD	$S_2(1.117)$, $A(0.990)$, $D_2(0.763)$, $D_3(0.490)$, $D_5(0.793)$

Table 4.2: Overview of the significant single variables and their type of effect for the Cox models with linear terms

In Table 4.2, an overview of the significant parameters for each of the 16 models is presented. For the categorical parameters, each category is tested separately. Later in Table 4.3, the parameters are tested as a whole which means that the model is compared with the model without all categories of the particular parameter. For instance, the full model is tested against the full model excluding all categories of diagnosis.
For each significant parameter i, the hazard ratio HR = $\exp(\beta_i)$ is given in brackets. It can be seen that the parameters A, D_2, D_3 and D_5 are significant in all models in which they are included. The parameter S_2 is significant in all models in which it is included except for the model SL. The parameter L is only significant in some models. The hazard ratio is lower than one for significant parameters A, D_2, D_3 and D_5 and greater than one for S_2 and L, when they are significant.

Model	Significant parameters
Null	
S	S
A	A
L	L
D	D
SA	S, A
SL	L
SD	S, D
AL	A, L
AD	A, D
LD	D
SAL	S, A, L
SAD	S, A, D
SLD	S, D
ALD	A, D
SALD	S, A, D

Table 4.3: Overview of the significant parameters for the Cox models with linear terms

In Table 4.3, the significant overall parameters are presented. Table 4.3 shows that the parameter D is also significant in all models in which it appears as a whole. The results for the other parameters do not change.

Lasso method

In this subsection, the results of the Lasso method for the linear Cox model are presented. In Figure 4.6, the coefficients of the linear Cox model are plotted over the l_1-norm of the

coefficients vector. On the upper x-axis the number of non-zero coefficients is displayed. The coefficient that vanishes at last is the coefficient of the parameter D_3, right after A and D_2. The other parameters are eliminated earlier.

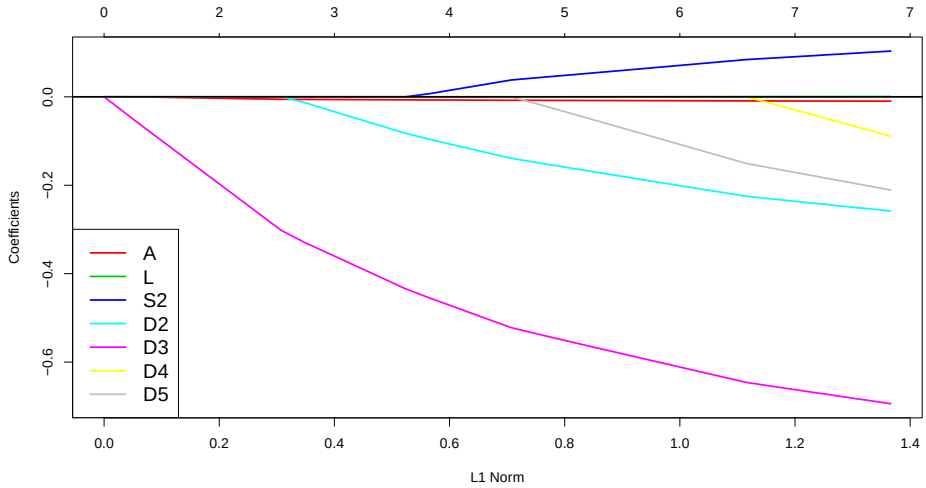


Figure 4.6: Lasso-plot for the full linear Cox model

A tenfold cross-validation (CV) is performed for the linear Cox model. In Figure 4.7, the partial likelihood deviance is plotted over the logarithm of λ with confidence intervals.

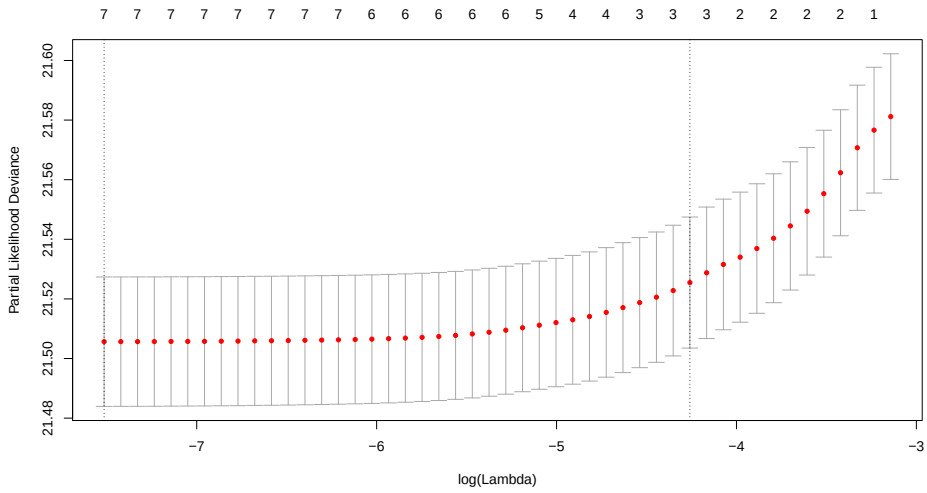


Figure 4.7: Tenfold cross-validation deviance for Lasso-method for full linear Cox model

The left vertical line in the plot shows where the CV-error curve hits its minimum. This minimum error occurs for $\lambda = 0.0005$. The coefficients for this value of λ are all non-zero. The right vertical line shows the most regularized model with CV-error within

one standard deviation of the minimum and it is roughly at $\lambda = 0.0141$ with non-zero coefficients $b_A = -0.0061$, $b_{D_2} = -0.0331$ and $b_{D_3} = -0.3594$.

4.1.4 Cox model with interaction terms

In this scenario, interaction terms between the parameters also are considered. So, a set of parameters enters the model like in the previous setting, but additionally to the single parameters, also all possible products of the parameters enter the model as parameters. For example, for two parameters S and A with values s and a, the model is $\lambda(t) = \lambda_0(t) \cdot \exp(b_S \cdot s + b_A \cdot a + b_{SA} \cdot s \cdot a)$.
Results for single variables are displayed in Table 4.4 and for whole parameters in Table 4.5.

Model	Significant variables (HR)
SA	$S_2(1.206)$, $A(0.993)$
SL	
SD	$D_2(0.653)$, $D_3(0.478)$, $D_4(0.281)$, $D_5(0.751)$, $S_2{\cdot}D_2(1.204)$, $S_2{\cdot}D_4(4.424)$, $S_2{\cdot}D_5(1.392)$
AL	$A(0.993)$, $L(1.005)$, $A{\cdot}L(0.999)$
AD	$A(0.988)$, $D_2(0.674)$, $D_3(0.390)$, $A{\cdot}D_3(1.006)$
LD	$D_2(0.678)$, $D_3(0.459)$, $D_4(1.355)$, $D_5(0.800)$, $L{\cdot}D_2(1.003)$, $L{\cdot}D_3(1.005)$, $L{\cdot}D_4(0.990)$, $L{\cdot}D_5(1.005)$
SAL	$S_2(1.262)$, $A(0.996)$, $L(1.007)$, $S_2{\cdot}A(0.996)$, $A{\cdot}L(0.999)$
SAD	$A(0.991)$, $D_2(0.644)$, $D_3(0.369)$; $S(1.320)^*$, $D_4(0.051)^*$, $S_2{\cdot}A(0.994)^*$
SLD	$D_2(0.601)$, $D_3(0.419)$, $D_5(0.700)$, $S_2{\cdot}D_2(1.224)$, $S_2{\cdot}D_4(5.868)$, $L{\cdot}D_3(1.006)$; $D_4(0.281)^*$, $S_2{\cdot}D_5(1.331)^*$
ALD	$A(0.989)$, $D_2(0.609)$, $D_3(0.346)$, $D_4(2.020)$; $A{\cdot}D_3(1.006)^*$
SALD	$D_2(0.609)$, $D_3(0.321)$, $S_2{\cdot}A(0.991)$; $S_2{\cdot}A{\cdot}D_5(1.023)^*$

Table 4.4: Overview of the significant single variables and their type of effect for the Cox models with interaction terms

Like in Table 4.2, the second column shows all significant parameters with significance level 0.05. Parameters with p-value between 0.05 and 0.07 are marked with *. For each significant parameter i, the hazard ratio $\text{HR} = \exp(\beta_i)$ is given in brackets.
Again, the parameters D_2 and D_3 are significant in all models and their effect on the hazard is decreasing. S_2 and L are significant in only two models. Parameter A is significant in all models except for the full model and the effect on the hazard is decreasing. The parameters A and D_2 have a decreasing impact on the hazard for their own in

models AD and ALD, but the product of these two parameters has an increasing impact on the hazard.

Model	Significant variables
SA	S, A
SL	S, L
SD	S, D
AL	A, L
AD	A, D
LD	L, D
SAL	S, A, L
SAD	S, A, D
SLD	S, L, D
ALD	A, L, D
SALD	S, A, L, D

Table 4.5: Overview of significant parameters for the Cox models with interaction terms

In Table 4.5, the results of the *Wald* tests for the whole parameters show that for the interaction models all included parameters are significant.

Lasso method

In this section, results of the Lasso method for the model with interaction are presented.

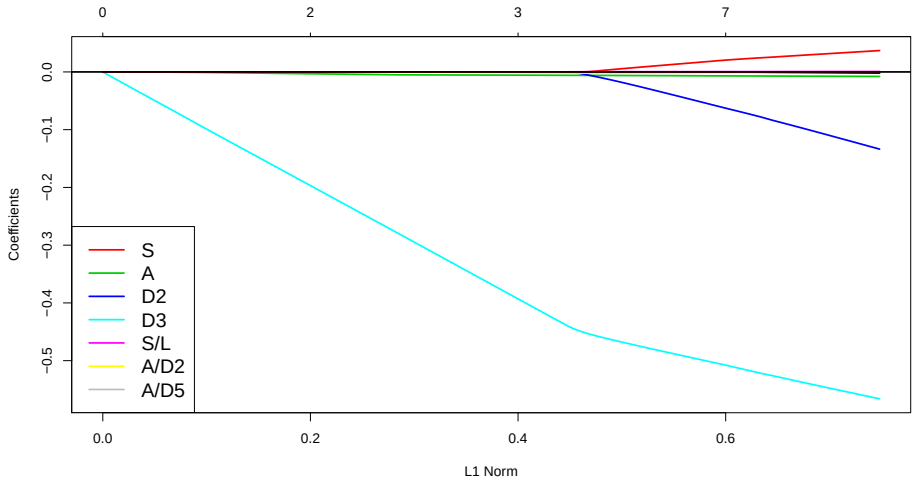


Figure 4.8: Lasso-plot for full model with interaction terms

In Figure 4.8, the coefficients of the Cox model with interaction terms are plotted over the l_1-norm of the coefficients vector. On the upper x-axis, the number of non-zero

coefficients is displayed. A range for the norm from 0 to roughly 0.8 is shown in the plot.
From the overall 39 coefficients, only seven are still non-zero, when the norm is 0.8. The coefficient that vanishes at last is the coefficient of parameter D_3, right after $A{\cdot}D_2$ and A. The other parameters are eliminated earlier.
A ten-fold cross validation (CV) is performed for the Cox model. In Figure 4.9, the partial likelihood deviance is plotted over the logarithm of λ with confidence intervals.

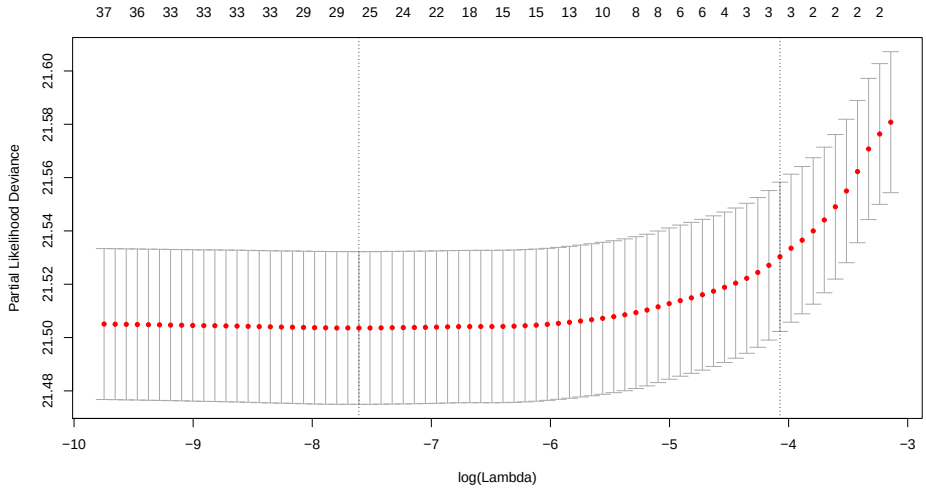


Figure 4.9: Tenfold cross-validation deviance for Lasso-method for full model with interaction terms

The left vertical line in the plot shows where the CV-error curve hits its minimum. This minimum error occurs for $\lambda = 0.0005$. For this value of λ 26 coefficients are non-zero. The right vertical line shows the most regularized model with CV-error within one standard deviation of the minimum and it is roughly at $\lambda = 0.0170$ with non-zero coefficients $b_A = -0.0052$, $b_{D_3} = -0.3128$ and $b_{A{\cdot}D_2} = -0.0052$.

4.1.5 Comparison of Cox models with and without interaction terms

The AIC and the AIC differences are calculated for all Cox models with and without interaction. In Table 4.6 it can be seen that for the models without interaction the full model and the model SAD have the lowest AIC by far. All the other models can be excluded from the set of plausible models according to this analysis.
For models with interaction terms, the full model has the lowest AIC followed by the models SAD and ALD within a reasonable range. Again, all the other models can be excluded from the set of plausible models. The full model with interaction also has the lowest AIC overall.

Model	AIC lin	AIC diff lin	AIC inter	AIC diff inter
Null	-			
SA	128515.5	353.3	128515.1	362.9
SL	128650.9	488.7	128652.1	499.9
SD	128332.2	170.0	128318.8	166.6
AL	128510.7	348.5	128506.1	353.9
AD	128178.9	16.7	128176.3	24.1
LD	128338.2	176.0	128323.3	171.1
SAL	128520.2	358.0	128497.9	345.7
SAD	128162.2	0	128155.5	3.3
SLD	128333.4	171.2	128311.2	159
ALD	128179.3	17.1	128161.6	9.4
SALD	128162.7	0.5	128152.2	0

Table 4.6: Rating of the models with and without interaction with AIC

In Table 4.7, a comparison of the significant variables of the models with and without interaction terms is shown.
In general, the parameters that are significant in the linear setting are also significant in the setting with interaction given the same set of parameters, but it occurs that parameters that are significant in the linear model are not significant anymore in the model with interaction, for example the parameter S_2 is significant in almost all linear models but only in two models with interaction. There are also parameters like D_4 that are not significant in any linear model, but are significant in some models with interaction.

Model	Linear	Interaction
Null		
S	$S_2(+)^*$	$S_2(+)^*$
A	$A(-)$	$A(-)$
L	$L(+)$	$L(+)$
D	$D_2(-)$,$D_3(-)$,$D_5(-)^*$	$D_2(-)$,$D_3(-)$,$D_5(-)^*$
SA	$S_2(+)$,$A(-)$	$S_2(+)$,$A(-)$
SL	$L(+)$	
SD	$S_2(+)$,$D_2(-)$,$D_3(-)$,$D_5(-)^*$	$D_2(-)$,$D_3(-)$,$D_4(-)$,$D_5(-)$,$S_2{\cdot}D_2(+)$, $S_2{\cdot}D_4(+)$,$S_2{\cdot}D_5(+)$
AL	$A(-)$,$L(+)$	$A(-)$,$L(+)$,$A{\cdot}L(-)$
AD	$A(-)$,$D_2(-)$,$D_3(-)$,$D_5(-)$	$A(-)$, $D_2(-)$, $D_3(-)$, $A{\cdot}D_3(+)$
LD	$D_2(-)$,$D_3(-)$,$D_5(-)^*$	$D_2(-)$,$D_3(-)$,$D_4(+)$,$D_5(-)$,$L{\cdot}D_2(+)$, $L{\cdot}D_3(+)$,$L{\cdot}D_4(-)$,$L{\cdot}D_5(+)$
SAL	$S_2(+)$,$A(-)$,$L(+)$	$S_2(+)$,$A(-)$,$L(+)$,$S_2{\cdot}A(-)$;$A{\cdot}L(-)$
SAD	$S_2(+)$,$A(-)$,$D_2(-)$,$D_3(-)$,$D_5(-)$	$A(-)$,$D_2(-)$,$D_3(-)$;$S_2(+)^*$, $D_4(-)^*$,$S_2{\cdot}A(-)^*$
SLD	$S_2(+)$,$D_2(-)$,$D_3(-)$,$D_5(-)^*$	$D_2(-)$,$D_3(-)$,$D_5(-)$,$S_2{\cdot}D_2(+)$, $S_2{\cdot}D_4(+)$,$L{\cdot}D_3(+)$;$D_4(-)^*$,$S_2{\cdot}D_5(+)^*$
ALD	$A(-)$,$D_2(-)$,$D_3(-)$,$D_5(-)$	$A(-)$,$D_2(-)$,$D_3(-)$,$D_4(+)$,$A{\cdot}D_3(+)^*$
SALD	$S_2(+)$,$A(-)$,$D_2(-)$,$D_3(-)$,$D_5(-)$	$D_2(-)$,$D_3(-)$,$S_2{\cdot}A(-)$;$S_2{\cdot}A{\cdot}D_5(+)^*$

Table 4.7: Comparison of the significant variables and their type of effect of the Cox models without and with interaction terms

4.1.6 Different diagnosis groups

Additional diagnosis group

After receiving feedback from experts who indicate that diagnosis $F30$ and $F31$ may differ in behavior from the other diagnoses in group D_2, this group is split up in two groups. One group consists of the diagnoses $F30$ and $F31$, the other consists of diagnoses $F32$ to $F39$.
So, the diagnoses are divided into six groups:

- D_1: F20-F29
- D_2: F32-F39
- D_3: F40-F48
- D_4: F50-F59

- D_5: F60-F69
- D_6: F30-F31

In Table 4.8, the significant variables and the AIC for the univariate Cox model with parameter D and the full models with and without interaction terms for this setting are shown.
It can be seen that D_2 and D_3 are significant in all three models. The full model with interaction has the lowest AIC.

Model	Significant variables (HR)	AIC
D	$D_2(0.682)$,$D_3(0.512)$;$D_5(0.895)^*$	128266
SALD	$S_2(1.120)$,$A(0.990)$,$D_2(0.715)$,$D_3(0.490)$,$D_5(0.080)$	128104
SALD interaction	$D_2(0.517)$,$D_3(0.321)$,$S_2{\cdot}A(0.991)$;$S_2{\cdot}A{\cdot}D_5(1.023)^*$	128093

Table 4.8: Overview of the significant variables and their type of effect and the AIC for the Cox models with six diagnosis groups

Single diagnosis groups

Another scenario is to categorize every subgroup Fxy separately. This leads to 36 factors for the parameter D in the analysis, because there are 37 different diagnoses in the data set. Table 4.9 shows the results for this model.

Model	Significant variables (HR)	AIC
D	$D_{22}(-)$,$D_{23}(-)$,$D_{32}(-)$,$D_{33}(-)$,$D_{34}(-)$,$D_{39}(+)$,$D_{40}(-)$,$D_{41}(-)$, $D_{43}(-)$,$D_{44}(-)$,$D_{45}(-)$,$D_{48}(-)$,$D_{51}(-)$,$D_{60}(-)$,$D_{61}(-)$,$D_{63}(-)$	128158
SALD	$S_2(+)$,$A(-)$,$D_{22}(-)$,$D_{23}(-)$,$D_{32}(-)$,$D_{33}(-)$,$D_{34}(-)$,$D_{39}(+)$, $D_{40}(-)$,$D_{41}(-)$,$D_{43}(-)$,$D_{44}(-)$,$D_{45}(-)$, $D_{48}(-)$,$D_{51}(-)$,$D_{60}(-)$,$D_{61}(-)$,$D_{63}(-)$	127800

Table 4.9: Overview of the significant variables and their type of effect and the AIC for the Cox models with single diagnosis groups

4.2 Multiple readmissions

4.2.1 Cox model

A patient can have multiple readmissions to hospital. A Cox model with multiple, consecutive events is used to model this scenario. The conditional model approach is used. That means the events are distinguishable from each other, so, they are stratified by the number of the event.
Table 4.10 shows the results for the Cox model with multiple events.

Model	Significant variables (HR)	AIC
S	$S_2(+)$	208175
A	$A(-)$	207796
L	$L(-)$	208179
D	$D_2(-)$, $D_3(-)$, $D_4(+)$	207942
SA	$S_2(+)$, $A(-)$	207950
SL	$S_2(+)$,$L(-)$	208166
SD	$S_2(+)$, $D_2(-)$, $D_3(-)$, $D_4(+)$	207926
AL	$A(-)$, $L(-)$	207972
AD	$A(-)$, $D_2(-)$, $D_3(-)$, $D_5(-)$;$D_4(+)^*$	207723
LD	$L(-)$, $D_2(-)$, $D_3(-)$, $D_4(+)$	207920
SAL	$S_2(+)$, $A(-)$, $L(-)$	207946
SAD	$S_2(+)$, $A(-)$, $D_2(-)$, $D_3(-)$, $D_5(-)$	207689
SLD	$S_2(+)$, $L(-)$, $D_2(-)$, $D_3(-)$, $D_4(+)$	207903
ALD	$A(-)$, $L(-)$, $D_2(-)$, $D_3(-)$, $D_5(-)$	207704
SALD	$S_2(+)$, $A(-)$, $L(-)$, $D_2(-)$, $D_3(-)$, $D_5(-)$	208798

Table 4.10: Overview of the significant variables and their type of effect and the AIC for the Cox models with multiple events

S_2, A, D_2 and D_3 are significant in all models, D_4 and D_5 in some of the models. S_2 and D_4 have an increasing effect on the hazard, while all the other parameters have a decreasing effect.
The comparison of the AIC values shows that the model SAD has the lowest AIC value.

4.3 Overview of significant variables

An overview of the significant variables of the linear Cox model, the Cox model with interaction terms and the Cox model with multiple events is presented.
Figure 4.10 shows in how many models each of the single parameters is significant. Every

single parameter appears in 23 models. The parameters D_2 and D_3 are significant in all of the 23 models, A in all but one model.

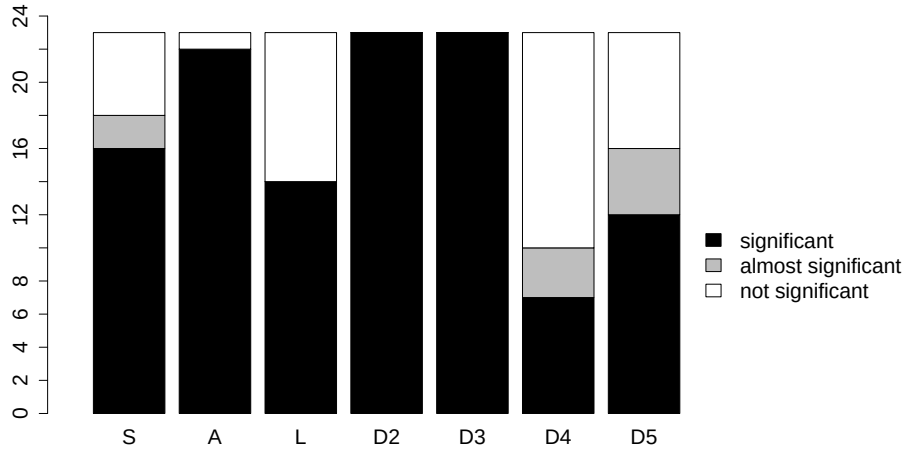


Figure 4.10: Counts of significant appearances of single parameters

Every product of two parameters appears in four models. In Figure 4.11, all parameters that are significant in at least one model are represented by a bar. Parameter S_2/A is significant the most often with two significant and one almost significant appearances.

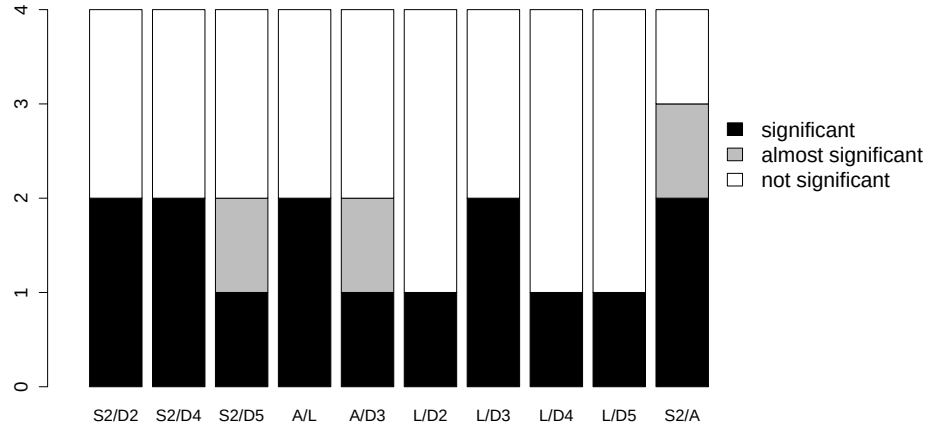


Figure 4.11: Counts of significant appearances of interaction parameters

5 Model

5.1 Microsimulation models

These models for health care issues are thoroughly described in [11].
Microsimulation models are mostly used to compare and evaluate different intervention strategies or scenarios by modeling a decision problem on a micro-unit level. Micro-units are, for example, single individuals or small groups of individuals.
Microsimulation or first-order Monte Carlo models follow a bottom-up approach, so they are individual-based. These models do not capture interactions between individuals and model a closed cohort. Microsimulation models are well suited to represent heterogeneity among the population since the covariates of an individual represent the specific attributes.
Formally, there are states, transitions between the states, transition probabilities, a cycle length and logical tests performed at the beginning of each cycle to determine the transitions. The states are mutually exclusive. A modeled individual must be in exactly one state at any time. The transitions are evaluated using first-order Monte Carlo simulation by comparing a random number with the given transition probability. A cycle is a predefined interval of time on which the transition probabilities are based. The transition probabilities are also depending on the covariates of the individuals. Microsimulation models generally have a fixed time horizon, for example, an average lifetime.
So-called tracker variables keep track of the path of each individual through the states in order to be able to analyze and visualize it afterwards. It is possible to do cross-sectional and longitudinal analyses on the results of microsimulation models.

5.2 General description

A simulation model for the prognosis of the development of the treatment status of a cohort of patients with chronic mental diseases is built. Different scenarios and policies are examined. Cross-sectional as well as longitudinal studies are of interest. Thus, the evolution of aggregate numbers is analyzed as well as the pathways of individuals through the system in time. The course of the events of a patient is modeled independently from other persons.
The chosen model type is a microsimulation model. That means that it follows the

bottom-up approach and every single individual is modeled. This approach is chosen because not only the cross-sectional analysis is important but also the longitudinal pathways of single individuals. Furthermore, this approach is suitable for the analysis of different policies and scenarios. Another reason is that the characteristics of the individuals are manageable with a bottom-up approach.
The events of a patient are expressed by state changes. The possible ways through the states are described by a transition matrix which can be interpreted as a directed acyclic graph. Every individual starts in state R (released after the first admission to hospital). If the most recent event of the patient was the ith readmission, it is in state A_i, if the most recent event was the ith ambulant psychiatrist visit, the patient is in state P_i and if the patient died, it is in state D. From initial state R transitions are possible to the states P_1, A_1 and D. From P_i the patient can go to A_i and D and from A_i to P_{i+1}, A_{i+1} and to D. From death no transitions are possible since it is an absorbing state.
In order to calculate the times of the events respectively the probabilities for the events to occur the hazard and survival functions have to be modeled. The methods from Chapter 3.1 are applied to determine the according statistical models, especially the Cox model, the Nelson-Aalen estimate and model selection methods. A stratified Cox model is calculated to estimate the hazard function for the transitions. The strata represent the transitions.
The model has a stochastic aspect since random numbers are used to decide if an event occurs.
The individuals are processed sequentially. This is possible since there is no interaction between the individuals in the model. The overall simulation time is fixed. The simulation starts for every patient with the day of the release from the first stay in a psychiatric department of a hospital. The simulation is executed in discrete time steps of one day.
The starting population is sampled from real data described in Chapter 2. It is modeled as a closed cohort, so there is no change in the size of the population except for deaths. Data is available for times of readmissions, ambulant visits to a psychiatrist and deaths. It depends on the chosen scenario which events are actually considered in the model. Patient parameters are age, sex, length of stay in the psychiatric department and the diagnosis made during the initial stay at the hospital.

5.3 Technical description and implementation

The model is fully implemented in R. The advantage of R is that many common methods of survival analysis are already implemented in packages. The package *survival* contains functions for the Cox regression (*coxph*), the Kaplan-Meier estimate (*survfit*) and the

Nelson-Aalen estimate (*bazehaz*). Additionally, a package *mstate* for the multi-state extension of the Cox model is available. This allows to transform data from wide format with one line per subject into long format with one row for each observation. This is needed to put the data into the function *coxph* which computes the Cox regression coefficients for each possible transition. The possible transitions are assigned to a transition matrix which can be generated via the function *msprep*. A positive integer in the (i, j)-element of the matrix indicates a possible transition.
At first the data has to be prepared. The data files with the patient data and readmission times and with the data of the visits to the psychiatrist are imported as data frames *dataaut* and *datapsy* from csv-files. After filtering out incomplete data sets, they are converted to data tables and merged. All dates are converted to integers beforehand manually. Although R can handle dates, this makes it much easier to handle the data since the function for the data preparation for the Cox model only takes integers as input. Also, the maximum number of readmissions z must be specified in advance.
For the multi-state extension of the Cox model, which is performed during the simulation, four matrices are needed: a covariate matrix, a transition matrix, a event time matrix and a status matrix. The creation of this matrices is described in the following. The actual design of the matrices always depends on the given scenario that is simulated. The covariate matrix can be extracted directly from the file *dataaut*.
The transition matrix is generated using the function *transMat* which takes a list of vectors with possible transitions from each state as input. The matrix is a $z \times z$ matrix and the possible transitions are numbered.
The event times are gathered from the two different data tables *dataaut* and *datapsy*. Every column in the event times matrix corresponds to a specific state. The times of each patient are chronologically ordered and put into the correct columns. There are z possible readmissions and z possible visits to the psychiatrist, one is possible before each readmission. There is also an initial state in which a patient is before any event happens and the state of death. So, there are $2z+2$ states in total. The next psychiatrist contact after the last recorded admission is also included, if there is one, and multiple contacts between two consecutive admissions are dismissed. Missing times from the data sets are denoted by NA in the first place and then replaced by the time of the censoring event which is either the end of the follow-up or death for readmissions and can additionally be the following readmission for visits to the psychiatrist. Finally, the date of the initial release from hospital is subtracted from all dates to get the times since the initial release. Negative values resulting from data errors are left out.
The status of the events can easily be derived since all entries in the event matrix before the adding of the censoring times which are indicated by NA are censored and therefore have status 0.
Depending on the chosen scenario, the transition matrix, the event time matrix and the status matrix are created. The covariate matrix is the same for all scenarios. When

those four matrices are prepared, they are transformed into long format by the function *msprep*. Then, the cumulative hazards are estimated by a Cox regression stratified by transition number.
Before the simulation starts, the number of runs and the duration of the simulation are specified. Then, the starting cohort for the simulation is generated. Also, the categorical covariates are converted from factors to multiple columns of indicator variables. The baseline hazard functions for each stratum and the hazard differences are also calculated in advance.
The actual simulation is implemented by two nested for-loops. The outer one loops over the individuals and the inner one over the time steps. In the outer for-loop for each subject, the hazard differences for the particular individual are calculated by multiplying the baseline hazard differences with $\exp(X\beta_i)$. Then, in each time step, the actual state of the individual is determined and for all possible transitions the transition probabilities are calculated. If a transition takes place is determined with the use of random numbers. Two matrices keep track of the visited states and the according state entry times of each subject to analyze and visualize the results after the simulation.

For the model selection, the Lasso-method is implemented in the *glmnet* package. It also provides a plotting routine for the results. The *waldtest* function of the *lmtest* package is used for $Wald$ tests for nested models. For reasons of speed and practicability, the package *data.table* is used. Most of the data is stored in the data table format. Data tables still also are data frames but there are slight differences in handling and faster methods for data manipulation. The *ggplot2* package is used for generating more sophisticated plots such as stacked area plots.
All calculations and plots are performed by RGui 3.1.0 and RStudio 0.98.953.

5.4 Validation and determination of sample size

5.4.1 Validation

The validation of the model is an important part of model building process. The validation is performed for the most extended scenario $3a$. The simulated time span is two years respectively 730 days.
The patient sample is partitioned into two disjoint sets. One fitting set for parametrization of the model with the Cox model and one set for the simulation and validation. The fitting set is created by sampling randomly a certain percentage of the data sample. Three runs are performed. The first run $F60/V40$ is executed with 60 percent of the data belonging to the fitting set, the second run $F50/V50$ with equally sized sets and the last run $F40/V60$ with 40% fitting set.
Four quantities are defined to measure the goodness-of-fit of the model:

- Overall number of events O
- Distribution of the numbers of actual events per patient at the end of the simulation num
- Number of patients that visited each state during the simulation vis
- Number of patients in each state at the end of the simulation $hist$

The comparison of the overall number of events O is a first rough estimate if the results of the simulation and the data are approximately of a similar scale. The number of events per patient num shows in more detail if there are too many or too less events per patient. Even if the overall numbers of events match there could be too many patients with few events and too less with many events or vice versa. The number of visits to a certain state vis shows, if the events are frequented in a similar way. It can be determined if specific events are over- or under frequented. So, the particular event or its hazard function can be investigated in detail. The combination of these quantities is reflected in the fourth quantity $hist$. A combination of an adequate number of events and a correct way through the states leads to a correct distribution of the patients at the end of the simulation. These four quantities are calculated for the results of the simulation and for the real data. Then, the errors regarding these quantities are calculated.
In Table 5.1, the differences in the overall number of events between the simulation and data are presented. For all three runs, the error is about 2%, so it does not depend on the actual choice of the sets.

Error	$F60/V40$	$F50/V50$	$F40/V60$
O	$700(2.5\%, n = 27622)$	$507(2.2\%, n = 23162)$	$421(2.3\%, n = 18373)$

Table 5.1: Differences of the overall numbers of events for all three validation runs

In Table 5.2, an overview of the maximum deviation and the mean deviation in percent is given for the vectorial quantities num, vis and $hist$ for each validation run. For num, only up to five events are considered since the number of patients with more than five events is very low. Also, for vis only the first four readmission states, the first four psychiatrist contact states and death are taken into account since too few patients visited the other states and every patient visited state R, so the error for state R is always zero. For $hist$ the same states as for vis, except for the fourth readmission and fourth psychiatrist contact states which are visited by a low number of patients are considered. The mean deviation for all three quantities shows no significant variations.

	$F60/V40$	$F50/V50$	$F40/V60$
num	8.5/3.4	19.2/6.5	11.7/7.0
vis	13.1/4.8	10.4/3.6	20.2/5.0
hist	13.1/7.2	10.1/6.1	20.2/5.4

Table 5.2: Maximum and mean errors for *num*, *vis* and *hist* for all three validation runs in percent

Figure 5.1 shows the mean deviation in percent for each run from Table 5.2 visually. The mean errors are not differing more than two percent points between the runs for each quantity except for *num* where the value of the run $F60/V40$ is lower. This indicates that the errors do not depend on the particular partition into fitting set and validation set and also not on the ratio of the sizes of the two sets.

Figure 5.1: Mean errors in percent for all runs for quantities *num*, *vis* and *hist*

In Figure 5.2, the quantity *vis* for the simulation results and for the data sample for the run $F50/V50$ is compared. The height of each bar shows the number of patients that visited the certain state during the simulation.
The numbers of patients that visited the readmission states is slightly higher in the simulation results except for the fourth readmission A_4. Also, more deaths occur in the simulation. The number of ambulant contacts to the psychiatrist is lower in the simulation. The differences are below five percent of the values for the data for each

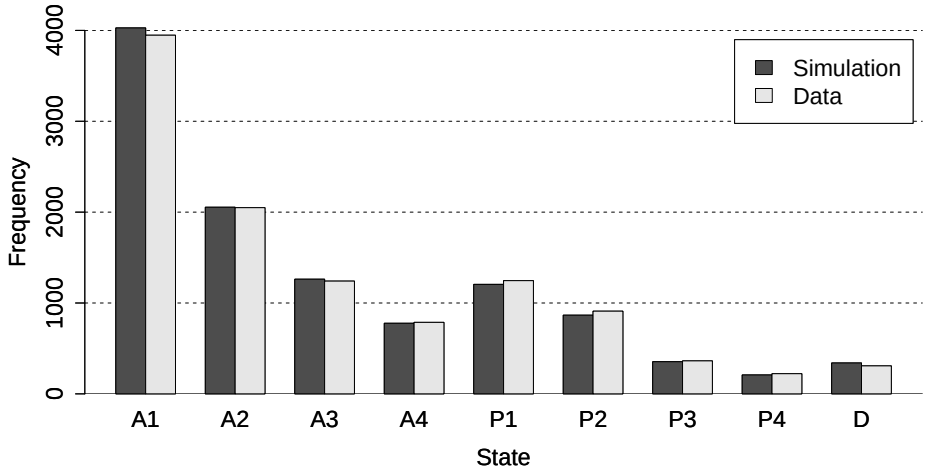


Figure 5.2: Comparison of numbers of patient entries per state between the results of the simulation for run $F50/V50$ and the data set

state except for deaths.

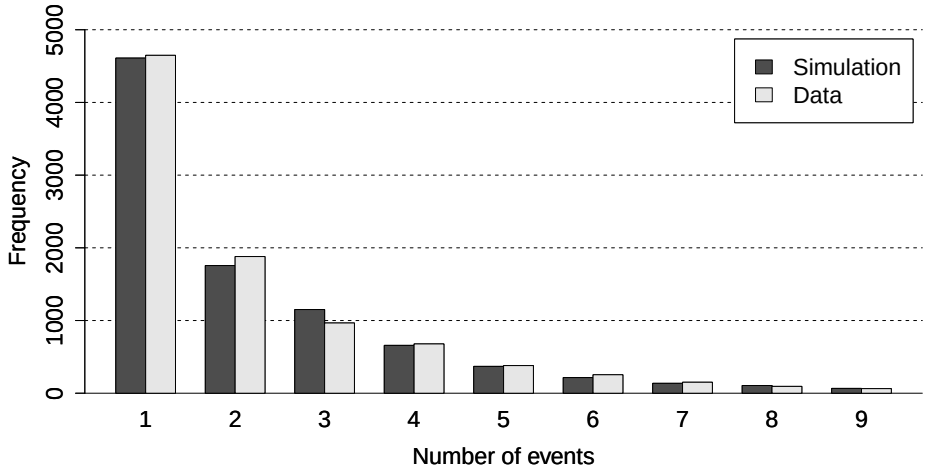


Figure 5.3: Comparison of the distribution of numbers of events over the patients between the simulation results for run $F50/V50$ and the data set

In Figure 5.3, the quantity *num* of the simulation results and of the data sample for the run $F50/V50$ are compared. The height of each bar shows the number of patients that had a certain number of events during the simulation. Every subject has at least one event since the initial release from hospital at time 0 is also counted as an event. So, the first bar shows the number of patients that had no more events after the initial release.

Their number is 0.8% higher in the data than in the simulation. Except for patients with three and eight events the numbers are always slightly higher in the data.
The mean deviation of the number of events per patient is slightly above five percent for all runs. The mean deviation of the visited states is between five and nine percent which is a tolerable, especially because the number of readmissions is overestimated. When the model results are used for planning it is more secure if the results tend to the worse case. It was also shown that the errors are stable independent of the fitting set which is necessary to provide reliable results and predictions.

5.4.2 Determination of the sample size

Random numbers are used to determine the state changes of the individuals in the simulation. In order to reduce the effect of the random numbers on the overall results the size of the population sample has to exceed a certain number. Since there is no interaction between the individuals the required number of patients can be simulated in a single run. A procedure to calculate the required number of samples is presented in [12].
The quantity that is taken into account for the calculations is the overall number of events. The variance of the overall number of events over several runs is calculated as well as the variance of the individuals events number. From the calculations follows that a sample of 7000 patients will provide a sufficiently small standard deviation of 30 events per run.

6 Simulations

6.1 Definition of scenarios

The given datasets contain full records of patients over a time span of up to two years. Often the data at hand is incomplete or contains only information about specific events. This can happen due to data protection issues, loss of data and many other reasons. In order to examine the consequences of using differently detailed patient-level data on result quality, three scenarios are defined. These scenarios only differ in the number and order of the readmissions that are used for the Cox model. So, the scenarios only differ in terms of the parametrization. This leads to different transition probabilities between the states for the different scenarios. The simulation itself however is identical for all scenarios.
Each scenario is executed with and without ambulant contacts to psychiatrists. All simulations are based on the same set of data to ensure comparability of the results.

6.1.1 First readmission only (scenario 1)

In scenario 1, only the first readmission of each patient is considered and all the other readmissions that are available in the data are dismissed. In the simulation, the transition rates from any readmission state A_i to states A_{i+1}, P_{i+1} and D are assumed to be equal for all i.
There are two versions of this scenario considering the visits to the psychiatrist, one with ambulant contacts and one without.

6.1.2 Readmissions without order (scenario 2)

In scenario 2, the first z readmissions of each patient are considered. The same number is used in scenario 3. All readmissions are regarded independently from each other, even if they belong to the same subject. So, there is no order of the readmissions and every readmission is regarded as first readmission. Like in scenario 1, the transition rates from any readmission state A_i to states A_{i+1}, P_{i+1} and D are assumed to be equal for all i in the simulation.
For all events, the event time is the time since the last admission. Between two consecutive admissions up to one contact to a psychiatrist is considered. After the last recorded

admission of a patient one additional psychiatrist contact is considered, if there is one present in the data and if less than z readmissions are recorded. Otherwise, each psychiatrist contact would be followed by a readmission. This would bias the transition rates from any psychiatrist contact state P_i to a readmission state A_i. The contacts are also regarded independently from each other like the readmissions and the event time is the time since the last readmission. Deaths enter the dataset z times, once for each readmission with the time since the particular readmission as event time, in order not to underestimate the number of deaths. Again, there are two versions of this scenario, one with contacts to the psychiatrist ($2a$) and one without any contacts ($2b$).

6.1.3 Readmissions with order (scenario 3)

In scenario 3, the first z readmissions of each patient are considered with the same number z as in scenario 2 but in contrast to scenario 2 the readmissions are ordered. That means that for the first z readmissions the transition rates from a readmission state are independent. From the $(z+1)$th readmission on, the rates are assumed to be equal to the transition rates starting from state A_z. In scenario $3a$, at most one contact to a psychiatrist between two consecutive admissions is possible. Therefore, also the psychiatrist contacts are ordered. In scenario $3b$, no psychiatrist contacts are considered.

In Figure 6.1, the utilization of data in the three scenarios for a time line with three readmissions is presented.

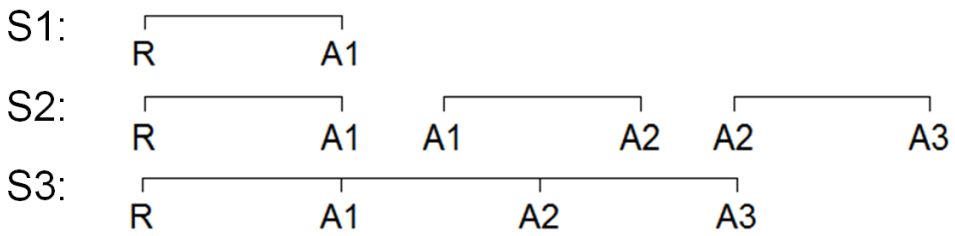


Figure 6.1: Schematic representation of the information needed in the three scenarios

6.2 Results - Austria

In this section, results of simulations for the data from whole Austria are presented. Results for single scenarios and comparisons of the matching scenarios that only differ in the inclusion of psychiatrist contacts and comparisons of all scenarios with and without contacts to a psychiatrist are shown.
A simulation time of 2 years is chosen, because the majority of the readmissions, especially of the first readmissions, which are the most crucial events, happen within this period, but also due to a lack of data for a longer time span. The maximum number of readmissions z is 4. A population of 18638 individuals is sampled from data set *dataaut* containing data of patients from whole Austria. The results of this section are based on a particular population sample denoted by population AT. These settings are valid for all following simulations in this section.
For every scenario, the evolution of the distribution of the patients over the states is shown in a stacked area plot. On the x-axis time is plotted, on the y-axis the percentage of each state is plotted on top of each other. The area under each curve is filled with a distinctive color. The states are coded with colors. Dark green represents always state R, the readmission states are assigned to lighter shades of green, the psychiatrist states have shades of red and dark red represents the state death D.
The percentage of patients in state R is continuously decreasing since this state can only be left but not reentered, state D on the other hand is an absorbing state. So, it can only be entered but not left.
Different subpopulations are considered for each scenario. Male and female patients, three age groups: younger than 45 years, between 45 and 64 years and older than 64 years, and two diagnosis groups, patients with psychotic diseases represented by the ICD-codes $F2x$ and $F30 + F31$ and patients with non-psychotic diseases, are compared. Only selected results are presented in this section. The results are qualitative equivalent for the other scenarios except noted otherwise.
In Table 6.1, the sizes of each subpopulation are presented.

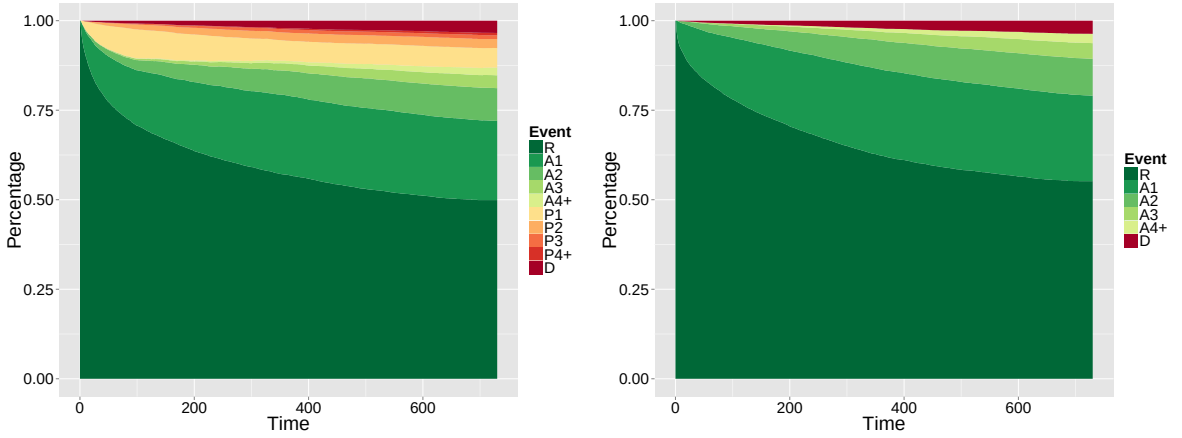
Category	Number	Percentage
male	7796	41.8
female	10842	58.2
< 45	10085	54.1
45-64	6625	35.5
> 64	1928	10.3
psychotic	6995	37.5
non-psychotic	11643	62.5

Table 6.1: Overview of subpopulations of data sample *dataaut*

6.2.1 First readmission only (scenario 1)

In this section, results for the whole population in scenarios 1a and 1b as well as a comparison between psychotic and non-psychotic patients are shown.

Stacked area plots of the evolutions of the patients distribution over the states are presented in Figure 6.2.



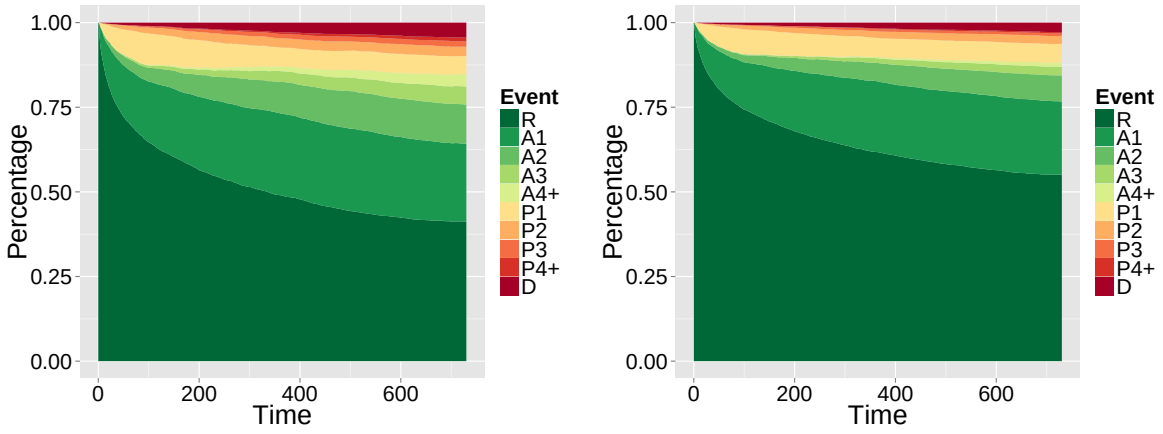
(a) Scenario 1a

(b) Scenario 1b

Figure 6.2: Evolution of the patients distribution over the states in scenarios 1a and 1b

The share of the patients in state R has a similar evolution in both scenarios and decreases almost exponentially. After two years about 50 percent are remaining in state R in scenario 1a, about 55 percent in scenario 1b. The percentage of deaths increases almost linearly for both scenarios. At the end of the simulation around 3.4% of the population died in scenario 1a and around 3.7% in scenario 1b. In scenario 1b always more patients are in states A_i than in scenario 1a which can mostly be explained by the fact that in scenario 1a where more competing events exist.
In the first months many patients enter state P_1 in scenario 1a, after about half a year the number decreases and stays about five percent until the end of the simulation.

The psychotic and the non-psychotic subpopulation are compared in Figure 6.3 in scenario 1a. The percentage of psychotic patients still remaining in state R is decreasing much faster than for non-psychotic. At the end of the simulation 41% of the psychotic patients are still in state R while about 55% of the non-psychotic patients are in state R. The percentages of patients in states P_i run very similarly in both subpopulations.



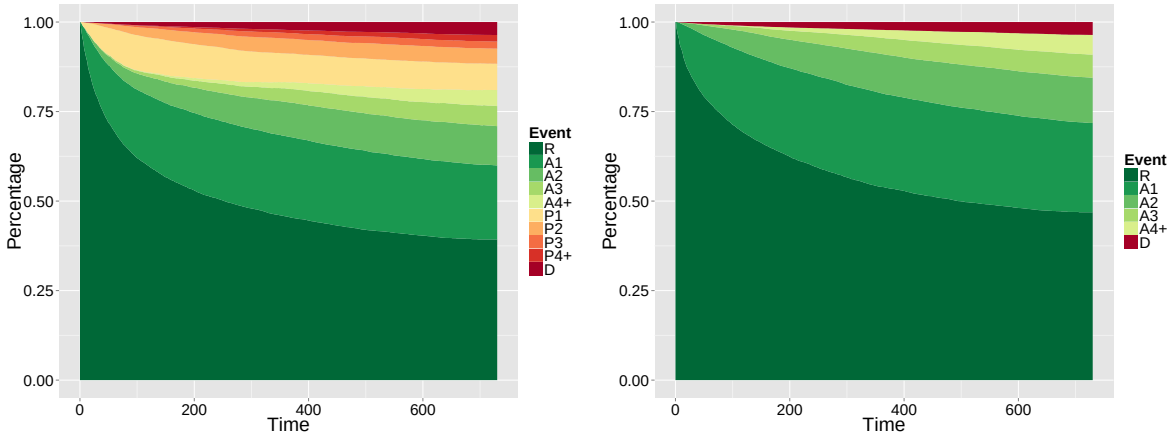
(a) Psychotic patients

(b) Non-psychotic patients

Figure 6.3: Evolution of the patients distribution over the states for psychotic and non-psychotic patients in scenario 1*a*

6.2.2 Readmissions without order (scenario 2)

Results for the whole population in scenarios 2*a* and 2*b* as well as a comparison between female and male patients are presented in this section.
Stacked area plots of the evolutions of the patients distribution over the states are presented in Figure 6.4.



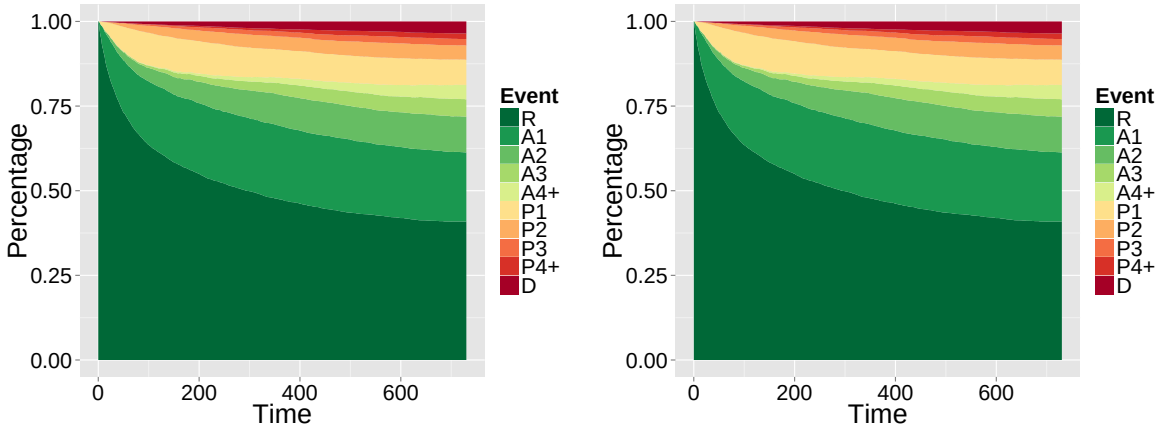
(a) Scenario 2*a*

(b) Scenario 2*b*

Figure 6.4: Evolution of the patients distribution over the states in scenarios 2*a* and 2*b*

The share of the patients in state R has a similar evolution for both scenarios and decreases almost exponentially to about 39% in scenario 2*a* and about 47% in scenario 2*b* after two years. At the end of the simulation around 3.6% of the population died in both scenarios. State P_1 has its maximum number of patients within the first half

year of the simulation. The numbers of states A_1 and P_2 increase until about 18 months and start to decrease afterwards. The numbers of all other states are monotonically increasing during the simulation.
The evolution of the patients distribution over the states for the male and female population is compared in Figure 6.5. The male population has a slightly higher share of patients with no event remaining in state R after two years with about 41% compared to 38%. This difference is due to the fact that a higher percentage of the female population has readmissions while the percentage in the states P_i is roughly the same for both sexes. The percentage of deaths during the simulation is similar with around 3.6% for male and around 3.7% for female patients.



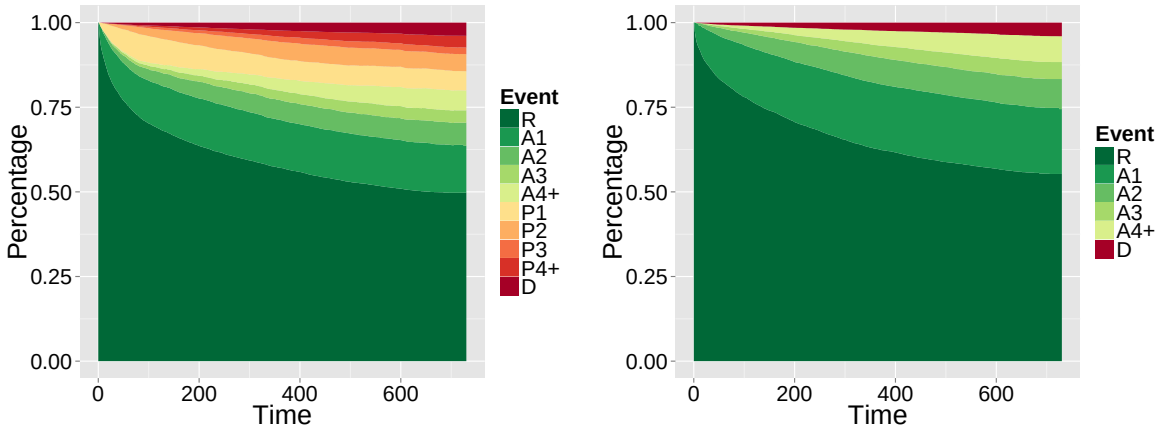
(a) Male patients

(b) Female patients

Figure 6.5: Evolution of the patients distribution over the states for male and female patients in scenario 2a

6.2.3 Readmissions with order (scenario 3)

In this section, results for the whole population in scenarios 3*a* and 3*b* as well as a comparison between patients under 45 and over 64 years are shown.
Stacked area plots of the evolutions of the patients distribution over the states are presented in Figure 6.6. The share of the patients in state R has a similar evolution for both scenarios and decreases almost exponentially. In scenario 3*a*, about 50 percent are remaining in state R at the end of the simulation, in scenario 3*b* about 55 percent. After two years, around 4% of the population died in both scenarios.
The evolutions of the distribution over the states for patients younger than 45 years and older than 64 years are compared in Figure 6.7. The decrease of the numbers of patients in state R goes slightly faster for the younger patients. After two years 49% of the younger patients are in state R and about 53% of the older ones. Also, the number of

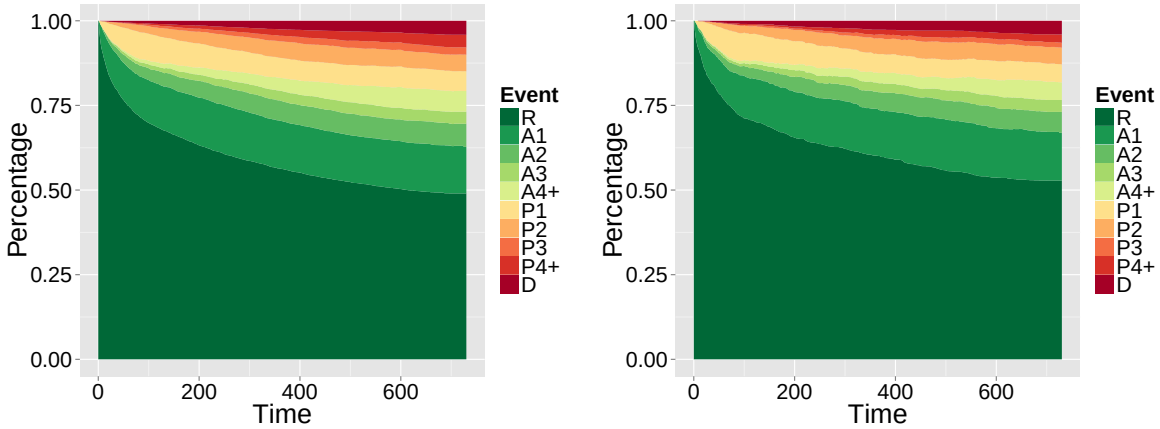


(a) Scenario 3*a*

(b) Scenario 3*b*

Figure 6.6: Evolution of the patients distribution over the states for scenarios 3*a* and 3*b*

deaths at the end of the simulation are similar for both subpopulations with a percentage of about 4.1.



(a) Patients under 45

(b) Patients over 64

Figure 6.7: Evolution of the distribution of the patients over the states for patients under 45 and over 64 in scenario 3*a*

6.2.4 Comparison of scenarios

In this section, numbers and times of events for all scenarios are examined and compared. In Table 6.2, the numbers of patients with readmissions are presented. It can be seen that scenarios $2a$ and $2b$ have a higher percentage of readmissions. The reason is that the transition probability from state R to state A_1 is higher for scenario 2, because in scenario 1 only the first readmissions from the data are used to fit the rate from state R to A_1 while in scenario 2 all readmission times are treated as first readmission times. Since the times for the later readmissions are shorter in average, the median of the first readmission times drops from 75 days for scenarios 1 and 3 to 63 days for scenario 2. This leads to a higher probability for entering state A_1 and having a readmission. Also, the percentage of actual events is higher for scenarios $2a$ and $2b$ in comparison to censored events.

Scenario	$1a$	$1b$	$2a$	$2b$	$3a$	$3b$
Readmissions (%)	42	43	51	51	42	42

Table 6.2: Percentage of patients with readmissions

In the following, statistics for the times of various events are compared for all scenarios.

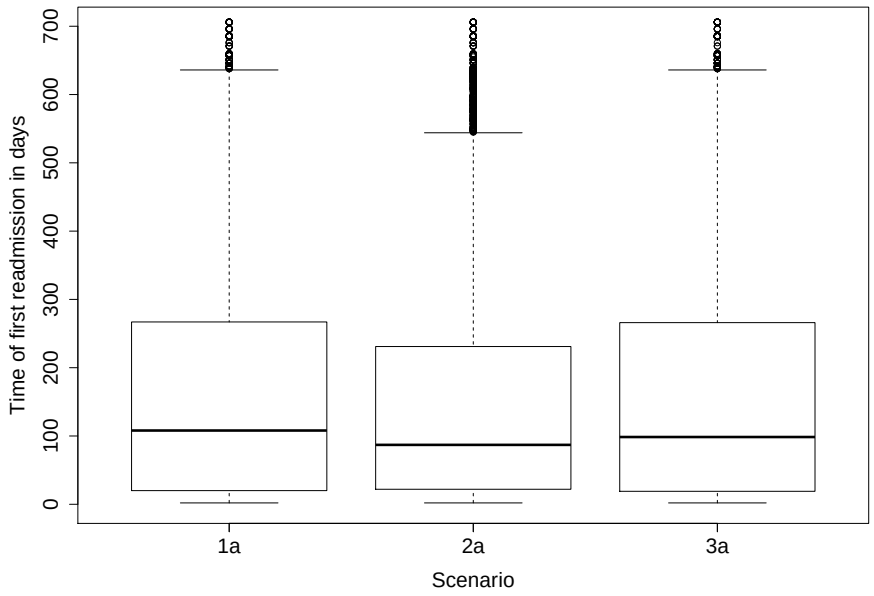


Figure 6.8: Boxplots of the first readmission times for scenarios with psychiatrist contacts

The times of the first readmissions are compared in Figure 6.8 for scenarios with contacts to the psychiatrist and in Figure 6.9 the scenarios without contacts to the psychiatrist. From the former scenarios $2a$ has the lowest median with 87 days followed by $3a$ with 98.5 days and $1a$ with 108 days, from the latter scenarios $2b$ has the lowest median with 84 days followed by $3b$ with 99 days and $1b$ with 100 days.

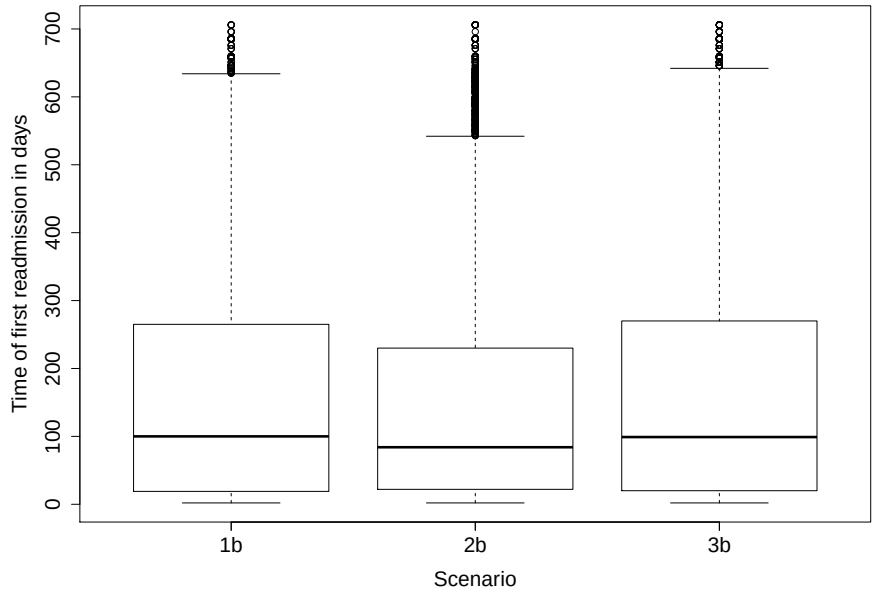


Figure 6.9: Boxplots of the first readmission times for scenarios without psychiatrist contacts

In Table 6.3, the numbers of patients with psychiatrist contacts are presented.

Scenario	$1a$	$2a$	$3a$
Psychiatric contacts (%)	20	28	28

Table 6.3: Percentage of patients with psychiatrist contacts

It can be seen that scenarios $2a$ and $3a$ have a higher percentage of readmissions with 28% than scenario $1a$ with 20%. Again, this can be attributed to the definition of the scenarios and the differing usage of data of those. In scenario $2a$ all ambulant contact times are used to calculate the rates for transitions to state P_1. The median of all psychiatrist times is higher than the median of the first contact times, but also the relative number of actual events is higher for all contacts compared to censored events. So, these two effects level each other regarding scenarios $2a$ and $3a$. In scenario $1a$ are less ambulant contacts, because the transition probabilities for the states P_i after the first readmission are the same as for P_1 in scenario $1a$, but higher in scenario $3a$. So,

patients are more likely to have the first contact to a psychiatrist after some readmissions in scenario 3a.
The times of the first contacts to the psychiatrist are compared in Figure 6.10. For all three scenarios, more than 75% of the first contacts happen during the first 100 days after the initial release. The medians are all around 40 days with the median of scenario 2a being higher with 45 days.

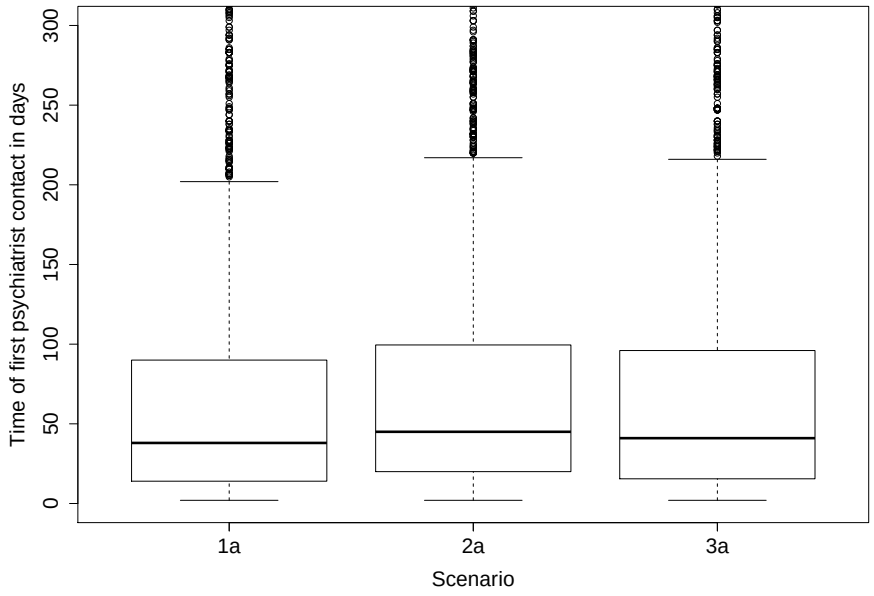


Figure 6.10: Boxplot of the times of the first contact to a psychiatrist

In Table 6.4, the numbers of dead patients are presented. It can be seen that scenarios 3a and 3b have the highest percentage but all scenarios are within a range of 0.6%.

Scenario	1a	1b	2a	2b	3a	3b
Deaths (%)	3.4	3.7	3.6	3.6	3.9	4.0

Table 6.4: Percentage of dead patients

Death times are compared in Figure 6.11. The medians range between 273 days for scenario 2b and 296 days for scenario 1b. In general, more than half of the deaths occur within the first year after the initial release.
In Table 6.5, an overview of the medians of the times of the first readmissions, the first ambulant contacts to a psychiatrist and death is given.
The patients with ambulant contacts to a psychiatrist (OPC) are compared with those without ambulant contacts (non-OPC). In Figure 6.12, the percentage of patients with

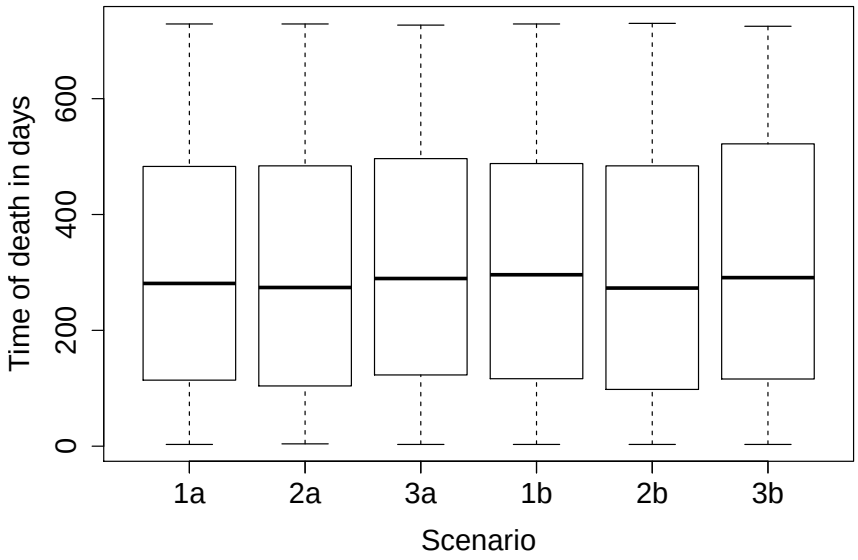


Figure 6.11: Boxplots of the death times for all scenarios

Scenario	$1a$	$1b$	$2a$	$2b$	$3a$	$3b$
First readmission	108	87	98.5	100	84	99
First ambulant contact	38	45	41			
Death	281	296	274	273	289.5	291

Table 6.5: Medians of times of first readmission, the first psychiatrist contact and death for population AT in days

readmissions is shown for both groups. The percentage for the patients with ambulant contacts is much higher, for scenarios $2a$ and $3a$ even twice as much as for patients without ambulant contacts.

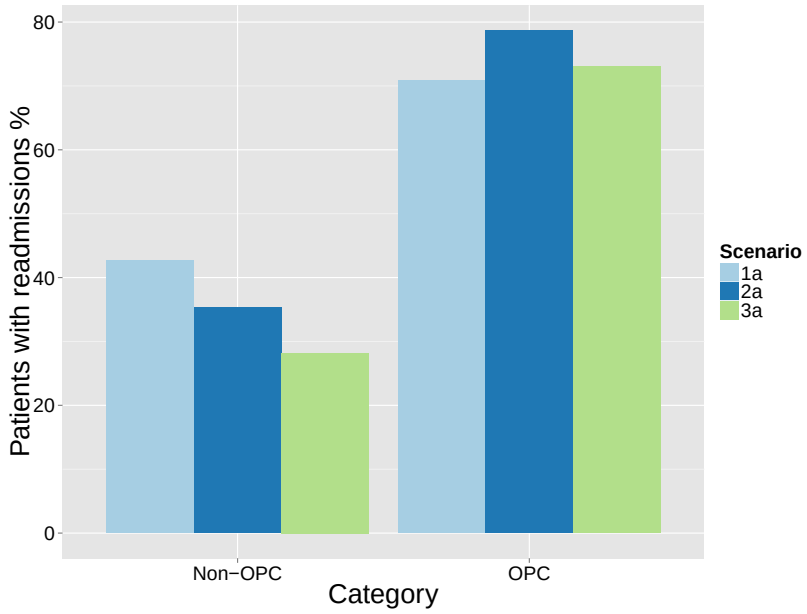


Figure 6.12: Comparison of the percentages of patients with readmissions between patients with and without ambulant treatment

6.2.5 Pathways of patients

Typical pathways of patients during the simulation are characterized in this section to be able to analyze the results of the simulation in more detail. For this purpose, in addition to the number of readmissions of a patient also the times of the readmissions are taken into account to classify the pathways.
The classification for the pathways is based on the given data sets *dataaut* and *datapsy*. Nine typical, distinctive pathways are chosen to split the population in roughly equally sized classes except for the class of patients with no readmission. This class is much bigger than the others.
Nine classes are defined and described in the following. Class 1 consists of patients with no readmissions. Classes 2 to 5 consist of patients with one readmission and differ by the time of the only readmission of the patients. In class 2, patients have their readmissions within the first month after the initial release, in class 3 between the second and sixth month, in class 4 in the second half of the first year and in class 5 in the second year. Classes 6 to 8 consist of the patients with two to four readmissions and differ by the time of the first readmission. In class 6, patients have the first readmission within the first month after the initial release, in class 7 between the second and sixth month and in class 8 after the first half year. The time of the next readmissions is not specified. Class 9 collects all individuals with more than four readmissions.
For the scenarios with psychiatrist contacts, each class is split into the patients with am-

bulant contacts and without any ambulant contact. This doubles the number of classes to a total of 18. The additional classes are numbered consecutively from 10 to 18. Class 10 corresponds to class 1, class 11 to class 2 and the other classes correspond analogously.

The defined pathways for patients are illustrated in Figure 6.13. The red crosses mark readmissions and the time unit is months. For classes 6 to 9, there could also be more readmissions as described in the paragraph above and seen in Table 6.6.

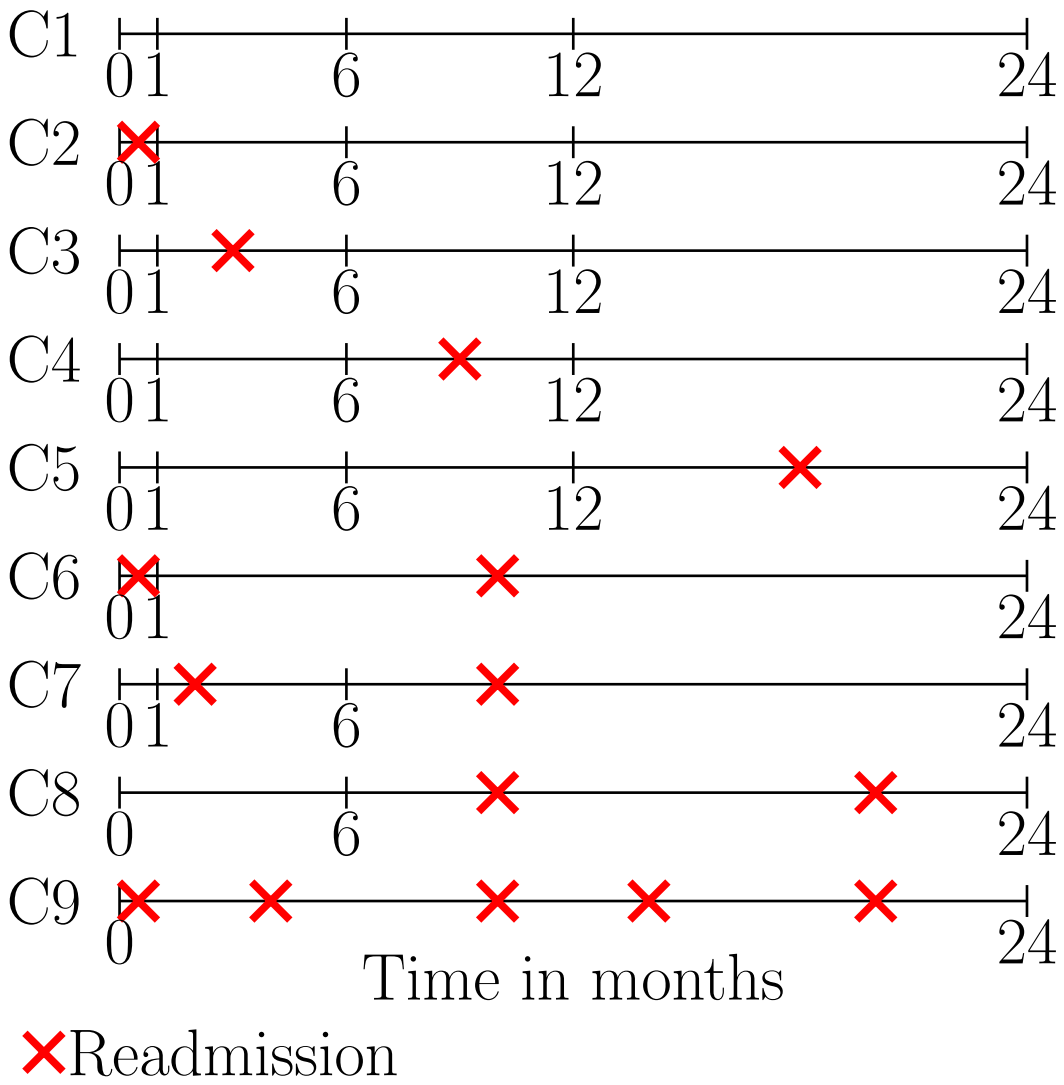


Figure 6.13: Schematic overview of the 9 typical pathways for patients characterized by the time in months and number of the readmissions

In Table 6.6, an overview in tabular form of the classification without ambulant contacts is given.

Class	Number of readmissions	Month of first readmission
1	0	–
2	1	1
3	1	2-6
4	1	7-12
5	1	13-24
6	2-4	1
7	2-4	2-6
8	2-4	7-24
9	> 4	any

Table 6.6: Classification of patient pathways

In Figure 6.14, the sizes of the classes for scenarios $1a$, $2a$ and $3a$ are presented. Classes 1 and 10 are not shown in the plot, because the number of patients without readmission has already been analyzed and the focus is on the patients with readmissions.

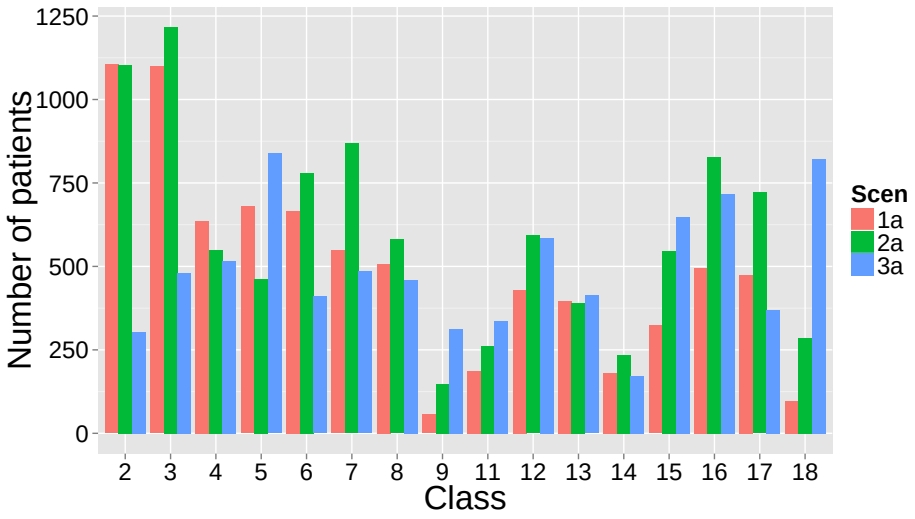


Figure 6.14: Sizes of classes for the scenarios with ambulant contacts

In scenarios $1a$ and $2a$ are more than three times as many patients in classes 2 and 3 than in scenarios $3a$. That means more individuals have exactly one readmission shortly after the release. This can be explained by the fact that in scenario $3a$ more patients have ambulant contacts, so more patients are in classes 10 to 18, and in scenario $2a$ more patients have readmissions, so the overall number of patients in the presented classes is higher. Scenario $3a$ has the most patients with one late readmission after the first year. In scenario $3a$ the number of patients with more than four readmissions is higher than in the other scenarios. In all scenarios, the number of patients with psychiatrist contacts

have a higher number of readmissions in average and the first readmission later.
In Figure 6.15, the sizes of the classes for scenarios 1*b*, 2*b* and 3*b* are presented. Again, class 1 is not shown in the plot, because the number of patients without readmissions has already been analyzed and the focus is on the patients with readmissions.

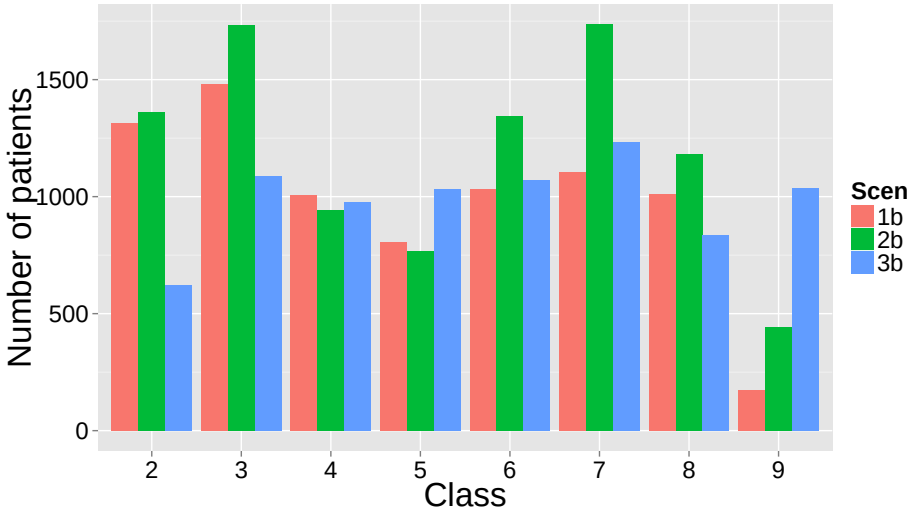


Figure 6.15: Sizes of classes for the scenarios without ambulant contacts

The trends are similar to the scenarios with ambulant contacts. In scenarios 1*b* and 2*b* are more than twice as many patients in class 2 than in scenarios 3*b* and class 3 is also bigger. This can be explained by the fact that in scenario 3*b* patients in average have more readmissions and in scenario 2*b* more patients have readmissions in general. Also, the number of individuals with more than four readmissions is much higher in scenario 3*b*.

6.2.6 Intervention

In order to reduce the number of readmissions, an intervention strategy is examined in this section. According to this strategy a compulsory visit to an ambulant psychiatrist 30 days after every admission to hospital is implemented. The question is, can this strategy reduce the number of readmissions to hospital?
In Table 6.7, the percentages of patients with readmissions, ambulant psychiatrist contacts (OPC) and deaths are compared for scenario 3*a*. The percentage of patients with readmissions is much higher with the intervention strategy. This leads to the conclusion that an ambulant contact increases the probability for a readmission. This could already be seen in the analysis of the pathways. In the intervention scenario almost every patient visits a psychiatrist during the simulation. Only patients who die within the first month

Type of Event	No intervention	Intervention
Readmissions	42.2	67.8
OPC	27.7	99.7
Deaths	3.9	3.6

Table 6.7: Comparison of percentages of the occurrence of events for scenario 3a with and without intervention

have no contact. The number of deaths is slightly lower with the intervention strategy. So, this strategy does not succeed in reducing the number of readmissions.

6.3 Sensitivity analysis

In this section, the influence of the composition of the population on the number and times of readmissions is examined.
Firstly, a base case with a random subpopulation sampled from dataset *dataaut* is considered. Starting from that population various other populations are generated by changing one parameter of all patients at a time while leaving the other parameters of the patients unchanged. 12 populations are generated: all male/female (M, F), all five years younger/older ($A-$,$A+$), length of stay 50% shorter/longer ($L-$,$L+$) and all with each of the six diagnosis groups (D_1 to D_6).

For the 13 populations, the medians of the times of the first readmissions are compared in Figure 6.16. The base case has a median of 105 days. The populations with single diagnosis groups $F4x$ and $F5x$ have the highest medians with 114.5 and 112 days. Population with diagnosis group $F6x$ has the lowest median with 94 days. The female population has more than half of the first readmission within the first 100 days in contrast to the male population. The results for the other populations differ hardly from the base case.
The diagnosis has the biggest influence on the distribution of the first readmission times. Depending on the diagnosis group the times are shifted to earlier or later times in average. The female population has more early first readmissions than the male population. Length of stay and age have very similar distributions to the base case.

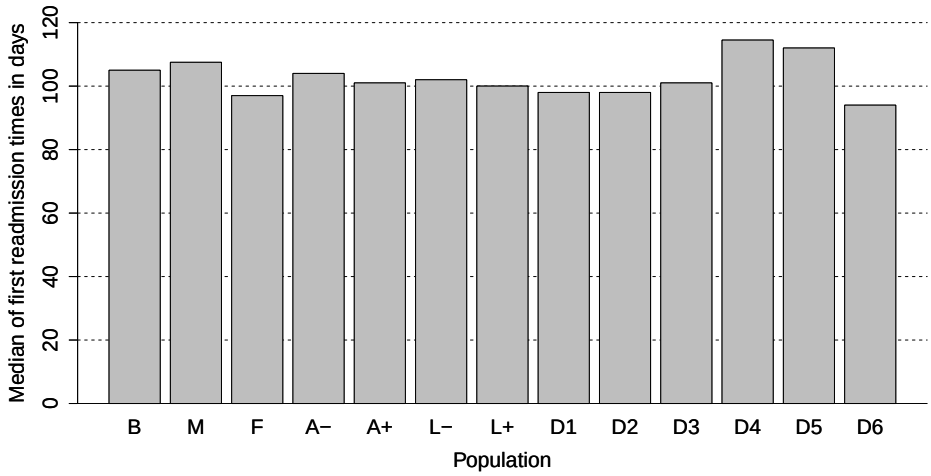


Figure 6.16: Medians of first readmission times for all considered populations

The number of patients with readmissions and the deviation of the number from the base case is also calculated. In Figure 6.17, a tornado plot for the numbers of patients

with readmissions for the given populations is presented. This diagram is a type of a bar chart. The bars are listed vertically and ordered by length. The vertical line at 0 marks the base case with no deviation. The bar for each parameter reaches from the deviation of the highest value to the deviation of the lowest value of the populations where that particular parameter is changed.

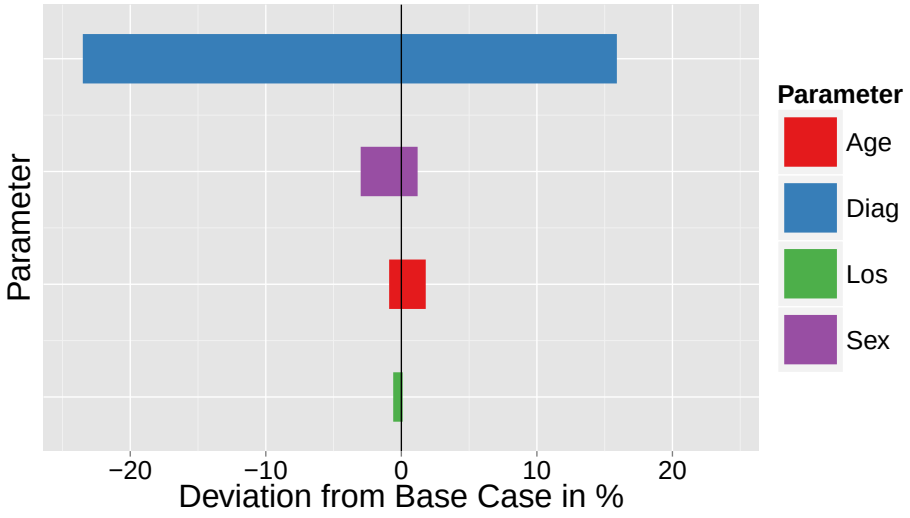


Figure 6.17: Tornado plot for numbers of patients with readmissions

The populations with single diagnosis groups have the highest deviations. The population with diagnosis $F4x$ has about 23% less patients with readmissions than the base case, while the population with diagnosis $F2x$ has about 15% more. Also, the other populations with single diagnosis groups have deviations between 6 and 13 which is not displayed in the tornado diagram, but in Table 6.8. The male population has about 3% less readmissions than the base case, the female has 1.2% more. Also, the older population and the population with longer stays in hospital have a slightly higher number of readmissions compared to the younger patients and the population with shorter stays. All of these deviations are lower than 2% as seen in Table 6.8.

Population	B	M	F	$A+$	$A-$	$L+$	$L-$
Patients with readmissions	4264	4134	4316	4341	4225	4269	4238
Deviation from base case(%)	0	−3.0	1.2	1.8	−0.9	0.1	−0.6
Population		D_1	D_2	D_3	D_4	D_5	D_6
Patients with readmissions		4943	4858	4008	3260	3860	4769
Deviation from base case(%)		15.9	13.9	−6.0	−23.5	−9.5	11.8

Table 6.8: Numbers of patients with readmissions for different populations and the deviation from the base case in percent

In Table 6.8, the numbers of patients with readmissions for all different populations and the deviation from the base case are displayed in percent. The biggest deviations for the variation of each of the parameters are also shown in Figure 6.17.
Again, the parameter diagnosis has the biggest influence on the results with deviations up to 23%. Each population with uniform diagnosis group has a bigger influence on the outcome than any of the considered populations. Sex is the parameter with the second most influence with female population tending to more readmissions and the male population to less. The parameters age and length of stay have little effect on the outcome.

6.4 Results - Lower Austria

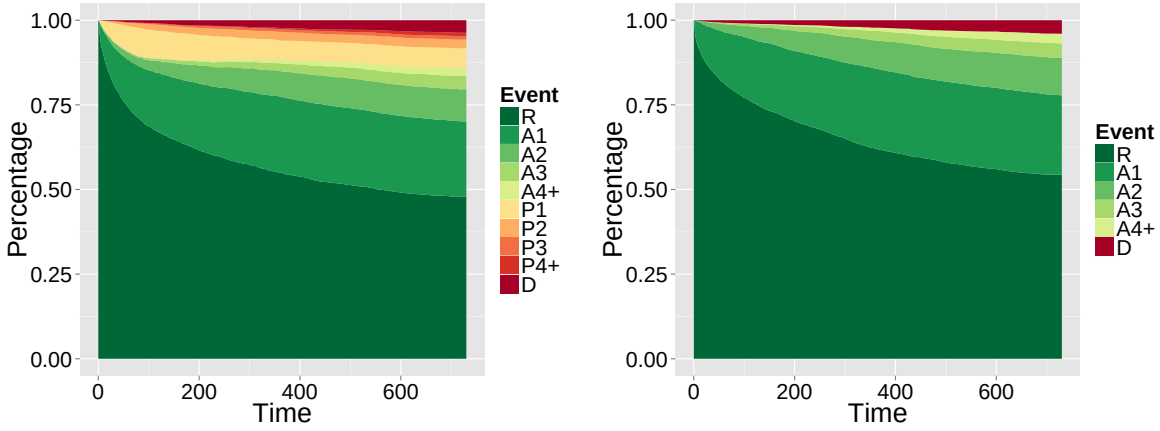
In this section, results of the simulation for data from Lower Austria are presented. Results for single scenarios and comparisons of the matching scenarios that only differ in the inclusion of psychiatrist contacts and comparisons of all scenarios with and without contacts to the psychiatrist are shown.
The general simulation time is 2 years, the population size is 6822 individuals and the maximum number of readmissions z is 4. The population is sampled from the data set *datanoe* containing data from Lower Austria and is denoted by LA. These settings are valid for all following simulations in this section.
For every scenario, the evolution of the distribution of the patients over the states is shown in a stacked area plot. On the x-axis time is plotted, on the y-axis the percentage of each state is plotted on top of each other. The area under each curve is filled with a distinctive color. The states are coded with colors. Dark green represents always state R, the readmission states are assigned to lighter shades of green, the psychiatrist states have shades of red and dark red represents the state death D.
The percentage of the patients in state R is continuously decreasing since this state can only be left but not reentered. State D is an absorbing state and can only be entered but not left.
Different subpopulations are considered for each scenario. Male and female patients, three age groups: younger than 45 years, between 45 and 64 years and older than 64 years, and two diagnosis groups, patients with psychotic diseases represented by the ICD-codes $F2x$ and $F30 + F31$ and patients with non-psychotic diseases, are compared. Only selected results are presented in this section. The results are qualitative equivalent for the other scenarios except noted otherwise.
In Table 6.9, the sizes of each subpopulation of population LA are presented.

Category	Number	Percentage
female	3982	58.4
male	2840	41.6
< 45	3861	56.6
45-64	2235	32.8
> 64	726	10.6
psychotic	2894	42.4
non-psychotic	3928	57.6

Table 6.9: Overview of subpopulations for population LA

6.4.1 First readmission only (scenario 1)

In this section, results for population LA in scenarios $1a$ and $1b$ as well as a comparison between psychotic and non-psychotic patients are shown.
Stacked area plots of the evolutions of the patients distribution over the states are presented in Figure 6.18. The share of the patients in state R has a similar evolution for both scenarios and decreases almost exponentially. After two years about 48 percent are remaining in state R in scenario $1a$, about 54 percent in scenario $1b$. The percentage of deaths increases almost linearly for both scenarios. At the end of the simulation around 3.7% of the population died in scenario $1a$ and around 4.0% in scenario $1b$. In scenario $1b$ always more patients are in states A_i than in scenario $1a$ which can mostly be explained by the fact that in scenario $1a$ there are more competing events.
In the first months many patients enter state P_1 in scenario $1a$, after about half a year the number decreases and stays at about 5.7% until the end of the simulation.

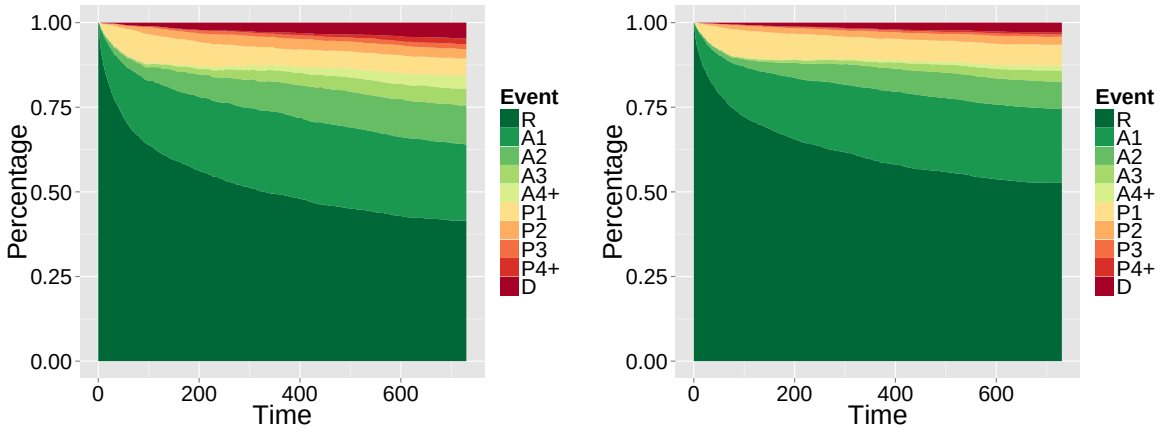


(a) Scenario $1a$

(b) Scenario $1b$

Figure 6.18: Evolution of the patients distribution over the states in scenarios $1a$ and $1b$ for population LA

The psychotic and the non-psychotic subpopulation are compared in Figure 6.19 for scenario $1a$. The percentage of psychotic patients still remaining in state R is decreasing much faster than for non-psychotic. At the end of the simulation 41.5% of the psychotic patients are still in state R while about 53% of the non-psychotic patients are in state R. At the beginning the shares of patients in state P_1 are almost equal for both subpopulations. Towards the end of the simulation relatively more non-psychotic patients are in state R_1 than psychotic patients. Since the psychotic patients tend to have more readmissions, a part of those has a readmission and goes to state A_1.



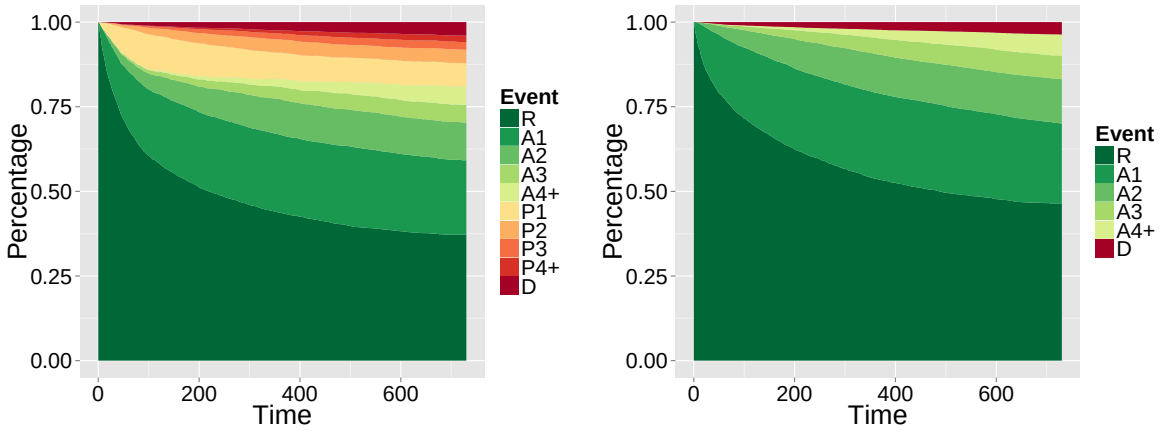
(a) Psychotic patients

(b) Non-psychotic patients

Figure 6.19: Evolution of the patients distribution over the states for psychotic and non-psychotic patients in scenario 1a for population LA

6.4.2 Readmissions without order (scenario 2)

Results for population LA in scenarios 2a and 2b as well as a comparison between female and male patients are presented in this section.
Stacked area plots of the evolutions of the patients distribution over the states are presented in Figure 6.20.



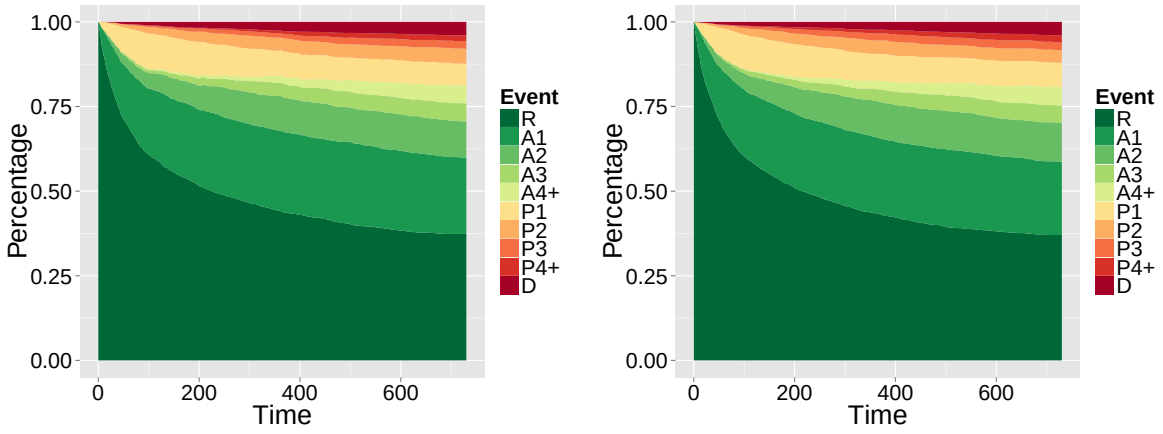
(a) Scenario 2a

(b) Scenario 2b

Figure 6.20: Evolution of the distribution of the patients over the states in scenarios 2a and 2b for population LA

The share of the patients in state R has a similar evolution for both scenarios and decreases almost exponentially to about 37% in scenario 2a and about 46% in scenario 2b after two years. At the end of the simulation around 4.0% of the population are dead

in scenario 2*a* and around 3.7% in scenario 2*b*. State P_1 has its maximum number of patients within the first half year of the simulation. The number of state A_1 increases until about 18 months and then starts to decrease in both scenarios. The number of all other states are monotonically increasing during the simulation.
The evolution of the patients distribution over the states for the male and female population is compared in Figure 6.21.



(a) Male patients

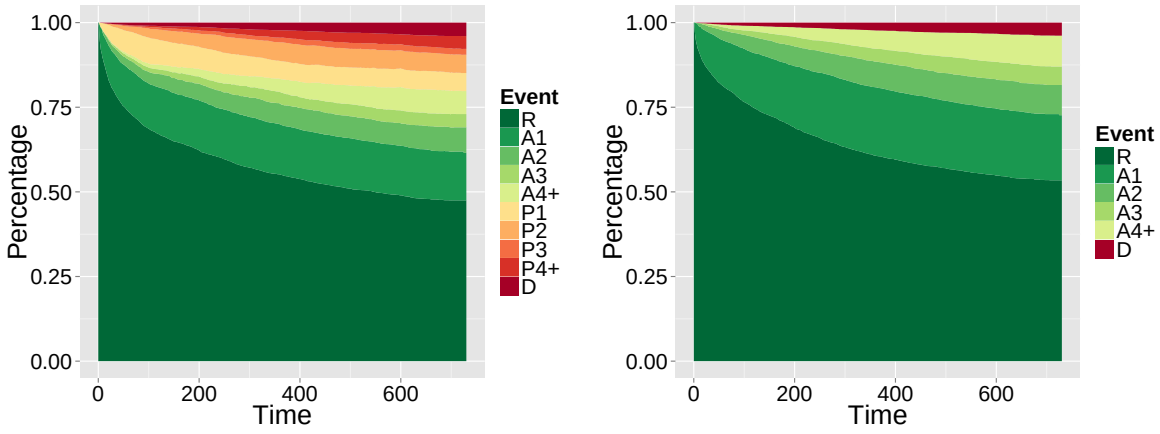
(b) Female patients

Figure 6.21: Evolution of the patients distribution over the states for male and female patients in scenario 2*a* for population *LA*

The share of the patients in state R has a similar evolution for both scenarios and decreases almost exponentially to about 37% for both subpopulations after two years. The percentage in the states P_i is roughly the same for both sexes. The percentage of deaths during the simulation is similar with around 4.0% for both subpopulations.

6.4.3 Readmissions with order (scenario 3)

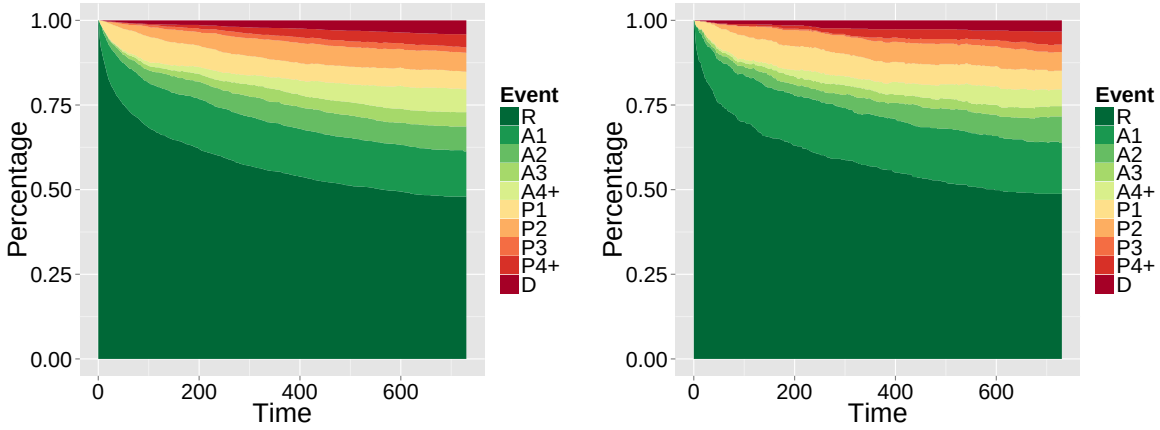
In this section, results for population *LA* in scenarios 3*a* and 3*b* as well as a comparison between patients under 45 and over 64 years are shown.
Stacked area plots of the evolutions of the patients distribution over the states are presented in Figure 6.22. The share of the patients in state R has a similar evolution for both scenarios and decreases almost exponentially. In scenario 3*a*, about 47 percent are remaining in state R at the end of the simulation, in scenario 3*b* about 53 percent. After two years, around 4.0% of the population died in scenario 3*a* and around 3.9% in scenario 3*b*.
The evolutions of the distribution over the states for patients younger than 45 years and older than 64 years are compared in Figure 6.23. The decrease of the numbers of patients in state R goes slightly faster for the younger patients. After two years 48% of



(a) Scenario 3*a* (b) Scenario 3*b*

Figure 6.22: Evolution of the distribution of the patients over the states in scenarios 3*a* and 3*b* for population *LA*

the younger patients are in state *R* and about 49% of the older ones. In percent, more younger patients than older one die during the simulation.



(a) Patients under 45 (b) Patients over 64

Figure 6.23: Evolution of the patients distribution over the states for patients under 45 and over 64 in scenario 3*a* for population *LA*

6.4.4 Comparison of scenarios

In this section, numbers and times of events for all scenarios are examined and compared.

In Table 6.10, the numbers of patients with readmissions are presented. It can be seen that scenarios $2a$ and $2b$ have a higher percentage of readmissions. This is due to the definition of scenarios as already explained in Section 6.2.4 about the simulation for population AT.

Scenario	$1a$	$1b$	$2a$	$2b$	$3a$	$3b$
Readmissions (%)	44	43	54	52	45	45

Table 6.10: Percentage of patients with readmissions for population LA

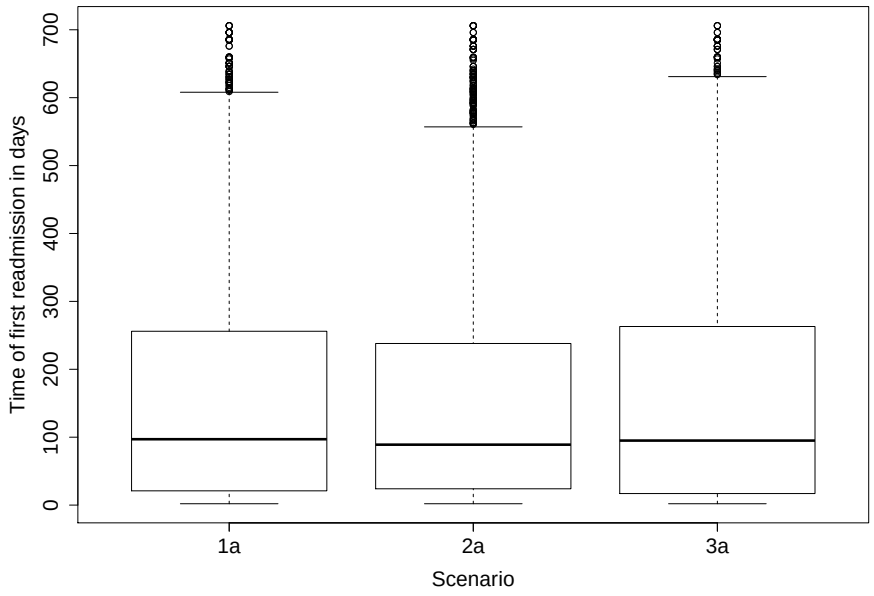


Figure 6.24: Boxplots of the first readmission times for scenarios with psychiatrist contacts for population LA

The times of the first readmissions are compared in Figure 6.24 for scenarios with contacts to the psychiatrist and in Figure 6.25 the scenarios without contacts to the psychiatrist. From the former scenarios $2a$ has the lowest median with 89 days followed by $3a$ with 95 days and $1a$ with 97 days, from the latter scenarios $2b$ has the lowest median with 85 days followed by $1b$ with 95 days and $3b$ with 96 days.
In Table 6.11, the numbers of patients with psychiatrist contacts are presented. It can be seen that scenarios $2a$ and $3a$ have a higher percentage of readmissions with 28%

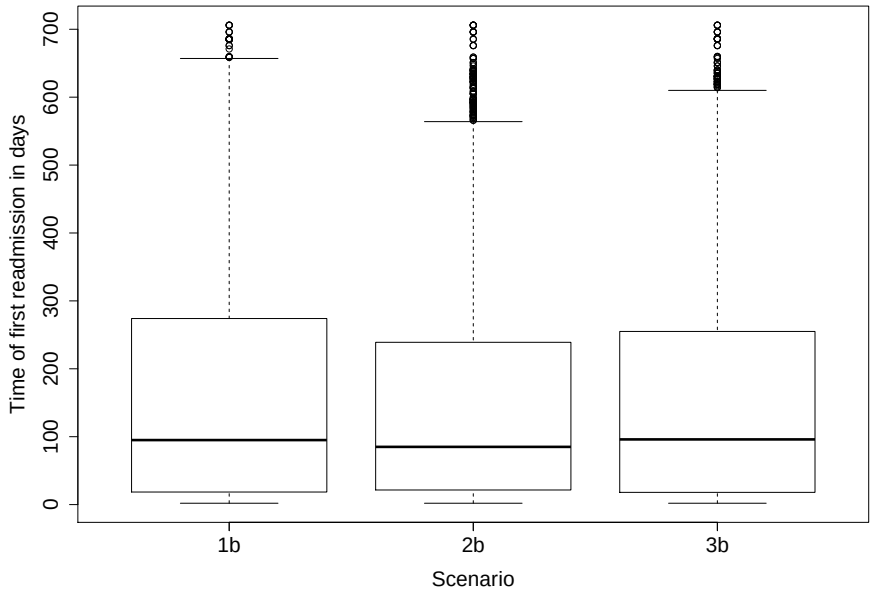


Figure 6.25: Boxplots of the first readmission times for scenarios without psychiatrist contacts for population LA

respectively 29% than scenario $1a$ with 22%. This is due to the definition of scenarios as already explained in Section 6.2.4 about the simulation for population AT.

Scenario	$1a$	$2a$	$3a$
Psychiatric contacts (%)	22	28	29

Table 6.11: Percentage of patients with psychiatrist contacts for population LA

The times of the first contacts to the psychiatrist are compared in Figure 6.26. For all three scenarios, more than 75% of the first contacts happen during the first 100 days after the initial release. The medians are all around 40 days with the median of scenario $2a$ being a litte higher with 44 days.
In Table 6.12, the numbers of dead patients are presented. It can be seen that scenarios $3a$ and $3b$ have the highest percentage but all scenarios are within the range of 0.3 percent points.

Scenario	$1a$	$1b$	$2a$	$2b$	$3a$	$3b$
Deaths (%)	3.7	4.0	4.0	3.7	4.0	3.8

Table 6.12: Percentage of dead patients

The death times are compared in Figure 6.27. The medians range from 265 days in

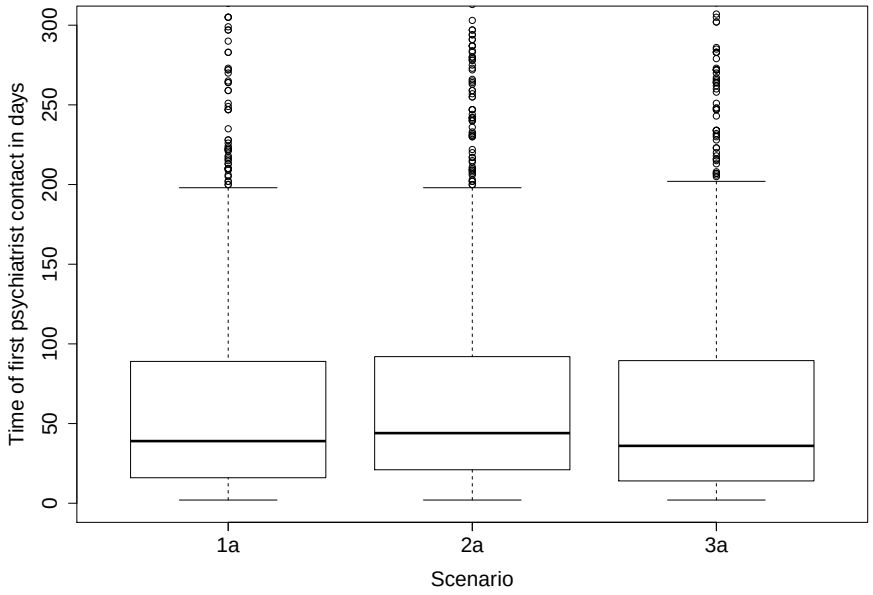


Figure 6.26: Boxplot of the times of the first contact to a psychiatrist for scenarios with psychiatrist contacts for population LA

scenario $2b$ to 323 days in scenario $1b$. In general, more than the half of the deaths occur within the first year after the initial release.

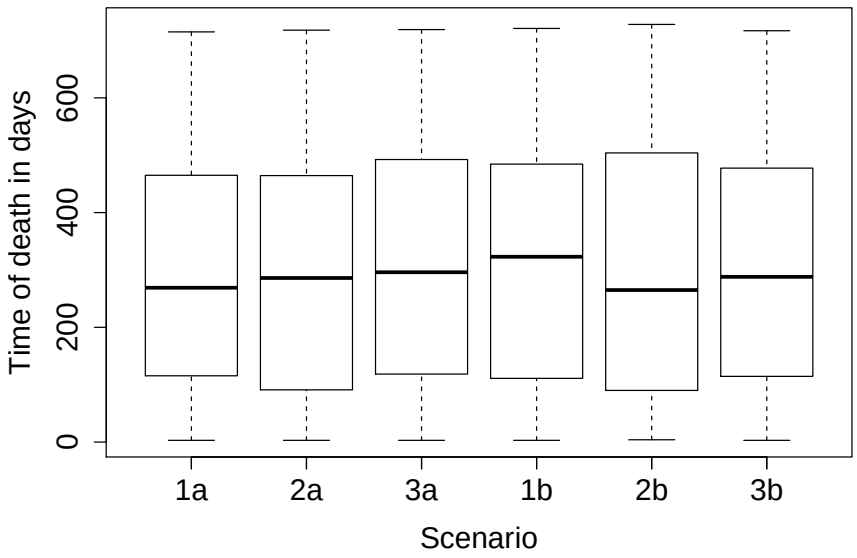


Figure 6.27: Boxplots of the death times for all scenarios for population LA

In Table 6.13, an overview of the medians of the times of the first readmissions, the first ambulant contacts to a psychiatrist and death is given.

Scenario	$1a$	$1b$	$2a$	$2b$	$3a$	$3b$
First readmission	97	89	95	95	85	96
First ambulant contact	39	44	36			
Death	269	323	286	265	296	288

Table 6.13: Medians of times of first readmission, the first psychiatrist contact and death for population LA

The patients with ambulant contacts to a psychiatrist (OPC) are compared to those without ambulant contacts (non-OPC). In Figure 6.28 the percentage of patients with readmissions is shown for both groups. The percentage for the patients with ambulant contacts is much higher, for scenarios $2a$ and $3a$ even twice as much as for patients without ambulant contacts.

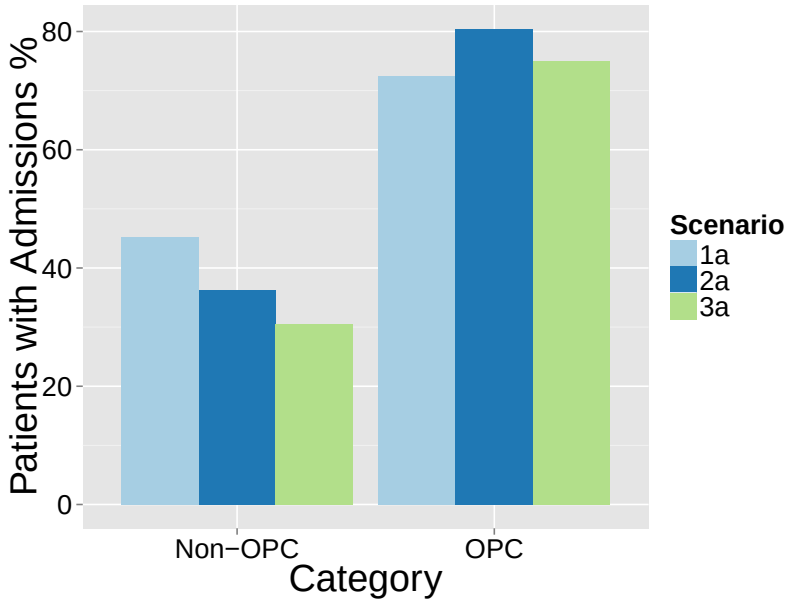


Figure 6.28: Comparison of the percentages of patients with readmissions between patients with and without ambulant treatment for population LA

6.4.5 Pathways of patients

In this section, the pathways of patients for population LA are presented. A detailed definition and explanation of the pathways are given in Section 6.2.5. In Table 6.14, an

overview of the classification without ambulant contacts is given.

Class	Number of readmissions	Month of first readmission
1	0	–
2	1	1
3	1	2-6
4	1	7-12
5	1	13-24
6	2-4	1
7	2-4	2-6
8	2-4	7-24
9	> 4	any

Table 6.14: Classification of patient pathways

In Figure 6.29, the sizes of the classes for scenarios $1a$, $2a$ and $3a$ are presented. Classes 1 and 10 are not shown in the plot, because the number of patients without readmission has already been analyzed and the focus is on the patients with readmissions.

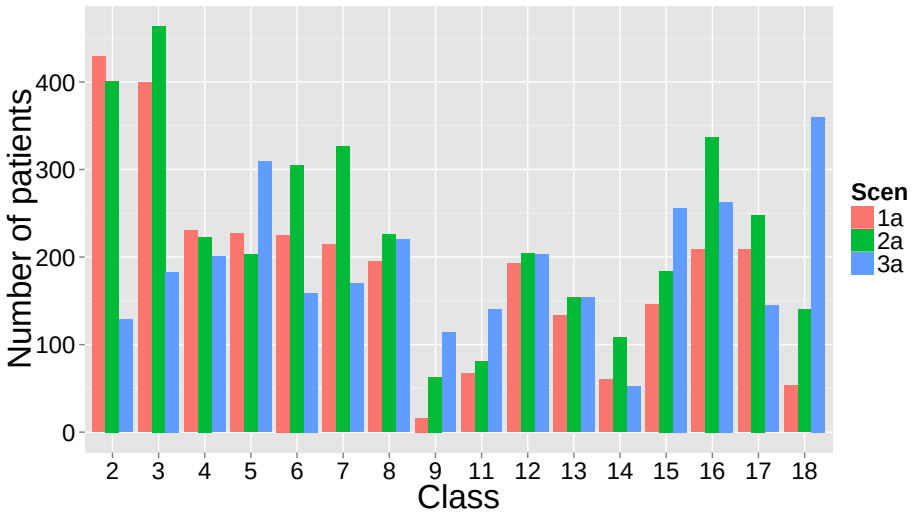


Figure 6.29: Sizes of classes for the scenarios with ambulant contacts for population LA

In scenarios $1a$ and $2a$ are more than three times as many patients in the classes 2 and 3 than in scenarios $3a$. That means more individuals have exactly one readmission shortly after the release. This can be explained by the fact that in scenario $3a$ more patients have ambulant contacts, so more patients are in classes 10 to 18, and in scenario $2a$ more patients have readmissions, so the overall number of patients in the presented classes is higher. Scenario $3a$ has the most patients with one late readmission after the first year.

In scenario 3*a* the number of patients with more than four readmissions is higher than in the other scenarios. In all scenarios, the number of patients with psychiatrist contacts have a higher average number of readmissions and the first readmission later.

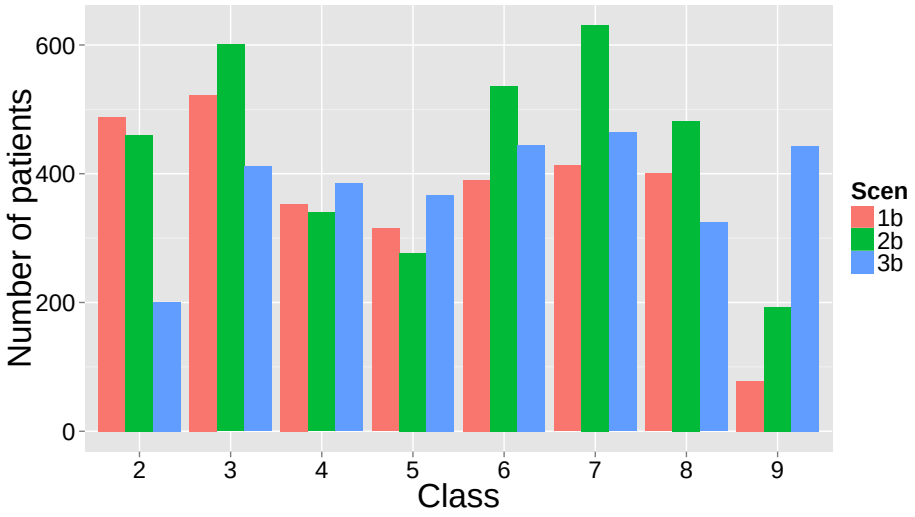


Figure 6.30: Sizes of classes for the scenarios without ambulant contacts for population *LA*

In Figure 6.30, the sizes of the classes for scenarios 1*b*, 2*b* and 3*b* are presented. Again, class 1 is not shown in the plot, because the number of patients without readmission has already been analyzed and the focus is on the patients with readmissions.
The trends are similar to the scenarios with ambulant contacts. In scenarios 1*b* and 2*b* are more than twice as many patients in class 2 than in scenarios 3*b* and class 3 is also bigger. This can be explained by the fact that in scenario 3*b* patients have more readmissions in average and in scenario 2*b* more patients have readmissions in general. Also, the number of individuals with more than four readmissions is much higher in scenario 3*b*.

6.4.6 Intervention

The intervention strategy introduced in Section 6.2.6 is also executed for population *LA*. In Table 6.15, the percentages of patients with readmissions, ambulant psychiatrist contacts and deaths are compared for scenario 3*a*. The percentage of patients with readmissions is much higher with the intervention strategy. This leads to the conclusion that an ambulant contact increases the probability for a readmission. This could already be seen in the analysis of the pathways. In the intervention scenario almost every patient visits a psychiatrist during the simulation. Only those patients who die within the first month have no contact. The number of deaths is slightly lower with the intervention

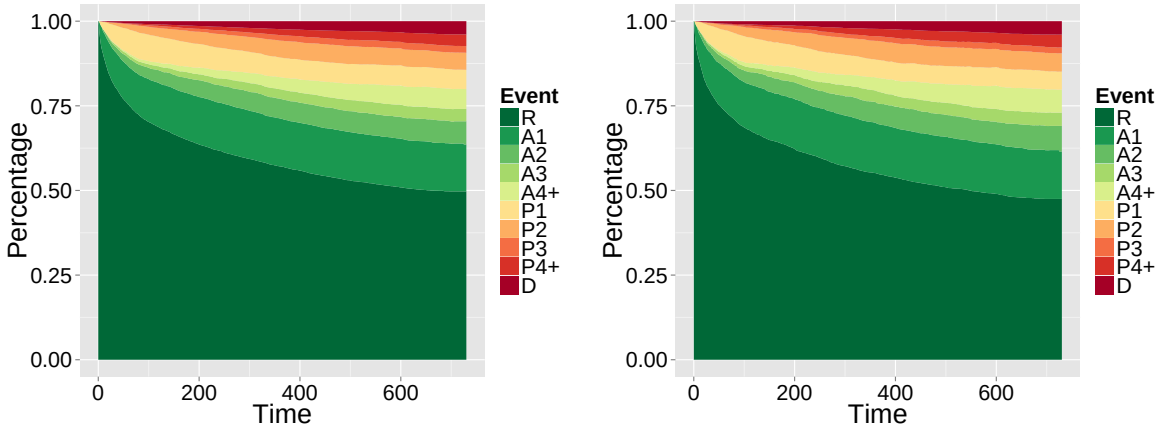
strategy.
So, this strategy does not succeed in reducing the number of readmissions.

Type of Event	No intervention	Intervention
Readmissions	44.8	68.4
OPC	28.6	99.7
Deaths	3.9	3.2

Table 6.15: Comparison of percentages of the occurrence of events for scenario $3a$ with and without intervention for population LA

6.5 Comparison of simulations for Austria and Lower Austria

The results of the simulations for populations AT and LA are compared in terms of numbers and times of events and the distribution over the classes defined by the pathways for scenario $3a$.
The evolutions of the patients distribution over the states for the two populations are presented in Figure 6.31. The share of the patients in state R has a similar evolution for both simulations and decreases almost exponentially. About 50 percent of the patients from Austria are remaining in state R at the end of the simulation, about 47 percent of the other population. So, patients in population LA have more readmissions, since the numbers for the states P_i and D are very similar for both populations.



(a) Austria

(b) Lower Austria

Figure 6.31: Evolution of the patients distribution over the states in scenario $3a$ for populations LA and AT

The proportions of the two populations with readmissions, ambulant psychiatrist contacts (OPC) and deaths are displayed in Table 6.16. Population LA has more events of every type. This can be linked to different compositions of the populations with a higher percentage of psychotic patients in population LA.
The influence of ambulant contacts on the number of readmissions is investigated. It is already known from previous sections that patients with OPC have more likely readmissions and that the patients in population LA have slightly more readmissions. In Figure 6.32, the percentage of patients with readmissions with and without ambulant contacts(OPC) is shown and it can been seen that for both groups (OPC and non-OPC) the latter effect is present.
The diagnosis always has a significant influence on the results. In Figure 6.33, the percentage of patients with readmissions for both populations split into the diagnosis

Type of Event	Austria	Lower Austria
Readmissions	42.2	44.8
OPC	27.7	28.6
Deaths	3.9	4.0

Table 6.16: Comparison of the proportions of the populations with readmissions, ambulant contacts and deaths for populations LA and AT

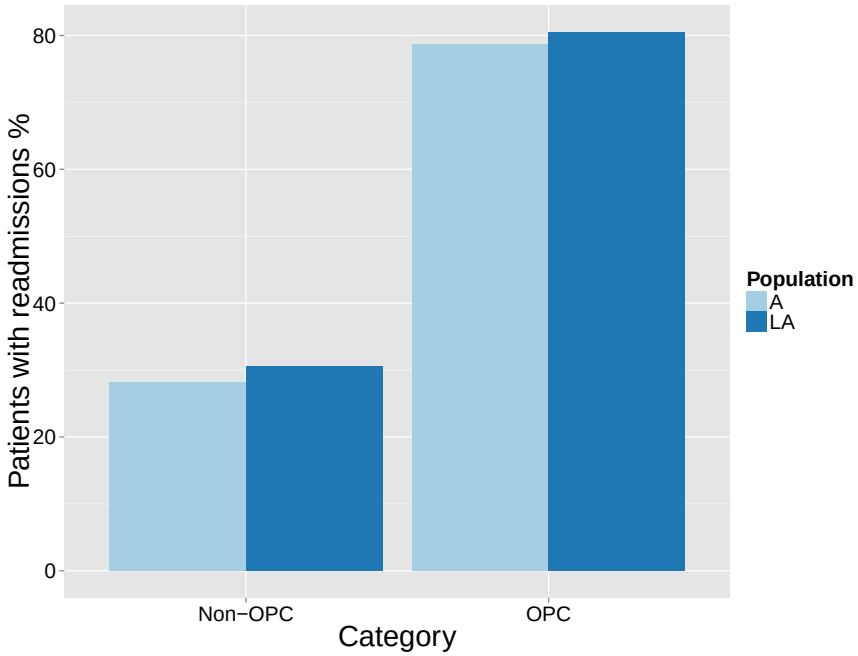


Figure 6.32: Comparison of the percentages of patients with readmissions between patients with and without ambulant treatment for populations LA and AT

groups is shown. The psychotic diagnosis groups D_1 and D_2 have relatively the most patients with readmissions. The highest difference is in group D_5, but this results are not very reliable since the number of patients with this diagnosis is very small in both populations.

The pathways of patients are also analyzed for the two populations. The numbers of the patients that follow each path are presented in Figure 6.34. Classes 1 and 10 are not shown in the plot, because the number of patients without readmission has already been analyzed and the focus is on patients with readmissions. For population LA, more patients without ambulant contacts have more than four readmissions (class 18). For all of the other classes the difference between the two populations is almost negligible.

The comparison of simulations results for populations LA and AT shows that the patients of the former have more events, especially more readmissions. This results from the bigger proportion of psychotic patients in the population of Lower Austria. The

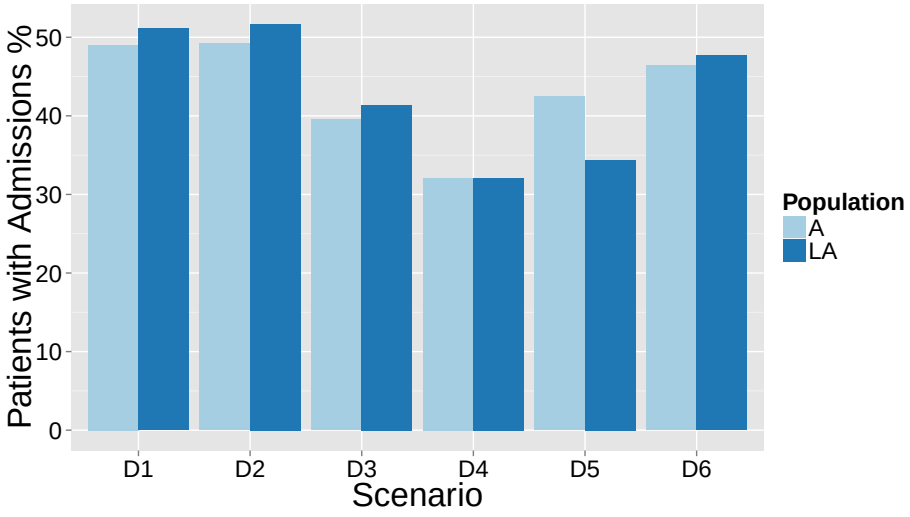


Figure 6.33: Comparison of the percentages of patients with readmissions for the six diagnosis groups for populations LA and AT

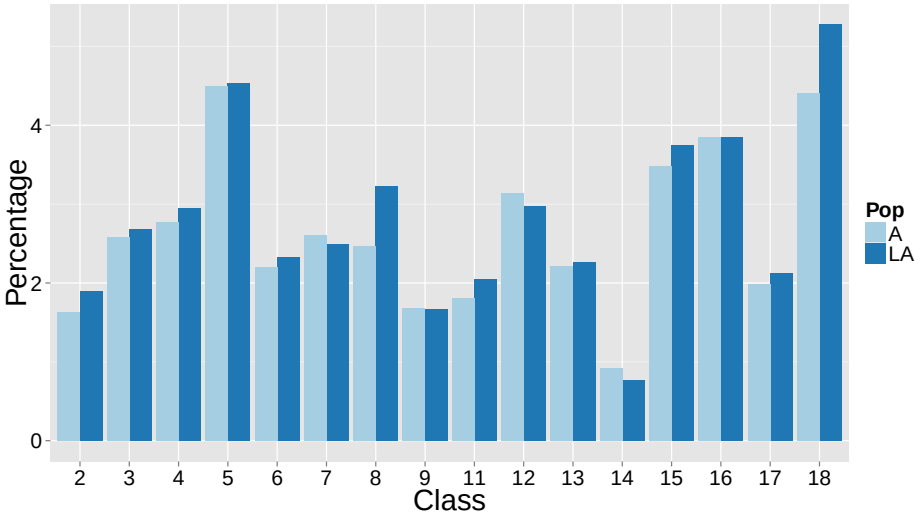


Figure 6.34: Comparison of the sizes of the classes for populations LA and AT

main observations for the population of Austria can be transferred to the population of Lower Austria, so no other substantial differences can be found in the results.

7 Conclusions

From the various methods of survival analysis the Cox model is chosen as the primary instrument for the data analysis regarding significant variables for the readmission times and also for the parametrization of the simulation model since it provides the features that are required for the given tasks. The main benefits of the Cox model are the inclusion of patient parameters and the flexibility for extensions for multiple events. For the parametrization, a multi-state extension is used. A drawback is the proportional hazards assumption. This assumption can be violated quite easily. However, the analysis of the data with the methods of survival analysis shows that the Cox model can be applied to the given data.
From the model selection methods Akaike's Information Criterion and the Lasso-method proved to be useful to assess the different models and parameters. The AIC is very easy interpretable. The issue with the Lasso-method is the difficult determination of the constraint for the regression coefficients.
The results of these analyses of the data combined with considerations about the complexity and running time of the model lead to the choice of the full linear Cox model to estimate the hazard rates for the events in the microsimulation model.
The microsimulation model is an appropriate tool to model the pathways of the patients. Both the longitudinal analyses of the single patients as well as the cross-sectional analyses can be carried out with little effort. The model is implemented in R which proved to be useful, because of the existing packages for survival analysis, especially the Cox model and even its extensions. Also, the visualization of the results can be done efficiently with the *ggplot*-package. For large populations the running time of the model can become an issue.

In general, the results show an exponential decrease of the number of patients with no event. Nevertheless, about half of the patients have no readmissions during the simulation. The number of patients with one ambulant contact and no readmission has its peak after a half of a year and declines afterwards. So, many of these patients have a readmission soon after the visit to the psychiatrist. The percentage of patients with a particular number of readmissions is indirectly proportional to the number of readmissions.
The comparison of the results of the different scenarios shows that in scenarios $2a$ and $2b$ the patients have more readmissions than in the other scenarios. This is due to an

overestimation of the number of readmissions because the order of the readmissions is not considered in these scenarios. Scenarios $2a$ and $2b$ as well as scenarios $3a$ and $3b$ have a higher proportion of psychiatrist visits among the population which is mainly caused by the definition of the scenarios again. So, the results of the scenarios with a lower level of data detail show significantly varying results from scenario $3a$ which uses the most detailed level of data. However, scenario $3a$ requires data of entire patient histories which is rarely available due to data protection issues.

For a more detailed analysis, the population is split into classes defined by typical pathways of patients. The pathways are defined by the time and number of readmissions. In comparison to the other scenarios, the readmissions of patients with only one readmission are later and the average number of readmissions per patients is higher in scenarios $3a$ and $3b$.

The diagnosis is the parameter with the biggest influence on the number and times of the readmissions and therefore also on the simulation results. The sensitivity analysis shows that changing the diagnosis of the population has a dramatic influence on the number of readmissions. The psychotic patients have considerably more readmissions than non-psychotic patients.

The proportion of patients with readmissions is much higher for patients with previous ambulant psychiatrist visits. Thus, ambulant contacts increase the probability for readmissions and are in most cases an indicator for a worsening of the condition of the patient. This also leads to the fail of the reduction of readmissions by the intervention strategy of compulsory visits to a psychiatrist after a certain time after the last admission.

The comparison of the populations of whole Austria and Lower Austria shows that more patients of the latter have readmissions and also ambulant contacts. This can be the result of the differing compositions of the populations regarding the parameter distributions.

A possible continuation of this is to perform the simulation with another model than the linear Cox model for the parametrization of the model, for example a model with interaction terms. Also, additional patient parameters can be included into the model.

List of Figures

List of Tables

Bibliography

[1] W. H. O., *International statistical classification of diseases and related health problems*, vol. 1. World Health Organization, 2004.

[2] T. M. Therneau and P. M. Grambsch, *Modeling survival data: extending the Cox model*. Springer Science & Business Media, 2000.

[3] D. Collett, *Modelling Survival Data in Medical Research*. CRC Press, 2003.

[4] B. R. Kirkwood and J. A. C. Sterne, *Essential Medical Statistics*. Essentials, Wiley, 2003.

[5] G. Rodríguez, *Lectures notes about generalized linear models*. 2008.

[6] D. R. Cox, *Regression models and life-tables*. Journal of the Royal Statistical Society. Series B (Methodological), p. 187–220, 1972.

[7] J. Beyersmann and H. Putter, *A note on computing average state occupation times*. Demographic Research, vol. 30, no. 62, pp. 1681–1696, 2014.

[8] R. Tibshirani, *The lasso method for variable selection in the cox model*. Statistics in Medicine, 1997.

[9] K. P. Burnham and D. R. Anderson, *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2002.

[10] D. W. Hosmer, S. Lemeshow, and S. May, *Applied survival analysis*. NJ, Wiley, 2008.

[11] U. Siebert, O. Alagoz, A. M. Bayoumi, B. Jahn, D. K. Owens, D. J. Cohen, and K. M. Kuntz, *State-transition modeling: a report of the ISPOR-SMDM modeling good research practices task force-3*. Value in Health, vol. 15, no. 6, pp. 812–820, 2012.

[12] A. O'Hagan, M. Stevenson, and J. Madan, *Monte Carlo probabilistic sensitivity analysis for patient level simulation models*. University of Sheffield. Department of Probability and Statistics, 2005.