Thea Lovise Leikvoll Fagerli

# Analysing Associations with Hospital Length of Stay

Master's thesis in Applied Physics and Mathematics
Supervisor: Andreas Asheim
June 2023

Master's thesis

**NTNU**
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Thea Lovise Leikvoll Fagerli

# Analysing Associations with Hospital Length of Stay

Master's thesis in Applied Physics and Mathematics
Supervisor: Andreas Asheim
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

# Abstract

This thesis aimed to analyse the association between hospital length of stay (LOS) and two distinct outcomes: mortality risk and visits to the general practitioner (GP). To accomplish this, Cox Regression was used to assess the association between LOS and mortality risk, while Poisson Regression was utilized to examine the association between LOS and number of visits to the GP, both within 60 days following admission.

Initially, a simulation study was conducted to explore the potential presence of immortal time bias in the analysis of association between LOS and mortality. Immortal time bias arises when patients who die at the hospital are excluded as they do not get a proper LOS. By testing various scenarios, the study aimed to determine if the estimated hazard ratio (HR) for LOS was influenced. The results of the simulation study revealed a HR of approximately 3%, providing valuable insight in the magnitude of immortal time bias.

The study population consisted of admissions to Norwegian health trusts with acute heart failure as the primary diagnosis between 2010 and 2021. The data set contained information about age at admission, sex, time of admission, education level and number of visits to multiple emergency medical services within the 60 days before admission.

Furthermore, individual Cox models with varying levels of strata were computed. The primary focus was on the hazard ratio of LOS, which was 4% for the models without strata, and 5% when strata were added. These findings suggested the existence of additional factors, beyond immortal time bias, that contributed to the analysis. It was reasonable to believe that the severity of the patients acted as a confounding factor, impacting both LOS and the mortality risk. Consequently, this could lead to incorrect conclusions regarding the association between LOS and mortality. Although the severity of the patients is challenging to measure quantitatively, it is an essential part of the analysis.

Lastly, Poisson Regression was used to investigate the association between LOS and visits to the GP. The incidence rate ratio (IRR) indicated that the expected number of visits increased by 2% for each additional day in the hospital.

By using these regression models, this thesis supplied insight into the occurrence of immortal time bias and confounding in an analysis where the severity of the patients is unknown. The findings may contribute to a better understanding of the relationship between LOS, mortality risk and visits to the GP, shedding light on the complex factors in the health care system.

# Sammendrag

Denne masteroppgaven analyserer sammenhengen mellom sykehusopphold (LOS) og to forskjellige utfall: dødelighetsrisiko og besøk hos fastlegen. For å oppnå dette ble Cox regresjon brukt til å vurdere sammenhengen mellom LOS og dødelighetsrisiko, mens Poisson regresjon ble brukt til å undersøke sammenhengen mellom LOS og antall besøk hos fastlegen, begge innen 60 dager etter innleggelse.

Innledningsvis ble det gjennomført en simuleringsstudie for å utforske den potensielle tilstedeværelsen av udødelighetsskjevhet i analysen av sammenheng mellom LOS og dødelighet. Udødelighetsskjevhet oppstår når pasienter som dør på sykehus blir ekskludert fordi de ikke får en skikkelig LOS. Ved å teste ulike scenarier var målet å avgjøre om den estimerte relative risikoen (HR) for LOS ble påvirket. Resultatene av simuleringsstudien viste en HR på omtrent 3%, noe som gir verdifull innsikt i omfanget av udødelighetsskjevhet.

Studiepopulasjonen besto av innleggelser i norske helseforetak med akutt hjertesvikt som hoveddiagnose mellom 2010 og 2021. Datasettet inneholdt informasjon om alder ved innleggelse, kjønn, tidspunkt for innleggelse, utdanningsnivå og antall besøk til flere akuttmedisinske tjenester i 60 dager før innleggelse.

Videre ble individuelle Cox modeller med varierende nivåer av strata beregnet. Hovedfokuset var på HR av LOS, som var 4% for modellene uten strata, og 5% når strata ble lagt til. Disse funnene antydet eksistensen av andre faktorer, i tillegg til udødelighetsskjevhet, som bidro til analysen. Det var rimelig å tro at pasientens alvorlighetsgrad virket konfunderende og påvirket både LOS og dødelighetsrisikoen. Dette kan derfor føre til feilaktige konklusjoner om sammenhengen mellom LOS og dødelighet. Selv om alvorlighetsgraden av pasientene er utfordrende å måle kvantitativt, er det en viktig del av analysen.

Til slutt ble Poisson regresjon brukt for å undersøke sammenhengen mellom LOS og besøk hos fastlegen. Insidensraterisikoen (IRR) indikerte at forventet antall besøk økte med 2% for hver ekstra dag på sykehuset.

Ved å benytte disse regresjonsmodellene ga denne oppgaven innsikt i forekomsten av udødelighetsskjevhet og konfundering i en analyse der alvorlighetsgraden til pasientene er ukjent. Funnene kan bidra til en bedre forståelse av sammenhengen mellom LOS, dødelighetsrisiko og besøk hos fastlegen, og belyse de komplekse faktorene i helsevesenet.

# Preface

This thesis is the final part of a Master of Science (M.Sc.) in Applied Physics and Mathematics with a specialization in Industrial Mathematics, completed at the Norwegian University of Science and Technology (NTNU). It constitutes the course TMA4900-Master thesis and is written during the spring semester 2023.

I would like to thank my supervisor Andreas Asheim for excellent guidance, informative discussions and feedback throughout this past year. His expertise and encouragement have been instrumental in this process. I would also like to thank Regforsk, and project manager Johan Håkon Bjørngaard, for the possibility to use the data in this thesis. Regforsk is a research group financed by The Research Council of Norway.

In addition, I would like to thank my fellow students, friends and family for their support and motivation. I want to dedicate a special thanks to my extraordinary mother, whose support has been the cornerstone of my education. I am forever grateful for her unwavering presence and belief in me. Lastly, I would like to thank my partner, who has supported me throughout this journey. His belief in my abilities has been an immense factor and bolstered my confidence, which I am incredibly grateful for.

Thea Lovise Leikvoll Fagerli<br>
*Trondheim, Spring 2023*

# List of abbreviations

| Abbreviation | Definition |
|---|---|
| AIC | Akaike information criterion |
| EMS | Emergency medical service |
| ER | Emergency room |
| GP | General practitioner |
| HF | Heart failure |
| HR | Hazard ratio |
| HT | Health trust |
| IRR | Incidence rate ratio |
| LOESS | Locally estimated scatterplot smoothing |
| LOS | Hospital length of stay |
| MPLE | Maximum partial likelihood estimator |
| OOH | Out-of-hours primary health care |

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

It is estimated that roughly 2% of the population in Norway lives with heart failure (HF). For people over 75 years, this percentage is estimated to be 10% (Skogli et al. 2020). HF is a clinical syndrome consisting of cardinal symptoms involving breathlessness and fatigue, where the heart is not able to pump enough blood out to the body (McDonagh et al. 2021).

The disease is possible to live with, however it is described as a public health issue due to the negative consequences it brings on the quality of life for the patients. In addition, the health care costs related to the disease are huge and are estimated to further increase in the next couple of years. A study done by Menon estimated that heart failure inflicted the Norwegian society cost of 48 billion NOK in 2018 (Skogli et al. 2020). They estimated that this will further increase up to 63 billion NOK in 2030 due to the increasing number of elderly people in society. The largest part of these costs is related to the burden of disease and loss of economic growth due to the disease, but it also includes costs for the health care system.

Heart failure is the most common cause of hospital admissions for people older than 65 years (Skogli et al. 2020). With the increasing number of elderly people that is happening, this will cause even more pressure on the health care system in the years to come. However, the detection of heart failure is much better today than it was decades ago. Both the number of acute events of heart failure and hospital length of stay (LOS) have gone down since 2005, by 45% and 58% respectively. One possible reason for the reduction in LOS can be the Coordination Reform ("Samhandlingsreformen"), which was implemented in Norway from 2012 (Helse- og Omsorgsdepartementet 2011). Its primary goal was to improve the help given to the patients, and to give the municipalities more responsibility for patients that are ready to be discharged. After the reform was implemented, studies were done to see if the reform had any impact on LOS and the readmission rate. One study showed that the LOS was generally shortened by 0.1 day by the reform, where the patient group of interest was patients admitted to the hospital with heart failure, hip fracture, stroke or chronic obstructive pulmonary disease (COPD) (Melberg and Hagen 2016). One of the main results in this study was that for patients that were defined as ready to be discharged, the LOS was shortened significantly with 4.6 days for patients admitted with hip fracture. Patients that were defined as ready to be discharged had also a higher readmission rate than the patients that were not defined as this, however they could not conclude that this was not a direct result of the Coordination Reform.

Several studies have been done on the association between LOS and readmission and mortality for given diseases. In a study done by the American College of Cardiology on patients in Canada, they looked at the association between short or long LOS and a 30-day readmission and mortality for patients admitted with heart failure. They found that a shorter LOS was associated with a higher rate of readmission for heart failure and cardiovascular related diseases (Sud et al. 2017). However, a long LOS was also related to a increasing rate of readmission of all types of diseases. In addition, they found that a long LOS gave the highest 30-day mortality risk. In this study, a short length of stay included 1-2 days, while a long length of stay was 9-14 days. From the results of the study it was clear that the answer to such a problem is complex and depends on several factors.

Patients suffering from heart failure will often be admitted into the hospital more than once due to the disease. A study done by the American Heart Association found that 61.3% of patients were readmitted due to HF within the first year after discharge. The first readmission occurred within the first and last deciles of the survival time after discharge, where each decile was a median of 63 days in length. Looking at any readmission, taking into consideration that each patient could have multiple hospitalizations, the largest proportion happened close to death, followed by the first period after being discharged. This could be due to treatment after a diagnosis is set, or that patients are in general sicker close to death. The sickness and mortality risk could also be due to other factors than the HF itself.

## 1.1 Heart Failure

The most common cause of HF is high blood pressure, hypertension, and coronary artery disease, which could be found in 75-80% of patients diagnosed with the disease in 2007 (Aarønæs et al. 2007). Other causes can be diabetes or congenital heart defect (CHD). HF can happen when the heart is not able to pump with enough force to cover the amount of blood that the body needs.

The heart is composed of four cambers, the left atrium and the right atrium, and the left ventricle and the right ventricle (Better Health Channel 2023). The atrium is a collecting chamber, while the ventricle is a large pumping chamber. The left side of the heart is larger than the right side as it does the main work by pumping oxygen-rich blood to all parts of the body. The right side's task is to collect blood from the body, which is low in oxygen, and to pump it into the lungs to be filled with oxygen again.

The most common type of HF happens in the left ventricular of the heart. A HF in the left ventricular means it has to work harder than normal to pump out enough blood to the body. If a heart failure happens in the right side, it is usually a result of failure in the left ventricle. When this happens, there is an increase in fluid pressure which results in a loss in pumping power for the right ventricular (American Heart Association 2023). If this happens in a brief time span, with acute symptoms, the patient may need urgent medical help often involving unplanned hospital admissions or visits to the emergency room (ER).

## 1.2 Hospital length of stay

The average LOS has gone done for approximately one third of the countries in the EU when comparing 2015 to 2020 (Eurostat 2023a). In Norway, the average length of stay was 5.2 days in 2020, which is a reduction of 0.3 days compared to 2015. In 2011 the average length of stay was 6.0 days, making it a reduction of nearly a day across the span of 10 years (Eurostat 2023b). This trend of reduced LOS may be a result of hospitals improving treatment time but also due to an incentive to reduce it. As mentioned, the Coordination Reform was implemented in Norway in 2012. As a way of giving the municipalities more responsibility, they were fined if the could not take care of patients that still needed help after discharge from the hospital (Hagen et al. 2013). This may have affected the LOS as this became expensive for the municipalities to ignore.

The active reduction in LOS and the economic expense that came with the Coordination Reform for the municipalities may have led to a change in routines for the different hospital trusts. There is a possibility that the average LOS has changed more for some hospital trusts, while others have the same trend over the years. The different routines make it hard to compare patients across hospital trust, as the average can vary substantially. This can also be said for trends over the years. The reduction of the average LOS in Norway may also be a result of an improvement in the health care sector, which needs to be taken into consideration when analyzing data spanning a decade. Another factor is that the LOS is sometimes short because the patient is at considerable risk of dying, and hospitals want to avoid in-hospital deaths. This may skew the data as the patient would probably get a higher LOS if the risk of death was lower. Other times patients may be discharged early because they will be readmitted within a short period of time due to treatment. In total, the LOS may say something about how sick the patient is, but exactly how is not straightforward.

Another aspect of the LOS is the fact that the process of discharging is comprehensive. Organizational factors have an impact on the LOS, where departments under pressure may struggle to discharge patients at the right time. This could skew the LOS and make it longer for some admissions where the patient was ready to be discharged but stayed longer because of the department.

For an admission to have a length of stay, there is a need to condition on survival trough the admission. Admissions where the patient died on the hospital must be filtered out, which gives the potential of immortal time bias. This means that to be included in the analysis, the patients are seen as "immortal" while they are in the hospital, i.e., the outcome of interest cannot occur in this time interval (Yadav and Lewis 2021). The removal of patients that died in the hospital, introduces a skew in the data and can potentially lead to overestimation of the outcome of interest, which needs to be considered when looking at the association between LOS and the 60 day mortality.

There is also a possibility of confounding in observational studies were the aim is to estimate the causal inference between treatment and outcome. In this thesis, this may occur if LOS depends on one or multiple components measured before admission, which are not adjusted for in the analysis. How sick the patient is at admission could be such a component, as it can have an impact on discharge of the patient and hence the LOS.

## 1.3 Analysis plan

In this thesis we will look at associations with hospital length of stay for patients admitted with heart failure. This involves mortality and visits to the general practitioner (GP), where the interest lies on the 60 days after a hospital admission. First, we look at the data given by the Norwegian Patient Registry and the Norwegian Cause of Death Registry, and the general survival rate based on the hospital admissions. Later we use this to generate data in a simulation study where the primary aim is to investigate the possibility of immortal time bias. In addition, we want to see what different scenarios based on the real survival rate would say about the relationship between LOS and mortality.

Further, Cox Regression models with different features and levels of strata will be computed in order to estimate the hazard ratio of LOS and other features on mortality. In addition, we aim to make Poisson models for the rate of visits to a GP in the 60-day window after an admission.

This thesis consists of 6 chapters. Chapter 2 presents the available data used, and an explanatory data analysis. In Chapter 3 the theory behind the statistical methods is described in detail, where the focus lies on the Cox Proportional Hazards Model and Poisson Regression. The chapter also includes an introduction to survival analysis. Chapter 4 presents the simulation study, and the different models that were fitted to the data. The results from these approaches are presented in Chapter 5. A discussion of the results and a conclusion can be found in Chapter 6.

# Chapter 2

# Data

This chapter presents the data that was available for the thesis. It holds a description and exploration of the features used in the models. The use of the data was approved by the Regional Committee of Ethics in Medical Research (2016/2159). All of the analyses were done on HUNT Cloud, which is a secure environment for storage and analysis, where only relevant data was made available for our analyses (HUNT Research Centre 2021).

## 2.1 Available data

The data was made available from the Norwegian Patient Registry and gave information about a nationwide cohort of $52,887$ patients admitted to Norwegian hospitals with heart failure, between January 1st, 2010, and December 31st, 2021. In Norway, all hospital trusts are obliged to give information about their clinical activity to the national registry. From this registry it was possible to obtain the sex of the patient, as well as age at admission and time of admission and discharge.

Additionally, the education level for each patient was included in the data set, where the levels are defined according to Statistics Norway (SSB) (Statistisk sentralbyrå 2023). Furthermore, the data set included information about the number of visits to the GP, out-of-hours emergency primary health care (OOH), emergency treatment in hospital and the total LOS before and after admission. The data also contained which hospital trust the admission was at, and a patient-ID such that it was possible to cluster admissions by patient. However, individual patients could not be directly identified.

The focus of this thesis is on acute admissions of HF, and hence we only included hospital stays with ICD-10 codes I50 as primary diagnosis. In addition, date of death was collected from the Norwegian Cause of Death Registry to know when or if patients died during the time period of study. If a patient had multiple registered hospital stays with less than 8 hours in-between, these stays where merged together as one assuming this is due to in-hospital moving or transfer between hospitals. This was also done for overlapping stays, where the admission date of the first admission and discharge date of the last admission were used as the time interval for hospitalization. In total, the resulting data set consisted of 84, 527 hospitalizations.

## 2.2 Features of interest

### 2.2.1 Survival rate

The data contained the days until death from the day of admission for approximately 70% of the admissions. The other 30% of the admissions were patients that survived until at least December 31rd 2021. The admissions made it possible to compute a 60-day survival curve which is shown in Figure 2.1. After 60 days, the average survival rate from the admissions was approximately 88%.

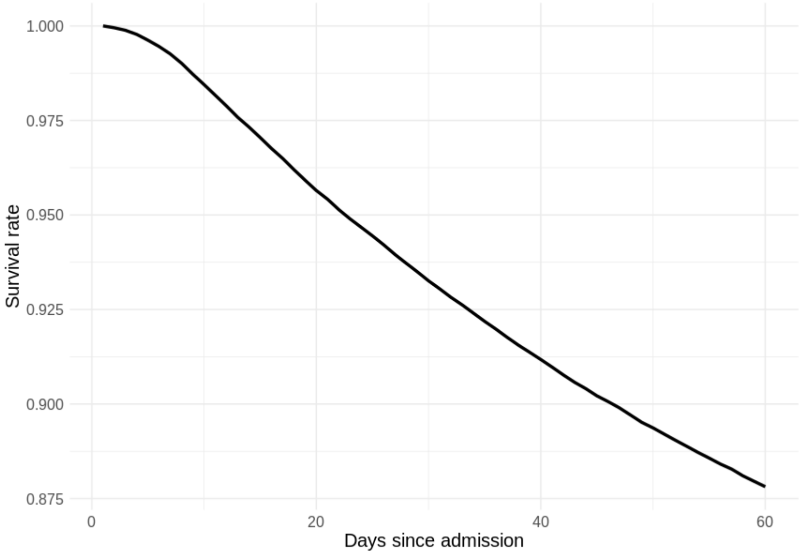
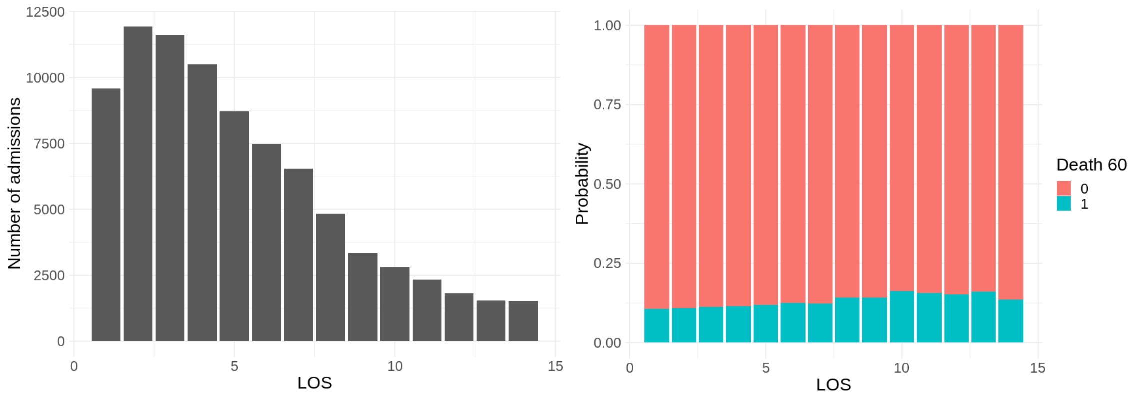


Figure 2.1: The figure shows the computed survival curve using the admissions in the data set.

### 2.2.2 Hospital length of stay (LOS)

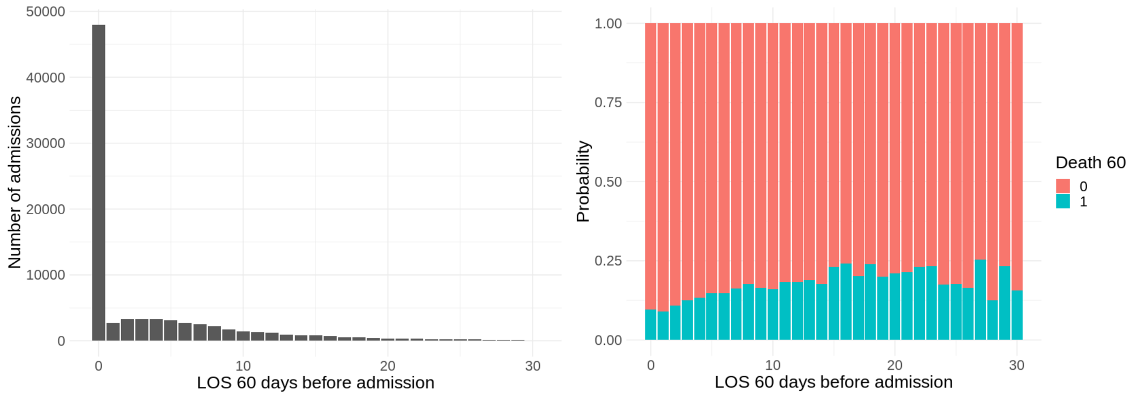
Figure 2.2a shows the distribution of the LOS for the admissions in the data set. The range of the feature is 1 to 14 days. Hospital stays of 0 days were removed as this mainly included short stays, and were thought to be planned visits related to treatment and not acute admissions. Hospitalizations of longer than 14 days were removed as it was only a few of them. The average LOS was 5.1 days for the full data set. The LOS is defined as the total number of days spent at the hospital and can include visits to multiple departments, if they are within 8 hours of each other. The percentage of 60-day mortality for each LOS is shown in Figure 2.2b. It shows a slight increase for longer stays compared to stays below the average. The increase in mortality risk is on average 0.2% per day increase in LOS.



(a) (b)

Figure 2.2: The distribution of length of stay over the admissions, and the distribution of death within 60 days for the length of stay.

Moreover, the number of days spent in the hospital in the 60 days before admission was also of interest. Figure 2.3a, shows that most admissions had no prior visits in the 60-days window before an admission. However, there were admissions registered with LOS in this time window, with some having up to 30 days in the hospital prior to a new admission. This could give an indication of the sickness of the patients as the admissions are only acute admissions. Figure 2.3b shows the mortality risk for the 60 days before admission.

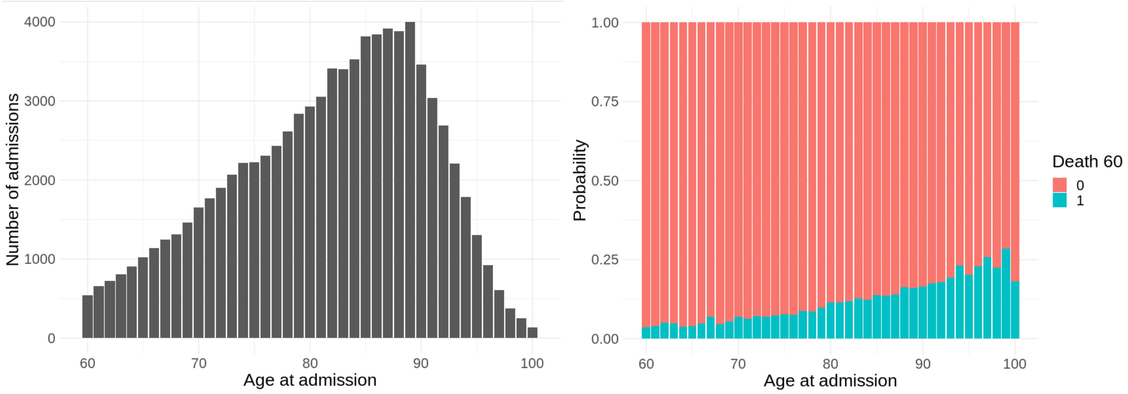


(a) (b)

Figure 2.3: The distribution of length of stay in the 60 days before admission, and the distribution of death within 60 days for the length of stay.

### 2.2.3 Age at admission

The feature *age* gave the age of the patient at admission. This is a feature that changed over time for patients with multiple admissions, if they were from different years. Figure 2.4a shows the distribution of age at admission, where the number of admissions increased with age and the mean age at admission was 81.6 years. The high age at admission could indicate that the patients were quite sick when they were admitted as the life expectancy for men and women in Norway was 80.9 and 84.4 years in 2021 respectively (Haug 2023). As there were only a few admissions of patients over the age of 100 years, these were left out of the plots to ensure full anonymity. The percentage of death within 60 days increased with age as seen in Figure 2.4b.



(a)

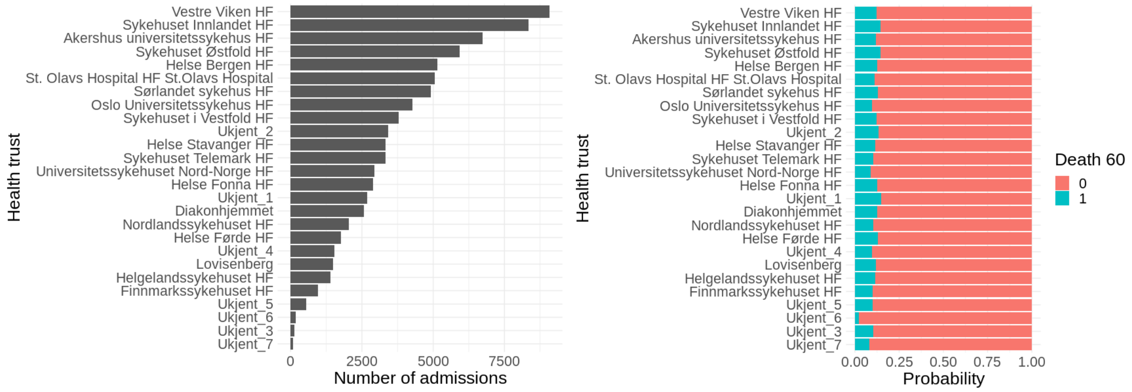
(b)

Figure 2.4: The distribution of age at admission, and the distribution of death within 60 days per year of age.

### 2.2.4 Health trust (HT)

The data was collected from 26 different health trusts in Norway. For some hospital trusts there were only a few hospital admissions in the data set while others had thousands of admissions, resulting in a wide range of admissions. The distribution of admissions can be seen in Figure 2.5a. The trusts are scattered all over Norway, with different routines and abilities, hence we would assume that there are systematic differences between them when it comes to the average LOS and other factors such as capacity and demand. Even though the number of admissions varies across the health trust, it was an insignificant variation in the probability of dying within 60 days as seen in Figure 2.5b.

Several admissions had no registration of health trust, and these were separated into 7 different unknown groups. As Figure 2.5a shows, 4 of them had very few admissions compared to the other HT's.

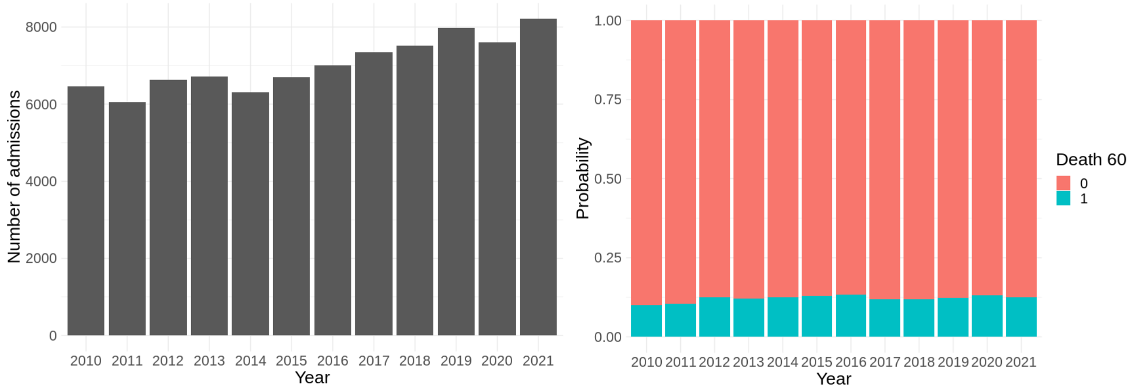


(a) (b)

Figure 2.5: The distribution of admissions across the health trusts, and the distribution of death within 60 days per health trust.

### 2.2.5 Time of admission

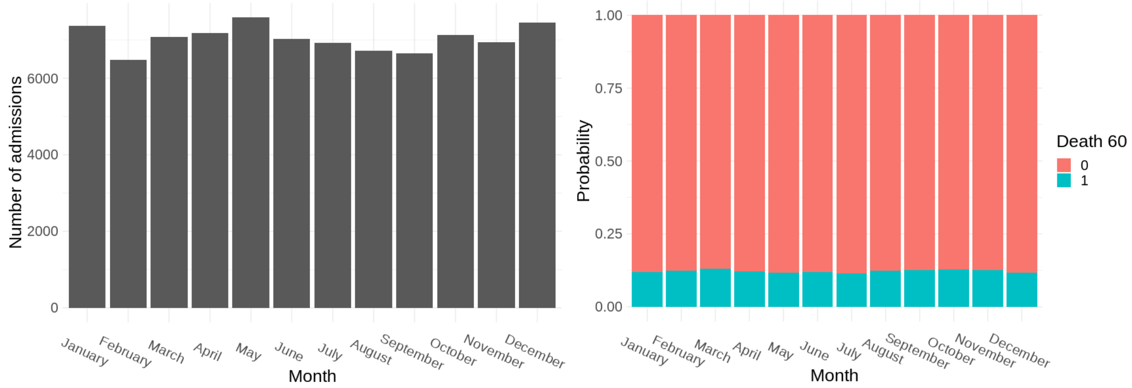
The data includes the time stamp of admission and discharge, which made it possible to extract the year, the month, the day and the hour of admission. Figure 2.6a shows the distribution of admissions over the years. The number of hospital admissions was even over the different years, but with a increasing trend. The mortality was, however, constant over the years.



(a) (b)

Figure 2.6: The distribution of admissions across the years, and the distribution of death within 60 days per year.

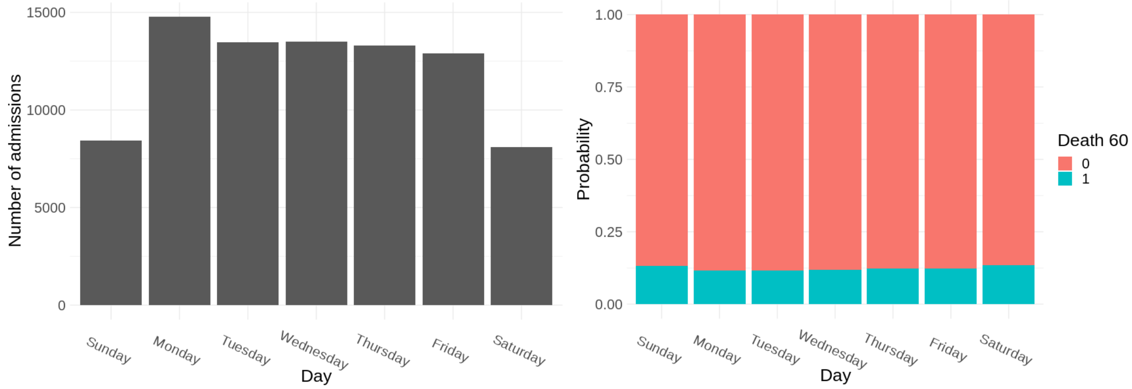
The admissions were also evenly spread over the months. There was a slight higher rate of admissions in January, May and December compared to the other months, however the differences were not significant. Figure 2.7b shows that the mortality was the same for all months.



(a) (b)

Figure 2.7: The distribution of admissions across the different months, and the distribution of death within 60 days per month.

In the plots for days of the week, shown in Figure 2.8a there is a higher variability in admissions for the different days. There were fewer admissions on the weekends compared to the rest of the week, with a peak on Mondays. During the week, patients may visit their general practitioner (GP) if the are feeling sick, and the GP can decide if they need to be admitted to a hospital. This is not possible during the weekends, resulting in fewer admissions. The peak of admissions on Mondays may be a result of this, as patients that are not acutely sick delay their visit to the GP over the weekend. However, the mortality was constant over the week.

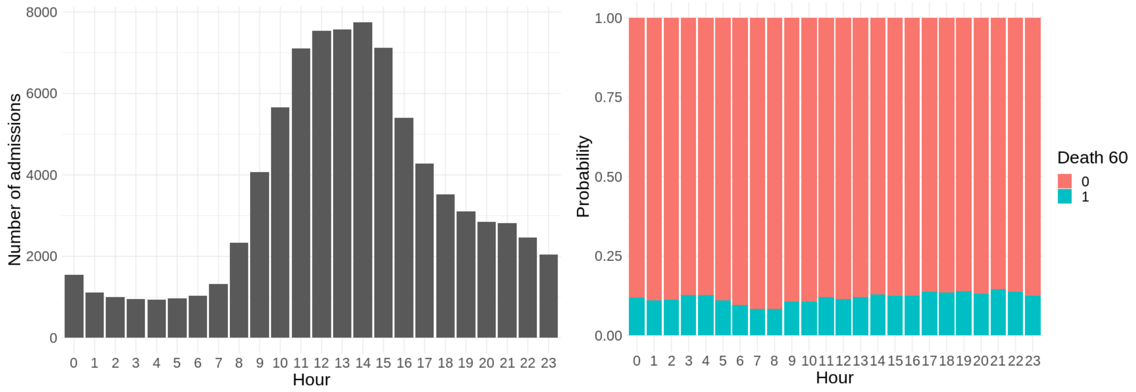


(a) (b)

Figure 2.8: The distribution of admissions across the days of the week, and the distribution of death within 60 days per day.

For the hours in the day, the distribution of admissions is shown in Figure 2.9a. The figure shows that the hospital was busiest during the day with a high peak from 11 : 00 to 15 : 00. The early hours, between 00 : 00 and 08 : 00, had fewer admissions where these were the most acute events with patients needing help immediately. The argument with the GP is also valid here. It is only possible to visit the GP during the day, leading to no referral from the GP during the late afternoon and night. There were also some admissions that were on a holiday. These admissions were also acute admissions as the GP is not available on holidays. There was some variation in the 60-day mortality rate for the different hours, with lowest probability around $07:00-08:00$ which can be seen in Figure 2.9b.

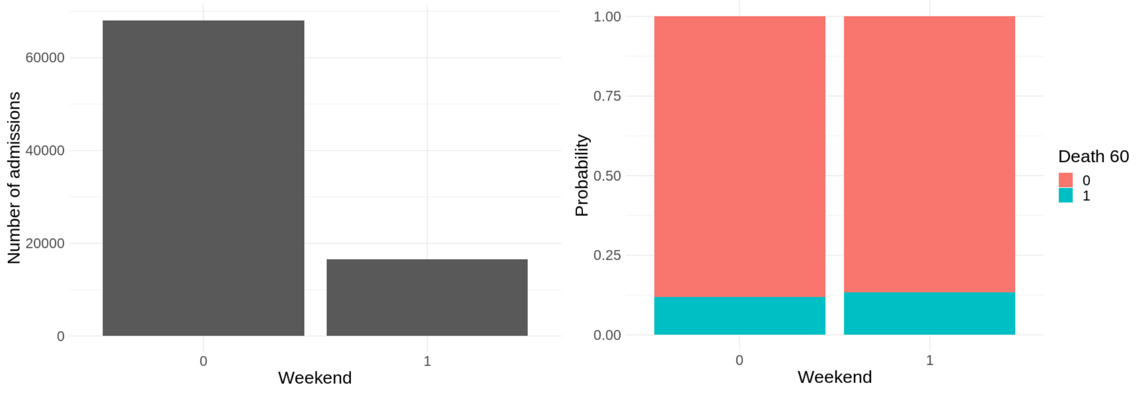
Grouping the days after the weekend or not shows that in total there was barely any difference in the mortality as seen in Figure 2.10b. Some days of the year are also holidays, where the routines and number of staff are similar to the ones during the weekend. However, the mortality was not affected by this.



(a)

(b)

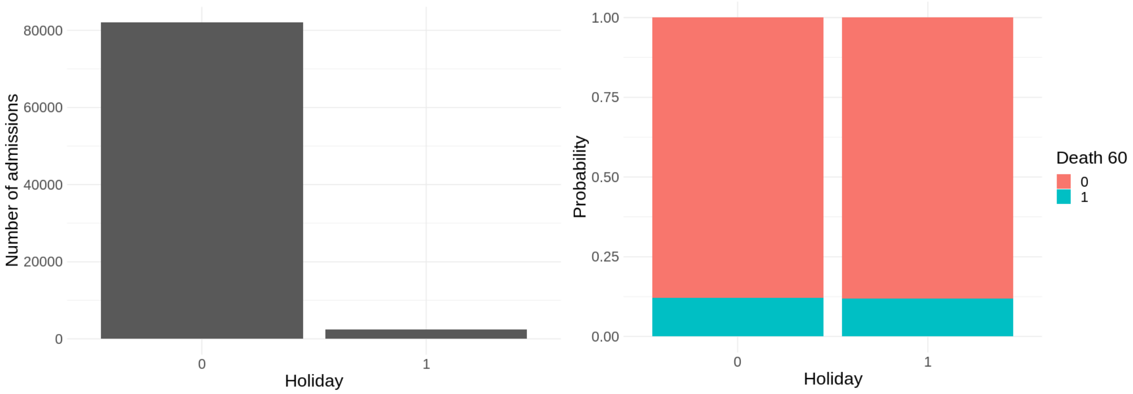
Figure 2.9: The distribution of admissions per hour of the day, and the distribution of death within 60 days per hour.



(a)

(b)

Figure 2.10: The distribution of admissions for weekday vs. weekend, and the distribution of death within 60 days grouped by weekday or weekend.



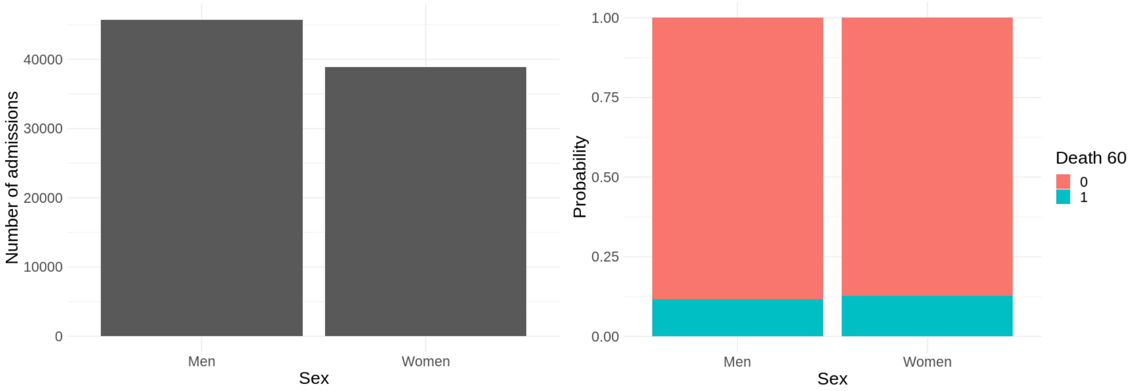
(a)

(b)

Figure 2.11: The distribution of admissions for regular days vs. holidays, and the distribution of death within 60 days grouped by regular day or holiday.

### 2.2.6 Sex

The data consisted of 45675 admissions of men and 38852 of women, which gave a fraction of 0.54 and 0.46 respectively. Figure 2.12a shows this distribution, and Figure 2.12b shows that there was a slightly higher risk of dying within 60 days for women compared to men.



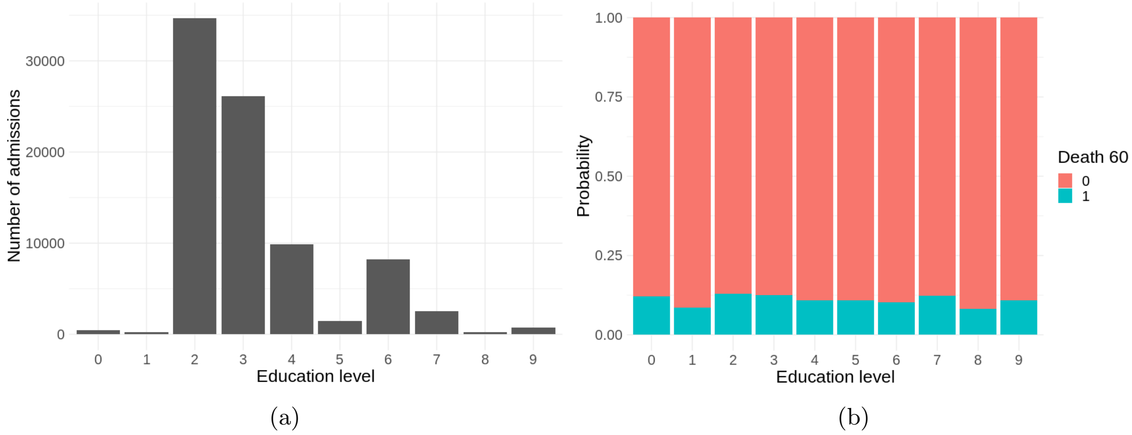
(a)

(b)

Figure 2.12: The distribution of admissions for men and women, and the distribution of death within 60 days for each sex.

### 2.2.7 Education level

The highest level of education for each patient was also included in the data set, where these levels follow the definition from SSB (Statistisk sentralbyrå 2023). Level 0 corresponds to no education, level 1 is primary education and so fort. The highest education is level 8 and represents postgraduate education. Admissions where the education level was unspecified, are represented by level 9. From Figure 2.13a it is clear that most of the patients belong to level 2 and 3, which represents lower secondary education and upper secondary education with basic education respectively. Education level is an important feature when modelling mortality as there are differences in life expectancy between people of high and low education. People with high education has a life expectancy which is $5-6$ years higher than people with low education following the Norwegian Institute of Public Health (Syse et al. 2023). Figure 2.13b shows some variability in the risk of dying, with level 1 and 8 having the lowest percentage. However, this may not be representative as these levels were the least represented levels in the data.



(a)

(b)

Figure 2.13: The distribution of admissions per education level, and the distribution of death within 60 days per level. Level 0 represents no education, level 1 is primary education and so forth. Admissions with unknown education level is represented by 9.

### 2.2.8 General practitioner (GP)

The data also included visits to the general practitioner (GP), both before and after admission. From Figure 2.14a it is possible to see that several patients visited their GP multiple times before admission to the hospital. A fraction of patients did not visit their GP, and Figure 2.14b shows that these patients had a slight higher probability of dying within 60 days after admission compared to patients that had just 1 visit before admission. Figure 2.14b shows an increasing trend in mortality as the number of visits increased, even though few admissions had over 10 visits to the GP beforehand which should be taken into consideration.

An outcome of interest was the number of visits to the GP in the 60 days after admission. The distribution of visits after admission followed the same shape as the distribution for visits before

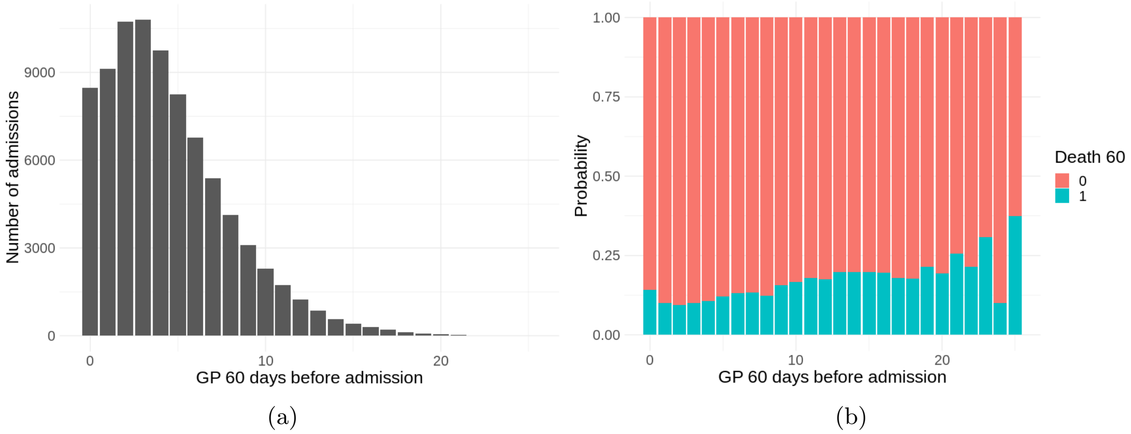


Figure 2.14: The distribution of admissions for visits to the GP in the 60 days before admission, and the distribution of death within 60 days for these visits.

admission, however the mortality rate for visits after had a decreasing trend. The mean rate of visits to the GP after admission was 5 visits, while the median was 4. As stated, a Poisson model with this outcome will be estimated in later chapters.

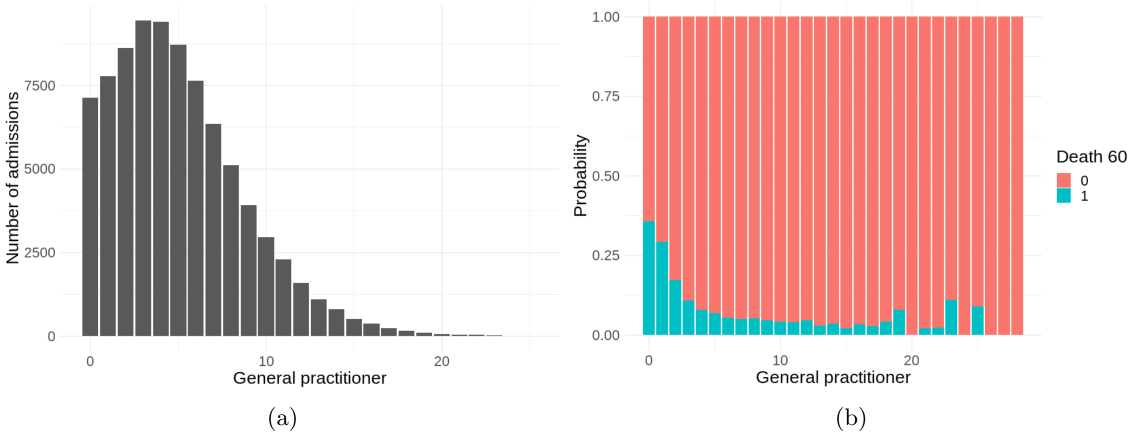
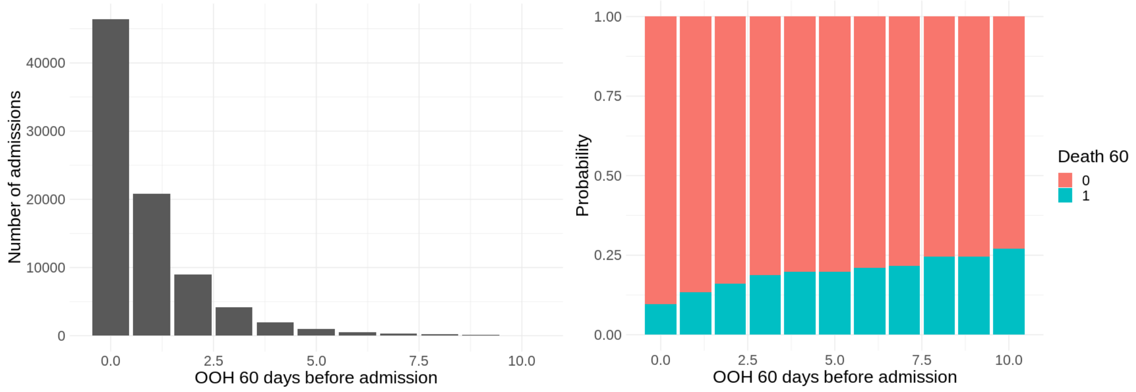


Figure 2.15: The distribution of admissions for visits to the general practitioner in the 60 days after admission, and the distribution of death within 60 days for these visits.

### 2.2.9 Out-of-hours primary health care (OOH)

In addition to visits to the GP, patients may have visited the out-of-hours primary health care (OOH), most likely when the GP was not available such as on weekends or in the evenings. Nearly half of the admissions in the data set did not have visits to the OOH in the 60 days before admission, and others had multiple visits in this time window. The mortality increased with the number of visits as in Figure 2.16b, however, there were only few visits to the OOH in general which could influence the distribution of the mortality rate.



(a)

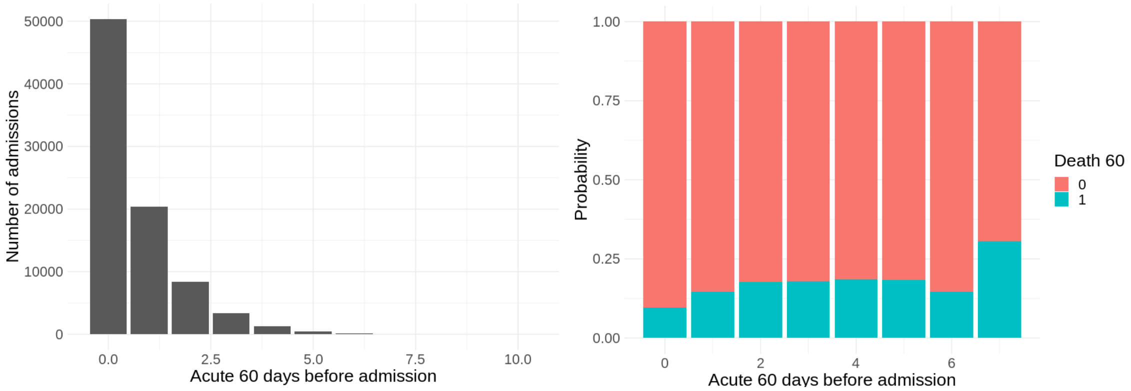
(b)

Figure 2.16: The distribution of admissions for visits to the out-of-hours primary health care in the 60 days before admission, and the distribution of death within 60 days for these visits.

### 2.2.10 Acute visits

There was also a possibility to visit the emergency room (ER) in the days before admission. As Figure 2.17a shows, most admissions had no visits to the ER prior, but for a small quantity of the admissions patients had visited the ER. As for the visits to the OOH, Figure 2.17b shows an increasing risk in mortality as the number of visits increased, and the same argument about uncertainty in the distribution could be used here.

All of these features of visits to different emergency medical services (EMS) were included to give indication of the sickness of the patients when they were hospitalized and if they had been in bad health in the months prior to an acute admission.



(a)

(b)

Figure 2.17: The distribution of admissions for visits to the emergency room in the 60 days before admission, and the distribution of death within 60 days for these visits.

## 2.3 Data Exploration

In addition to the distribution of each feature, there was a need for assessment of the relationship between them. In general, statistical models assume that there is little to no correlation between features used in the model making. The relationship between the different features and the outcome was also of interest as we wanted to assess if there were clear associations that could easily be seen without the need for regression models.

### 2.3.1 The relationship between length of stay and age

Figure 2.4b showed that the risk of dying increased with age. Hence, there was a need to evaluate the relationship between age and LOS as age was believed to be an important factor in the analysis in general. Figure 2.18 shows the smoothed relationship between LOS and age. It was constant for patients between 60 and 85 years. For elderly patients it decreased rapidly and nearly a day from 85 years to 100 years. However, this needs to be seen in context with the number of patients with this age. There were only a few admissions of patients over 95 years, which could give uncertainty in the estimation of the relationship. Additionally, it may be that the hospital chooses to discharge older patients earlier than normal, in order to avoid in-hospital deaths.

### 2.3.2 The relationship between length of stay and health trust

For the boxplots in the next sections, the mean and standard deviation of the logarithm of the LOS is used to estimate the width of the box and wiskers. The line is chosen to be $\exp(E(\log(LOS))$, while the range of the box is $\exp(E(\log(LOS))\pm \mathrm{sd}(\log(LOS)))$. The wiskers range from $\exp(E(\log(LOS)) - 3 \cdot \mathrm{sd}(\log(LOS)))$ to 14, as the upper limit goes way beyond the range in the data set which is not

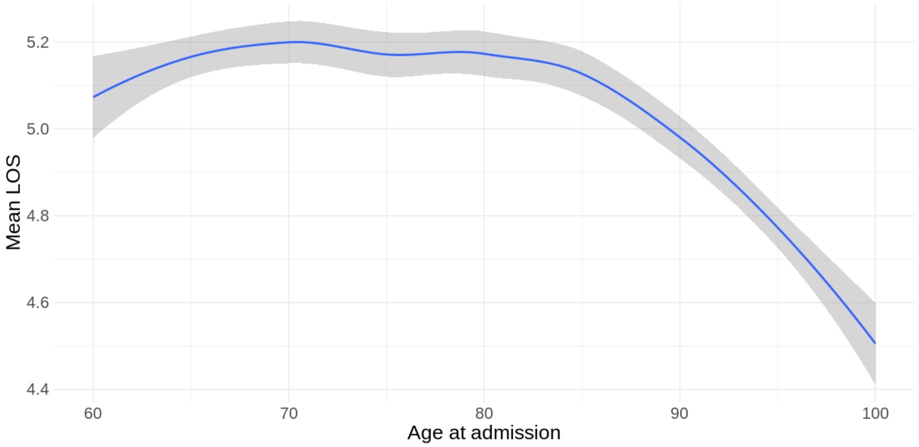


Figure 2.18: The plot shows the relationship between age and average length of stay.

of interest.

Subsection 2.2.4 supplies insight in how the data was spread between the different health trusts. In addition to the number of admissions, there was also important to look at distribution of LOS across the different health trusts. Figure 2.19 shows boxplots for the LOS for each hospital trust. The mean LOS varied across the different health trusts, where the lowest mean was 2.2 days and the highest mean was 5.3 days. These differences argue that the data must be grouped by health trusts in order to get comparable admissions in the models, such that systematic differences could not affect the influence of LOS on the outcome.

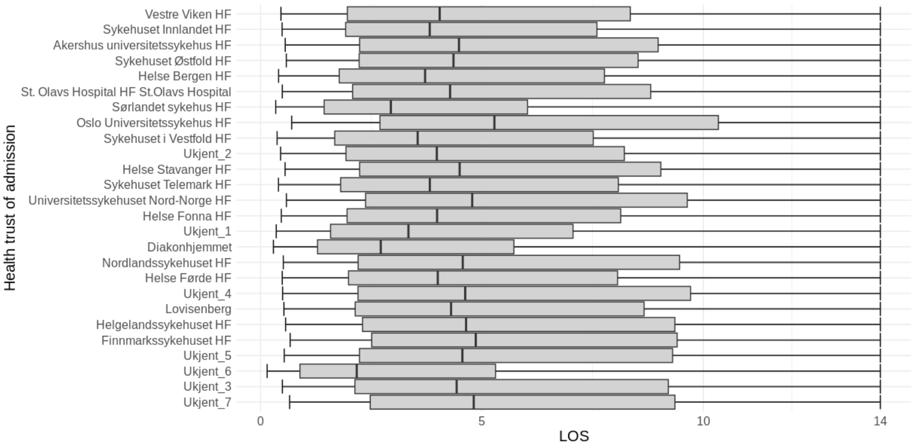


Figure 2.19: Boxplot of length of stay for each health trust. The logarithm of LOS was used to set the median and the ranges before the values where transformed back to a linear scale.

### 2.3.3 The relationship between LOS and time

There was also reason to believe that there were differences in the mean LOS for the different years, months and days of admissions. Figure 2.20, shows boxplots for the years of admissions. There was only minor differences in the mean LOS, and standard deviation. The same could be said for the month of admission, as seen in Figure 2.21. The months May, June, July and December had slightly lower mean compared to the other months, which could be because these months include more holidays compared to the other months.

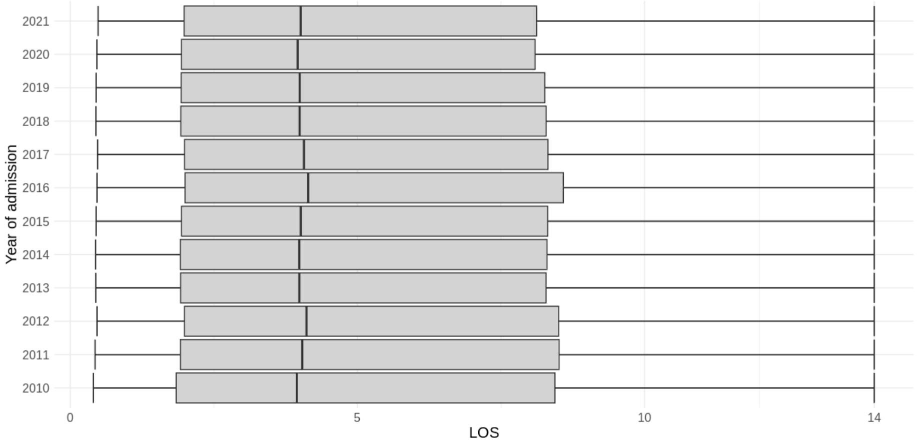


Figure 2.20: Boxplot of length of stay for each year.

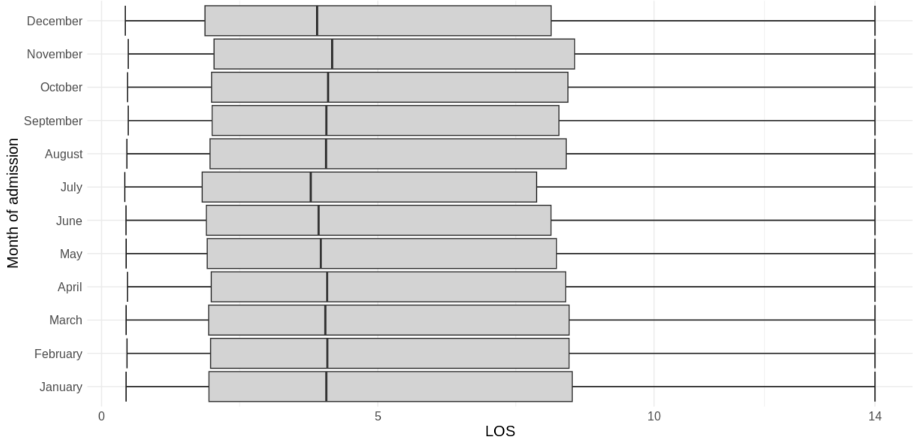


Figure 2.21: Boxplot of length of stay for each month.

Figure 2.8a shows that there were differences in number of admissions across the days of the week, with a clear peak on Mondays. However, Figure 2.22 shows that the mean LOS on Mondays was on the lower side, indicating that even though many patients were admitted on a Monday, they were discharged quite rapidly. As the days of the week went by, the mean LOS increased with a peak on Fridays. Comparing admissions for each day of the week, adjusted for the differences between the days.

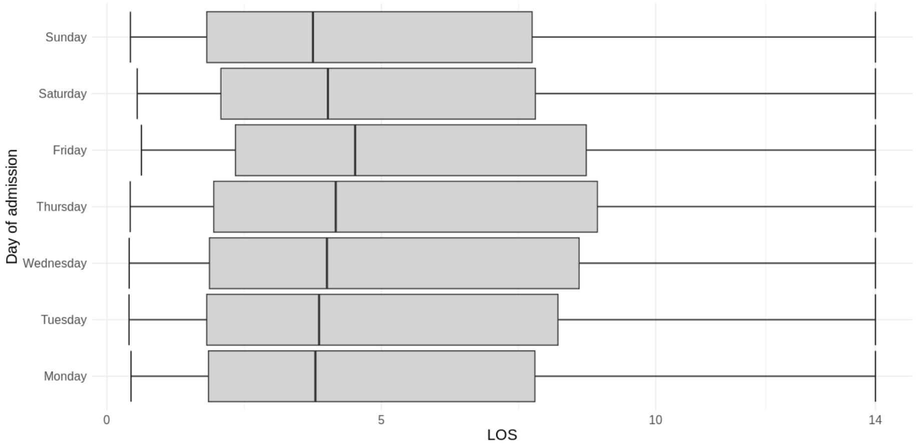


Figure 2.22: Boxplot of length of stay for each day of the week.

### 2.3.4 Correlation between features

The correlation between the numerical features is shown in Figure 2.23. The figure shows minimal correlation between LOS and the other features. The binary variable indicating whether or not a patient died within 60 days, called death_60, had a high negative correlation with days_til_death which was reasonable as these values were used to set the binary variable. The outcome variable death_60 had also some correlation with the outcome GP_60, which showed that patients with elevated risk of dying had a tendency to visit their GP before they died. There was also some negative correlation between the visits prior to admission and days_til_death, indicating that multiple visits to these EMS was associated with earlier death.

Lastly the acute_pre60 and LOS_pre60 had a high correlation, which is reasonable as patients that were so sick that they need to go to the ER have in general a higher risk of being admitted to the hospital.

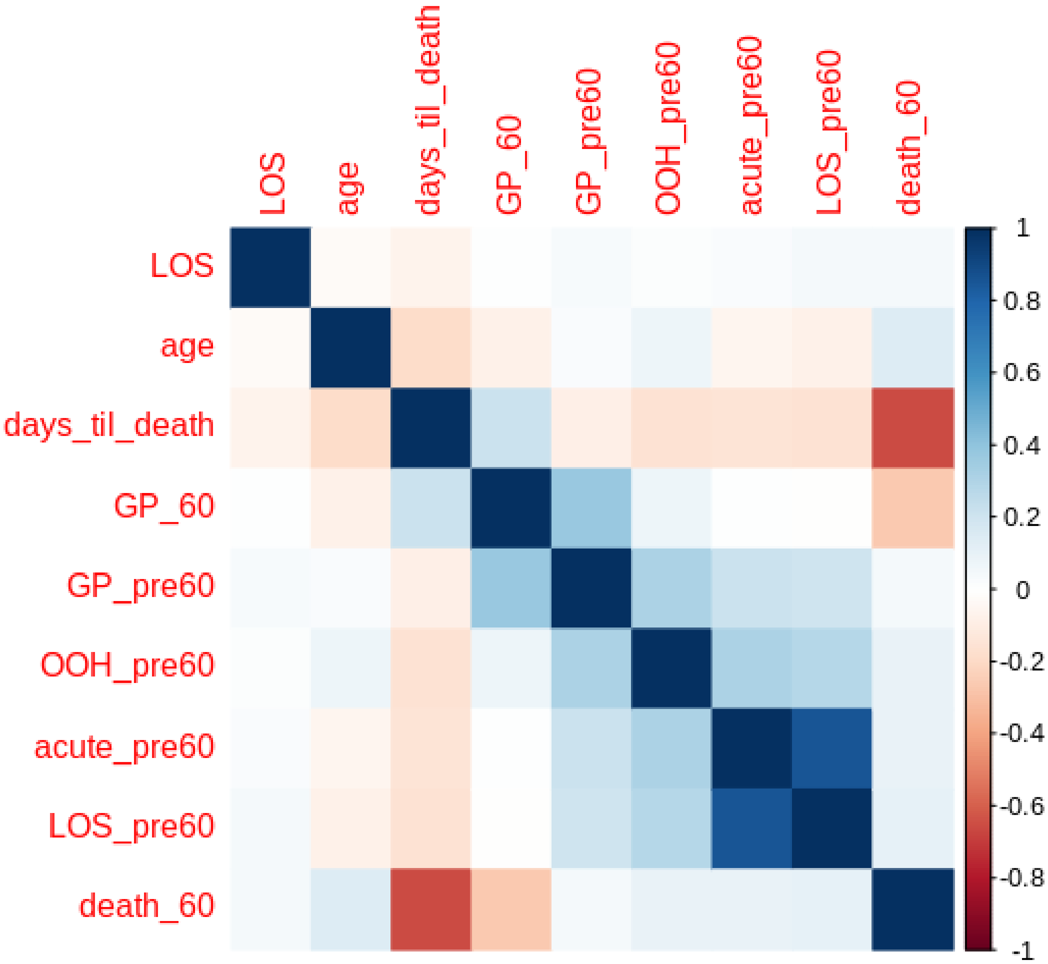


Figure 2.23: Correlation plot showing the correlation between the numerical features in the data.

# Chapter 3

# Statistical Theory

In this chapter we will present the basic ideas of survival analysis, and further the theory behind the statistical methods used in this thesis including Cox and Poisson Regression. In addition, evaluation methods used for model selection are presented.

## 3.1 Survival Analysis

For studies involving a survival time, the use of ordinary regression does not work because one must wait for the event of interest to happen and in addition it is not given that it will happen to all individuals during the study period. The event may not ever happen to some individuals, but it can also happen after the period. The main take is that this information will be unknown in some sense.

Survival analysis is a much-used part of statistics for the analysis of data involving the time to an event occurs. Often this implies time until death, but it can also be the time to a readmission to the hospital or time until cancer is detected. The time from the initiating event, this can be the entry time for a study or the time from when the individuals where born, to the event of interest is denoted *survival time* (Aalen et al. 2008, Ch. 1).

### Censoring and left truncation

An important aspect of survival analysis is the term *censored* or *right-censored.* The survival time of a individual is censored if the event of interest does not happen during the time of study (Aalen et al. 2008, Ch. 1). Such observations are incomplete since we will lack information about the individual. The event may happen later in life for the individual, but this information will be unknown. Censored survival times can also happen if the individual withdraws from the study or is lost to follow-up. Individuals can also enter a study at various times, for instance time zero can be when the patient is admitted to the hospital with HF or when a cancer tumor is detected. Then some patients will have a *delayed* entry compared to patients that entered at an earlier time. This is also called left truncation of the data. This affects the risk set, which is the set of individuals that have not yet experienced the event of interest or are censored at time $t$. The risk set will decrease as time passes, but new individuals entering the study will increase the size of the set.

## 3.2 Basic definitions

### 3.2.1 Survival Function

One of the most essential functions in survival analysis is called the *survival function*, which is a function representing the probability that the event of interest has not happened by time $t$. Denoting the random variable $T$, the time to the event of interest i.e., the survival time, we can write the survival function as:

$$S(t) = P(T > t), \tag{3.1}$$

where $S(t)$ is assumed to be continuous. In most settings, this function will go to zero as $t$ increases. However, for events that do not happen to all individuals, the random variable $T$ can be infinite and the survival function $S(t)$ will converge to a small positive value.

### 3.2.2 Hazard Rate

The other important function is the hazard rate, which is the probability of experiencing the chosen event in the small time interval $[t, t + \Delta t)$, for the individuals that have not yet experienced the event. The hazard rate is defined by a conditional probability, where the survival time $T$ is assumed to have a probability density $f(t)$. The hazard rate is defined by:

$$\alpha(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t) = \frac{f(t)}{S(t)}. \tag{3.2}$$

If the time interval is "infinitesimally small", i.e., the probability of something happening immediately conditionally on the survival time $t$, then the hazard rate $\alpha$ can be written as:

$$\alpha(t) = -\frac{S'(t)}{S(t)}. \tag{3.3}$$

The estimation of the hazard rate can be comprehensive, so the cumulative hazard rate is estimated instead as this is easier. The cumulative hazard rate is defined as

$$A(t) = \int_0^t \alpha(s) ds. \tag{3.4}$$

This can be rewritten as

$$A'(t) = \alpha(t) = -\frac{S'(t)}{S(t)}, \tag{3.5}$$

which results in

$$S(t) = \exp\left\{ -\int_0^t \alpha(s) ds \right\} = \exp\{-A(t)\} \tag{3.6}$$

by integration and using the fact that $S(0) = 1$ (Aalen et al. 2008).

### 3.2.3 Counting processes

The occurrence of events over time contributes to a set of point processes which can be described by a *counting process*, where the number of events that happens are counted as they happen. An example of such a process is the well-known Poisson process that will be described later. A counting process consists of the pair of functions $(N_i(t), Y_i(t))$ with

$$N_i(t) = \text{the number of events happened in } [0,t] \text{ for individual } i,$$
$$Y_i(t) = \begin{cases} 1 & \text{if individual } i \text{ is at risk for the event of interest just before the time } t \\ 0 & \text{otherwise.} \end{cases}$$

$N_i(t)$ is a right-continuous process and will in many problems be a binary variable with 0 until an event of interest happens e.g., death and then become 1 (Therneau and Grambsch 2000). The function $Y_i(t)$ is on the other hand left-continuous and is called an "at risk" indicator. This means that it indicates which individuals can give information about the events at a given time. It can also be written as $Y_i(t) = I(t \leq \tau_i)$, with $I$ being the usual indicator function and where $\tau_i$ is the time at the end of the observation.

We also have the notations $\bar{Y}(t) = \sum_i Y_i(t)$ and $\overline{N}(t) = \sum_i N_i(t)$, which, in a small time interval $(t - \epsilon, t]$, is the number of individual at risk and the total number of events respectively.

### 3.2.4 Nelson-Aalen estimator

The estimation of the cumulative hazard rate given in Eq. (3.4) for a non-parametric model can be done with the *Nelson-Aalen* estimator. The estimator is given by

$$\widehat{A}(t) = \sum_{j:t_j \leq t} \frac{\Delta\overline{N}(t_j)}{\overline{Y}(t_j)}, \tag{3.7}$$

with $\Delta\overline{N}(t) = \overline{N}(t) - \overline{N}(t-)$ being the number of events occurring at time $t$ and $t_j$ the $j$-th failure time. The estimate can be interpret as the slope which estimates the hazard rate.

### 3.2.5 Kaplan-Meier estimator

The survival function in Equation (3.1) can be estimated using the Kaplan-Meier estimator, for censored data (Aalen et al. 2008, Ch. 3). Assuming we only have right-censored survival times with no left-truncation or no tied survival times, let $N(t)$ and $Y(t)$ be as described in Section 3.2.3. The times are ordered $T_1 < T_2 < ...$ for when an occurrence of the event is observed. The time interval $[0, t]$ is usually partitioned into $k$ sub-intervals where $0 = t_0 < t_1 < t_2 < ... < t_K = t$. Then, using the multiplication rule for conditional probabilities one gets

$$S(t) = \prod_{k=1}^{K} S(t_k|t_{k-1}) = \prod_{k=1}^{K} \frac{S(t_k)}{S(t_{k-1})}. \tag{3.8}$$

Each partition $S(t_k|t_{k-1})$ is the conditional probability that the event of interest will happen later than $t_k$ given that it has not happened by time $t_{k-1}$. The assumption of no tied events makes it possible to divide the time interval $[0, t]$ into such small sub-intervals that each interval only includes one observed event and with only censoring at the right side of an interval. Thus we can estimate the conditional probability by

$$S(t_k|t_{k-1}) = \begin{cases} 1 & \text{, if no event is observed in} \, (t_{k-1}, t_k] \\ 1 - \frac{1}{Y(t_{k-1})} = 1 - \frac{1}{Y(T_j)} & \text{, if an event is observed at time } T_j \in (t_{k-1}, t_k] \, . \end{cases} \tag{3.9}$$

This gives the estimate for Eq. (3.8):

$$\hat{S}(t) = \prod_{k=1}^{K} S(t_k|t_{k-1}) = \prod_{T_j \leq t} \left\{ 1 - \frac{1}{\overline{Y}(t_j)}, \right\}, \tag{3.10}$$

the Kaplan-Meier estimator. If however there are tied survival times in the data, the estimator is given by

$$\hat{S}(t) = \prod_{k=1}^{K} S(t_k|t_{k-1}) = \prod_{T_j \leq t} \left\{ 1 - \frac{d\overline{N}(t_j)}{\overline{Y}(t_j)} \right\}, \tag{3.11}$$

where $d\overline{N}(t) = \Delta\overline{N}(t)$.

### 3.2.6 Martingale

The Nelson-Aalen estimator has statistical properties that can be derived from martingale theory. A discrete-time stochastic process is defined as a family of random variables $\{X_t : t \in T\}$ with $T$ being discrete (Pinsky and Karlin 2010, Ch. 1). The discrete-time stochastic process $M = M_0, M_1, ...$ is a *martingale* if

$$E(M_n|M_0, ..., M_{n-1}) = M_{n-1}, n \geq 1, \tag{3.12}$$

with $M_0 = 0$. In words, the martingale property states that the conditional expectation of a random variable in a process, with the past given, equals the previous value (Aalen et al. 2008, Ch. 2).

## 3.3 The Cox Proportional Hazard Model

In most studies, the main goal is to assess the effect of one or several explanatory variables on some outcome of interest. When the outcome of interest is survival or some time interval of interest, with censored data, the most used regression model is the *Cox Proportional Hazard Model.* The hazard rate of individual $i$ is assumed to take the form

$$\alpha_i(t|\mathbf{x}_i) = \alpha_0(t)\exp{(\mathbf{x}_i^T\boldsymbol{\beta})}, \qquad \text{for } i = 1, ..., n. \tag{3.13}$$

The first term, $\alpha_0$, is the *baseline hazard* which describes the shape of the hazard rate as a function of time and is assumed to be identical for all individuals. The feature values for individual $i$ is denoted $x_{i1}, x_{i2}, ..., x_{ip}$ and collected in the feature vector $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$. These can either be fixed over time or vary as time passes. The vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$, consists of $p$ unknown coefficients to be determined by the model. The term $\exp{(\mathbf{x}_i^T\boldsymbol{\beta})}$ is the *hazard ratio* or *relative risk* which describes the impact of the features on the size of the hazard rate. This is because if we take the ratio between the hazard for two different individuals $i$ and $j$, where the features are assumed to be constant over time, we get

$$\frac{\alpha_i(t|\mathbf{x}_i)}{\alpha_j(t|\mathbf{x}_i)} = \frac{\alpha_0(t)\exp{(\mathbf{x}_i^T\boldsymbol{\beta})}}{\alpha_0(t)\exp{(\mathbf{x}_j^T\boldsymbol{\beta})}} = \frac{\exp{(\mathbf{x}_i^T\boldsymbol{\beta})}}{\exp{(\mathbf{x}_j^T\boldsymbol{\beta})}},$$

which is independent of time (Therneau and Grambsch 2000, Ch. 3).

As with regular regression, the aim is to find a model that estimates the values in the $\boldsymbol{\beta}$ vector. This estimation is based on the partial likelihood function which has the form:

$$PL(\boldsymbol{\beta}) = \prod_{i=1}^{n}\prod_{t\geq 0}\left\{\frac{Y_i(t)r_i(\boldsymbol{\beta}, t)}{\sum_j Y_j(t)r_j(\boldsymbol{\beta}, t)}\right\}^{dN_i(t)}. \tag{3.14}$$

The term $r_i(\boldsymbol{\beta}, t)$ denotes the *risk score* for individual $i$, and is defined as $r_i(\boldsymbol{\beta}, t) = \exp{(\mathbf{x}_j^T\boldsymbol{\beta})}$. $dN_i(t)$ represents the increment in $N_i$ in the time interval $[t, t+\Delta t)$, where $\Delta t$ infinitesimal. Taking the logarithm of this equation, gives the log partial likelihood written as

$$l(\boldsymbol{\beta}) = \log[PL(\boldsymbol{\beta})] = \sum_{i=1}^{n}\int_0^\infty \left[Y_i(t)\mathbf{x}_i^T\boldsymbol{\beta} - \log\left(\sum_j Y_j(t)r_j(t)\right)\right]dN_i(t). \tag{3.15}$$

The score vector $U(\boldsymbol{\beta})$ is a $p \times 1$ vector defined as the differentiation of this equation with respect to $\boldsymbol{\beta}$:

$$U(\boldsymbol{\beta}) = \frac{\partial}{\partial\boldsymbol{\beta}}l(\boldsymbol{\beta}) = \sum_{i=1}^{n}\int_0^\infty [\mathbf{x}_i(s) - \bar{x}(\boldsymbol{\beta}, s)]dN_i(s), \tag{3.16}$$

where

$$\bar{x}(\boldsymbol{\beta}, s) = \frac{\sum Y_i(s)r_i(s)\mathbf{x}_i(s)}{\sum Y_i(s)r_i(s)} \tag{3.17}$$

is the weighted mean of $\mathbf{x}$, the observations still at risk at time $s$ and $Y_i(s)r_i(s)$ is the weights.

The negative differentiation of the score function with respect to $\boldsymbol{\beta}$ gives a $p \times p$ matrix called the *observed information matrix*, and is given by

$$\mathcal{I}(\boldsymbol{\beta}) = -\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^\infty V(\boldsymbol{\beta}, s) dN_i(s), \tag{3.18}$$

with

$$V(\boldsymbol{\beta}, s) = \frac{\sum Y_i(s) r_i(s) [\mathbf{x}_i(s) - \bar{x}(\boldsymbol{\beta}, s)]^T [\mathbf{x}_i(s) - x(\bar{\boldsymbol{\beta}}, s)]}{\sum_i Y_i(s) r_i(s)} \tag{3.19}$$

being the weighted variance at time $s$. The maximum partial likelihood estimator (MPLE) $\hat{\boldsymbol{\beta}}$ is found by solving the equation

$$U(\hat{\boldsymbol{\beta}}) = 0,$$

and the solution $\hat{\boldsymbol{\beta}}$ is asymptotically normal distributed with mean $\boldsymbol{\beta}$ and variance $[\epsilon\mathcal{I}(\boldsymbol{\beta})]^{-1}$.

The equation is impossible to solve analytically, so the *Newton-Raphson* algorithm is used so solve the equation iteratively (Therneau and Grambsch 2000, Ch. 3). The algorithm computes the estimates of $\boldsymbol{\beta}$ following the equation:

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + \mathcal{I}(\hat{\boldsymbol{\beta}}^{(t)})^{-1} U(\hat{\boldsymbol{\beta}}^{(t)}),\ t = 0, 1, 2, ... \tag{3.20}$$

where it starts with an initial guess $\hat{\boldsymbol{\beta}}^{(0)}$ and computes new estimates until convergence.

### 3.3.1 Stratified Cox Regression

The regular Cox regression assumes that the baseline hazard is the same for all individuals. If there are reasons to believe that this is not true one can stratify the model, where the individuals are split into disjoint subsets or *strata*. Then each strata gets a unique baseline hazard, but the coefficient vector $\boldsymbol{\beta}$ is the same for all strata. The model takes the form:

$$\alpha(t|\mathbf{x}_i) = \alpha_{s0}(t) \exp(\mathbf{x}_i^T \boldsymbol{\beta}), \tag{3.21}$$

where $s = 1, .., k$ represents the different strata giving the distinct baseline hazards for each strata, $\alpha_{s0}$.

The unknown coefficient vector $\boldsymbol{\beta}$ is computed in the same way as for the regular Cox, but now the log partial likelihood given in Eq. (3.15) is computed for each strata and then summed up giving the overall log likelihood:

$$l(\boldsymbol{\beta}) = \sum_{s=1}^{k} l_s(\boldsymbol{\beta}).$$

This approach is also taken for the score vector and the information matrix used in Eq. (3.20).

The different strata are treated as a categorical variable in the model and if several strata are included, each unique combination of these gets a unique baseline hazard. The disadvantage of doing this compared to including it as a normal categorical feature is that the model does not produce any estimate of the importance of the different strata, i.e., it does not give any p-value. In addition, if the number of strata is large the precision of the estimated coefficient vector may be reduced and give a worse hypothesis test. On the other hand, it is the easiest way of adjusting for confounding features. For categorical features with many levels, it may be easier to include them as strata instead, since it gives a model that is easier to interpret and with less estimated coefficients.

The use of strata could affect the power of the Cox model as it affects the sample size. As the number of strata increases, the number of individuals in each strata decreases which can reduce the power of the model greatly. However, in the presence of censoring, with a high survival rate the stratified models may end up not losing much power.

### 3.3.2 Important residuals for the Cox Regression

There are several residuals which are interesting when modelling the data with a Cox model, the first of them being the martingale residuals. These are mostly used to assess the functional form of the features used in the models (Therneau and Grambsch 2000). The second residual of interest is the deviance, which is a normalization transform of the martingale residuals. This residual was mostly made to identify individual outliers when plotting. The score residuals and the Schoenfeld residuals are the other two residuals of interest. The score residuals are used to assess individual influence and for robust variance estimation while the Schoenfeld residuals are used to check the assumption of proportional hazard curves.

#### Martingale Residuals

The martingale residual is one of the most important residuals for Cox Regression, as it assesses the functional form of the numerical features (Therneau and Grambsch 2000, Ch. 4). These residuals are defined as

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) r_i(\boldsymbol{\beta}, s) \alpha_0(s) ds, \tag{3.22}$$

for individual $i$, or with words the difference between the observed and expected number of events for each individual over the full study time.

For a model that is fit to data, the martingale residual process is defined as

$$\widehat{M_i}(t) = N_i(t) - \int_0^t Y_i(s) r(\hat{\boldsymbol{\beta}}, s) d\hat{A}_0(s), \tag{3.23}$$

where $\hat{\boldsymbol{\beta}}$ is the estimate of the MPLE and $\hat{A}_0(t)$ is the estimation of the baseline cumulative hazard given by:

$$\hat{A}_0(t) = \int_0^t \frac{d\overline{N}(s)}{\sum_{j=1}^n Y_j(s) r_j(\hat{\boldsymbol{\beta}}, s)}, \tag{3.24}$$

where $d\overline{N}(t) = \Delta\overline{N}(t)$.

For a Cox model, where there are only time-independent features the martingale residual reduces to

$$\hat{M}_i(t) = \delta_i - \hat{A}_0(\tau_i) r_i(\boldsymbol{\beta}, s), \tag{3.25}$$

where $\delta$ is 0 for censored observations and 1 for uncensored observations and $\tau_i$ denotes the observation time for individual $i$ (Therneau, Grambsch and Fleming 1990). The residuals take values in the interval $(-\infty, 0]$ for uncensored observations, and in $(-\infty, 1]$ for censored observations. This means that for plots of continuous features versus the residuals, a locally estimated scatterplot smoothing (LOESS) curve should be parallel to a constant line in zero.

### Deviance Residuals

A disadvantage with the martingale residuals, especially when the cox is a model with single events, is the skewness of the model (Therneau, Grambsch and Fleming 1990). To have a more normal shaped distribution of the residuals, it can be helpful to transform them. The deviance residuals are inspired by the deviance residuals for generalized linear models, which is defined as:

$$D = 2(l(\beta_{sat}) - l(\hat{\beta}_0)),$$

where $\beta_{sat}$ is the estimated coefficients for a model with as many coefficients as number of observations $n$, giving $\mu_i = y_i$ and $\hat{\beta}_0$ is the coefficient for the estimated model (Fahrmeir et al. 2022).

The deviance residuals are defined as

$$d_i = \text{sign}(\widehat{M_i}) \cdot \sqrt{-2(\widehat{M_i} + \delta_i \log(\delta_i - \widehat{M_i}))}$$

for the Cox models, where

$$\text{sign}(x) = \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases} \tag{3.26}$$

### Score Residuals

The third important residual is the score residual. This is used to assess each data point's impact on the fit of the model (Therneau and Grambsch 2000, Ch. 7). The most straightforward way of measuring this impact is the jackknife value

$$J_i = \hat{\beta} - \hat{\beta}_{(i)},$$

where $\hat{\beta}_{(i)}$ is the result from a fitted model where observation $i$ is left out. This way of assessing the impact is not effective as it requires a high amount of computation, and hence other methods are used to find an approximate of this value. For a Cox model one uses the score residuals to assess influence. The score process for the $i$-th individual is defined by

$$U_i(\boldsymbol{\beta}, t) = \int_0^t [\mathbf{x}_i(s) - \bar{x}(\boldsymbol{\beta}, s)] dM_i(s),$$

with $\bar{x}(\boldsymbol{\beta}, s)$ given by Eq.(3.17). The score process is a $p \times 1$ vector with the components $U_{ij}(\boldsymbol{\beta}, t)$, $j = 1, .., p$ and is written as $U_i(\boldsymbol{\beta}, t) = (U_{i1}(\boldsymbol{\beta}, t), ..., U_{ip}(\boldsymbol{\beta}, t))^T$.

In addition to the individuals and the features, the set of score processes is also related to time. When looking at the time points which include one or more events, the time becomes discrete and denoting the $k$th event time $t_k$ we get the components

$$U_{ijk}(\boldsymbol{\beta}) = \int_{t_k}^{t_k} [x_{ij}(s) - \bar{x}_j(\boldsymbol{\beta}, s)] dM_i(s). \tag{3.27}$$

We define the *score residual* as

$$U_{ij} = U_{ij}(\boldsymbol{\beta}, \infty),$$

which in total forms a $n \times p$ matrix.

A way of estimating the jackknife value for a Cox model is to use the Newton-Raphson iteration. Eq. (3.20) can be rewritten as

$$\hat{\boldsymbol{\beta}}^{(n+1)} - \hat{\boldsymbol{\beta}}^{(n)} = \Delta\boldsymbol{\beta} = \mathbf{1}^T U \mathcal{I}^{-1} \equiv \mathbf{1}^T D, \tag{3.28}$$

where $U$ is the $n \times p$ matrix composed of the score residuals given above. The matrix $D$ is called the matrix of *dfbeta* residuals, and the calculations necessary for this matrix are the same calculations used to fit the model hence this approach does not require any additional computations.

In addition to assessing influential observations, the jackknife can be used to compute a robust estimate of the variance for a Cox model. Let $J$ be a $n \times p$ matrix, with the $i$-th row being the jackknife value $J_i$. Then the jackknife estimate of the variance can be written as

$$V_J = \frac{n-1}{n}(J - \bar{J})^T (J - \bar{J}),$$

with $\bar{J}$ being a matrix with the column means of $J$. This matrix can be approximated with $D^T D = \mathcal{I}^{-1} U^T U \mathcal{I}^{-1}$, which is a sandwich type of estimator $ABA$ where $A$ is the ordinary variance and $B$ is the correction term.

In an ordinary Cox model, one assumes that all observations are independent when estimating the variance for $\hat{\boldsymbol{\beta}}$ (Therneau and Grambsch 2000, Ch. 8). When a model has multiple observations per individual, the jackknife value must be calculated differently. In such a model the choice is a grouped jackknife estimate, which leaves out one individual at a time instead of one single observation.

### Schoenfeld Residuals

One of the most important aspects with the Cox model is the assumption of proportional hazard curves (Therneau and Grambsch 2000, Ch. 6). If the number of strata is small, this can be assessed with plots of the survival curves. When the hazard curves are proportional, the log of the survival curves will have the same shape and decay with the same rate. However, if the levels in the strata are many or one has continuous features, plotting the survival curves will not be as interpretable.

For such problems, the *Schoenfeld residuals* will be a better choice. At the $k$th event time, this residual is defined as

$$s_k = \int_{t_{k-1}}^{t_k} \sum_{i=1}^{n} [x_i(s) - \bar{x}(\hat{\boldsymbol{\beta}}, s)] dN_i(s), \tag{3.29}$$

where $\bar{x}(\hat{\boldsymbol{\beta}}, s)$ is given by Eq. (3.17). The Schoenfeld residuals is a sum over the score process array, which gives a process that varies over time. In total they form a $k \times p$ matrix with one row per event time. It is possible to plot the event time against these residuals for each feature in the model, where a LOESS curve fitted to the data should be a straight curve with mean and gradient zero. This is an indication that the feature satisfies the proportional hazards assumption and is independent of time.

It is also possible to check the proportionality assumption with a hypothesis test using the chi-square distribution. This can either be done for each feature or for all at once with either 1 or $p$ degrees of freedom, respectively. Nonetheless, the hypothesis test can be significant and indicate non-proportionality even though the LOESS curve shows small variation of $\hat{\beta}(t)$ vs. $\hat{\beta}$, where $\hat{\beta}(t)$ is the estimated effect of the feature at time $t$ while $\hat{\beta}$ is the best "overall" effect.

### 3.3.3 Tied event times

In most of the material presented, there is an assumption that each event time corresponds to only one event. This is not always the case, patients can die after the same number of days included in the study, and the deaths are independent of each other. For the Cox model there are three ways of dealing with such problems, the *Breslow approximation*, the *Efron approximation* and the *exact partial likelihood* (Therneau and Grambsch 2000). The Breslow approximation is the easiest to compute, but the Efron approximation is more accurate. The exact partial likelihood involves heavy computation if there are many events at the same event time. Let $r_i(t_i)$ be the risk score of individual $i$ at time $t_i$, and assume $k$ out of $n$ individuals die at the same time. Then the Breslow approximation of the $k$ first terms of the likelihood would be

$$\left(\frac{r_1}{r_1 + r_2 + ... + r_m}\right)\left(\frac{r_2}{r_1 + r_2 + ... + r_m}\right)...\left(\frac{r_k}{r_1 + r_2 + ... + r_m}\right). \tag{3.30}$$

With this approach individuals which died would be counted multiple times, producing bias in addition to a less accurate approximation. The Efron approximation on the other hand, weights the deaths in the denominator resulting in the terms

$$\left(\frac{r_1}{r_1 + r_2 + ... + r_m}\right)\left(\frac{r_2}{\frac{k-1}{k}\sum_{i=1}^{k} r_i + \sum_{i=k+1}^{m} r_i}\right)...\left(\frac{r_k}{\frac{1}{k}\sum_{i=1}^{k} r_i + \sum_{i=k+1}^{m} r_i}\right). \tag{3.31}$$

## 3.4 The Poisson Model

When the outcome of interest is to count the number of times an event happens in a time-specific interval, the Poisson regression model is the most widely used model (Fahrmeir et al. 2022, Ch. 5). Poisson distributed variables are discrete, counting variables where the possible outcomes are $\{0, 1, 2, ...\}$. The probability density function of a Poisson distributed variable gives the probability of a specific number of events happening in a fixed time interval, with a constant rate $\lambda$, and is given by

$$f(y|\lambda) = \frac{\lambda^y}{y!} \exp(-\lambda), \quad \lambda > 0, \quad y = 0, 1, 2, ... \tag{3.32}$$

The Poisson model has the property $\mathrm{E}(Y) = \mathrm{Var}(Y) = \lambda$.

### 3.4.1 Poisson Regression Model

In most situations one considers regression models where the response variable is assumed to follow a Poisson distribution, instead of focusing on a single variable. Let $Y_1, Y_2, ..., Y_n$ be random variables and assume that the observed values $y_1, ..., y_n$ are Poisson distributed with rate $\lambda_i$ and that they are conditionally independent. Let $\mathbf{x}_i = (1, x_{i1}, ..., x_{ip})^T$ denote the feature vector for observation $i$, and assume that the unknown coefficient vector is denoted by $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$. The observed values of the features can be collected in the $n \times (p + 1)$ matrix called the design matrix $X$,

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{pmatrix}. \tag{3.33}$$

The observations are no longer assumed to be identically distributed, hence the rate $\lambda_i$ is given by

$$\lambda_i = E(y_i|\mathbf{x}_i) = \exp(\mathbf{x}_i^T \boldsymbol{\beta}) = \exp(\eta_i),$$

with the linear predictor $\eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_1 x_{ip}$. As for the Cox regression, and regression models in general, the optimal values for $\boldsymbol{\beta}$, denoted $\hat{\boldsymbol{\beta}}$, can be calculated by solving the equation

$$U(\hat{\boldsymbol{\beta}}) = 0,$$

where $U$ is the score function given by differentiating the log-likelihood function. The likelihood for Poisson distributed data is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i}}{y_i!} \exp(-\lambda_i), \tag{3.34}$$

which gives the log-likelihood

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \log(\lambda_i) - \lambda_i - \log(y_i!)] = \sum_{i=1}^{n} [y_i \mathbf{x}_i^T \boldsymbol{\beta} - \exp(\mathbf{x}_i^T \boldsymbol{\beta}) - \log(y_i!)]. \tag{3.35}$$

Differentiating this equation with respect to $\boldsymbol{\beta}$ gives the score vector

$$U(\boldsymbol{\beta}) = \frac{\partial l}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \mathbf{x}_i^T - \mathbf{x}_i^T \exp(\mathbf{x}_i^T \boldsymbol{\beta})] = \sum_{i=1}^{n} \mathbf{x_i}^T [y_i - \lambda_i]. \tag{3.36}$$

Solving this equation for $\hat{\beta}$ gives a nonlinear system of equation, which can be solved numerically using the Fisher scoring algorithm (Fahrmeir et al. 2022). The algorithm performs the computations

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + F^{-1}(\hat{\boldsymbol{\beta}}^{(t)}) U(\hat{\boldsymbol{\beta}}^{(t)}),\ t = 0, 1, 2, ..., \tag{3.37}$$

where the starting value $\hat{\boldsymbol{\beta}}^{(0)}$ is known. The matrix $F$ is the expected Fisher information matrix, $F(\hat{\boldsymbol{\beta}}) = E(\mathcal{I}(\hat{\boldsymbol{\beta}}))$ which normally is easier to compute than the observed information matrix. The $n \times n$ observed information matrix $\mathcal{I}$, is given by

$$\mathcal{I}(\boldsymbol{\beta}) = -\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \lambda_i \tag{3.38}$$

for the Poisson distribution. In this case, the two matrices are identical, and the algorithm is identical to the Newton-Raphson algorithm given in Eq. (3.20).

The convergence criterion for both algorithms have several different options, with one of them being

$$\|\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t)}\| / \|\hat{\boldsymbol{\beta}}^{(t)}\| \leq \epsilon$$

where $\epsilon$ is some small number. When the criterion is met the estimate for $\hat{\boldsymbol{\beta}}$ is set to $\hat{\boldsymbol{\beta}}^{(t)}$. To be able to converge to a solution, the algorithm needs an information matrix that is invertible for all $\boldsymbol{\beta}$. This is the case when the design matrix X, has full rank $p + 1$. When $n \to \infty$, the estimated coefficient vector has the distribution

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} N(\boldsymbol{\beta}, F^{-1}(\hat{\boldsymbol{\beta}})).$$

### 3.4.2 Offset term

For rate data which is a count of events divided on some measure of the unit's exposure, Poisson regression is also useful. This exposure is handled with an *offset* term, where log(exposure) is the offset. The Poisson rate is then given by

$$\lambda_i = \frac{E(y_i|\mathbf{x}_i)}{\text{exposure}_i} = \exp(\boldsymbol{x}_{ij}^T \boldsymbol{\beta}) \quad \text{or} \quad \log(\lambda_i) = \mathbf{x}_i^T \boldsymbol{\beta} + \log(\text{exposure}_i). \tag{3.39}$$

### 3.4.3 Mixed Poisson Models

For data which are believed to be clustered, either if there are repeated observations per individual or per cluster, random effects are introduced to handle differences between clusters (Fahrmeir et al. 2022, Ch. 7). For data sets consisting of $n_i$ repeated observations $(y_{i1}, ..., y_{in_i}, \mathbf{x}_{i1}, ...\mathbf{x}_{in_i})$ where

$i = 1, ..., m$ denotes the cluster and $j = 1, ..., n_i$ denotes observation $j$ in cluster $i$, the rate for the Poisson model becomes

$$\lambda_{ij} = \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta} + \gamma_{0i}), \tag{3.40}$$

where $\gamma_{0i} \sim N(0, \tau_0^2)$ is a random deviate from the fixed intercept $\beta_0$ and called the random, or cluster-specific, intercept.

### 3.4.4 Deviance residuals

For a Poisson model, deviance is used to assess the goodness-of-fit of the model (Roback and Legler 2021, Ch. 4). The deviance measures how much the predictions deviates from the observed data, so the deviance residual is defined as

$$d_i = \text{sign}(y_i - \hat{\lambda}_i) \cdot \sqrt{2\left[y_i \log\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i)\right]}, \tag{3.41}$$

with $\hat{\lambda}_i = \exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}})$ and $\text{sign}(y_i - \hat{\lambda}_i)$ is as defined in Eq. (3.26).

The deviance will be small for observations that are well fitted by the model, and larger for observations were the model struggles with the fit. When the model fit is perfect, we have $y_i = \hat{\lambda}_i$, and $d_i = 0$. Hence, observations that fits perfectly will not contribute to the sum of squared deviance's, which is called the residual deviance of the model.

## 3.5 Hypothesis testing

### 3.5.1 The significance of the coefficient vector

In addition to finding the optimal values for the $\boldsymbol{\beta}$ vector, one aims to find out if these coefficients are significant or not. If they are significant, this means that the feature is significant for the outcome of interest and has an influence on it. To test the significance of a single coefficient $\beta_j$, we test the hypothesis

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0,$$

with the test statistic

$$t_j = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)}, \tag{3.42}$$

where $\text{se}(\hat{\beta}_j) = \widehat{\text{Var}(\hat{\beta}_j)}^{1/2}$ denotes the estimated standard deviance or standard error of the coefficient. The test statistic is t-distributed with $n - p$ degrees of freedom, and the null hypothesis is rejected if

$$|t_j| > t_{1-\alpha/2}(n - p),$$

with $\alpha$ being the significance level. If one wishes to test several coefficients simultaneously, the hypothesis test will be for the subvector $\boldsymbol{\beta}_1 = (\beta_1, ..., \beta_r)^T$, with

$$H_0 : \boldsymbol{\beta}_1 = \mathbf{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\beta}_1 \neq \mathbf{0}.$$

The test statistic in this case is

$$F = \frac{1}{r}\hat{\boldsymbol{\beta}_1}^T \widehat{\text{Cov}(\hat{\boldsymbol{\beta}}_1)}^{-1} \hat{\boldsymbol{\beta}}_1 \sim F_{r,n-p}. \tag{3.43}$$

The null hypothesis is then rejected if $F > F_{r,n-p}$.

These are both special cases of tests for general linear hypotheses (Fahrmeir et al. 2022, Ch. 3), denoted

$$H_0 : C\boldsymbol{\beta} = \mathbf{d} \quad \text{vs.} \quad H_1 : C\boldsymbol{\beta} \neq \mathbf{d}.$$

The matrix $C$ is a $r \times p$ matrix with $\text{rank}(C) = r \leq p$, the vector $\boldsymbol{\beta}$ is the usual coefficient vector and $\mathbf{d}$ is a $r \times 1$ vector. The hypothesis test tests $r$ linearly independent conditions at once, with the test statistic

$$F = \frac{1}{r}(C\hat{\boldsymbol{\beta}} - \boldsymbol{d})^T(\hat{\sigma}^2 C(X^TX)^{-1}C^T)^{-1}(C\hat{\boldsymbol{\beta}} - \boldsymbol{d}) \sim F_{r,n-p}, \tag{3.44}$$

which is rejected if it is larger than $F_{r,n-p}(1-\alpha)$.

### 3.5.2 Likelihood-based testing

For the hypothesis test $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, where $\boldsymbol{\beta}_0$ is a known initial value, the most common test statistics are the *likelihood ratio test statistic, the Wald statistic* and *the score statistic* (Fahrmeir et al. 2022). The likelihood ratio test statistic is defined as

$$\chi^2_{LR} = 2\{l(\hat{\boldsymbol{\beta}}) - l(\boldsymbol{\beta}_0)\}. \tag{3.45}$$

If the estimate $l(\hat{\boldsymbol{\beta}})$ is much larger than $l(\boldsymbol{\beta}_0)$ resulting in a large value for $\chi^2_{LR}$, the null hypothesis is rejected in favor of the alternative hypothesis $H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ (Fahrmeir et al. 2022, Appendix B). The second test statistic possible to use is the Wald test statistic,

$$\chi^2_W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \hat{\mathcal{I}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \tag{3.46}$$

with the observed information matrix $\hat{\mathcal{I}} = \mathcal{I}(\hat{\boldsymbol{\beta}})$ given by Eq. (3.18). If the model only includes one feature, this statistic reduces test statistic given in Eq. (3.42). The Wald statistic is easier to compute compared to the likelihood statistic, however it is considered the least reliable statistic.

In addition to finding the estimates for $\boldsymbol{\beta}$, the score vector can be used to assess the significance of the estimates. The third test statistic is the *score test statistic.* It is given by

$$\chi^2_{SC} = U(\boldsymbol{\beta}_0)^T \mathcal{I}(\boldsymbol{\beta}_0)^{-1} U(\boldsymbol{\beta}_0), \tag{3.47}$$

and can be computed with the first iteration of the Newton-Rapshon algorithm where

$$\mathcal{I}(\boldsymbol{\beta}_0)^{-1} U(\boldsymbol{\beta}_0) = \boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}_0$$

and $\boldsymbol{\beta}^{(1)}$ is the first iteration of Eq.(3.20).

All three test statistics are asymptotically equivalent and asymptotically $\chi^2$-distributed with $p$ degrees of freedom under $H_0$. They are rejected if the estimated values of the test statistics are larger than $\chi^2_p(1-\alpha)$. For the Cox models, these test statistics are used to test if the final estimate for the coefficients differs from the initial values.

The p-values are calculated using the $\chi^2_p$ distribution, where

$$p = P(X^2_p \geq \chi^2 | H_0 \text{ is true}) = 1 - F_{\chi^2_p}(\chi^2)$$

and $F_{\chi^2_p}(\chi^2)$ can be found in tables. If the $p$-value is less than some significance level $\alpha$, usually $0.01, 0.05$ or $0.1$, then the test is significant, and the null hypothesis is rejected.

## 3.6 AIC

To compare statistical models with various levels of strata or random intercept, or distinctive features in general the Akaike's information criterion (AIC) is a much-used evaluation method. It is defined as

$$\text{AIC} = -2l(\hat{\beta}) + 2p, \tag{3.48}$$

using the log-likelihood of the estimated values of the coefficient vector from the model, and where the last term penalizes complex models. In general, a lower value corresponds to a better fit of a model. The AIC is a compromise between a good fit to the data and model complexity, as too many features often result in overfitting.

# Chapter 4

# Methods

In this Chapter we will first present how the simulation study was done. Further, the different Cox models for mortality with different levels of strata are presented. Lastly the Poisson models for visits to the GP are presented. The results from these approaches are presented in the next chapter.

## 4.1 Simulation study

Based on the survival rate from the data set shown in Figure 2.1, we wanted to simulate data with similar shape with the purpose of analysing the association between mortality and LOS and to see how immortal time bias could be a factor in such a problem. The simulation consisted of a fixed probability of being discharged each day given by

$$p_{out} = 0.5 \cdot \frac{d_s}{14}, \tag{4.1}$$

where $d_s = 1, ..., 13$ was the number of days in the stay. The maximum number of days a patient could be hospitalized, was set to 14 and hence for $d_s = 14$ this probability was set to 1. The patients were followed for 60 days after admission, and several scenarios for the risk of dying, $p_{dead}$, were tested. The scenarios are given by:

$$\begin{aligned}
\text{High acute risk rapid decrease} &= 0.03 + 0.01 \cdot \exp{(-5 \cdot \frac{d_f}{60})} \\
\text{High risk in hospital phase} &= \frac{1}{(10 + \exp((d_f/14)^2 + 1))} + 0.02 \\
\text{Constant rate} &= 0.05 \\
\text{Increasing risk in hospital phase} &= \frac{\frac{d_f}{15} \cdot \exp(-d_f/15)}{7.5} + d_f \cdot 0.0004
\end{aligned}$$

Where $d_f = 1, ..., 60$ is the day of follow-up after admission.

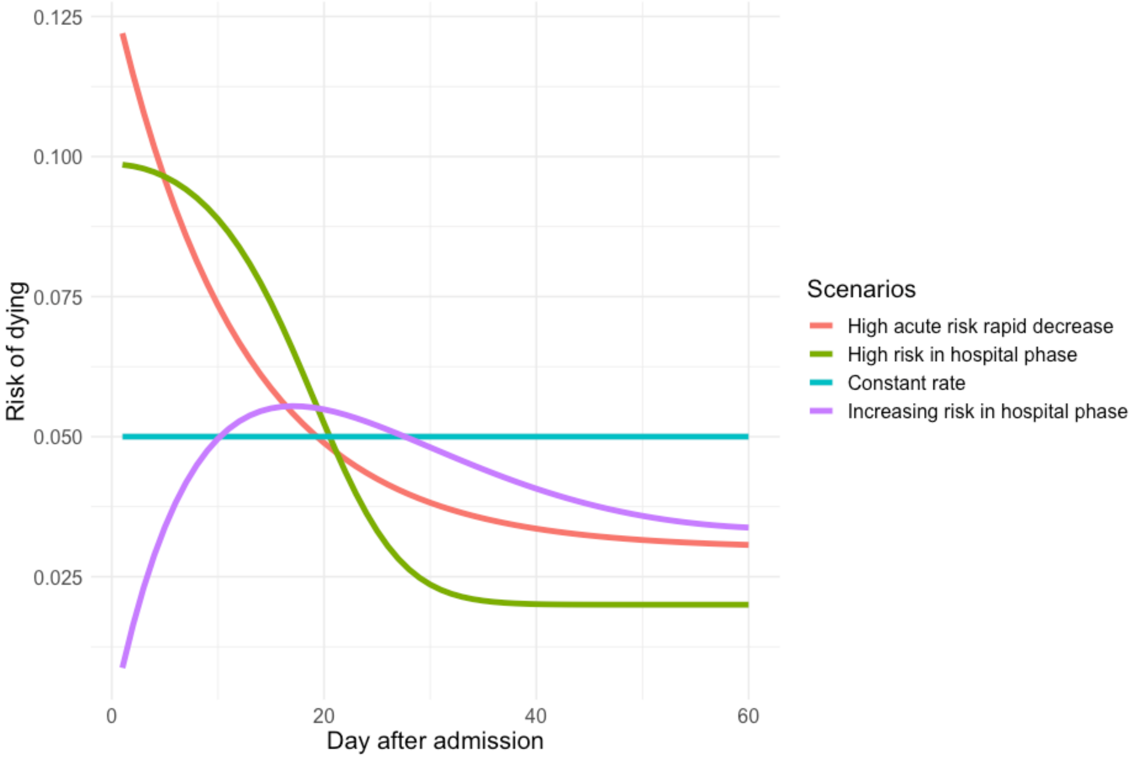


Figure 4.1: The figure shows the risk of dying in the 60 days after admission, for the various scenarios.

Figure 4.1, shows that the scenarios had different shapes and starting point, however they all flattened out after some time. Two of the scenarios had a high initial risk, which decreased fast for the first one, while the other one was at a higher value in the hospital phase before it decreased rapidly. The third scenario was a constant value for each day after admission, while the fourth scenario had an increasing risk in the hospital phase with a peak around 17 days after admission, before it decreased and flattened out.

The simulation was done with 5000 iterations per scenario to make the results more robust. For each iteration and scenario, LOS and day of death was simulated for 10,000 patients. The LOS was chosen to be the first day where the probability of being discharged, $p_{out}$, was larger than a randomly generated number $p$, between 0 and 1 following the standard uniform distribution (Wikipedia 2023), given by

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

The day of death for each patient was set to be the first day $p_{dead}$ was larger than a randomly generated number which followed the exponential distribution with rate $\lambda = 0.07$, with the probability density function

$$p(x) = \lambda e^{-\lambda x}, x \geq 0. \tag{4.3}$$

Patients that survived to day 60 and further, had the day of death set to *Inf* as it became unknown. A Cox model was estimated in each iteration and each scenario, with the 60 day survival as outcome.

## 4.2 Cox Regression on mortality

To study the survival time for patients admitted with heart failure Cox Models where used. For patients with multiple admissions, the study time was defined from the first admission date and until they died or the censor date December 31st 2021. Patients with a single admission were followed from admission date and until death or the censor date. The primary goal was to see if LOS had an influence on the 60 day mortality, and hence the outcome of interest was if they died within 60 days or not. This time interval was chosen to capture a period when HF was the most likely cause of death. For all of the Cox Regression models fitted in this thesis we used the *coxph* function from the package *survival* (Therneau and Grambsch 2000) in the statistical software R (R Core Team 2023). In the data set there were multiple admissions per patients which were clustered together using the *cluster* function from same package.

### Models with LOS as the only feature

Similar to the simulation, we started with regression models with LOS as the only feature in the model. This was done to see the direct influence of LOS on the 60-day mortality, without any adjustments from other features. The hazard rate for admission $i$ could then be written as

$$\alpha_i(t|\mathbf{x}_i) = \alpha_0(t) \exp(\mathbf{x}_i^T \boldsymbol{\beta}), \tag{4.4}$$

where the relative risk was given by

$$\exp(\mathbf{x}_i^T \boldsymbol{\beta}) = \exp(\beta_1 \cdot \mathrm{LOS}_i). \tag{4.5}$$

### Models with multiple features

Further, we computed models including several features. The data included features such as age, sex, and others, which were registered before the LOS of an admission was set and hence not a consequence of it. These features could be included to remove differences between patients in order to get the patients to be as comparable as possible, such that the focus could be on the differences in the LOS. It was unlikely that these features could introduce bias or error, as long as they were registered correctly. Including the relevant features, the relative risk of Eq. (4.4) was now given by

$$\begin{aligned}\exp(\mathbf{x}_i^T\boldsymbol{\beta}) = \exp(&\beta_1 \cdot \mathrm{LOS}_i + \beta_2 \cdot \mathrm{age}_i + \beta_3 \cdot \mathrm{woman}_i + \beta_4 \cdot \mathrm{holiday}_i + \beta_5 \cdot \mathrm{GP\_pre60}_i \\ &+ \beta_6 \cdot \mathrm{OOH\_pre60}_i + \beta_7 \cdot \mathrm{acute\_pre60}_i + \beta_8 \cdot \mathrm{education1}_i + \beta_9 \cdot \mathrm{education2}_i \\ &+ \beta_{10} \cdot \mathrm{education3}_i + \beta_{11} \cdot \mathrm{education4}_i + \beta_{12} \cdot \mathrm{education5}_i + \beta_{13} \cdot \mathrm{education6}_i \\ &+ \beta_{14} \cdot \mathrm{education7}_i + \beta_{15} \cdot \mathrm{education8}_i + \beta_{16} \cdot \mathrm{education9}_i + \beta_{17} \cdot \mathrm{LOS\_pre60}_i \\ &+ \beta_{18} \cdot \mathrm{shiftEvening}_i + \beta_{19} \cdot \mathrm{shiftNight}_i).\end{aligned} \tag{4.6}$$

The features woman and holiday were binary, where the reference categories were man and normal day respectively. For the categorical feature education, level 0, representing no education, was set as the reference category. Thus, the estimated values for the other levels represented the relative difference between no education and the various levels of education. Instead of including all admission hours, these were split into three shifts with *Day* being the reference category for the admission hours from $8 - 15$, while *Evening* included the hours from $16 - 23$ and *Night* between $00 - 07$. As a sensitivity analysis, models with age as a categorical feature were also computed.

### Stratification

For both of the models described models without any strata were computed. This was to see the overall effect of the LOS on the mortality. Further we included strata to further adjust for differences between patients.

Due to systematic differences between hospitals we included strata on HT. Hospitals have in general different ways of solving issues, and also different capacities and number of admissions with acute heart failure. The use of strata could remove bias and noise arising from these differences. As seen in Section 2.2.4, there was 26 different hospital trust resulting in a model with 26 strata.

The implementation of the Coordination Reform gave reason to believe that there could be systematic differences between the years as well as the HT. In addition, improvement in treatment time could contribute to make admissions non-comparable across the years. There could also be the case that a hospital implemented a new system which could change the trends in the data set. Hence adding strata on both HT and year was reasonable to include in a model. This resulted in 280 different strata, as some health trusts did not have admissions from all of the years.

The choice of strata can be many, and the day of the week was also a possible candidate. The routines for the hospitals on a weekday could be different from the routines in the weekend, and also the amount of admissions. As we could see in Figure 2.8a, with the distribution of admissions across the days, there was a peak of admissions on Mondays and compared to the other days. The possibility of systematic differences between admissions for the different days was there, and hence it could be reasonable to add day of the week to the strata. Stratifying on HT, year and day of the week resulted in 1927 comparable groups of admissions.

There was also a possibility to add month to the stratification. Routines may shift across the months as the weather changes. However, introducing too many strata may give too many groups with few comparable admissions. Stratifying with month as well as the other ones gave 20283 levels, which on average was only 4 comparable admissions per strata.

The hazard rate for a model with strata is given by

$$\alpha_{is}(t|\mathbf{x}_i) = \alpha_{0s}(t)\exp(\mathbf{x}_i^T\boldsymbol{\beta}), \quad s = 1, ..., k, \tag{4.7}$$

with the same relative risk as in the previous models, given by Eq. (4.5) and (4.6). The number $k$ represents the number of strata, which were 26, 280, 1927 and 20283 for the different models respectively.

## 4.3 Poisson Regression for visits to the General Practitioner

In addition to study the 60-day mortality, it was interesting to look at visits to the GP in the 60 days after an admission. For this, Poisson regression was used. Patients are not able to visit the GP if they are still in the hospital, or if they die, hence an offset term of the exposure to the GP was added to the models. This was given as

$$\text{exposure} = \begin{cases} 60 - \text{LOS}, & \text{if the patient survived the 60 days,} \\ \text{days until death} - \text{LOS}, & \text{if the patient died within 60 days of admission.} \end{cases} \tag{4.8}$$

### Model with LOS as the only feature

To look at the direct association between LOS and visit to the GP a simple Poisson model with only LOS as the feature was computed. The rate for such a model was given by

$$\lambda_i = \exp(\beta_0 + \beta_1\text{LOS}_i + \log(\text{exposure}_i)), \tag{4.9}$$

for admission $i$ and with the offset term for exposure time.

### Model with multiple features

Further, multiple features were added in order to adjust for differences between patients in order to get their admissions to be more comparable. Then, the rate became

$$\lambda_i = \exp(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta} + \log(\text{exposure}_i)), \tag{4.10}$$

where

$$\begin{aligned} \mathbf{x}_i^T\boldsymbol{\beta} = {} & \beta_1 \cdot \text{LOS}_i + \beta_2 \cdot \text{age}_i + \beta_3 \cdot \text{woman}_i + \beta_4 \cdot \text{holiday}_i + \beta_5 \cdot \text{GP\_pre60}_i \\ & + \beta_6 \cdot \text{OOH\_pre60}_i + \beta_7 \cdot \text{acute\_pre60}_i + \beta_8 \cdot \text{education1}_i + \beta_9 \cdot \text{education2}_i \\ & + \beta_{10} \cdot \text{education3}_i + \beta_{11} \cdot \text{education4}_i + \beta_{12} \cdot \text{education5}_i + \beta_{13} \cdot \text{education6}_i \\ & + \beta_{14} \cdot \text{education7}_i + \beta_{15} \cdot \text{education8}_i + \beta_{16} \cdot \text{education9}_i + \beta_{17} \cdot \text{LOS\_pre60}_i \\ & + \beta_{18} \cdot \text{shiftEvening}_i + \beta_{19} \cdot \text{shiftNight}_i. \end{aligned} \tag{4.11}$$

### Random effects

As for the Cox regression, the data could be grouped by HT, year, day and month to get more comparable admissions. For this, we added random effects which estimated cluster-specific intercepts using the package *fixest* in R (Bergé 2018). Random intercepts where added to both a model with LOS as the only feature, and a model with multiple features.

The approach taken with the Cox models was also used for these models, with the data first grouped by HT, then HT per year, further HT per year and per day of the week, and lastly HT per year, day and month.

For models with several features, the rate was given by

$$\lambda_{ij} = \exp(\beta_0 + \mathbf{x}_{ij}^T\boldsymbol{\beta} + \log(\text{exposure})_j + \gamma_{0i}) \quad i = 1, ..., m. \tag{4.12}$$

with $m = 26, 280, 1925, 19893$ respectively and $\gamma_{0i}$ the cluster-specific intercept for cluster $i$. For these models, the linear predictor $\eta_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta}$ was given by

$$\begin{aligned}\mathbf{x}_{ij}^T\boldsymbol{\beta} = {} & \beta_1 \cdot \text{LOS}_j + \beta_2 \cdot \text{age}_j + \beta_3 \cdot \text{woman}_j + \beta_4 \cdot \text{holiday}_j + \beta_5 \cdot \text{GP\_pre60}_j \\ & + \beta_6 \cdot \text{OOH\_pre60}_j + \beta_7 \cdot \text{acute\_pre60}_j + \beta_8 \cdot \text{education1}_j + \beta_9 \cdot \text{education2}_j \\ & + \beta_{10} \cdot \text{education3}_j + \beta_{11} \cdot \text{education4}_j + \beta_{12} \cdot \text{education5}_j + \beta_{13} \cdot \text{education6}_j \\ & + \beta_{14} \cdot \text{education7}_j + \beta_{15} \cdot \text{education8}_j + \beta_{16} \cdot \text{education9}_j + \beta_{17} \cdot \text{LOS\_pre60}_j \\ & + \beta_{18} \cdot \text{shiftEvening}_j + \beta_{19} \cdot \text{shiftNight}_j,\end{aligned} \tag{4.13}$$

for the $j$-th admission in cluster $i$.

# Chapter 5

# Results

## 5.1 Simulation study

The different scenarios presented in Section 4.1 was used to generate data that was used to fit Cox proportional hazards models, where the patients were exposed to a length of stay in the hospital and then studied for 60 days after admission. The resulting survival curves are plotted in Figure 5.1.

Table 5.1: The resulting mean length of stay and hazard ratio for each scenario. Both values are the mean of the iterations.

| Scenario | LOS | HR |
|---|---|---|
| High acute risk rapid decrease | 6.243 | 0.960 |
| High risk in hospital phase | 6.240 | 0.950 |
| Constant rate | 6.268 | 0.979 |
| Increasing risk in hospital phase | 6.273 | 0.981 |

The curves show that on day 60 after admission, the survival rate was approximately 83% across the different scenarios, which is similar to the data set. The scenarios resulted in individual mean LOS as well, and these can be found in Table 5.1. The mean LOS was lowest for the scenario with a high risk in the hospital phase, and highest for the scenario with an increasing risk in the hospital phase. However, the differences are minimal. The table also includes the mean hazard ratio for LOS, with all of them being under 1, and a difference of 3% between the highest and the lowest value. Having said that, it is worth looking at the distribution of the HR over the iterations. Figure 5.2 shows the density of the HR for LOS, for each scenario, where it is clear that the constant rate and the increasing risk in hospital phase had a distribution closer to 1 with some iterations where the HR was estimated to be higher than 1.

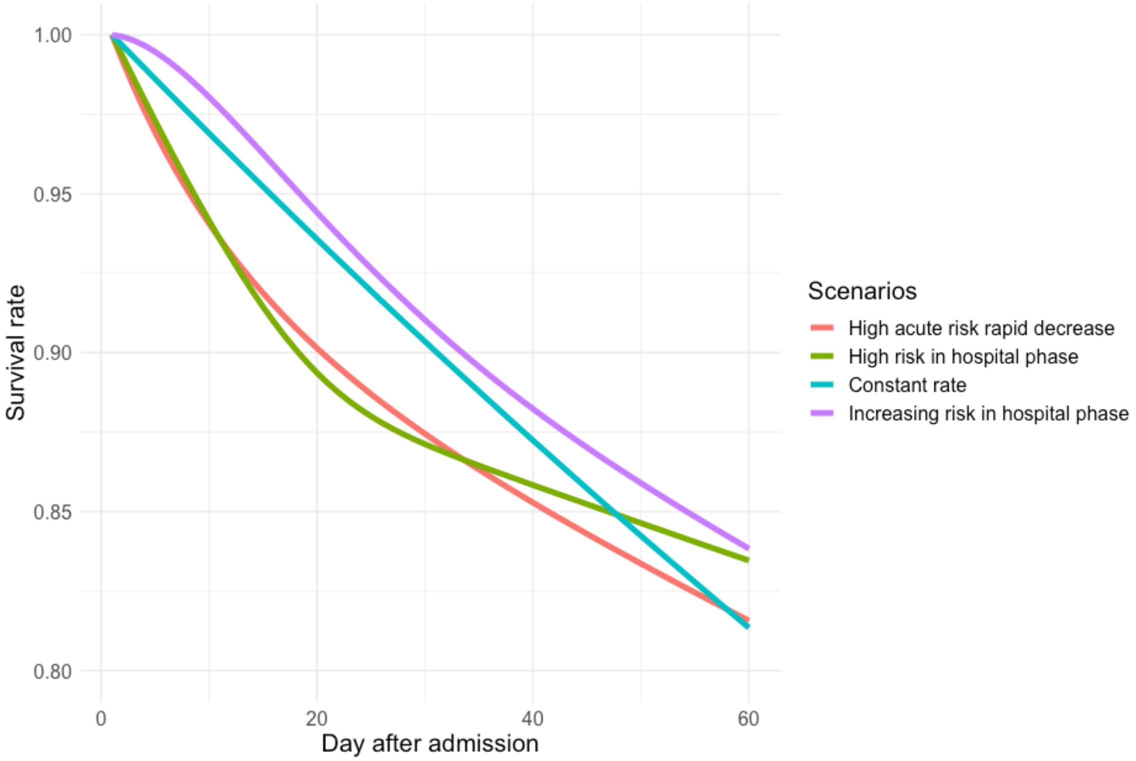


Figure 5.1: Figure showing the estimated survival rates from the simulation study, for each scenario. The mean survival rate after 60 days was 83%.

The assessment of the functional form between the outcome and LOS is done by checking the martingale residuals, found in Figure A.1a in Appendix A.1. The constant line in zero indicates a linear relationship, as wanted. The Schoenfeld residuals in Figure A.1b show a line constant over time, indicating that the proportional hazards assumption was met. As for the deviance residuals, shown in Figure A.2a, there is a cluster of observations below zero, which indicates individuals that ”lived too long”, and a larger cloud of observations between 1 and 4 representing patients that ”died to soon”. The smoothed line is below zero, which indicate that patients in general lived longer than expected. The last important residual is the score residual, shown in Figure A.2b. The overall line is a constant line in zero, indicating that none of the observations influenced the coefficient estimate for LOS.

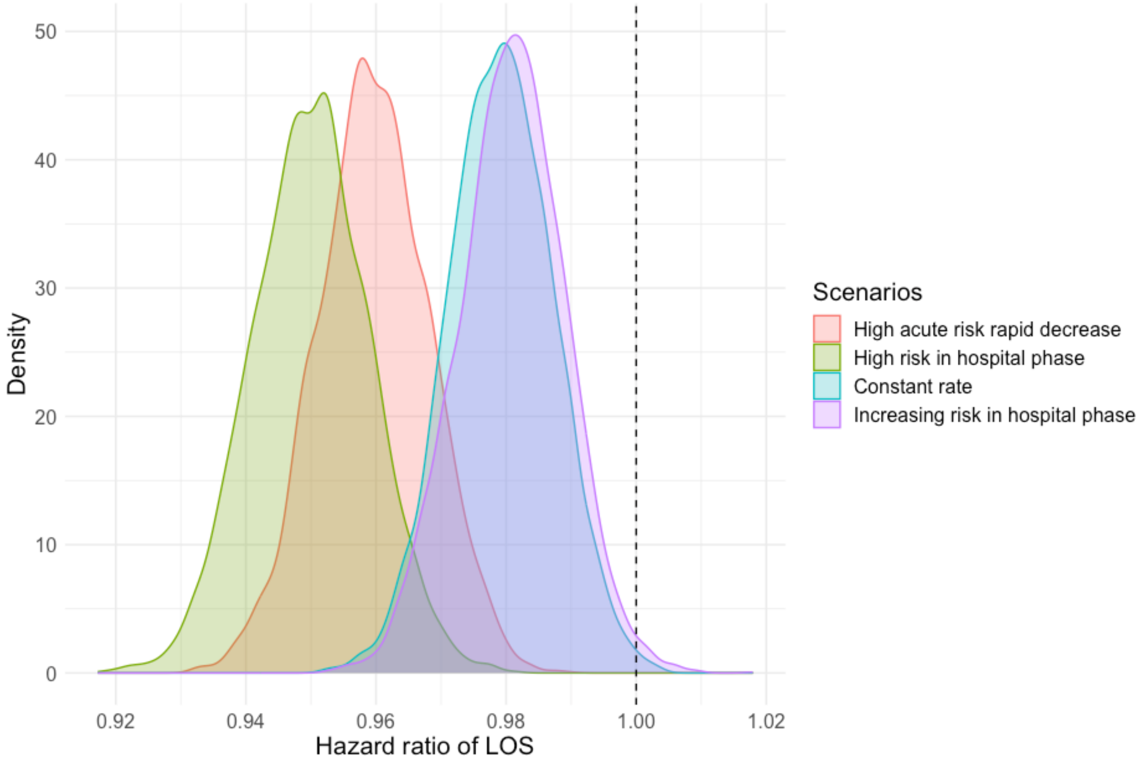


Figure 5.2: Figure showing the density of the hazard ratio for each scenario. The dashed line represents a HR of 1, which means no association between the HR for LOS and the outcome. Two of the scenarios estimated an HR higher than 1 in some iterations.

## 5.2 Cox Models on mortality

The main results from the models described in Section 4.2 are shown in Table 5.2. Firstly, the estimates for models with LOS as the only feature are presented, and further the estimates for models with multiple features. From the table we can see that all of the Cox models estimated the LOS feature to be highly significant for the outcome of 60 days mortality. In all of the different combination of strata, the estimated hazard ratio was 1.050 for models with multiple features and approximately 1.048 for models with LOS as the only feature. The models without any stratification had an estimated hazard ratio which was lower in comparison, with an increased risk of 4.1% per day longer stay. A sensitivity analysis was also done, where the estimated HR for LOS was 1.042 and 1.051 for a model without any stratification and a model with stratification on HT, year and day respectively. In these models, age was implemented as a categorical feature instead of a numerical feature.

Table A.2 in Appendix A.2, shows the estimated coefficients of the other features included in the models with multiple features. LOS, age, woman, OOH_pre60, acute_pre60, LOS_pre60 were highly

Table 5.2: Table displaying the estimated hazard ratio, 95% CI for the HR and the p-value for the LOS per day, for the Cox models described in Section 4.2.

| Without other features | | | | |
|---|---|---|---|---|
| | Method | HR | (95% CI) for HR | p-value |
| LOS | Without strata | 1.041 | (1.035-1.047) | <2e-16 |
| | Strata on HT | 1.047 | (1.041-1.053) | <2e-16 |
| | Strata on HT, year | 1.048 | (1.042-1.054) | <2e-16 |
| | Strata on HT, year, day | 1.048 | (1.042-1.054) | <2e-16 |
| | Strata on HT, year, day, month | 1.047 | (1.04+-1.053) | <2e-16 |
| With additional features | | | | |
| LOS | Without strata | 1.041 | (1.035-1.047) | <2e-16 |
| | Strata on HT | 1.049 | (1.043-1.056) | <2e-16 |
| | Strata on HT, year | 1.050 | (1.044-1.056) | <2e-16 |
| | Strata on HT, year, day | 1.051 | (1.045-1.057) | <2e-16 |
| | Strata on HT, year, day, month | 1.050 | (1.043-1.057) | <2e-16 |

significant in all choices of strata. All of them, except woman, had a hazard ratio higher than 1, indicating that higher values were associated with a higher risk of dying. GP_pre60 became less significant as the data was more stratified, but the hazard ratio was higher than 1 in all cases. For education, none of the levels had an impact on the outcome of 60 day mortality, education level 6 had the lowest p-value across the models, but it was only significant on a 0.1 significance level. Which shift the admission fell on was highly significant when stratas were added to the models, and the night shift was not significant when the model did not include strata. Looking at the hazard ratio of these factor levels, they are higher than 1 indicating that admission on evening or night was associated with a higher risk of dying within 60 days compared to being admitted during the day.

Further, in Table A.2, the estimated $\chi^2$ statistics for the hypothesis testing of proportional hazards are shown with corresponding p-value. The LOS became more proportional as several stratas were added, even though the p-value was significant in all cases. In addition, Figure and A.3b and A.4b show that the estimated coefficient for LOS had a slight increase as a function of time, however this increase was small. The Cox model assumes further that the relationship between the log of the hazard and the continuous features are linear. The straight lines in zero for the martingale residuals shown in Figure A.3a and A.4a indicate that this assumption was met for both the model with no stratification, and for the model with stratification on health trust. These plots are not included for further stratification as the line was constant in zero in all cases. Additionally the deviance residuals are shown in Figure A.3c and A.4c. The plots shows that there were some outliers with a high absolute value, indicating that the model struggled with the prediction for these. However most admissions had a small absolute value of the deviance. To assess the influence of the individual admissions we looked at the scaled score residuals, called dfbeta residuals. These can be found in Figure A.3d and A.4d. For the LOS feature, there were some outliers which produced change in the estimated coefficients when they were left out, even though the change was small in value. In total, these admissions did not influence the estimate as the fitted line is a constant line in zero.

Table 5.3: The table displays the estimated AIC values for the Cox models described in Section 4.2. The column in the middle shows the estimates when LOS is the only feature in the models, and the column to the right present the values when the models include multiple features.

| AIC values | | |
|---|---|---|
| | Without other features | With multiple features |
| Without strata | 227129 | 222664 |
| Strata on HT | 166321 | 161798 |
| Strata on HT, year | 115069 | 111120 |
| Strata on HT, year, day | 75665 | 71873 |
| Strata on HT, year, day, month | 29458 | 26714 |

From the Cox models, the AIC values could be calculated and these are found in Table 5.3. It is clear that stratifiying the models, reduced the log-likelihood and hence the AIC. The AIC was also lower for the models with multiple features, compared to the models with LOS as the only feature. Figure 5.3 shows the estimated survival curve for the Cox model without stratification, and with multiple features. After 60 days, the estimated survival rate was approximately 93%.

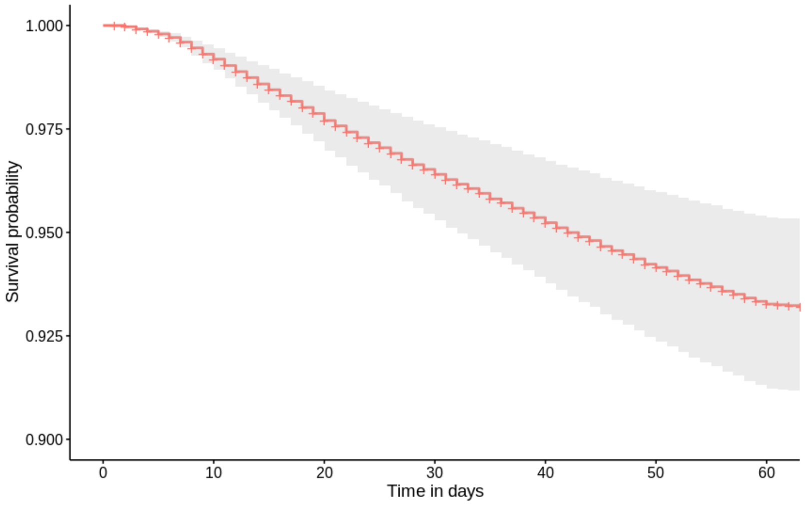


Figure 5.3: Figure displaying the survival curve for the Cox model without any stratification, and with multiple features. The plot also includes the confidence interval for the point estimates of the curve.

## 5.3 Poisson regression for visits to the General Practitioner

The last model of interest was the Poisson regression on number of visits to the GP in the 60 days after an admission. The results from the models described in Section 4.3 are shown in full detail in Appendix A.3 and the main results are presented in Table 5.4 and 5.5. The incidence rate ratio (IRR) for LOS, when this was the only feature in the models, was estimated to 1.020. When the admissions were adjusted for differences in hospital trusts, years and so forth, the IRR of LOS on the number of visits to the GP increased with approximately 0.1%. However, the estimate did not change as the number of random intercept effects were increased.

Furthermore, adding multiple features to the models decreased the IRR for all models. For the model without any random intercept effects, the expected number of visits to the GP increased with 1.7% per day longer stay, while for models with random intercepts the increase was 1.9%.

Table A.3 in Appendix A.3, shows the estimated IRR for all features used in the models, the 95% for the IRR and the p-value of the estimate. The visits to the GP in the 60 days before the admission, was highly significant for the number of visits after the admission with an IRR higher than 1. Increasing the number of visits before admission with one, gave an increased expected number of visits after the admission of 7.3%.

The other features were of less significance, and in general with a IRR lower than one. Being a woman, reduced the expected number of visits with approximately 3%. This was also the case if an admission happened on a holiday. However, the CI is wide for this feature, ranging from 7% to 0.06% in reduction. For education, the reference category was set to level 0, representing no education. The other categories had an IR below 1 indicating that having an education in general

Table 5.4: The estimated incidence rate ratio (IRR), 95% CI for the IRR and the p-value for LOS per day, for the Poisson models described in Section 4.3.

| Without other features | | | | |
|---|---|---|---|---|
| | Method | IRR | 95% for IRR | p-value |
| LOS | Without strata | 1.020 | (1.018-1.021) | <2e-16 |
| | Random intercepts for HT | 1.021 | (1.019-1.022) | <2e-16 |
| | Random intercepts for HT, year | 1.021 | (1.020-1.023) | <2e-16 |
| | Random intercepts for HT, year, day | 1.021 | (1.019-1.023) | <2e-16 |
| | Random intercepts for HT, year, day, month | 1.021 | (1.019-1.023) | <2e-16 |
| With multiple features | | | | |
| LOS | Without fixed effects | 1.017 | (1.016-1.019) | <2e-16 |
| | Random intercepts for HT | 1.018 | (1.017-1.020) | <2e-16 |
| | Random intercepts for HT, year | 1.019 | (1.017-1.020) | <2e-16 |
| | Random intercepts for HT, year, day | 1.019 | (1.017-1.020) | <2e-16 |
| | Random intercepts for HT, year, day, month | 1.019 | (1.017-1.020) | <2e-16 |

decreased the mean number of visits to the GP after an admission. This was also the case for admissions where the education was unknown. Being admitted in the evening or night was highly significant, with a decrease of 4% and 9% respectively compared to being admitted during the day.

Further, the addition of random intercepts gave only small changes in the IRR for LOS, for both models. For the models with random effects, the mean of the estimated random intercept effects are shown in Table 5.5. The table also shows the mean IRR for each of the models, where the simple models with only LOS as a feature, had a higher value in general compared to the models with multiple features. In addition, the mean values decreased as the data was grouped further with additional random intercepts.

In the model without any random effects, there is an intercept term $\beta_0$ representing the expected number of visits if all features are zero. For models with random effects, this term is included in the random intercept effect estimates when using the function *fepois* from the *fixest* package (Bergé 2018). The values for $\beta_0$ were $-2.409$ and $-2.508$, which gave an IRR of 0.09 and 0.081, for the model with a singular feature and the model with multiple features, respectively. Comparing these with the mean value of the random effects, we see that adding random effects decreased this value. However, the range of the estimated value across the effects increased as we added extra effects. When the admissions were clustered on health trusts, the difference between the highest and the lowest value for expected number of visits giving that all features were equal to 0, was 0.041. Adding year, day and month to the random effects gave a difference of 1.955.

Table 5.6 shows that the AIC value was higher for the models with a singular feature, and that increasing the number of random effects decreased the value. However, it did not decrease as much as for the Cox models when stratas were included. To assess the model fit, we also looked at the deviance residuals of the models. In Figure A.5a and A.5b, this is shown for the models without any random intercept effects. For the model with LOS as the only feature, the relationship between the log of the rate and LOS was linear as the fitted line is a constant zero line. However, the model did not fit any values to be over 6 visits in the 60 days after admission. The model with multiple features managed this, but the fitted line deviates from zero for higher values of the outcome indicating a less linear relationship.

Table 5.5: Table presenting the resulting number of fixed effects, the estimated mean value of the fixed effects, log of the mean and the range of these estimates.

| Without other features | | | | |
|---|---|---|---|---|
| Method | $m$ | E( $\gamma_{0i}$) | log($E(\gamma_{0i})$) | [Min,Max] |
| Random intercepts for HT | 26 | -2.42 | 0.089 | [0.063, 0.104] |
| Random intercepts for HT, year | 280 | -2.43 | 0.088 | [0.046, 0.172] |
| Random intercepts for HT, year, day | 1925 | -2.45 | 0.086 | [0.017, 0.216] |
| Random intercepts for HT, year, day, month | 19893 | -2.51 | 0.081 | [0.004, 1.959] |
| With multiple features | | | | |
| Method | $m$ | E( $\gamma_{0i}$) | log($E(\gamma_{0i})$) | [Min, Max] |
| Random intercepts for HT | 26 | -2.54 | 0.079 | [0.060, 0.088] |
| Random intercepts for HT, year | 280 | -2.55 | 0.078 | [0.047, 0.170] |
| Random intercepts for HT, year, day | 1925 | -2.58 | 0.076 | [0.018, 0.230] |
| Random intercepts for HT, year, day, month | 19893 | -2.62 | 0.073 | [0.003, 2.526] |

Table 5.6: The table presents the estimated AIC values for the Poisson models described in Section 4.3. The middle column shows the estimates when the models only include LOS as a feature, while the column to the right shows the values when the models include several features.

| AIC values | | |
|---|---|---|
| | Without other features | With multiple features |
| Without random intercepts | 485964 | 452311 |
| Random intercepts for HT | 481805 | 450405 |
| Random intercepts for HT, year | 480249 | 449596 |
| Random intercepts for HT, year, day | 478552 | 448587 |
| Random intercepts for HT, year, day, month | 464260 | 441382 |

# Chapter 6

# Discussion and conclusion

In this thesis we have analysed the relationship between hospital length of stay and mortality, and visits to the general practitioner. The association between LOS and the two outcomes were studied in a 60-day time window after discharge, to capture a period where mortality was likely to be related to the hospital visit. Data from the Norwegian Patient Registry and the Norwegian Cause of Death Registry was used, where the focus was on acute admissions of heart failure. The analysis was split into three parts; one simulation study examining the association between LOS and mortality, one approach where this association was examined for heart failure patients and lastly an examination of the visits to the General Practitioner for these patients.

From the simulation study, the mean hazard ratio was estimated to 0.97, which showed that immortal time bias was a factor in the analysis. The LOS and the death were generated independently of each other for each patient, such that the overall HR should have been 1. Multiple mortality curves were tested to see how it affected the outcome, but there was minimal difference between them.

To analyse the 60-day survival after a hospital admission due to heart failure, Cox Regression was used where models with and without strata were estimated. Initially, models with LOS as the only feature were computed to assess the hazard ratio without adjustment from other features. The hazard ratio for a model without any strata was estimated to be 1.041, meaning that for one day increase in LOS, the risk of dying increased with 4.1%. Moreover, when strata were added to adjust for systematic differences in health trust, year, day and month, the increasing risk of dying was estimated to be 4.8%. Furthermore, other features were added to the models to adjust for differences between short and long LOS to obtain a causal estimate if possible. These features were measured before admission, and hence did not introduce any error to the models. The other features increased the risk with 0.2% for the models with strata to 5.0% but did not change the estimate for the models without any strata.

The final part looked at the association between LOS and number of visits to the general practitioner in the 60 days after admission. Initially, a simple Poisson model with LOS as the single feature was computed to assess the IRR, where one day increase in LOS was expected to increase the expected number of visits to the GP by 2.0%. Random intercept effects were added to group admissions by health trust, year, day and month. However, this did not affect the estimated IRR for the length of stay significantly. Adding multiple features decreased the IRR by 0.2%, indicating that there are factors that had an impact on the LOS and further the outcome. Similar to the association between LOS and mortality, the sickness of the patients could give a confounding bias where it looks like a longer hospital stay is associated with more visits to the GP, when it may be that patients that have more visits are sicker and in need of more help or treatment.

The simulation study showed that immortal time bias takes part in an analysis like this, due to the conditioning on survival through the stay to be a part of the analysis. In the two scenarios with a high risk in the hospital phase, the mean hazard ratio was approximately 2.6% lower compared to the other two scenarios. This made sense as with these approaches, patients were more likely to die earlier in the time window and possibly in the hospital resulting in a higher immortal time bias. For the scenario with an increasing risk in the hospital phase and peak risk after discharge, patients were less likely to die in the hospital and hence less patients were removed due to this. The effect of the immortal time bias was overall 3%, and further we expected this to be a part of the analysis of the heart failure admissions. However, in these models there were also other factors influencing the estimates as the effect of one day longer LOS was estimated to be $4 - 5\%$.

The factor that is believed to be the most important for the assessment of LOS is the severity of the patients when they are admitted. Two patients could have identical values for all features, but one of them could still be sicker than the other and hence be at higher risk of dying and most likely get a longer LOS. However, this is hard to measure, and it introduces the possibility of confounding as the severity of the patients affects both the LOS and the risk of dying. In the data set, there was only information about the length of stay for each admission. This is chosen by several factors, and we had no information about why the patients got the exact length of stay or if the severity of the patients affected this length. There could also be the case that a patient got a longer LOS due to high pressure on the hospital and died shortly after discharge even though the admission was not severe.

Age was thought to be an indicator for the severity, however, in the sensitivity analysis where we implemented age as a categorical feature in two models, the estimated HR for LOS did not change in either. As Figure 2.18 showed, there was no difference in average LOS for patients between $60 - 90$ years, thus it is reasonable that the estimate kept constant.

Adjustment for health trust year and day was necessary to remove potential systematic differences. A disadvantage of using strata is that it was not possible to estimate the importance of each stratum and assess if there were significant differences between them. The statistical power of the model could also be reduced as the admissions were not well-balanced across the health thrusts as seen in Figure 2.5a. Adding several strata reduced the number of admissions in each group even further, with some of them having admissions where everyone survived. However, the use of strata reduced the possibility of noise in the models and as stated, the HR for LOS increased as the data became more stratified. This may be because when the data is stratified into groups, it can be easier to discover associations within each group that affect the overall outcome. Organizational factors can also be confounding, as the severity can affect which hospital the patient is sent to, and further the LOS. The capacity of the hospitals may also affect where the patient is being admitted, and as the boxplots in Figure 2.19 showed, there were clear differences in average LOS between them.

If we were interested in the effect of each health trust or year or the combination of these two on the LOS and the risk of dying, they could have been added to the models as categorical features instead. The clear disadvantage with this is the number of levels that this would result in when adding combinations of these features.

In the Poisson models, this was possible with random intercept effects. Each random intercept had an estimated value in each model, and hence it was possible to assess if there were significant differences between them. For some clusters, the estimates were significantly higher than others resulting in a higher expected number of visits for those clusters. Further, the measurement error became large when we introduced many levels of the fixed effects. For a cluster with only a few admissions, each admission could have a significant impact on the estimated value compared to a cluster with many admissions.

The use of strata on health trusts, to have more comparable patients, was expected to give a higher effect on the LOS for mortality than for visits to the general practitioner. The reason for this is that the GP works independently of the hospitals, thus it was not expected to differ across the health trusts. The same could be said for years, as there was no reason to believe that the use of GP had changed over the years.

For both the Cox and the Poisson models, time of admission was a significant feature for the outcome. In the Cox models, admissions in the evening compared to the day had an increased risk of death of 10% per day longer stay, while for the Poisson the expected number of visits to the GP decreased with 9.6% per day longer stay. Being admitted during the night was only significant for the Poisson models. Time of admission could also say something about the severity of the patient being admitted, as admission outside normal working hours are in general of severely sick patients in acute need of help.

## 6.1 Concluding remarks

The main issue in this analysis was the severity of the patients. Our findings revealed an association between hospital length of stay and both mortality and visits to the general practitioner. However, it is important to note that the the severity of the disease potentially acted as a confounding factor, influencing both LOS and outcomes. Although we lacked specific information regarding these relationships, there was no indication to suggest an increased risk of mortality associated with an extended hospital stay, contrary to the results. Additionally, the analysis shed light on how the issue of immortal time bias affected the estimation of hazard ratios when considering hospital length of stay as a numerical feature.

## 6.2 Future work

There are several alternative approaches for this type of analysis done on observational data. One of the issues in this thesis was the influence of immortal time bias on the estimation of the hazard ratios. A potential way of solving this issue is to reformulate the analysis to a target trial (Hernán et al. 2022). In this type of approach, the analysis would be on the outcome for patients that were discharged on a given day compared to the outcome for the patients that had a longer stay, for each possible length of stay in the hospital. Survival through the stay would still be a necessity to be a part of the analysis, but the immortal time bias could potentially be avoided when the analysis is done on each day in the hospital instead of the length of stay.

The other issue was the existence of confounding by indication in the analysis, as the severity of the patients was not measured. Quasi experimental methods such as instrumental variable analysis, could potentially solve this issue given certain conditions. In addition, it could assess if organizational factors affect the outcome of interest. For patients that are ready for or close to discharge, the decision could be affected by pressure in the clinical department. This could be if there is a need for more beds in the hospital, or if it is a Friday where the discharge tendency will be higher compared to the other days of the week.

Both approaches have been tested in a study which were done on patients admitted with hip fracture (Nilsen et al. 2020). In this study, they analysed the effect of organizational pressure to discharge on 60 day mortality. The results showed that organizational factors affecting the discharge could increase the risk of death. However, there is some potential for improvement for these types of analyses in general.

# Bibliography

Aalen, Odd, Ornulf Borgan and Hakon Gjessing (2008). *Survival and event history analysis: a process point of view.* Springer Science & Business Media.

Aarønæs, Marit et al. (2007). 'Kronisk hjertesvikt–etiologi og diagnostikk'. In: *Tidsskrift for Den norske legeforening.*

American Heart Association (2023). *Types of Heart Failure.* URL: https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure/types-of-heart-failure (visited on 23rd Feb. 2023).

Bergé, Laurent (2018). 'Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm'. In: *CREA Discussion Papers* 13.

Better Health Channel (2023). *Heart explained.* URL: https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/heart (visited on 27th Feb. 2023).

Eurostat (2023a). *Hospital discharges and length of stay statistics.* URL: https://ec.europa.eu/eurostat/statistics-explained/index.php? (visited on 30th Apr. 2023).

— (2023b). *In-patient average length of stay (days).* URL: https://ec.europa.eu/eurostat/databrowser/view/hlth_co_inpst/default/table?lang=en (visited on 30th Apr. 2023).

Fahrmeir, L. et al. (2022). *Regression Models, Methods and Applications.* Second edition. Springer.

Hagen, Terje P, David P McArthur and Trond Tjerbo (2013). *Kommunal finansiering av utskrivningsklare pasienter. Erfaringer fra første året.* Tech. rep. University of Oslo, Health Economics Research Programme.

Haug, Magnus (2023). *Forventet levealder falt i 2022.* URL: https://www.ssb.no/befolkning/fodte-og-dode/statistikk/dode/artikler/forventa-levealder-falt-i-2022 (visited on 24th Apr. 2023).

Helse- og Omsorgsdepartementet (2011). 'Samhandlingsreformen–Lovpålagte samarbeidsavtaler mellom kommuner og regionale helseforetak/helseforetak'. In: *Nasjonal veileder. Oslo.*

Hernán, Miguel A, Wei Wang and David E Leaf (2022). 'Target Trial Emulation: A Framework for Causal Inference From Observational Data'. In: *JAMA* 328.24, pp. 2446–2447.

HUNT Research Centre (2021). *HUNT Cloud Services 1.9. We believe it should be a simple thing to get elegant and secure environments for your scientific computing.* NTNU. URL: https://assets.hdc.ntnu.no/assets/ebook-hunt-cloud-services.pdf#page=8.

McDonagh, Theresa A et al. (2021). '2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) With the special contribution of the Heart Failure Association (HFA) of the ESC'. In: *European heart journal* 42.36, pp. 3599–3726.

Melberg, Hans Olav and Terje P Hagen (2016). 'Liggetider og reinnleggelser i somatiske sykehus før og etter Samhandlingsreformen'. In: *Tidsskrift for omsorgsforskning* 2.2, pp. 143–158.

Nilsen, Sara Marie et al. (2020). 'Hospitals discharge tendency and risk of death-An analysis of 60,000 Norwegian hip fracture patients'. In: *Clinical Epidemiology*, pp. 173–182.

Pinsky, Mark and Samuel Karlin (2010). *An Introduction to stochastic modeling*. Academid press.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Roback, Paul and Julie Legler (2021). *Beyond multiple linear regression: applied generalized linear models and multilevel models in R*. CRC Press.

Skogli, Erland, Ole Magnus Stokke and Siri Vikøren (2020). 'VURDERING AV TILTAK FOR Å REDUSERE SAMFUNNSKOSTNADENE KNYTTET TIL HJERTESVIKT'. In: *Menon Economics*.

Statistisk sentralbyrå (2023). *Standard for utdanningsgruppering (NUS)*. URL: https://www.ssb.no/klass/klassifikasjoner/36/ (visited on 11th May 2023).

Sud, Maneesh et al. (2017). 'Associations between short or long length of stay and 30-day readmission and mortality in hospitalized patients with heart failure'. In: *JACC: Heart Failure* 5.8, pp. 578–588.

Syse, Astri et al. (2023). *Sosiale helseforskjeller i Norge*. URL: https://www.fhi.no/nettpub/hin/samfunn/sosiale-helseforskjeller/ (visited on 12th May 2023).

Therneau, Terry M and Patricia M Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. Springer.

Therneau, Terry M, Patricia M Grambsch and Thomas R Fleming (1990). 'Martingale-based residuals for survival models'. In: *Biometrika* 77.1, pp. 147–160.

Wikipedia (2023). *Continuous Uniform Distribution*. URL: https://en.wikipedia.org/wiki/Continuous_uniform_distribution (visited on 28th Apr. 2023).

Yadav, Kabir and Roger J Lewis (2021). 'Immortal time bias in observational studies'. In: *Jama* 325.7, pp. 686–687.

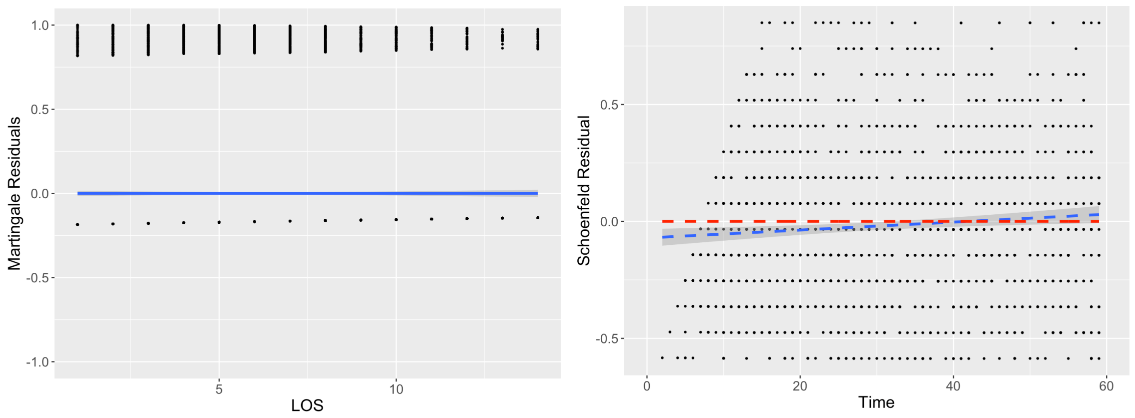
# Appendix A

# Additional figures and tables

The appendix includes residuals plots from some of the models that were estimated, the estimated hazard ratio/incidence rate ratio for the additional features used in the models, 95% confidence interval for these estimates and the corresponding p-value.

## A.1 Simulation

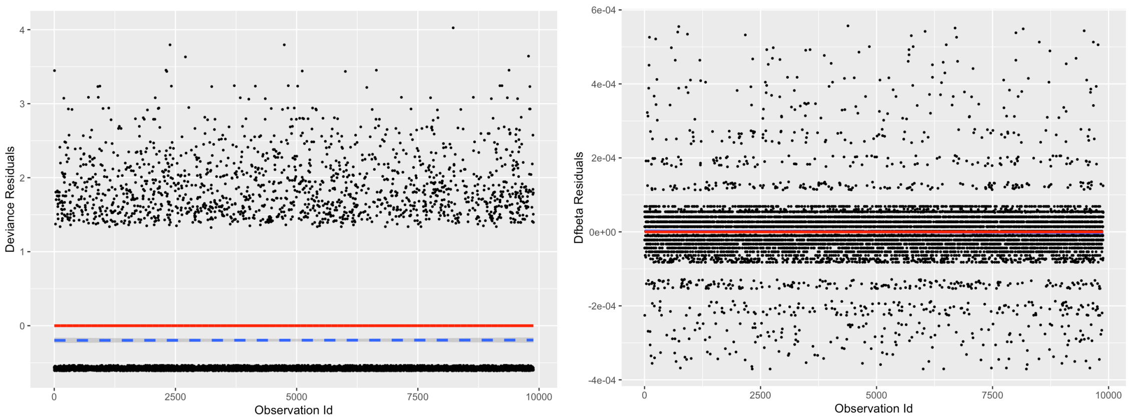
The various residuals for the Cox models are shown in Figure A.1 and A.2. The martingale, Schoenfeld and dfbeta residuals are plotted for the LOS feature in the model, while the deviance residual plot is estimated from the model it self.



(a)

(b)

Figure A.1: The plots show the Martingale residual for LOS to the left, and the Schoenfeld residuals for LOS to the right for the scenario with an increasing risk in the hospital phase.



(a) (b)

Figure A.2: The figure shows the deviance residuals in the left plot, and the scaled score residuals, to the right.

## A.2 Cox Regression

The estimated hazard ratios (HR) for all features included in the models with different levels of strata are shown in Table A.1. The table also includes the 95% confidence interval for these estimates, and the corresponding p-value. Table A.2 shows the estimated $\chi^2$ statistics for the proportional hazards assumption with corresponding p-value.

Table A.1: The table displays the estimated hazard ratio, 95% CI for the HR and the p-value per day, for the estimated values from the Cox regression models.

| | Method | HR | 95 % CI for HR | p-value |
|---|---|---|---|---|
| LOS | Without strata | 1.042 | (1.036-1.049) | <2e-16 |
| | Strata HT | 1.050 | (1.044-1.056) | <2e-16 |
| | Strata HT, year | 1.051 | (1.045-1.057) | <2e-16 |
| | Strata HT, year, day | 1.052 | (1.045-1.058) | <2e-16 |
| | Strata HT, year, day, month | 1.051 | (1.044-1.058) | <2e-16 |
| Age | Without strata | 1.073 | (1.070-1.077) | <2e-16 |
| | Strata HT | 1.074 | (1.071-1.078) | <2e-16 |
| | Strata HT, year | 1.070 | (1.067-1.074) | <2e-16 |
| | Strata HT, year, day | 1.070 | (1.067-1.074) | <2e-16 |

Continued on next page

**Table A.1 – continued from previous page**

| | Method | HR | 95 % CI for HR | p-value |
|---|---|---|---|---|
| | Strata HT, year, day, month | 1.072 | (1.068-1.058) | <2e-16 |
| Woman | Without strata | 0.868 | (0.828-0.911) | 4.7e-09 |
| | Strata HT | 0.871 | (0.830-0.914) | 1.6e-08 |
| | Strata HT, year | 0.896 | (0.854-0.941) | 1.1e-05 |
| | Strata HT, year, day | 0.903 | (0.860-0.947) | 3.2e-05 |
| | Strata HT, year, day, month | 0.902 | (0.857-0.949) | 7.1e-05 |
| Holiday | Without strata | 0.920 | (0.819-1.034) | 0.155 |
| | Strata HT | 0.912 | (0.812-1.024) | 0.119 |
| | Strata HT, year | 0.921 | (0.820-1.034) | 0.164 |
| | Strata HT, year, day | 0.900 | (0.799-1.013) | 0.081 |
| | Strata HT, year, day, month | 0.926 | (0.804-1.066) | 0.282 |
| GP_pre60 | Without strata | 1.019 | (1.013-1.026) | 4.4e-09 |
| | Strata HT | 1.017 | (1.010-1.023) | 4.4e-07 |
| | Strata HT, year | 1.012 | (1.005-1.019) | 3.7e-04 |
| | Strata HT, year, day | 1.011 | (1.005-1.018) | 9.4e-04 |
| | Strata HT, year, day, month | 1.007 | (1.000-1.014) | 0.048 |
| OOH_pre60 | Without strata | 1.076 | (1.061-1.092) | <2e-16 |
| | Strata HT | 1.080 | (1.066-1.095) | <2e-16 |
| | Strata HT, year | 1.076 | (1.062-1.091) | <2e-16 |
| | Strata HT, year, day | 1.079 | (1.064-1.094) | <2e-16 |
| | Strata HT, year, day, month | 1.091 | (1.073-1.109) | <2e-16 |
| acute_pre60 | Without strata | 1.044 | (1.023-1.065) | 6.7e-05 |
| | Strata HT | 1.041 | (1.020-1.062) | 1.1e-04 |
| | Strata HT, year | 1.035 | (1.013-1.058) | 1.5e-03 |
| | Strata HT, year, day | 1.032 | (1.009-1.055) | 5.9e-03 |
| | Strata HT, year, day, month | 1.041 | (1.015-1.068) | 2.1e-03 |
| LOS_pre60 | Without strata | 1.040 | (1.036-1.043) | <2e-16 |
| | Strata HT | 1.041 | (1.037-1.044) | <2e-16 |

Continued on next page

**Table A.1 – continued from previous page**

| | Method | HR | 95 % CI for HR | p-value |
|---|---|---|---|---|
| | Strata HT, year | 1.040 | (1.036-1.044) | <2e-16 |
| | Strata HT, year, day | 1.041 | (1.037-1.045) | <2e-16 |
| | Strata HT, year, day, month | 1.043 | (1.039-1.048) | <2e-16 |
| education1 | Without strata | 0.876 | (0.454-1.691) | 0.693 |
| | Strata HT | 0.854 | (0.444-1.643) | 0.637 |
| | Strata HT, year | 0.852 | (0.436-1.664) | 0.639 |
| | Strata HT, year, day | 0.883 | (0.452-1.728) | 0.717 |
| | Strata HT, year, day, month | 1.001 | (0.512-1.958) | 0.997 |
| education2 | Without strata | 0.879 | (0.641-1.206) | 0.425 |
| | Strata HT | 0.846 | (0.614-1.164) | 0.304 |
| | Strata HT, year | 0.880 | (0.631-1.226) | 0.449 |
| | Strata HT, year, day | 0.863 | (0.624-1.192) | 0.370 |
| | Strata HT, year, day, month | 0.828 | (0.571-1.201) | 0.320 |
| education3 | Without strata | 0.861 | (0.628-1.183) | 0.356 |
| | Strata HT | 0.834 | (0.606-1.149) | 0.268 |
| | Strata HT, year | 0.869 | (0.623-1.212) | 0.409 |
| | Strata HT, year, day | 0.849 | (0.614-1.175) | 0.324 |
| | Strata HT, year, day, month | 0.843 | (0.581-1.225) | 0.371 |
| education4 | Without strata | 0.810 | (0.587-1.119) | 0.202 |
| | Strata HT | 0.803 | (0.580-1.113) | 0.188 |
| | Strata HT, year | 0.832 | (0.594-1.167) | 0.288 |
| | Strata HT, year, day | 0.821 | (0.599-1.142) | 0.242 |
| | Strata HT, year, day, month | 0.786 | (0.538-1.148) | 0.213 |
| education5 | Without strata | 0.885 | (0.617-1.270) | 0.507 |
| | Strata HT | 0.880 | (0.611-1.266) | 0.490 |
| | Strata HT, year | 0.905 | (0.622-1.318) | 0.604 |
| | Strata HT, year, day | 0.882 | (0.610-1.275) | 0.503 |
| | Strata HT, year, day, month | 0.875 | (0.578-1.326) | 0.530 |

Continued on next page

**Table A.1 – continued from previous page**

| | Method | HR | 95 % CI for HR | p-value |
|---|---|---|---|---|
| education6 | Without strata | 0.738 | (0.534-1.021) | 0.066 |
| | Strata HT | 0.724 | (0.522-1.005) | 0.054 |
| | Strata HT, year | 0.755 | (0.537-1.061) | 0.105 |
| | Strata HT, year, day | 0.747 | (0.536-1.041) | 0.085 |
| | Strata HT, year, day, month | 0.702 | (0.481-1.026) | 0.067 |
| education7 | Without strata | 0.908 | (0.647-1.277) | 0.580 |
| | Strata HT | 0.903 | (0.640-1.273) | 0.558 |
| | Strata HT, year | 0.941 | (0.660-1.343) | 0.738 |
| | Strata HT, year, day | 0.925 | (0.654-1.308) | 0.658 |
| | Strata HT, year, day, month | 0.883 | (0.595-1.310) | 0.537 |
| education8 | Without strata | 0.732 | (0.391-1.370) | 0.329 |
| | Strata HT | 0.768 | (0.410-1.440) | 0.411 |
| | Strata HT, year | 0.796 | (0.425-1.490) | 0.475 |
| | Strata HT, year, day | 0.790 | (0.438-1.427) | 0.435 |
| | Strata HT, year, day, month | 0.544 | (0.291-1.019) | 0.057 |
| education9 | Without strata | 0.809 | (0.541-1.212) | 0.305 |
| | Strata HT | 0.775 | (0.516-1.164) | 0.218 |
| | Strata HT, year | 0.797 | (0.525-1.210) | 0.287 |
| | Strata HT, year, day | 0.783 | (0.518-1.182) | 0.244 |
| | Strata HT, year, day, month | 0.755 | (0.477-1.195) | 0.231 |
| shiftEvening | Without strata | 1.102 | (1.056-1.149) | 7.1e-06 |
| | Strata HT | 1.107 | (1.061-1.155) | 2.5e-06 |
| | Strata HT, year | 1.103 | (1.058-1.151) | 5.4e-06 |
| | Strata HT, year, day | 1.106 | (1.060-1.154) | 3.7e-06 |
| | Strata HT, year, day, month | 1.121 | (1.069-1.176) | 2.9e-06 |
| shiftNight | Without strata | 0.949 | (0.886-1.017) | 0.139 |
| | Strata HT | 0.959 | (0.896-1.028) | 0.239 |
| | Strata HT, year | 0.960 | (0.896-1.029) | 0.253 |

Continued on next page

**Table A.1 – continued from previous page**

| | Method | HR | 95 % CI for HR | p-value |
|---|---|---|---|---|
| | Strata HT, year, day | 0.956 | (0.892-1.154) | 0.211 |
| | Strata HT, year, day, month | 0.944 | (0.874-1.019) | 0.140 |

Table A.2: The table includes the estimated $\chi^2$ statistic and the corresponding p-value used for the proportional hazards tests.

| | Model | chisq | p-value |
|---|---|---|---|
| LOS | Without strata | 25 | 5.8e-07 |
| | Strata HT | 16 | 5.3e-05 |
| | Strata HT, year | 22 | 2.6e-06 |
| | Strata HT, year, day | 19 | 1.1e-05 |
| | Strata HT, year, day, month | 5 | 0.025 |
| age | Without strata | 0 | 0.69 |
| | Strata HT | 0 | 0.92 |
| | Strata HT, year | 32 | 1.3e-08 |
| | Strata HT, year, day | 30 | 4.8e-08 |
| | Strata HT, year, day, month | 26 | 3.4e-07 |
| woman | Without strata | 125 | <2e-16 |
| | Strata HT | 127 | <2e-16 |
| | Strata HT, year | 91 | <2e-16 |
| | Strata HT, year, day | 87 | <2e-16 |
| | Strata HT, year, day, month | 68 | <2e-16 |
| holiday | Without strata | 2 | 0.12 |
| | Strata HT | 2 | 0.15 |
| | Strata HT, year | 1 | 0.22 |
| | Strata HT, year, day | 1 | 0.22 |

Continued on next page

**Table A.2 – continued from previous page**

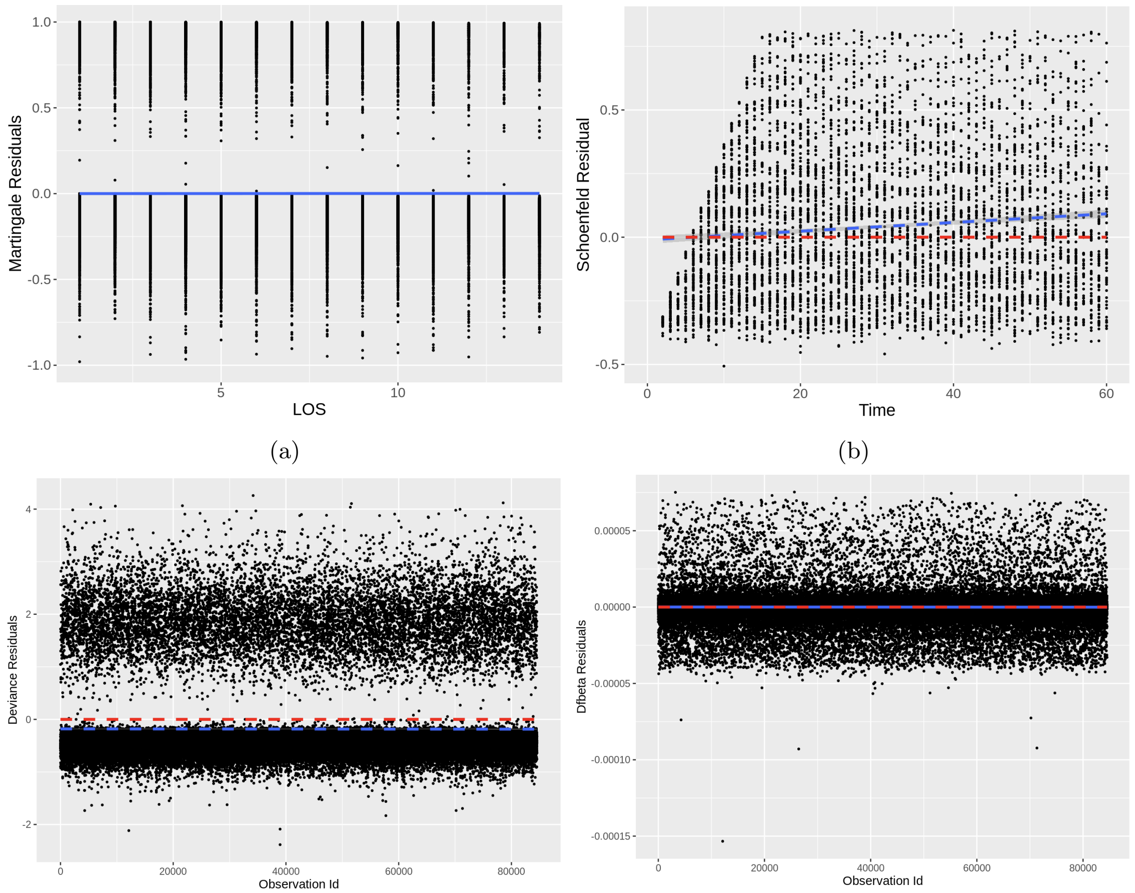
| | Model | chisq | p-value |
|---|---|---|---|
| | Strata HT, year, day, month | 0 | 0.96 |
| GP_pre60 | Without strata | 309 | <2e-16 |
| | Strata HT | 315 | <2e-16 |
| | Strata HT, year | 166 | <2e-16 |
| | Strata HT, year, day | 158 | <2e-16 |
| | Strata HT, year, day, month | 112 | <2e-16 |
| OOH_pre60 | Without strata | 66 | <2e-16 |
| | Strata HT | 83 | <2e-16 |
| | Strata HT, year | 36 | 1.9e-09 |
| | Strata HT, year, day | 31 | 3.3e-08 |
| | Strata HT, year, day, month | 32 | 1.3e-08 |
| acute_pre60 | Without strata | 199 | <2e-16 |
| | Strata HT | 188 | <2e-16 |
| | Strata HT, year | 134 | <2e-16 |
| | Strata HT, year, day | 137 | <2e-16 |
| | Strata HT, year, day, month | 121 | <2e-16 |
| education | Without strata | 10 | 0.34 |
| | Strata HT | 11 | 0.26 |
| | Strata HT, year | 10 | 0.37 |
| | Strata HT, year, day | 10 | 0.39 |
| | Strata HT, year, day, month | 12 | 0.19 |
| LOS_pre60 | Without strata | 175 | <2e-16 |
| | Strata HT | 153 | <2e-16 |
| | Strata HT, year | 111 | <2e-16 |
| | Strata HT, year, day | 135 | <2e-16 |
| | Strata HT, year, day, month | 130 | <2e-16 |
| shift | Without strata | 2 | 0.34 |
| | Strata HT | 2 | 0.34 |

Continued on next page

**Table A.2 – continued from previous page**

| | Model | chisq | p-value |
|---|---|---|---|
| | Strata HT, year | 3 | 0.21 |
| | Strata HT, year, day | 3 | 0.18 |
| | Strata HT, year, day, month | 2 | 0.48 |
| Global | Without strata | 610 | <2e-16 |
| | Strata HT | 591 | <2e-16 |
| | Strata HT, year | 369 | <2e-16 |
| | Strata HT, year, day | 368 | <2e-16 |
| | Strata HT, year, day, month | 292 | <2e-16 |

Figure A.3 shows the residuals plot for the Cox model without stratification. Subfigure A.3a, A.3b and A.3d are plotted for the feature LOS, while the deviance residual in Figure A.3c is estimated from the Cox model. Figure A.4 shows the same plots, but for the model with stratification on health trust.



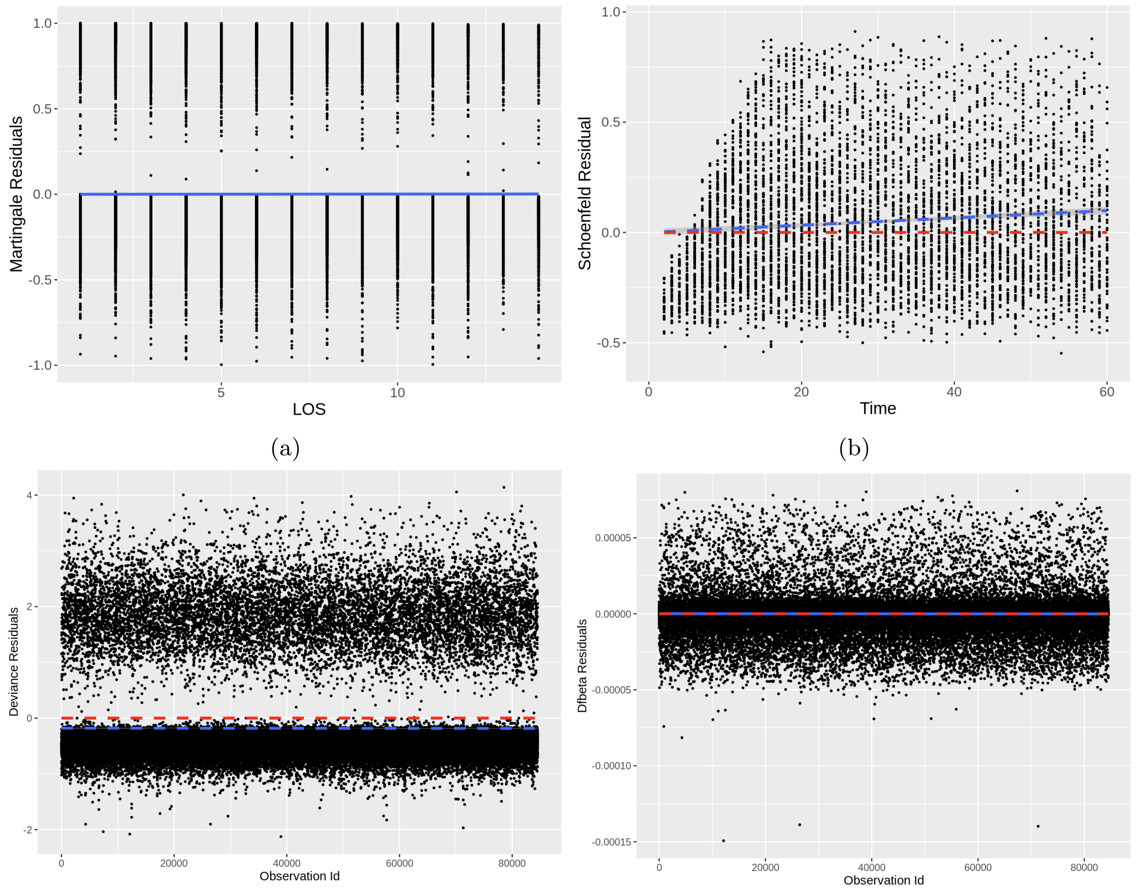
(a)

(b)

(c)

(d)

Figure A.3: The figure shows the Martingale residuals in the top left plot, the Schoenfeld residuals in the top right plot, the deviance residuals in the bottom left plot and the scaled score residuals in the bottom right plot. They are all results from the Cox model without any stratification, for the model with multiple features.



(a)

(b)

(c)

(d)

Figure A.4: The figure shows the Martingale residuals in the top left plot, the Schoenfeld residuals in the top right plot, the deviance residuals in the bottom left plot and the scaled score residuals in the bottom right plot. They are all results from the Cox model with stratification on health trust, for the model with multiple features.

## A.3 Poisson Model

The estimated incidence rate ratios (IRR) for all features included in the Poisson models with different number of fixed effects are shown in Table A.3. The table also includes the 95% confidence interval for this estimate, and the p-value of the estimated value.

Table A.3: The estimated incidence rate ratio (IRR), 95% CI for the IRR and the p-value per day, for the estimated values from the Poisson regression models.

| | Method | IRR | 95 % CI for IRR | p-value |
|---|---|---|---|---|
| LOS | Without Fixed effects | 1.017 | (1.016-1.019) | <2e-16 |
| | Fixed effects on HT | 1.018 | (1.016-1.019) | <2e-16 |
| | Fixed effects on HT, year | 1.018 | (1.017-1.020) | <2e-16 |
| | Fixed effects on HT, year, day | 1.019 | (1.017-1.020) | <2e-16 |
| | Fixed effects on HT, year, day, month | 1.019 | (1.017-1.020) | <2e-16 |
| Age | Without Fixed effects | 0.998 | (0.997-0.999) | 1.5e-08 |
| | Fixed effects on HT | 0.999 | (0.998-0.999) | 5.3e-06 |
| | Fixed effects on HT, year | 0.999 | (0.998-0.999) | 1.1e-05 |
| | Fixed effects on HT, year, day | 0.999 | (0.998-0.999) | 9.4e-05 |
| | Fixed effects on HT, year, day, month | 0.999 | (0.998-0.999) | 5.0e-05 |
| Woman | Without Fixed effects | 0.969 | (0.958-0.979) | 1.3e-08 |
| | Fixed effects on HT | 0.974 | (0.962-0.983) | 2.7e-07 |
| | Fixed effects on HT, year | 0.973 | (0.961-0.981) | 4.8e-08 |
| | Fixed effects on HT, year, day | 0.975 | (0.962-0.983) | 2.9e-07 |
| | Fixed effects on HT, year, day, month | 0.976 | (0.964-0.988) | 7.2e-05 |
| Holiday | Without Fixed effects | 0.968 | (0.941-0.994) | 1.8e-02 |
| | Fixed effects on HT | 0.967 | (0.942-0.994) | 1.6e-02 |
| | Fixed effects on HT, year | 0.967 | (0.932-0.984) | 1.8e-03 |
| | Fixed effects on HT, year, day | 0.971 | (0.938-0.991) | 9.4e-03 |
| | Fixed effects on HT, year, day, month | 0.992 | (0.959-1.027) | 6.5e-01 |
| GP_pre60 | Without Fixed effects | 1.074 | (1.073-1.076) | <2e-16 |
| | Fixed effects on HT | 1.072 | (1.071-1.073) | <2e-16 |

Continued on next page

**Table A.3 – continued from previous page**

| | Method | IRR | 95 % CI for IRR | p-value |
|---|---|---|---|---|
| | Fixed effects on HT, year | 1.072 | (1.070-1.073) | <2e-16 |
| | Fixed effects on HT, year, day | 1.072 | (1.070-1.073) | <2e-16 |
| | Fixed effects on HT, year, day, month | 1.074 | (1.072-1.076) | <2e-16 |
| OOH_pre60 | Without Fixed effects | 1.001 | (0.997-1.004) | 9.1e-01 |
| | Fixed effects on HT | 1.001 | (0.997-1.005) | 7.4e-01 |
| | Fixed effects on HT, year | 1.001 | (0.996-1.003) | 8.7e-01 |
| | Fixed effects on HT, year, day | 1.002 | (0.997-1.004) | 7.2e-01 |
| | Fixed effects on HT, year, day, month | 1.002 | (0.997-1.006) | 4.5e-01 |
| acute_pre60 | Without Fixed effects | 0.978 | (0.971-0.985) | 8.0e-10 |
| | Fixed effects on HT | 0.983 | (0.978-0.991) | 1.1e-05 |
| | Fixed effects on HT, year | 0.984 | (0.977-0.991) | 5.8e-06 |
| | Fixed effects on HT, year, day | 0.985 | (0.979-0.992) | 3.5e-05 |
| | Fixed effects on HT, year, day, month | 0.988 | (0.981-0.996) | 2.4e-03 |
| LOS_pre60 | Without Fixed effects | 0.999 | (0.998-1.001) | 4.3e-01 |
| | Fixed effects on HT | 0.999 | (0.998-1.001) | 5.0e-01 |
| | Fixed effects on HT, year | 0.999 | (0.998-1.001) | 3.9e-01 |
| | Fixed effects on HT, year, day | 1.000 | (0.998-1.001) | 5.6e-01 |
| | Fixed effects on HT, year, day, month | 0.999 | (0.998-1.000) | 1.4e-01 |
| education1 | Without Fixed effects | 0.835 | (0.740-0.941) | 3.3e-03 |
| | Fixed effects on HT | 0.861 | (0.765-0.967) | 1.2e-02 |
| | Fixed effects on HT, year | 0.862 | (0.768-0.965) | 1.0e-02 |
| | Fixed effects on HT, year, day | 0.869 | (0.772-0.975) | 1.7e-02 |
| | Fixed effects on HT, year, day, month | 0.846 | (0.735-0.973) | 1.9e-02 |
| education2 | Without Fixed effects | 0.972 | (0.915-1.035) | 3.9e-01 |
| | Fixed effects on HT | 0.934 | (0.880-0.996) | 3.7e-02 |
| | Fixed effects on HT, year | 0.938 | (0.883-1.000) | 5.3e-02 |
| | Fixed effects on HT, year, day | 0.940 | (0.884-1.003) | 6.3e-02 |
| | Fixed effects on HT, year, day, month | 0.927 | (0.857-1.002) | 5.7e-02 |

Continued on next page

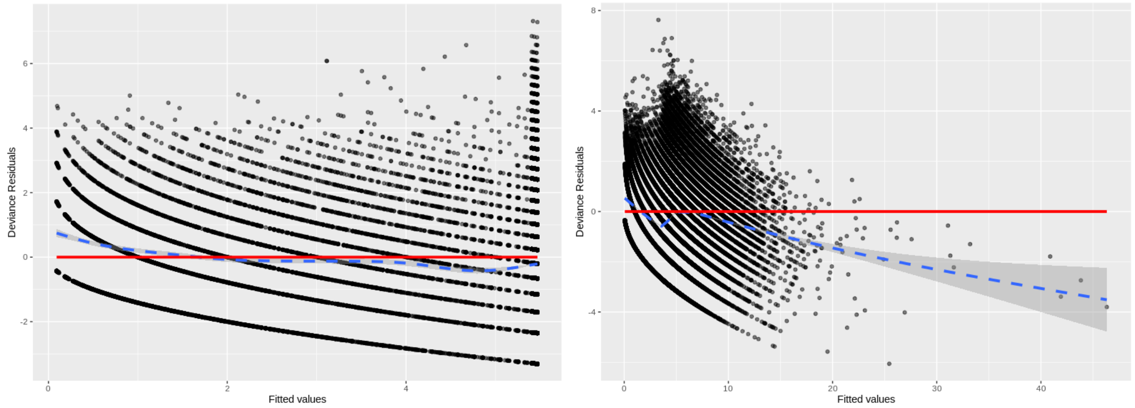
**Table A.3 – continued from previous page**

| | Method | IRR | 95 % CI for IRR | p-value |
|---|---|---|---|---|
| education3 | Without Fixed effects | 0.950 | (0.894-1.119) | 1.1e-01 |
| | Fixed effects on HT | 0.919 | (0.864-0.979) | 8.6e-03 |
| | Fixed effects on HT, year | 0.923 | (0.868-0.984) | 1.3e-02 |
| | Fixed effects on HT, year, day | 0.925 | (0.870-0.987) | 1.78e-02 |
| | Fixed effects on HT, year, day, month | 0.914 | (0.845-0.988) | 2.3e-02 |
| education4 | Without Fixed effects | 0.919 | (0.864-0.981) | 1.1e-02 |
| | Fixed effects on HT | 0.898 | (0.845-0.959) | 1.2e-03 |
| | Fixed effects on HT, year | 0.903 | (0.849-0.965) | 2.6e-03 |
| | Fixed effects on HT, year, day | 0.905 | (0.851-0.968) | 3.1e-03 |
| | Fixed effects on HT, year, day, month | 0.893 | (0.825-0.967) | 5.1e-03 |
| education5 | Without Fixed effects | 0.950 | (0.884-1.020) | 1.6e-01 |
| | Fixed effects on HT | 0.924 | (0.859-0.993) | 3.1e-02 |
| | Fixed effects on HT, year | 0.928 | (0.863-0.998) | 4.3e-02 |
| | Fixed effects on HT, year, day | 0.925 | (0.860-0.996) | 3.8e-02 |
| | Fixed effects on HT, year, day, month | 0.902 | (0.826-0.986) | 2.2e-02 |
| education6 | Without Fixed effects | 0.921 | (0.864-0.981) | 1.0e-02 |
| | Fixed effects on HT | 0.903 | (0.847-0.962) | 1.6e-03 |
| | Fixed effects on HT, year | 0.908 | (0.851-0.967) | 2.8e-03 |
| | Fixed effects on HT, year, day | 0.910 | (0.852-0.970) | 3.9e-03 |
| | Fixed effects on HT, year, day, month | 0.896 | (0.828-0.970) | 5.9e-03 |
| education7 | Without Fixed effects | 0.894 | (0.834-0.958) | 1.5e-03 |
| | Fixed effects on HT | 0.886 | (0.826-0.950) | 6.4e-04 |
| | Fixed effects on HT, year | 0.889 | (0.830-0.954) | 1.0e-03 |
| | Fixed effects on HT, year, day | 0.891 | (0.830-0.956) | 1.4e-03 |
| | Fixed effects on HT, year, day, month | 0.889 | (0.816-0.968) | 6.4e-03 |
| education8 | Without Fixed effects | 0.885 | (0.775-1.008) | 6.6e-02 |
| | Fixed effects on HT | 0.902 | (0.788-1.033) | 1.4e-01 |
| | Fixed effects on HT, year | 0.899 | (0.786-1.029) | 1.2e-01 |

Continued on next page

**Table A.3 – continued from previous page**

| | Method | IRR | 95 % CI for IRR | p-value |
|---|---|---|---|---|
| | Fixed effects on HT, year, day | 0.887 | (0.774-1.016) | 8.4e-02 |
| | Fixed effects on HT, year, day, month | 0.869 | (0.745-1.012) | 6.9e-02 |
| education9 | Without Fixed effects | 0.942 | (0.862-1.028) | 1.8e-01 |
| | Fixed effects on HT | 0.941 | (0.862-1.027) | 1.7e-01 |
| | Fixed effects on HT, year | 0.949 | (0.869-1.037) | 4.4e-01 |
| | Fixed effects on HT, year, day | 0.957 | (0.876-1.045) | 3.2e-01 |
| | Fixed effects on HT, year, day, month | 0.916 | (0.827-1.015) | 9.3e-02 |
| shiftEvening | Without Fixed effects | 0.958 | (0.948-0.968) | <2e-16 |
| | Fixed effects on HT | 0.955 | (0.945-0.965) | <2e-16 |
| | Fixed effects on HT, year | 0.955 | (0.945-0.964) | <2e-16 |
| | Fixed effects on HT, year, day | 0.963 | (0.953-0.972) | 1.4e-13 |
| | Fixed effects on HT, year, day, month | 0.966 | (0.955-0.977) | 8.9e-09 |
| shiftNight | Without Fixed effects | 0.910 | (0.896-0.925) | <2e-16 |
| | Fixed effects on HT | 0.909 | (0.895-0.924) | <2e-16 |
| | Fixed effects on HT, year | 0.910 | (0.896-0.925) | <2e-16 |
| | Fixed effects on HT, year, day | 0.922 | (0.907-0.937) | <2e-16 |
| | Fixed effects on HT, year, day, month | 0.914 | (0.898-0.931) | <2e-16 |



(a)

(b)

Figure A.5: The figure shows the deviance residuals for the Poisson model with a singular feature to the left, and with multiple features to the right. Both of them are results from the models without any fixed effects.

NTNU

Norwegian University of
Science and Technology