EXPLAINABLE ANALYTICS IN MEDICINE

A new approach to investigate individual patient data for the health economic evaluation of medical interventions

Saskia Schiele

Vollständiger Abdruck der von der Fakultät für Wirtschafts- und Organisationswissenschaften der Universität der Bundeswehr München zur Erlangung des akademischen Grades eines

Doktors der Wirtschafts- und Sozialwissenschaften (Dr. rer. pol.)

genehmigten Dissertation.

Gutachter:

1. Univ.-Prof. Dr. Andreas Brieden
2. Univ.-Prof. Dr. Claudius Steinhardt

Die Dissertation wurde am 01.10.2020 bei der Universität der Bundeswehr München eingereicht und durch die Fakultät für Wirtschafts- und Organisationswissenschaften am 07.01.2021 angenommen. Die mündliche Prüfung fand am 25.01.2021 statt.

EXPLAINABLE ANALYTICS IN MEDICINE

A new approach to investigate individual patient data for the health economic evaluation of medical interventions

Saskia Schiele

Acknowledgments

This thesis is the result of the effort and support of several people to whom I am extremely grateful. Some of them I want to thank particularly.
First and foremost, I thank my doctoral advisor Prof. Dr. Andreas Brieden, without whom this work would not have been possible. I am especially grateful for the opportunity to work on such a fascinating topic applying mathematics to the world of medicine. Moreover, I want to thank him for his unrivaled support and continuously creating a friendly and pleasant working atmosphere.
I would also like to offer my special thanks to Prof. Dr. Stephan Heres and Prof. Dr. Stefan Leucht for the good cooperation as well as their medical expertise.
Furthermore, I want to express my gratitude to my colleagues and friends Vienna, Vanessa, Martin, Matthias and Sebastian, and of course to all my other colleagues at the university, without whom completing this thesis would not nearly have been as much fun as it was. In particular, I want to thank Martin, not only for proofreading this thesis, but for having my back, especially during the recent months. I cannot begin to describe how grateful I am for that.
Additionally, I want to thank my friends, most importantly Lisa, Stefan, and Tanja, for their support as well as for providing necessary distractions when I needed it.
Finally, I want to thank my family. I am extremely grateful for growing up with endlessly supporting parents and an amazing brother. Especially without my father's encouragement, unwavering belief in me, and always valuable advice I would not be the person I am today.
Above all I would like to thank my fiancé Sebastian for his endless love and constant support, and for keeping me sane over the past few months. Thank you for being my tower of strength, muse, and most of all, for being my best friend.

Abstract

With the increased possibilities to collect and store data ('Big Data'), numerous popular methods have been presented for their analysis. For instance in the medical sector, data analysis is used to provide valuable data for health economic evaluations. However, it was widely overlooked, that most of the newly presented methods are accompanied by non-explainable results, due to so-called 'black box' models. This is not acceptable in medical data analytics, including but not limited to the General Data Protection Regulation (GDPR) of the European Union introduced in 2018. Therefore, the importance of explainable analytics in general and in medicine in particular has been discussed and investigated. While one possibility is trying to explain the workings of a black box model, another approach is to introduce advanced analytical methods generating explainable results in the first place.

In this thesis we present a new method for explainable analytics to analyze individual patient data in order to generate new findings for an improved future patient care. Furthermore, the new method provides reliable data for health economic evaluations. The new approach is based on an endpoint-oriented clustering approach, developed by Brieden and Gritzmann, forming sufficiently large clusters of patients with similar combinations of their characteristic values. We present a method for the cluster-based analysis of individual patient data to reliably predict the outcome of a patient (e.g. efficacy of a medical intervention). Furthermore, we introduce the newly invented cluster-based survival analysis to predict the 'survival' of a patient (e.g. continuance of a treatment). Besides predicting what the outcome of a patient might be, the method provides a unique explanation for the specific prediction, based on individual patient characteristics. Finally, we show the success of the newly introduced explainable method on a real world data set originating from a clinical trial including patients suffering from schizophrenia.

Zusammenfassung

Mit zunehmenden Möglichkeiten der Datensammlung sowie der Speicherung dieser Daten („Big Data“) wurden bereits zahlreiche Methoden für deren Analyse vorgestellt. Mit Hilfe analytischer Methoden werden unter anderem wertvolle Daten für gesundheitsökonomische Bewertungen generiert. Allerdings führen viele dieser Methoden aufgrund sogenannter Black-Box-Modelle zu nicht erklärbaren Ergebnissen. In der Medizin sind nicht erklärbare Prognosen nicht hinnehmbar, unter anderem aufgrund der 2018 eingeführten Datenschutzgrundverordnung (DSGVO) der Europäischen Union. Aus diesem Grund nimmt die Bedeutung der erklärbaren Analytik, insbesondere in der Medizin, deutlich zu. Erklärbare Prognosen können dabei sowohl über das nachträgliche Erläutern eines Black-Box-Modells als auch die Entwicklung moderner, direkt erklärbarer Methoden generiert werden.

In dieser Arbeit wird eine neue erklärbare Methode für die Analyse patientenindividueller Daten vorgestellt, um neue Erkenntnisse für die Behandlung zukünftiger Patienten zu generieren. Des Weiteren wird durch die Methode eine verlässliche Datengrundlage für gesundheitsökonomische Evaluationen erzeugt. Der Ansatz basiert auf einem von Brieden und Gritzmann entwickelten endpunkt-orientiertem Clustering Ansatz, der ausreichend große Patientencluster mit ähnlichen Merkmalsausprägungen bildet. Neben der cluster-basierten Analyse zur Prognose des Behandlungserfolgs eines Patienten, wird eine neu entwickelte cluster-basierte Überlebenszeitanalyse vorgestellt, um das „Überleben“ eines Patienten (z.B. Fortsetzung einer Behandlung) vorherzusagen. Dabei sind alle von den Ansätzen erzeugten Prognosen mit individuellen Patientencharakteristika erklärbar. Abschließend werden mit Hilfe der neuen Methode anhand einer klinischen Studie neue Erkenntnisse für die Behandlung von Schizophreniepatienten generiert.

Table of contents

List of tables

List of figures

List of procedures

List of abbreviations

AIMS	Abnormal Involuntary Movement Scale
BARS	Barnes Akathisia Rating Scale
CLGRY	Calgary Depression Scale for Schizophrenia
CATIE	Clinical Antipsychotic Trials of Intervention Effectiveness
CGI-S	Clinical Global Impression Scale
DALY	Disability-adjusted life years
DSM	Diagnostic and Statistical Manual of Mental Disorders
G-BA	Federal Joint Comittee
GDPR	General Data Protection Regulation
HR	Hazard Ratio
IQWiG	Institute for Quality and Efficiency in Health Care
ITAQ	Insight & Treatment Attitudes Questionnaire
MLE	Maximum Likelihood Estimation
PANSS	Positive and Negative Syndrome Scale
PLE	Partial Likelihood Estimation
PSP	Personal and Social Performance
RMST	Restricted Mean Survival Time
SAS	Simpson-Angus-Scale
YLD	Years lived with disability
YLL	Years of life lost

Part I

Introduction

The desire to gain knowledge from data has been around since the beginning of civilization. For instance, the Roman Empire gathered census of the population and extensively collected data about the empire's geographical area and wealth [147]. In the last decades, the speed with which data is generated increased drastically and the amount of data therefore burst. This growth is far from reaching a stop. The International Data Corporations (IDC) Global DataSphere forecast predicts, that the amount of data created over the next three years will exceed the created data volume of the past 30 years [77].
Data as a term is broadly formulated. According to the Oxford English Dictionary [133] *data* are

> 'facts or information, especially when examined and used to find out things or to make decisions.'

Discovering new findings is of the utmost importance in the field of medicine, since besides the pure generation of knowledge, human lifes are at stake. Therefore, the analysis of large amounts of data provides a tremendous potential for medical decision-making [117]. New methods to analyze medical data can be used to meet the information needs of patients as well as clinicians, researchers and health policy makers [95].
With the increased possibilities to collect, process, and store data, in general and in medicine, numerous methods have been presented to analyze this data both in a descriptive and predictive manner. Achieving the most accurate prediction of an outcome of interest has become a major challenge for predictive analytics or so-called supervised learning. In the medical sector, many of these methods provide valuable data for the health economic evaluation of medical interventions in terms of evidence-based medicine. Health care institutions, like the Institute for Quality and Efficiency in Health Care (IQWiG), and pharmaceutical companies use this data to evaluate the benefit as well as economic implications of medical interventions. However, it was widely overlooked, that many of the newly invented methods, especially in the field of machine learning, are black box models generating non-explainable results. Therefore, even though many methods predict what might happen, a lot of

them are not able to explain why a certain prediction was made. Non-explainable methods and results are causing difficulties for high-stakes decision-making, for instance in banking and criminal justice, but also medicine and health care.
Most importantly, in medical decision-making, patients deserve an ethical right for explanation. It is not fair-minded to simply tell a patient that a specific treatment will fail with a high probability, without explaining why. Furthermore, non-explainable models are highly unlikely to be adopted in clinical practice and they do not allow for further implications in the form of prescriptive analytics. Additionally, there are legal concerns arising from the General Data Protection Regulation (GDPR) of the European Union introduced in 2018 as well as the international standards for transparency in health economic evaluations. Hence, there is a strong need for explainable analytics in general and in particular in medicine.
It is one possibility to explain the workings of a black box model. Among other things, this is motivated by the believe, that only machine learning models can predict an outcome accurately, which is arguably a myth [121]. Furthermore, if a black box model can be explained a posteriori, the explanation will most likely be a complicated decision pattern, not usable in the real world [121].
Another possibility is to generate explainable results in the first place. Unfortunately an often encountered problem in general and in patient data in particular is underlying heterogeneity. Therefore, it is usually impossible to use one 'average' model to describe the entire patient data. Hence, the desire for explainable analytics in medicine is accompanied by the need for advanced analytical methods also addressing the problem of underlying heterogeneity inside the patient data.

In this thesis we present explainable methods for the analysis of individual patient data to generate new findings for improved future patient care. The methods especially provide reliable data for the analysis of the benefit of medical interventions during their health economic evaluation. The approaches are able to evaluate and predict the outcome of a patient (e.g. efficacy of a medical intervention) solely based on the patient's baseline characteristics. The methods furthermore address underlying heterogeneity inside the patient data. By applying an innovative endpoint-

oriented clustering approach, developed by Brieden and Gritzmann [27], we form sufficiently large clusters of patients with similar combinations of characteristic values. The underlying assumption is, that patients inside a cluster show similar outcome, whereas the outcome varies across clusters.
Based on these homogeneous clusters we present a method for the cluster-based analysis to reliably predict the outcome of a patient. Furthermore, we introduce the newly invented cluster-based survival analysis to predict the 'survival' of a patient. Besides the actual survival of a patient, the continuance of a treatment is a frequently explored outcome in a cluster-based survival analysis. Thereby, a separate survival analysis is performed on each of the derived clusters. Besides providing a prediction of a patient's outcome, all presented methods give a unique explanation why the specific prediction was made, based on individual patient characteristics.
Finally, we generate new findings for a real life data set by applying the newly invented cluster-based survival analysis. The data set originates from a clinical trial including patients suffering from schizophrenia. Schizophrenia is a severe mental illness resulting in major psychological pain for the patient as well as a tremendous economic burden for society [146] [110]. Therefore, besides showing the excellent results of the new explainable method, we provide new information for an improved treatment of patients suffering from this serious mental illness, reducing the patients', their families' and society's burden.

We begin by motivating our investigation into new methods for explainable analytics in medicine in Part II. After providing an overview about the field of explainable analytics in Chapter 1, we highlight the importance of explainable methods in medicine and discuss the implications of the General Data Protection Regulation (GDPR) in the European Union. To highlight the relevance of new methods for the health economic evaluation of medical interventions, we give an overview about general evaluation methods as well as health economic evaluations in Germany. To get an understanding of the necessity of finding improved treatment possibilities for patients suffering from schizophrenia, we will give an overview about the disease, its assessment, and its burden - both for the individual and for society in Chapter 2.

In Part III of the thesis, we present and discuss a new method for explainable analytics in medicine. After an overview about common survival analysis in Chapter 3, we point out and discuss the often encountered problem of underlying heterogeneity in survival analysis. Chapter 4 provides an overview of the mathematical background to generate homogeneous clusters of patients. Besides discussing a transformation technique, we also present an automated approach to identify the most promising baseline characteristics for the explainable methods. In Chapter 5 we present the cluster-based analysis. We begin by discussing the statistical evaluation of the derived clusters and explain how to justify the division into those. Furthermore, we will discuss the use of confidence intervals instead of hypothesis testing throughout the methods. To conclude this part, we will introduce the cluster-based survival analysis in Chapter 6. Besides the literal 'survival', various other outcomes of a patient can be predicted by the cluster-based survival analysis (e.g. continuance of a treatment). After estimating both cluster-based survival functions and cluster-based survival models, we will extend the approach to stratified models. Furthermore, we will discuss how to evaluate differences regarding the outcome of patients, both across different treatments as well as different clusters.
In Part IV of this thesis, we generate new findings for an improved treatment of patients suffering from schizophrenia, by applying the cluster-based survival analysis. In Chapter 7 we investigate the clinical trial CATIE, a randomized clinical trial assessing the discontinuation of treatment. We provide an overview about the trial and previous findings, especially about possible predictors for the discontinuation of treatment. After deriving a research question, we will present the results of a common survival analysis. The following cluster-based survival analysis both describes the estimation of the cluster-based survival functions as well as cluster-based survival models. We will summarize the new findings by justifying different outcomes both across treatments and clusters. Furthermore, we will highlight the new findings and compare them with the results of the common survival analysis.
In Part V we will summarize the main results and discuss possible future work on explainable analytics in medicine.

Part II

Motivation

In Part II of this thesis we motivate our investigation into new methods for explainable analytics in medicine. We begin by motivating the need for methods for explainable analytics. Afterwards, we motivate the need of investigating optimized treatment plans for patients suffering from schizophrenia.

Chapter 1 begins with an overview about data analytics and the role of machine learning. Afterwards we will discuss explainability and interpretability. Besides generating a common understanding of both terms, we will illustrate the difference between explainable and non-explainable methods and motivate the use of explainable methods for data analytics. Based on that, we discuss two possibilities to generate explainable models and discuss their differences. Furthermore, we highlight the importance of explainable analytics in medicine. Besides emphasizing its importance, we examine the implications of the General Data Protection Regulation (GDPR) in the European Union. To conclude this chapter, we discuss the relevance of new methods for health economic evaluations.

Chapter 2 provides an overview of schizophrenia, its prevalence, and the associated burden both for the individual and society. Besides briefly reviewing its history, we discuss the assessment and treatment options of schizophrenia. Afterwards, we highlight the importance of improved possibilities to treat every patient in order to reduce his or her own pain, as well as the pain for his or her family. To conclude, we examine the economic burden put on society by schizophrenia worldwide and in particular in Germany.

1. Explainable analytics

The aim of this thesis is to gain additional knowledge about data by applying new methods for explainable analytics. The methods will be presented in Part III and applied to a real world medical data set in Part IV.
In this chapter we give an overview of the field of explainable analytics. Recently, many terms regarding the analysis of data have become very popular and are widely used ever since. Especially machine learning has become a major component in data analysis. After defining the terms analytics and data analytics in general, we will briefly discuss its relations and distinctions to machine learning and put it into context regarding supervised and unsupervised learning. Afterwards we will define and discuss explainability and interpretability in analytics and emphasize its importance especially in the field of medicine. To conclude, we give a brief overview of the consequences of the General Data Protection Regulation (GDPR) of the European Union introduced in 2018, especially for automated decision-making in the medical field. Furthermore, we highlight the relevance of new methods for health economic evaluations.

Throughout this chapter it will become clear, that there is not always one unique definition for many of the addressed terms. Many publications have discussed this phenomenon as well as given extensive overviews about the entire field. As this is not the aim of this thesis, we will also refrain from giving a universal definition for the discussed terms. We will however give plenty of references for the interested reader.

1.1 Analytics

Analytics and especially data analytics are widely used terms not just in the field of mathematics and statistics, but in many other fields of research as well. According to the Oxford Dictionary [133] *analytics* is

> 'the systematic computational analysis of data or statistics.'

Davenport and Harris define *data analytics* as follows [47]:

> 'Data analytics is the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions.'

Analytics can be divided into descriptive, predictive, and prescriptive analytics [47] [130]. [47] furthermore mention autonomous analytics as an additional branch of analytics. Descriptive analytics (or business intelligence) is mainly concerned with the reporting or summary of data. An example would be the reporting of the observed death rate of patients in a clinical trial. Predictive analytics uses historical data combined with quantitative techniques to predict the future. An example would be a confidence interval in which the death rate of future patients lies with a 95% chance, based on the data of a clinical trial. Prescriptive analytics additionally specifies actions and the prospective outcome of these actions. One example could be administering a certain treatment for future patients in order to reduce the death rate. Descriptive and predictive analytics directly address the underlying data, whereas prescriptive analytics specifies actions based on results from descriptive and predictive analytics. Since the main concern of this thesis is the generation of knowledge from data, we therefore focus on descriptive and especially predictive analytics.
It is important to notice, that predictive analytics can not really predict anything [130]. No form of data analytics has the ability to generate certain results for the future. Physicist Nils Bohr expressed the following in 1971 [104]:

> 'It is very difficult to predict - especially the future.'

However, predictive analytics can predict what might happen in the future with a certain chance, and is therefore a valuable instrument for decision-making.

Remark 1.1. *Lately the terms supervised and unsupervised learning became very popular and are frequently used instead of predictive and descriptive analytics, especially for machine learning methods. Supervised learning methods aim to predict a specific outcome based on available input information, similar to predictive analytics [81] [6]. Unsupervised learning methods do not aim to predict an outcome, but rather to describe and find patterns in data, similar to descriptive analytics [81] [6].*

In the following we will restrict ourselves to predictive analytics (or supervised learning), due to our dataset in Part IV including the outcome of the patients, that we wish to predict.

Machine learning and data analytics

With the increased collection and storage of data ('Big Data'), combined with the ability to process it in a fast and efficient way, new methods have been developed in the field of data analytics. Over a short period of time many terms like machine learning, statistical learning, deep learning, and artificial intelligence became very popular. Especially machine learning, as well as artificial intelligence, and deep learning have been part of numerous books and other publications with most of them pointing out the absence of a universal definition for all of those terms [76] [47] [108] [119] [96] [6].
In 1998 Poole, Mackworth, and Goeble defined artificial intelligence as the study of the design of intelligent agents. Agents act in an environment and therefore vary from animals, humans, to airplanes [115]. Machine learning is understood to be a part of artificial intelligence usually investigating large amounts of data. Simultaneously, machine learning is a part of data analytics in general. In the following we want to give a brief understanding of machine learning. It is neither the main part of this thesis, nor are we using its methods. However, some problems arise from it, motivating our investigation into explainable analytics.

The principal goal of machine learning is the same as in all data analytics, to gain knowledge or (in case of supervised learning) make predictions based on data. Machine learning methods are designed to make the most accurate prediction possible, without investigating the inference between the relations of the involved variables. [109] and [14] provide an introduction, [76] discuss the field of machine learning as well as further implications for so-called 'deep learning' - a part of machine learning - and many other books cover multiple aspects as well as methods from machine learning [39] [96] [108] [119] [129]. Some use the term statistical learning [65], which is often considered to be a synonym. However, we do not want to imply the latter. Furthermore, [47] and [38] provide a look into the history of machine learning for the interested reader.

Besides the investigation of machine learning methods, some people discussed the distinction to classical statistical methods for data analytics [79] [39] [51]. [38] provide a thorough but compact overview, with focus on applications in neuroscience. The main difference to classical statistical methods lies in the focus of the process to generate the knowledge. Classical statistical methods are concerned with identifying relations between involved variables, whereas (supervised) machine learning is solely concerned with the best prediction of the outcome.

An often overlooked disadvantage of machine learning (opposed to classical statistical methods) is the so-called black box phenomenon. Due to machine learning only aiming for the most accurate prediction possible, usually there is no information or explanation available for the made predictions. The model is therefore considered to be a black box. Despite its popularity, not just in research, but in general, this disadvantage of machine learning and its implications are often ignored, making the models and results non-interpretable and non-explainable. Even though this might not be severe in some scenarios, it is unacceptable in several fields of research with high-stakes decision-making, e.g. medicine or health care. Moreover, explainability and interpretability are not only important for machine learning methods, but for analytics in general. Therefore we want to generate a common understanding of both terms in the following section, before emphasizing the importance of explainability and interpretability in the field of medicine.

1.2 Explainability and interpretability in analytics

According to the Oxford Dictionary the verb *to interpret* is to 'explain the meaning of information or actions' [133]. Building on that the verb *to explain* is defined as 'making an idea or situation clear to someone by describing it in more detail or revealing relevant facts' [133]. Therefore interpretability builds on explainability, which is why we will motivate the importance of both for data analytics.
There is no mathematical definition of explainability or interpretability in general and especially not for machine learning or data analytics as a whole [106] [49]. Even a formal definition is elusive [49]. [49] [100] [52] [106] [46] all provide their understanding of interpretability and explainability, while giving plenty of references ranging from psychology to machine learning and artificial intelligence. Some furthermore discuss the degree of interpretability. We want to note, that many publications motivate interpretability and explainability from a machine learning point of view, as machine learning methods hardly ever provide either. However, the importance can be extended to data analytics in general leading to explainable analytics. Similar to many other publications, we will use explainable and interpretable synonymously, as we intend to both understand the workings of our model as well as its results [70]. In the following we will therefore use the terms explainable and non-explainable. For the interested reader we highly recommend [100] for a thorough discussion about interpretability and explainability.
Based on [106] and [49], we therefore consider a model explainable, if in addition to the prediction ('what') we (as humans) can understand the prediction by looking at the available input information ('why').

In data analytics, explainability is often illustrated by a (black) box generating the data as visualized in Figure 1.1 [24]. The available input information are considered to go in on one side, and on the other side some outcome leaves the box. Nature associates the available variables with the respective outcome, but prior to any analysis of the data this association is unknown.

Figure 1.1: Box model for analytics (see [24])

Explainable analytics, including but not limited to classical statistical approaches, estimates the model inside the box based on the available data, which is then used for predicting the outcome as visualized in Figure 1.2.

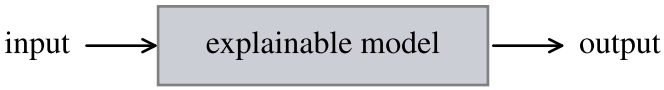


Figure 1.2: Box model for explainable analytics (see [24])

Hence, explainable analytics provide a look into the box, which is then no longer considered to be a black box.
Opposed to explainable analytics, many approaches do not seek to estimate the model inside the black box and consider it as complex and unknown [24]. The aim is to find an algorithm operating on the available input, that most accurately predicts the outcome as visualized in Figure 1.3. Especially machine learning methods fall into this category.

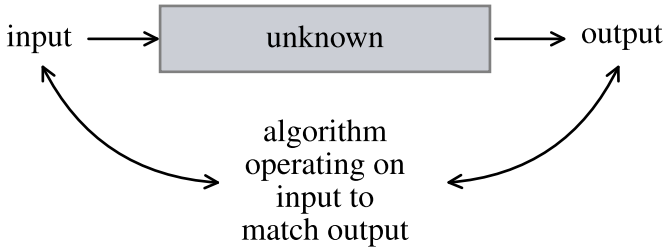


Figure 1.3: Box model for non-explainable analytics (see [24])

The focus does not lie on the model generating the data, but on the prediction and especially its accuracy. Therefore, non-explainable analytics do not provide a look into the box, which then remains to be a black box. We want to note, that even though [24] provide a decent introduction into these different kinds of modelling, they advocate for the algorithmic modelling culture. However Cox, the developer of the Cox proportional-hazards model, together with other researchers disagrees with the main statement of [24], which is included as comment in [24]. Furthermore, [24] only discuss the overall usage of algorithmic modelling and do not investigate differences for specific domains of applications, like medicine and health care.

Due to the opacity of many analytical approaches, especially in machine learning, many publications strongly advocate the necessity of explainable models [46] [49] [52] [100] [10] [106] [124] only to name a few. [5] mention the importance of explainable results in their white paper about challenges and opportunities of 'Big Data' from 2011. [13] also emphasize the importance of explainable models and that this area needs more attention. [49] constitute, that even the best black box model needs an explanation, because the used accuracy to develop such models is only a single metric and therefore an incomplete specification of almost every real world application.

There are two possibilities to explain a black box model. Recently the work on explainable machine learning (and also explainable artificial intelligence) proverbially exploded [121]. Most methods aim to create a second (post-hoc) model to explain the first black box model. However, it is already discussed, whether 'easy' explainability is even possible for all machine learning methods [70]. [111] and [139] show that marginal changes in an input (in this case an image) can lead to large changes in the respective prediction. Moreover [121] comprehensively argue against these approaches and show the excellent performance of an explainable model in [151] for recidivism prediction. At this point, we want to recite three of the arguments in [121] and underline some of them with additional references. First of all, there is not necessarily a trade-off between accuracy and explainability due to the good performance even of the 'simplest' models [48] [68] [69] [141]. Second of all, black box models often lead to complicated decision patterns, which are not usable for real world applications due to their strong dependency on human errors [122] [148]. Finally, black box models only appear to reveal hidden structures inside the data. If the structure was there, it could also be detected by a sufficiently flexible explainable model. Otherwise, the supposedly hidden structure was misleading or maybe just a random occurrence.
Instead of creating a post-hoc model explaining the black box model, it is possible to generate models that are explainable by definition, as they already provide their own explanations [121].

Besides the discussed advantages of explainable methods, their disadvantages have to be taken into account as well. One difficulty lies within generating explainable models in the first place, as 'simple' models like linear regression are often not appropriate due to underlying heterogeneity inside the data. Therefore, many real life problems require using and further investigating complex methods from various parts of mathematics. Furthermore, many explainable methods require inspecting assumptions about the underlying model before applying the method itself. Opposed to that, many non-explainable methods from machine learning can be applied immediately.

Nevertheless, explainable analytics are substantial and inevitable for many fields with high-stakes decision-making, like banking, insurance, criminal justice, and public policy. Another critical domain for explainable analytics is health care and medicine, as the made decisions deeply impact human lifes. Many of the previously mentioned publications already emphasize the importance of explainable models in the field of medicine. In the following section we want to give a more thorough understanding by providing additional references as well as an illustrative example.

1.3 Importance of explainable analytics in medicine

We already discussed the importance of explainable analytics in general. In this section, we want to take a closer look at analytics in medicine and the implication of non-explainable methods to further motivate our investigation into it. Furthermore, we will briefly discuss the implications of the General Data Protection Regulation (GDPR) of the European Union introduced in 2018 for analytics in medicine and discuss the relevance of new methods for health economic evaluations.

There are numerous publications emphasizing the importance of explainable models and results in medicine and health care. [54] state, that non-explainable models are one of the unintended consequences of machine learning in medicine. [144] motivate the importance of explainable models in medicine and health care. Even though they advocate explainable machine learning based on the creation of post-hoc models to explain the black box model, they provide plenty of arguments specifically for the medical domain and give numerous references to underline their statements. Furthermore they argue, that integrating medical experts in medical data analysis is equally important, which we strongly agree with. This is why the practical application presented in Part IV involved cooperating with medical experts. Besides highlighting the importance of explainable models [70] additionally point out the often encountered problem of heterogeneous data sources in medical data analytics. Therefore, new methods for explainable analytics in medicine are ideally supposed to overcome this heterogeneity. [79] provide an overview about machine learning

in health care and emphasize the challenge of explainability in machine learning in medical applications, even though the book covers and also recommends many non-explainable approaches for different domains of medicine. However, we want to note, that even though [79] recommend using machine learning methods (with its non-explainable models) in medicine, they never suggest those methods to replace a doctor but to support him in order to make decisions.

We want to illustrate the implications of explainable and non-explainable models for clinicians and patients with the following fictive example. Both explainable and non-explainable methods get the 'patient' as input. This input can include every information available about the patient such as age, severity of disease, or demographics. Since we are investigating predictive (or supervised) analytics, the outcome, e.g. probability of success of a specific treatment, has to be available as well. Both explainable and non-explainable methods will give a prediction for every patient. As illustrated in Figure 1.4, the non-explainable method might conduct, that one patient has a probability of success of 30%, whereas another patient has a higher probability of success of 60%. However, there is no explanation why those patients differ regarding the probability of success.
Opposed to that, explainable methods state, why the respective prediction was made, as illustrated in Figure 1.5. Hence, in the fictive example, the Body Mass Index (BMI) was responsible for the different predictions.

There are plenty of arguments for the investigation into explainable analytics in medicine. We want to put emphasis on the following four reasons.

1. **Explainable models will be accepted for clinical decision-making**
   Predictions made by a non-explainable method will most likely not be accepted by clinicians, as he or she is unable to explain it to the patient or other medical experts [70]. Clinicians want to understand the cause for a prediction before applying it in clinical practice.

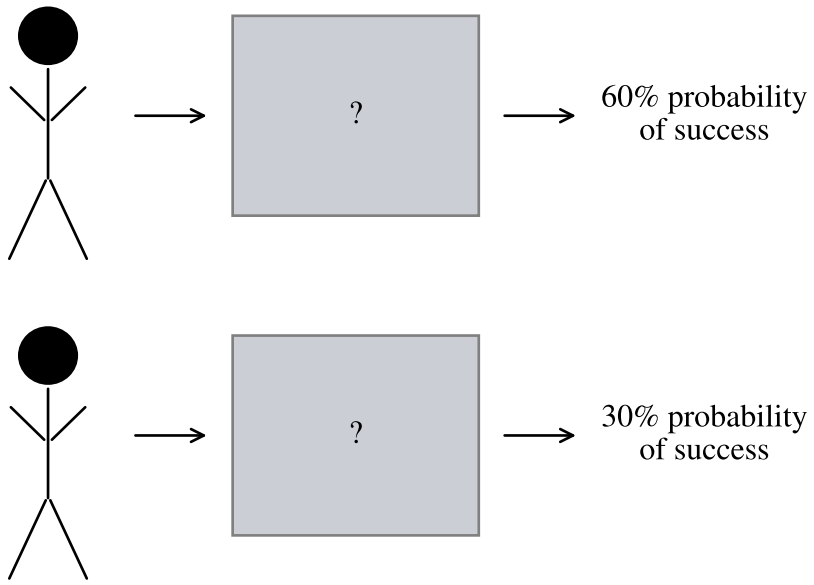


Figure 1.4: Non-explainable analytics in medicine

2. **Explainable models might lead to further implications**
   Medical experts might be able to draw further conclusions about the future treatment plan of a patient based on the findings of explainable analytics. Suppose the results show, that besides other characteristics, the Body Mass Index (BMI) of a patient is highly influential on the outcome of a patient. Besides prescribing medications, a clinician would recommend to loose (or gain) weight in order to achieve a better outcome. Recommendations like this are only possible, if the results are explainable in terms of individual patient characteristics. Of course there are other baseline characteristics (e.g. age of a patient), that can not be adjusted. However, explainable models provide the possibility for those baseline characteristics, that can be adjusted, to be identified in order to improve the future treatment of a patient.

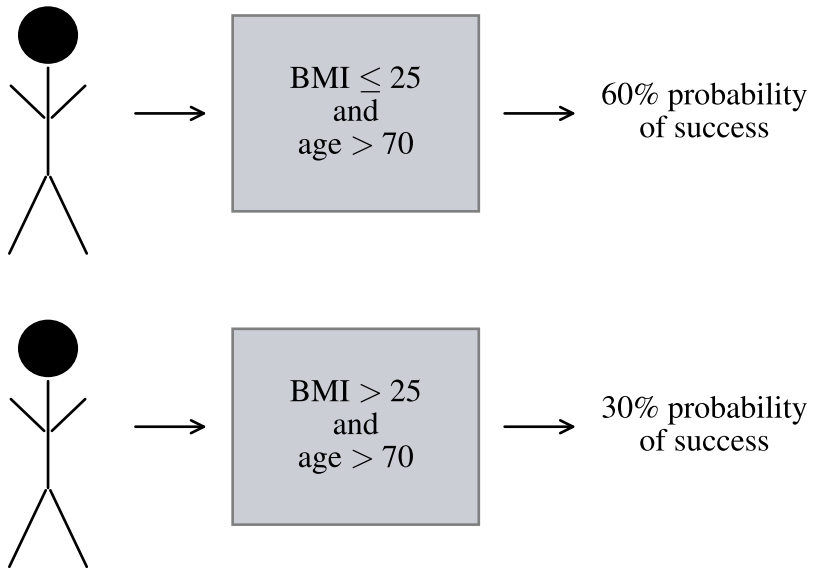


Figure 1.5: Explainable analytics in medicine

3. **No controlling authority in non-explainable methods**
   As we have seen in the previous section, even the most accurate analytical approaches (both explainable and non-explainable) are not faultless [122] [148]. Every method relies on the available data and the applied metric to evaluate the accuracy. When considering the medical domain, human lifes are at stake. A wrong decision might lead to non-optimal treatment plans and therefore in the worst case scenario the death of a patient. Even though wrong decisions can never be excluded, explainable methods always provide the opportunity for a last controlling authority. A clinician is able to understand the decision made by an explainable method. Therefore, before treating a patient, the clinician can adapt the decision if necessary. Decisions made by non-explainable methods do not provide this possibility for a last control.

4. **Ethical right for explanation**
   Even if the worst case scenario is not met, we strongly think, that every patient has a right for explanation. Simply stating, that the black box model generated the prediction of the patients' outcome is ethically dubious.

Due to the General Data Protection Regulation (GDPR) of the European Union introduced in 2018 the lastly mentioned ethical concerns also become legal concerns, which we will take a closer look at in the following section.

1.3.1 Consequences of the General Data Protection Regulation

With the introduction of the General Data Protection Regulation (GDPR) in the European Union, the implications for non-explainable analytics in medicine (and other domains) have been part of research.
According to articles 13(2)(f), 14(2)(g), and 15(1)(h) of the GDPR the controller is obligated to provide the data subject (i.e. human) with meaningful information about the logic involved [53]. Furthermore article 22(1) states, that the data subject has a right to not be subject to a decision based solely on automated processing [53]. As per an official briefing, non-explainable analytics like machine learning are part of the GDPR and great care must be taken in its development and deployment [55], raising the much-debated question about granting a 'right to explanation' to the data subject [34].
There are several publications discussing the impact both in general [34] and explicitly for medical decision-making [60]. It is believed, that non-explainable analytics will be difficult to use in practice [144] [70]. However, there are also publications questioning an overall effective 'right to explanation' [94]. Furthermore [145] constitute, that the GDPR is too vague regarding the transparency and accountability of automated decision-making and propose legislative steps to improve it.

Due to this thesis not being about the GDPR, we refrain from discussing the legal consequences ourselves. However, we want to emphasize, that regardless whether the

GDPR implicates a 'right to explanation', it all the more motivates the investigation into explainable analytics, especially for the analysis of medical data.

1.4 Relevance for health economic evaluations

To complete this chapter, we want to highlight the relevance of the newly invented explainable methods for health economic evaluations. Besides giving an overview about general evaluation methods, we will briefly discuss health economic evaluations in Germany.
The interdisciplinary field of health economics is concerned with the production, distribution, and consumption of health goods [143]. The efficiency principle included in article 12(1) of the Social Code Book V (SGB V) dictates, that all medical services in Germany have to be sufficient, appropriate, and economical [37]. Therefore a key task of health economics is balancing medical possibilities with their respective costs and quality [143]. Health economic evaluations deal with developing new methods for this task [143]. In order to perform health economic evaluations, clinical, economic, and epidemiological data is applied [127]. It is therefore of vital importance, that the provided data as well as the methods used to analyze the data are reliable. In this thesis we will introduce a new explainable method for the analysis of clinical data, addressing an often encountered problem of underlying heterogeneity. The introduced method provides reliable data for the health economic evaluation of medical interventions.

1.4.1 General evaluation methods

While evaluating medical interventions, the costs and the patients' outcome are compared. The investigated outcome can for example be the intervention's influence on the life expectancy of a patient [127]. There exist different methods to evaluate medical interventions, depending on the unit of the measured outcome. In this context, we only want to highlight the cost-efficiency analysis (CEA) and mention several alternatives. The following descriptions are mainly based on [127] and [88],

who also provide an extensive overview about the field for the interested reader.

The cost-efficiency analysis (CEA) compares the costs with the efficacy of a medical intervention by building a ratio out of them. Costs are represented in monetary units, whereas the patients' outcome is expressed in one-dimensional non-monetary units, e.g. life-years saved. Even though the CEA provides a simple way to include a patients' outcome in the health economic evaluation, an obvious disadvantage is the requirement of the same effect measure for the patients' outcome for all of the compared medical interventions. Furthermore the CEA can not be applied for multiple effect measures of a medical intervention.

A more simple alternative to the CEA is the cost-minimization analysis (CMA), comparing the net costs of the interventions. However, it can only be used, if the medical interventions show no difference in the patients' outcome. The cost-utility analysis (CUA) addresses the disadvantage of the CEA by replacing a single effect measure with a utility function merging multiple effect measures. If the utility of a medical intervention can furthermore be expressed in monetary values, a cost-benefit analysis (CBA) can be carried out. The CBA can especially be used to compare investments in the health sector with those in other sectors.

Among other things, those methods are used for the evaluation of medical interventions in Germany. We therefore want to give a brief overview about health economic evaluations in Germany.

1.4.2 Health economic evaluation in Germany

In Germany the responsibility for health economic evaluations falls to the scientific Institute for Quality and Efficiency in Health Care (IQWiG), which was established during the German Health Care Reform in 2004 by the Federal Joint Committee (G-BA). Its legal foundations can be found in the Social Code Book V (SGB V), which have been extended as consequence of further German Health Care Re-

forms [78]. IQWiG's main task is to evaluate medical interventions with their benefit, and harm, as well as their economic implications [78]. In the following we want to highlight two elements of the institute's regulations. The entire regulations with thorough explanations can be found in [78].

Evidence-based medicine

Evidence-based medicine (EBM) is an important part of IQWiG's work. It ensures health care for patients being based on evidence surveyed by scientific methods instead of personal opinions and conventions [123]. EBM prevents beneficial medical interventions from not being included in the provision of health care as well as interventions without use for the patients from being included in the provision of health care [78]. Furthermore EBM ensures, that only probabilities or predictions for groups of patients are possible to evaluate the benefit of a medical intervention [78]. The methods introduced in this thesis provide predictions for the outcome of groups of patients while providing an explanation for the predictions. That is why results generated from the new methods can be used as part of health economic evaluations.

Transparency in health economic evaluations

According to IQWiG's regulations, transparency is crucial in every aspect of the evaluation of medical interventions [78]. For instance, transparency regarding the conflict of interest and funding is required [82]. Following international standards for the health economic evaluation of medical interventions, it is also required, that the development and explanation of the model is transparent and especially comprehensive [78]. Especially when it comes to defining specific subgroups of patients, who might benefit more or less from a medical intervention, they require a unique definition of those subgroups [78]. Therefore, such subgroups can in particular not be identified by non-explainable methods.

Hence, explainable analytics are needed for providing valid and reliable data for the health economic evaluation of medical interventions, all the more motivating our investigation into new methods for explainable analytics.

2. Schizophrenia

Schizophrenia is a serious mental illness producing great suffering for patients as well as their families. It affects the brain and therefore the fundamental identity of a person. Schizophrenia interferes with a person's thinking, feeling, and behaviour. The early onset of schizophrenia between the late teen years and the early thirties results in the majority of this person's lifetime living with major psychological pain and disability, if left untreated [146] [110]. The estimated lifetime prevalence varies, but it seems to affect from 0.3% to 0.7% worldwide, making it an important public health problem [7] [146].
In this chapter we want to give an overview of the disease schizophrenia and especially its assessment. The introduced scales are used in the practical application in Part IV. Furthermore, we want to emphasize the burden resulting from this underestimated disease, both individual and economic. We want to strongly motivate our investigations into a clincal trial covering patients suffering from schizophrenia, to find improved possibilities to individually treat every patient to reduce his or her own pain as well as the pain for his or her family.
For a detailed description about various parts of schizophrenia, please refer to [146] as well as the information made available by the National Institute of Mental Health in the United States of America [110].

2.1 History, diagnosis, symptoms, and consequences

Schizophrenia (Greek: schizein = fragmenting, splitting; phren = mind) as a psychotic disorder was only characterised in the last century. However, it was far from being a 'new disease'. Psychoses like they occur in schizophrenia have been recognized since at least the first millennium BC [146]. Psychotic conditions were

even mentioned in the book of Samuel in the Old Testament, where King Saul becomes paranoid about David planning and even attempting to murder David himself (1 Samuel 10–12) [146]. In 1896 Kraeplin conceptualized psychotic disorders by calling them 'dementia praecox' (early onset dementia), which was translated and made broadly available in 1919 [92] [93]. In 1911 and widely available in 1950, Bleuler introduced the term schizophrenia to replace dementia praecox by stating that schizophrenia is a group of diseases instead of just one disease [15] [16] [146]. He furthermore introduced basic symptoms of schizophrenia. In 1959 Schneider proposed his symptoms classification [126], forming the base for standard diagnostic instruments, such as the Diagnostic and Statistical Manual of Mental Disorders (DSM).

According to latest edition of the DSM, there are five characteristic symptoms of schizophrenia [7]:

- delusion
- hallucinations
- disorganized speech (e.g., frequent derailment or incoherence)
- grossly disorganized or catatonic behaviour
- negative symptoms (i.e., diminished emotional expression or avolition)

A diagnosis of schizophrenia (or schizoaffective disorder or schizophreniform disorder) requires the presence of at least two of these symptoms for at least a month. The latest version of the DSM (DSM-V) requires one of the symptoms to be delusion, hallucinations or disorganized speech. The valid version of the DSM during the enrollment into the clinical trial of Part IV (DSM-IV) did not have this additional specification. Another condition for a diagnosis with schizophrenia are continuous signs of disturbance for at least 6 months [126]. There are several other requirements for a diagnosis with schizophrenia (or schizoaffective disorder or schizophreniform disorder), which can be found in [126].

The prevalence of schizophrenia is estimated to be 0.3% to 0.7% worldwide, with variations across ethnic and geographical backgrounds [7] [146]. Besides the symptoms directly related to schizophrenia, patients also face severe other con-

sequences regarding mortality, comorbidities, and social losses. Patients suffering from schizophrenia have an increased risk of committing suicide. According to [7] between 5% and 6% of all patients commit suicide and approximately 20% attempt suicide at least once. Furthermore, the life expectancy of patients suffering from schizophrenia is reduced due to comorbidities like diabetes or cardiovascular diseases. Those comorbidities are more prevalent under schizophrenics, compared with the general population [7].
Besides the increased mortality, patients suffering from schizophrenia are faced with social as well as occupational losses [7]. Additionally, the disease does not only put pain on the patient, but on the affected families as well [146] [135].
If left untreated, schizophrenia results in persistent symptoms possibly leading to life-long disability [110]. Hence, schizophrenia results in a high individual burden on the patient and his family. That is one of the reasons, why we investigate better treatment options for patients suffering from schizophrenia in Part IV.

2.2 Assessment and treatment of schizophrenia

In this section we give an overview of the assessment of schizophrenia and its treatments. Many of the mentioned scales are part of our investigation in Part IV and therefore need to be briefly discussed.

Assessment

There exist several scales to assess the severity of schizophrenia. One of the commonly used ones is the Positive and Negative Syndrome Scale (PANSS), which was also used in the clinical trial in Part IV. PANSS is a 30-item interview to rate the patient's positive (e.g. excitement), negative (e.g. emotional withdrawal), and general (e.g. disorientation) symptoms [85]. Every symptom is rated from 1 (absent) to 7 (extreme) and it can be divided into a positive, a negative, and a general subscale. In 1997 Marder proposed a different division into five factors, i.e. negative symptoms, positive symptoms, disorganized thought, uncontrolled hostility/excitement,

and anxiety/depression [102]. During our investigation in Part IV, all subscales and Marder factors were considered as possible predictors for the outcome of a patient.

Several possibilities exist to measure the outcome of a clinical trial. A typical one is symptomatic remission. Certain eight items of the PANSS scale (i.e. PANSS-8) have to be 1 (absent) or 2 (minimal) for a patient to reach symptomatic remission [9] [66]. Another typical outcome is response, which is measured by a certain reduction of the PANSS score (e.g. 50%).
CATIE, the clinical trial of Part IV, investigated the discontinuation of the treatment [135]. Therefore the primary outcome in CATIE was the amount of days until discontinuation.

Many other baseline characteristics of a patient are recorded in clinical trials. We will introduce the scales used in Part IV and give references for the interested reader. The following two scales are used to rate the severity of illness and depression of a patient:

- Clinical Global Impression – Severity scale (CGI-S)
  7-item scale to rate the severity of the patient's illness [64]
- Calgary Depression Scale (CLGRY)
  9-item interview scale developed to assess depression in schizophrenics [4]

The following three scales rate involuntary movements of patients, a common side effect observed in patients being treated for schizophrenia:

- Barnes Akathisia Rating Scale (BARS)
  4-item rating scale for drug-induced akathisia, a movement disorder characterised by inner restlessness and the inability to stay still [12]
- Abnormal Involuntary Movement Scale (AIMS)
  12-item scale to assess the severity of dyskinesia, the involuntary movement of muscles [64]

- Simpson-Angus-Scale (SAS)
  10-item rating scale for extrapyramidal side effects, i.e. physical symptoms, that are primarily associated with improper dosing of or unusual reactions to antipsychotic medications [132]

Finally, the following scale is used to rate the personal and social performance of a patient:

- Personal and Social Performance (PSP)
  4-item scale to assess patients' social functioning [107]

Other baseline characteristics commonly recorded in clinical trials are social parameters (e.g. marital status), demographics (e.g. age, sex), physical values (e.g. blood pressure), and also the abuse of alcohol and drugs. All of them are part of our investigation in Part IV and serve as potential predictors for the outcome of the patient.

Treatment

Antipsychotic medications play a major role in the treatment of schizophrenia. Additional forms of treatment include psychological treatment and other forms of support. Antipsychotics are prescribed to reduce symptoms and are crucial for the daily function of patients [110].
The first-generation antipsychotics (FGA), also known as 'typical antipsychotics', were developed in the 1950s. In the 1980s, second-generation antipsychotics (SGA), also known as 'atypical antipsychotics' emerged [3]. The interested reader is referred to [146] and [3] for a thorough overview about antipsychotics as treatment for schizophrenia.
There is an ongoing debate about the similarities and differences between FGAs and SGAs [146]. In fact, the main conjecture of the CATIE study, which we investigate in Chapter 7, was earlier discontinuation of the treatment with FGAs compared with SGAs. We do not intend to discuss the differences ourselves and refer the interested reader to the mentioned references.

2.3 Burden of schizophrenia

There are several measures for the burden of disease. Commonly used measures are years lived with disability (YLDs), years of life lost (YLL), and disability-adjusted life years (DALYs). YLDs record the prevalence of conditions leading to non-fatal health loss multiplied by the associated loss of health [58]. On the other hand, YLLs record the gap between the normative life expectancy and the observed mortality [58]. DALYs record the number of lost years due to the disease, related disability, or premature death [58]. DALYS are therefore calculated by the total of YLDs and YLLs [58].
According to the global burden of disease study in 2015, schizophrenia is in the top 10 causes for global age-specific YLD for ages 25 to 54, i.e. an enormous part of a person's adult life [59]. Simultaneously, it is associated with a high number of DALYs [58]. Hence, it is a dangerous disease with a relatively high prevalence.

In addition to the burden put on the patient and his or her family, a tremendous economic burden is related to schizophrenia. To conclude this chapter, we will therefore discuss the burden put on society and highlight current costs related to schizophrenia in Germany.

2.3.1 Economic burden of schizophrenia

There exist numerous publications about the economic burden of schizophrenia in different countries and regions all over the world. To conclude this overview of schizophrenia, we want to take a closer look at the global economic burden and specifically the economic burden in the United States of America and Germany.

Global

In their respective systematic reviews [74] and [40] highlight the tremendous economic burden of schizophrenia worldwide. Together they reviewed studies from 23 countries from all continents (e.g. Germany, USA, Japan, Australia, Nigeria).

Even though they found large methodological heterogeneity across the studies, all emphasize the high economic burden put on society by schizophrenia. In addition to direct medical and non-medical costs, they especially considered indirect costs due to morbidity and premature mortality. Productivity loss (e.g. sick leave, unemployment, early retirement) of the patients and family members or caregivers account for a large amount of the indirect health care costs. References to the numerous studies can be found in [74] and [40].

United States of America

The CATIE study investigated in Part IV of this thesis was conducted in the United States of America. Obviously there exist extensive studies about the economic burden of schizophrenia in the USA [120] [42] [149]. The most recent one (2013) called the economic burden significant and pointed out indirect and non-health care costs as strong contributors to this economic burden [42].

Germany

There are two studies directly investigating the economic burden of schizophrenia in Germany [91] [56]. Both acknowledge the high costs put on German society. Furthermore, in 2014 [56] found a considerable difference in the distribution of costs between younger and older patients. This is strongly confirmed by the most current data (2015) available from the federal government's health reporting based on data from the German Federal Statistical Office [134]. Schizophrenia, schizotypal and delusional disorders are responsible for 0.92% of all medical expenses for all ages. The share is much higher in some age groups. From ages 15 to 45 the share is twice as high, i.e. 2.07% of all medical expenses are due to schizophrenia. We want to set this share into context by comparing it with two diseases, that are believed to be rather common, diabetes mellitus and hypertension. Even though the overall share of medical expenses is higher for both disease, i.e. 2.18% and 2.99%, with 0.86% and 0.81% their share is considerably lower than the share of schizophrenia in the age group from 15 to 45. The data is visualized in Figure 2.1.

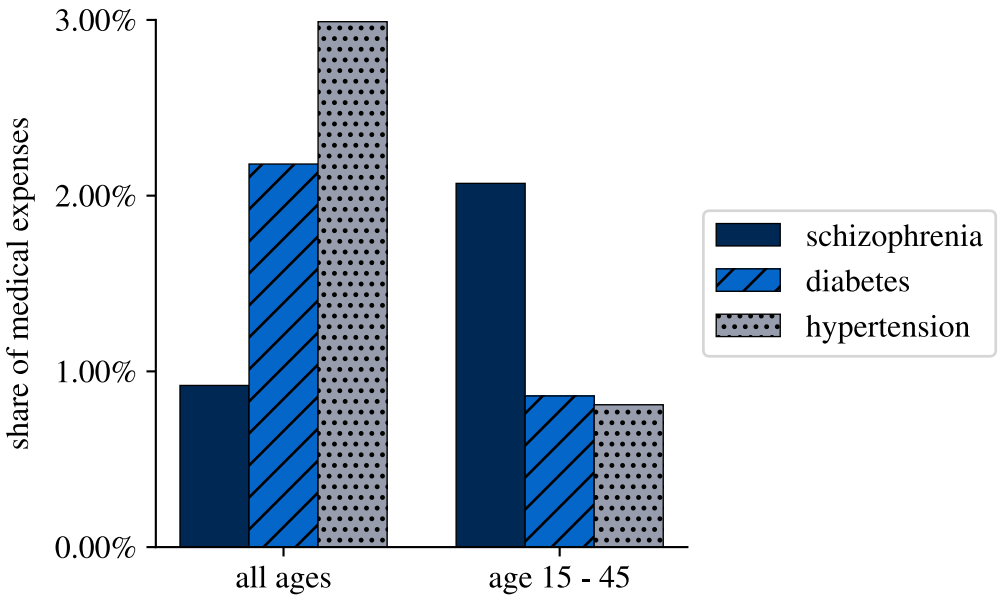


Figure 2.1: Medical expenses in Germany (2015) [134]

In conclusion, the importance of further investigations of data sets including patients suffering from schizophrenia is not only motivated by the immense individual burden put on patients and their families, but also by the tremendous economic burden put on society. Therefore, we investigate a data set from a clinical trial including schizophrenics in Part IV to help finding an improved individual treatment for each patient to reduce the patient's, his family's, and society's burden.

Part III

Methods

In Part III of this thesis we present a new method for explainable analytics in medicine.
We begin with an overview of common survival analysis in Chapter 3 and emphasize often encountered problems due to underlying heterogeneity. All subsequent methods are based on an endpoint-oriented clustering approach, developed by Brieden and Gritzmann.

In Chapter 4 we discuss the mathematical background in order to form sufficiently large clusters of patients with similar combinations of their characteristic values. This includes a transformation technique as well as an automated approach to identify the most promising baseline characteristics for the following methods.

Based on the geometric clustering, we will discuss the cluster-based analysis in Chapter 5. We will examine how to statistically evaluate the derived clusters and justify the division into those. To conclude, we will argue for the use of confidence intervals instead of hypothesis testing throughout the methods.

In the final chapter of this part, we will introduce the newly invented cluster-based survival analysis of individual patient data. We will present how to estimate cluster-based survival functions as well as cluster-based survival models and discuss their analysis. Furthermore, we will extend them to cluster-based stratified models. Finally, we will discuss how to evaluate differences regarding the outcome of patients, both across different treatments as well as different clusters.

The presented methods are both explainable and address heterogeneity inside the patient data, making them adoptable into clinical decision-making. All methods presented in this part have been developed in close cooperation with Prof. Dr. Andreas Brieden [26] [32].

3. Survival analysis

Survival analyses are statistical tools to analyze the time until an event of interest occurs [87]. Thereby time is usually measured in years, months or days from the beginning of a specific point in time, like the beginning of a treatment or the beginning of a clinical trial. The event of interest can be the decease of a patient. However, there are multiple other events, which might be of interest, like the dropout of a clinical trial or the discontinuation of the administered treatment. The latter is the case in CATIE, which we will analyze in Chapter 7. Since the discontinuation of a treatment is a one time event, we will only consider survival analyses with one event of interest. The interested reader is referred to [87] for further details on survival analysis with more than one event or recurrent events of interest.

In the following section we give an outline of survival functions - often referred to as survival curves - and the estimation of those with the Kaplan-Meier estimator as a representative. Furthermore, we want to give an overview about survival models using the Cox proportional-hazards model as representative. The following introduction is mainly based on [87] [71] [86]. Afterwards we will outline current possibilities to deal with heterogeneity in survival analysis and motivate the newly invented cluster-based survival analysis introduced in Chapter 6.

3.1 Survival function and its estimation

In survival analysis, the outcome variable is the time until a specific event of interest occurs, for instance death or discontinuation of treatment. The aim is to reliably estimate the probability of not reaching the event of interest in a specific time interval (i.e. the probability of survival or continuation of treatment). Opposed to other

outcomes, the time to the event is always positive and the underlying distributions are often skewed. Therefore, many other statistical procedures assuming normality of distributions do not apply. Furthermore, the true time to the event might not always be available due to incomplete data or the end of the clinical trial. We will focus on that in Section 3.1.3.

3.1.1 Survival data and survival function

Survival analyses are based on two kinds of information. Firstly, whether or not a patient experiences the event and secondly, the time until the event happens. Therefore, the first information describing the presence of the event of interest can be expressed as a Bernoulli distributed random variable

$$D \sim Ber(p)$$

with unknown parameter p and characteristics

$$d_j = \begin{cases} 0 & \text{event does not occur in observed period} \\ 1 & \text{event occurs in observed period} \end{cases}$$

for patient $j = 1, \ldots, n$.
The second information about the time until the event happens can be expressed as a random variable Y with characteristics $y_j > 0$ for patient $j = 1, \ldots, n$. Y can both follow a continuous or a discrete distribution, depending on the unit in which the time until the event happens is measured. For the sake of clarity and without loss of generality, the following derivations assume Y to follow a continuous distribution.

Remark 3.1. *In survival analysis, the random variable describing the time until the event happens is usually denoted by T instead of Y. However, as we combine common survival analyses with cluster-based analyses, we need a more general notation for the outcome, as it might not always be the time until an event happens.*

Definition 3.2 (survival data)**.** *The characteristics $(y_1, d_1), \ldots, (y_n, d_n)$ of the tuple of random variables (Y, D) are called survival data.*

In summarizing survival data, the most important information is given by the survival function describing the probability of survival.

Definition 3.3 (survival function)**.** *Let Y be a random variable denoting the time until the event of interest happens. Then*

$$S(y) := P(Y > y)$$

is called the corresponding survival function for $y > 0$.

Figure 3.1 shows a theoretical survival function.

Remark 3.4. *It holds that $S(0) = 1$ and $\lim_{y \to \infty} S(y) = 0$.*

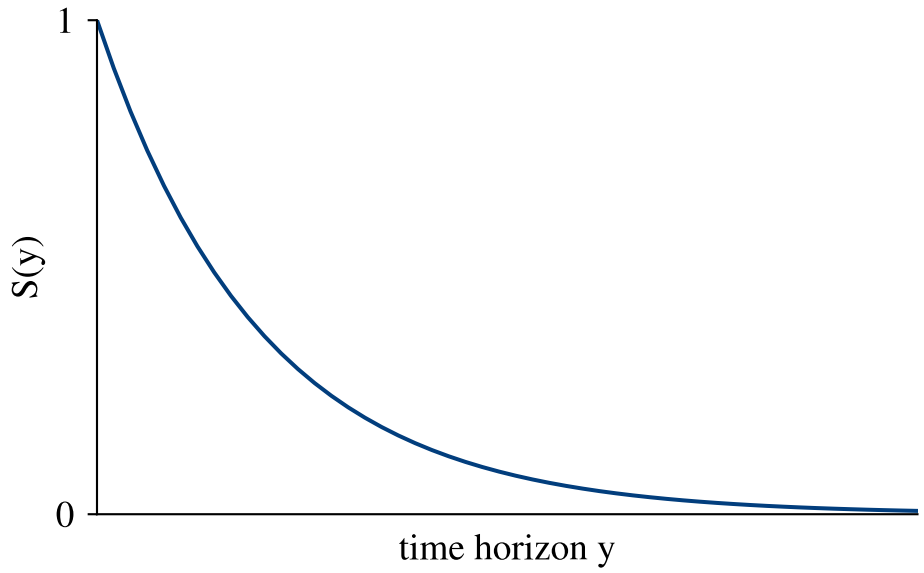


Figure 3.1: Illustration of a theoretical survival function

Remark 3.5. *Survival functions are often used to compare treatments with each other. In this case it is common to add an index for the treatment t to the respective survival function $S^t(y)$ in order to distinguish between multiple treatments.*

If the survival data is known to follow a specific distribution, it can be used as survival function. If however the true underlying distribution is unknown, the survival function has to be estimated.

3.1.2 Estimating the survival function

There are several possibilities to estimate the survival function $S(y)$. Many parametric methods make assumptions about the distribution of survival times Y. Common distributions include the exponential, Weibull and log-normal distribution [45]. The interested reader is referred to [71] [97] for details on parametric methods in survival analysis.

In contrast to parametric methods, non-parametric methods make no assumptions about the underlying distribution of the survival times [97]. Non-parametric methods aim to estimate the survival distribution and hence the survival function. One of the oldest methods for estimating the survival function is the life-table method [97]. Thereby survival times are organized into equally spaced intervals. Based on that the probabilities of surviving each of the defined intervals are estimated. Even though the approach is straight forward, an obvious disadvantage is the dependency on the prior defined intervals. The Kaplan-Meier estimator addresses this problem by re-estimating the survival probability each time an event occurs. Hence, the estimation of the probability of survival takes into account the size of the set at risk. Obviously the set at risk changes, once an event of interest occurs. However it might also change due to the loss of information on some patients. This phenomenon is called censoring. Another advantage of the Kaplan-Meier estimator is its ability to deal with some kind of censoring [71], which we will discuss in the following section.

3.1.3 Censoring

There are three different reasons for data being censored [87]:

- A patient does not experience the event of interest before the end of the observation period.

- A patient is lost to follow-up during the observation period.
- A patient withdraws from the study (e.g. due to adverse drug reaction).

In these cases the true survival time is unknown. However, it can be estimated based on the last observed survival time. Thereby the true survival time is assumed to be equal or greater than the observed survival time. This kind of censoring is called right censoring, as the information is missing on the right hand side of the observed time period. For further details on right censoring, censoring in general and examples for censored data please refer to [87] [71].

3.1.4 Kaplan-Meier estimator

Due to the true survival function $S(y)$ being unknown, it has to be estimated. Therefore let $\widehat{S}(y)$ denote the estimator of the true survival function $S(y)$, i.e. $\widehat{S}(y)$ estimates the probability of surviving longer than y. Many different estimators have been introduced to estimate the survival function [87] [86] [71]. However, we will always refer to the most commonly used estimator, the Kaplan-Meier estimator, when using $\widehat{S}(y)$. The estimator has originally been published in [84] and, due to its structure is often referred to as product limit approach.

The Kaplan-Meier method is based on the assumptions, that the survival probabilities are independent from the time of enrollment in a study, that the censoring times are independent from the event times, and that the events and censoring happened at the observed times [89].

The following derivations are based on [87] [86] [71].

Lemma 3.6. *Let $y_1, \ldots, y_n$ be the survival times, amongst which $m \leq n$ are times with an event, and $n-m$ censored values. We denote the rank-ordered survival times as $y_{(1)} < \ldots < y_{(m)}$. Then the survival function at the f-th ordered survival time can be obtained by*

$$\widehat{S}(y_{(f)}) = \prod_{j=1}^{f} \widehat{P}(Y > y_{(j)} | Y \geq y_{(j)})$$

Proof. A basic rule from probability is that the probability of a joint event (i.e. $P(A \cap B)$) is equal to the probability of one event (i.e. $P(A)$) times the conditional probability of the other event given the first event (i.e. $P(B|A)$). By defining the first event as $A = \{Y \geq y_{(f)}\}$ and the other one as $B = \{Y > y_{(f)}\}$ it immediately follows, that the joint event equals $B = \{Y > y_{(f)}\}$. Therefore:

$$\begin{aligned}\widehat{S}(y_{(f)}) &= \widehat{P}(Y > y_{(f)}) \\ &= \widehat{P}(Y > y_{(f-1)}) \cdot \widehat{P}(Y > y_{(f)} | Y \geq y_{(f)}) \\ &= \widehat{S}(y_{(f-1)}) \cdot \widehat{P}(Y > y_{(f)} | Y \geq y_{(f)})\end{aligned}$$

By recursively substituting $\widehat{S}(y_{(f-1)}), \widehat{S}(y_{(f-2)}), \ldots$ we get:

$$\widehat{S}(y_{(f)}) = \prod_{j=1}^{f} \widehat{P}(Y > y_{(j)} | Y \geq y_{(j)})$$

□

Corollary 3.7. *For any $y > 0$ and not only the observed survival times $y_{(1)} < \ldots < y_{(m)}$ Lemma 3.6 can be written as:*

$$\widehat{S}(y) = \prod_{y_{(f)} \leq y} \widehat{P}(Y > y_{(f)} | Y \geq y_{(f)})$$

Remark 3.8. *Lemma 3.6 and Corollary 3.7 are formulated for n patients experiencing n different survival times y_i. The survival function can also be estimated, if there are multiple occurrences at a specific survival time. The respective probabilities can be estimated taking the number of patients at risk and the number of patients with an event into account. Explicit representations will be derived hereafter.*

In order to compute the estimation, we need to estimate the probability of surviving past $y_{(f)}$ for any of the observed survival times. Let

$n_{(f)}$ be the number at risk at $y_{(f)}$ and
$d_{(f)}$ be the number of observed events at $y_{(f)}$.

Furthermore let $(y_1, d_1), \ldots, (y_n, d_n)$ be a sample of n independent observations (i.e. patients) from (Y, D). The number of patients at risk at time $y_{(f)}$ can be expressed by

$$n_{(f)} = \sum_{j=1}^{n} 1_{\{y_j \geq y_{(f)}\}}$$

and the number of patients with an event at exactly $y_{(f)}$ can be expressed by:

$$d_{(f)} = \sum_{j=1}^{n} 1_{\{y_j = y_{(f)}\}} \cdot 1_{\{d_j = 1\}}$$

Given survival until $y_{(f)}$, the probability of surviving past $y_{(f)}$ can then be estimated by:

$$\widehat{P}(Y > y_{(f)} | Y \geq y_{(f)}) = \frac{n_{(f)} - d_{(f)}}{n_{(f)}}$$

Therefore Lemma 3.6 simplifies to:

$$\widehat{S}(y) = \prod_{y_{(f)} \leq y} \frac{n_{(f)} - d_{(f)}}{n_{(f)}}$$

In conclusion, $\widehat{S}(y)$ measures the fraction of patients surviving at least until y. Therefore it has to be re-estimated if an event occurs ($n_{(f)}$ and $d_{(f)}$ change) or the set at risk changes due to censoring ($n_{(f)}$ changes).

The variance $\widehat{\sigma}^2_{\widehat{S}(y)}$ of the estimate $\widehat{S}(y)$ can be estimated by Greenwood's formula [61]:

Lemma 3.9. *Let $\widehat{S}(y)$ be a Kaplan-Meier estimator. The variance can then be estimated by [61]:*

$$\widehat{\sigma}^2_{\widehat{S}(y)} = \left(\widehat{S}(y)\right)^2 \cdot \sum_{y_{(f)} \leq y} \frac{d_{(f)}}{n_{(f)}(n_{(f)} - d_{(y)})}$$

3.2 Analysis of the estimated survival function

In this section we discuss how to analyze and compare the estimated survival functions. Besides a confidence interval for the Kaplan-Meier estimator, we take a closer look at the median survival time with respective confidence interval. In order to compare the estimated survival functions with each other, we will introduce the restricted mean survival time as a possibility to quantify the estimated survival function.
A popular method to compare two estimated survival functions with each other is performing a log-rank test, where the observed events are compared with the expected events. In a common survival analysis this is an appropriate approach. However, in this thesis we will introduce cluster-based survival analysis, during which specific subgroups of patients are determined within the method. Due to their determination being part of the method, we are not able to define the hypothesis for the log-rank test prior to the method. In Section 5.3.2 we will discuss this problem in detail and motivate the use of confidence intervals instead of hypothesis testing throughout this thesis.

3.2.1 Confidence interval for the survival function

The Kaplan-Meier estimator $\widehat{S}(y)$ is asymptotically normally distributed [97]. Using Lemma 3.9 we can therefore construct the following $(1-\alpha)$-confidence interval for the survival function $S(y)$ for every $y > 0$,

$$\mathtt{I}^{\alpha}\left(S(y)\right) \coloneqq \left[\widehat{S}(y) - z_{1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_{\widehat{S}(y)}, \widehat{S}(y) + z_{1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_{\widehat{S}(y)}\right]$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-percentile of the standard normal distribution.
The entire survival function is estimated, hence for every $y > 0$ the true probability of surviving longer than y, i.e. $S(y) = P(Y > y)$ is unknown. However, it holds that

$$P\left(S(y) \in \mathtt{I}^{\alpha}\left(S(y)\right)\right) = 1 - \alpha$$

and therefore in $(1-\alpha) \cdot 100\%$ of all cases, the true probability of surviving longer

than y lies in $\mathtt{I}^{\alpha}(S(y))$.

3.2.2 Median survival time

Sometimes the concern is to solely estimate the distribution of survival times in a single group. More often, the goal is to compare the survival times of two or more groups with each other. Usually the aim is determining, whether the survival time in a treatment group is longer than in the placebo group. Therefore, the estimated survival functions of two or more groups can be compared with each other. To quantify the difference between specific groups, a commonly used method is to compare the respective median survival times and their confidence intervals.

Definition 3.10 (median survival time)**.** *Let $S(y)$ be a survival function. If*

$$S(M) = 0.5$$

then M is called the median survival time of $S(y)$.

M represents the time, after which half of all patients discontinued their treatment. The true median survival time M is unknown and therefore has to be estimated based on the respective Kaplan-Meier estimations.

Remark 3.11. *$\widehat{S}(M)$ is the estimated survival probability at the true median survival time. We denote the estimated median survival time by $\widehat{M}$. It can be estimated by the argument, for which $\widehat{S}(y)$ falls below 0.5 for the first time.*

Lemma 3.12. *Let $\widehat{S}(M)$ be the estimated survival probability at the true median survival time M. Then according to [35]:*

$$\frac{\left(\widehat{S}(M) - 0.5\right)^2}{\widehat{\sigma}^2_{\widehat{S}(M)}} \sim \chi^2_1$$

Based on the distributional property from Lemma 3.12 we can construct the following $(1-\alpha)$-confidence interval for the median survival time.

Lemma 3.13.

$$\mathtt{I}^{\alpha}(M) \coloneqq [y_{lower}, y_{upper}]$$

is a $(1-\alpha)$*-confidence interval for the median survival time, with* y_{lower} *being the minimum and* y_{upper} *being the maximum y fulfilling*

$$\left(\widehat{S}(y) - 0.5\right)^2 < c_{\alpha} \cdot \widehat{\sigma}^2_{\widehat{S}(M)}$$

where c_{α} *denotes the* $(1-\alpha)$*-percentile of the* χ^2_1*-distribution.*

The median survival time has to be estimated, as M such that $S(M) = 0.5$ is unknown. However, it holds that

$$P\left(M \in \mathtt{I}^{\alpha}(M)\right) = 1 - \alpha$$

and therefore in $(1-\alpha) \cdot 100\%$ of all cases, the true median survival time lies in $\mathtt{I}^{\alpha}(M)$.

While comparing several groups with each other, the confidence intervals of their respective medians can be used as a measure. If their confidence intervals are found to be disjoint, i.e. non-overlapping, for a reasonable choice of α, the median survival time of the groups can be found to substantially differ from each other.

3.2.3 Restricted mean survival time

The median survival time is well established to compare estimated survival functions with each other. However, as the name suggests, only the median and not the entire estimated survival function is taken into account. To conclude this section we want to present a possibility to compare the entire estimated survival functions with each other.

Definition 3.14 (restricted mean survival time)**.** *Let* Y *denote the survival time,* $S(y)$ *the corresponding survival function and* y^* *the limited time horizon (e.g. end of the clinical trial). The restricted mean survival time (RMST) is the mean of the survival*

time limited to y^*, i.e. $E(min(Y,y^*))$. It equals the area under the survival function from $y = 0$ to $y = y^*$:

$$\widehat{RMST} = E(min(Y,y^*)) = \int_0^{y^*} S(y)dy$$

Remark 3.15. *It holds that*

$$\widehat{RMST} \in [0,y^*]$$

for all survival functions $S(y)$.

The restricted mean survival time describes the average survival time of patients from baseline to the limited time horizon y^*. It converges towards zero, with the time interval also converging towards zero. This is the case, if all events occur immediately after the starting point. The restricted mean survival time equals y^*, if no event occurs until the end of the time horizon. Both extreme cases are displayed in Figure 3.2.

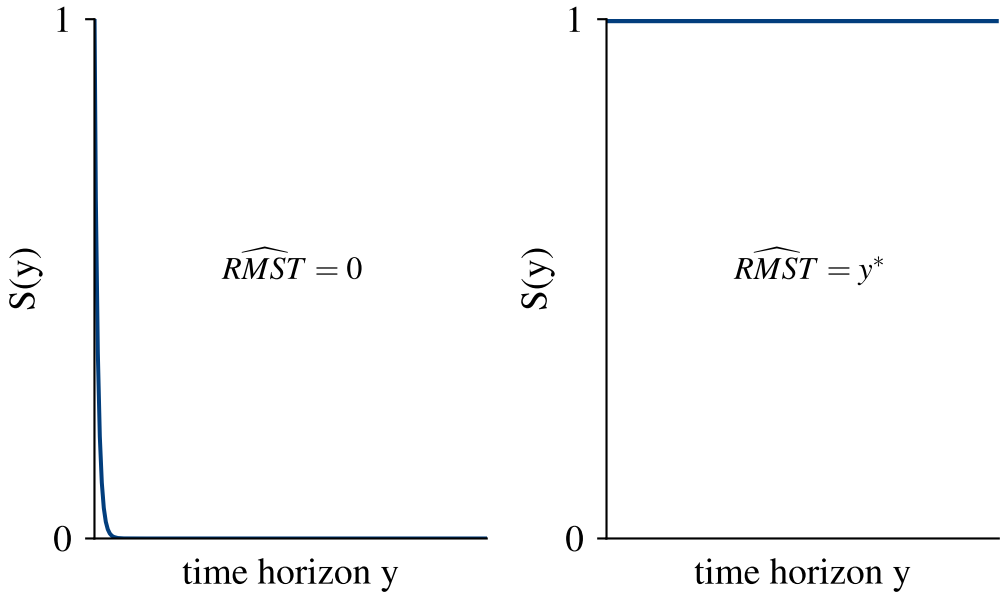


Figure 3.2: Extreme cases restricted mean survival time

The restricted mean survival time therefore takes the entire estimated survival function into account. Obviously the estimated survival functions may not overlap for the restricted mean survival time to be interpreted correctly.

In the following section we will discuss the Cox proportional-hazards model. One main assumption of this model are proportional-hazards between the compared groups. This assumption also leads to non overlapping survival functions, the above mentioned restriction.

3.3 Survival models

Based on survival functions, we will now present survival models. In addition to analyzing the time until the event of interest happens, survival models take additional covariates into account.
The Cox proportional-hazards model (Cox PH model) is a widely used survival model. Especially in medical research, it is often applied to relate independent covariates to the time until the event of interest happens. One of the typically investigated covariates is the administered treatment. Hence, the comparison of different treatments is one of the possibilities given by the Cox PH model. In contrast to sole survival functions, survival models like the Cox PH model provide an effect estimate by quantifying the difference between observed groups. Furthermore, the possibility of adjusting for covariates is provided. The method was originally introduced by Cox [44]. It is usually preferred over logistic regression models, which only include the occurrence of an event but neither the time until this event happens nor censoring. Both informations are used in Cox PH models. Furthermore, it will closely approximate the (obviously) unknown true model [87] [86]. If however the true underlying model is known for certain (i.e. the survival times follow an exponential distribution), it is preferred to use the respective parametric model [97]. Since this is often not the case, the Cox PH model offers the opportunity to closely approximate the true model and is therefore widely used, especially in medical research.

The following introduction is mainly based on [44] [87] [97] [86] [71], who also provide a deeper look into the topic for the interested reader. In contrast to survival functions, the measure of effect in Cox PH models is the hazard rate $h(t)$, which we will examine in the following section.

3.3.1 Hazard rate and hazard ratio

So far, we only used the survival function $S(y)$ to describe the probability of not having an event, given the survival up to a specific time point. As counterpart to $S(y)$ we want to introduce the hazard rate $h(y)$, describing the probability of having an event at a specific time point, again given the survival up to the specific time point [87].

Definition 3.16 (hazard rate)**.** *Let Y describe survival times for an event of interest. For any $y > 0$ the hazard rate $h(y)$ is defined as:*

$$h(y) := \lim_{\Delta y \to 0} \frac{P(y \leq Y < y + \Delta y | Y \geq y)}{\Delta y} \tag{3.1}$$

Remark 3.17. *Hazard rates are often used to compare treatments with each other. In this case it is common to add an index for the treatment t to the respective hazard rate $h^t(y)$ in order to distinguish between multiple treatments.*

Once either $S(y)$ or $h(y)$ is known, the other one can be derived immediately [87].

Lemma 3.18. *Let Y describe survival times for an event of interest and $S(y)$ and $h(y)$ the related survival function and hazard rate. $S(y)$ and $h(y)$ are connected as follows.*

$$S(y) = exp\left(\int_0^1 h(u)du\right)$$

and

$$h(y) = -\left[\frac{dS(y)/dy}{S(y)}\right]$$

In most situations the aim is to compare two groups with respect to their hazard rates. Usually the hazard ratio, i.e. the ratio of their respective hazard rates, is used as measure for the effect of the difference.

Definition 3.19 (hazard ratio)**.** *Let Y describe survival times for an event of interest and $h_1(y)$ and $h_2(y)$ the hazard rates for two groups of interest. Then the hazard ratio is defined as the ratio between those hazard rates:*

$$HR = \frac{h_2(y)}{h_1(y)}$$

Remark 3.20. *The hazard ratio is a measure for the probability of an event occuring in group 2 compared with the probability of an event occuring in group 1. Group 1 can therefore be referred to as reference group.*

$$HR \begin{cases} > 1 & \textit{probability of an event is higher in group 2} \\ < 1 & \textit{probability of an event is lower in group 2} \\ \approx 1 & \textit{probability of an event is approximately the same in both groups} \end{cases}$$

The true hazard rate and therefore all hazard ratios between any two groups are unknown. In the following we will show how the hazard rate and hazard ratio can be estimated by the Cox PH model in order to reliably compare different groups with each other.

3.3.2 Cox proportional-hazards model

We will begin by formulating the model, before briefly discussing how the coefficients of the model can be estimated.

Model formulation

Let $X = \left(X^1, \ldots, X^p\right)$ denote p independent covariates. Then for any $y > 0$ the Cox PH model can be written as

$$h(y,X) = h_0(y) \cdot exp\left(\sum_{i=1}^{p} \beta^i X^i\right) \tag{3.2}$$

where $h(y,X)$ is the expected hazard at time $y > 0$, given the covariates X, and $h_0(y)$ is the baseline hazard when all covariates are equal to zero. The Cox PH model therefore assumes the hazard at time y to be the product of the baseline hazard $h_0(y)$ and the exponential function of the linear combination of the covariates. The baseline hazard $h_0(y)$ does not depend on the covariates X, yielding to proportional hazards over time and thus a constant hazard ratio over time, explaining the name of the method. Another assumption of the Cox PH model is the independence of the observed survival times. Due to the mentioned assumptions, the Cox PH model is a semi-parametric model. However, as no assumption is made about the shape of the baseline hazard $h_0(y)$, it is only semi-parametric and not fully parametric. [71] provides an introduction to parametric survival models assuming the survival times to follow a specific distribution, like the exponential or Weibull distribution.

In the following we will restrict the covariates X to be fixed over time and thus independent from y, as we will later on use this method to predict a patient individual outcome by their baseline characteristics. Nevertheless, the Cox PH model can be extended to time varying covariates if needed [71] [87].

The interest now lies in the connection between each of the covariates X^i and the outcome. This can be quantified by the respective coefficient β^i for covariate X^i in (3.2). In the following we will briefly describe the estimation of the coefficients and present a confidence interval for it. Based on that we will explain how to compare different groups with each other using the respective estimations.

Partial likelihood estimation of β's

Usually, the maximum likelihood method is used to estimate parameters as in (3.2). However, this is not possible for the Cox PH model [71], which is thoroughly discussed in [83]. Cox therefore introduced a so-called partial likelihood function, that only depends on the parameter of interest. Partial likelihood estimators are proven to have the same distributional properties as full maximum likelihood estimators [71]. The interested reader is referred to [43] for the construction of the partial likelihood estimator.

Definition 3.21 (partial likelihood estimator)**.** *Let $X = \left(X^1, \ldots, X^p\right)$ denote p independent covariates and (3.2) the related Cox PH model. Then*

$$\widehat{\beta}^i$$

is called the partial likelihood estimator (PLE) of β^i for $i = 1, \ldots, p$.

The estimator $\widehat{\beta}^i$ quantifies the influence of covariate X^i on the hazard rate, i.e. the influence on the probability of having an event. The estimator of the variance of $\widehat{\beta}^i$ can be obtained as in most maximum likelihood estimations. The interested reader is referred to [71].

Definition 3.22 (variance of partial likelihood estimator)**.** *Let $\widehat{\beta}^i$ denote the partial likelihood estimation of β^i for $i = 1, \ldots, p$ in (3.2). Then*

$$\widehat{\sigma}^2_{\widehat{\beta}^i}$$

is called the estimated variance of $\widehat{\beta}^i$ for $i = 1, \ldots, p$.

Remark 3.23. *The actual estimations can be calculated using any common statistical software.*

3.4 Analysis of survival model

In the following section we want to discuss how a survival model like the Cox PH model can be analyzed. We will discuss, how two groups can be compared with each other by using the estimations from the Cox PH model to estimate their hazard ratio and present a confidence interval for the estimated hazard ratio.
Even though the Cox proportional-hazards model is a regression model and it is common to individually test the coefficients in such a model, we will refrain from doing so in this thesis. As mentioned, we will introduce cluster-based survival analysis during which specific subgroups of patients are determined within the method. Due to their determination being part of the method, we are not able to define the hypothesis for the tests of the coefficients prior to the method. In Section 5.3.2 we will discuss this problem in detail and motivate the use of confidence intervals instead of hypothesis testing throughout this thesis.

3.4.1 Estimated hazard ratio

By using the estimated Cox PH model, we are able to estimate the hazard ratio between two groups. Therefore let $X^1 = \left(x_1^1, \ldots, x_1^p\right)$, $X^2 = \left(x_2^1, \ldots, x_2^p\right)$ denote the characteristics of the covariates for group 1 and group 2 respectively. Then

$$\widehat{HR} = exp\left(\sum_{i=1}^{p} \widehat{\beta}^i \left(x_1^i - x_2^i\right)\right) \tag{3.3}$$

estimates the hazard ratio between group 1 and group 2. Hence, it is a measure for the probability of an event occuring in group 1 compared with the probability of an event occuring in group 2. If we want to estimate the effect of a single covariate X^i on the hazard rate, we assume the other covariates to have the same characteristics and therefore (3.3) simplifies to:

$$\widehat{HR} = exp\left(\widehat{\beta}^i \left(x_1^i - x_2^i\right)\right) \tag{3.4}$$

The level of scale of the involved covariates has to be taken into account for the hazard ratio in (3.3) and (3.4) to be interpreted correctly. Calculating the difference $\left(x_1^i - x_2^i\right)$ is reasonable, if the related covariate X^i is of cardinal or ordinal scale. X^i may also be of nominal scale, if this differences can be interpreted in a reasonable way, as we will see in the following derivations.
Often the aim of a survival analysis is to compare two treatments with each other or a treatment group with a placebo group. In that case one of the covariates in the Cox PH model, i.e. X^t, denotes the administration of a treatment. Thus, there are only two characteristics for X^t, namely $x_1^t = 1$ for patients receiving the treatment and $x_2^t = 0$ for patients receiving placebo. Then the exponential function of the estimated coefficient

$$\widehat{HR} = exp\left(\widehat{\beta}^t\right) \tag{3.5}$$

automatically yields the hazard ratio between the treatment and placebo group. It compares the probability of an event occurring for patients receiving the treatment with the probability of an event occurring for patients receiving placebo. When making comparisons like that, the Cox PH model is not restricted to the one covariate describing the administered treatment. If other covariates are included, the Cox PH model is said to be adjusted for those.

3.4.2 Confidence intervals for β's and hazard rates

In this section, we will show how to construct a confidence interval for the estimated coefficients $\widehat{\beta}^i$ and therefore the derived estimated hazard ratios. Since the partial likelihood estimator β^i has the same distributional properties as full maximum likelihood estimators, it is also asymptotically normally distributed [71]. Therefore we can construct the following $(1-\alpha)$-confidence interval for β^i in (3.2)

$$\mathtt{I}^{\alpha}\left(\beta^i\right) := \left[\widehat{\beta}^i - z_{1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_{\widehat{\beta}^i}, \widehat{\beta}^i + z_{1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_{\widehat{\beta}^i}\right]$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-percentile of the standard normal distribution.

The coefficients in (3.2) are estimated. However, it holds that

$$P\left(\beta^i \in \mathtt{I}^\alpha\left(\beta^i\right)\right) = 1-\alpha$$

and therefore in $(1-\alpha)\cdot 100\%$ of all cases, the true coefficient in (3.2) lies in $\mathtt{I}^\alpha\left(\beta^i\right)$.

When comparing two groups, i.e. two groups of treatment, with each other, we look at their hazard ratio and the respective confidence interval. Due to relation (3.5)

$$\mathtt{I}^\alpha\left(HR^t\right) \coloneqq \left[exp\left(\widehat{\beta}^t - z_{1-\frac{\alpha}{2}}\cdot\widehat{\sigma}_{\widehat{\beta}^t}\right), exp\left(\widehat{\beta}^t + z_{1-\frac{\alpha}{2}}\cdot\widehat{\sigma}_{\widehat{\beta}^t}\right)\right]$$

is a $(1-\alpha)$-confidence interval for the hazard ratio between the treatment and placebo group, if $\widehat{\beta}^t$ denotes the estimated coefficient for the variable describing the administration of a treatment. Instead of comparing a treatment group with a placebo group, it is also possible to compare two treatments with each other. If the interval does not include 1, even for small values of α, we can conclude the probability of an event in one group being substantially larger than in the other group.

3.5 Extension: The stratified Cox model

The main underlying assumption of the Cox PH model is, as the name suggests, the proportional behaviour of the hazards (PH assumption). However, some of the covariates might not satisfy this assumption. The stratified Cox model (SC model) provides a solution for this problem.

Let $X = \left(X^1,\ldots,X^p\right)$ denote p independent covariates satisfying the PH assumption and Z a categorical stratification covariate with values $1,\ldots,K$ not satisfying the PH assumption. Then for any $y>0$ the SC model can be written as

$$h(y,X,Z) = h_{Z0}(y)\cdot exp\left(\sum_{i=1}^{p}\beta^i X^i\right)$$

allowing the baseline hazard $h_{Z0}(y)$ to depend on the stratum, but assuming the effect of the covariates being the same for each stratum [86]. If more covariates are violating the PH assumption, a new stratification variable can be defined by forming combinations of the categories of these covariates and assigning those combinations as categories to the newly defined variable [87]. The coefficients β^i can be estimated in the same way as in the Cox PH model by partial likelihood estimations.

The main application of the SC model is to include variables for which the PH assumption is known to be violated. As the stratum-specific baseline hazard is now an essential part, the SC model can only be reliable if enough events are observed in each stratum [86]. Typical examples for stratification variables include the site in clinical trials (see Chapter 7). More details on the SC model can be found in [87].

In the final section of this chapter we will address the problem of underlying heterogeneity in survival analyses and briefly discuss current methods to deal with it.

3.6 Recommendation for the consideration of heterogeneity

The problem of underlying heterogeneity, both detected and undetected, and its influence on the estimated outcome has been critically discussed in several publications [72] [73] [101] [2] [1] [128]. Many of them present different methods of addressing heterogeneity in parametric survival models, i.e. survival models where the survival time is assumed to follow a specific distribution. Especially the mixture of distributions is proposed as one solution to deal with heterogeneity in parametric survival models [73] [2] [1]. The factors causing heterogeneity can thereby be modelled as covariates in the respective models. However, sometimes not all relevant factors are included as covariates due to the belief of them not being important. Thus, neglecting covariates leads to heterogeneity in survival analysis [73] and ambiguous results [1].

So far to the knowledge of the author, no method has been proposed for addressing heterogeneity in non-parametric models like the Kaplan-Meier method. In [128] heterogeneity in semi-parametric survival models is addressed by adjusting for other covariates and stratification, as introduced in Section 3.5. They point out the serious effect heterogeneity has on survival analyses, especially while comparing treatments with each other and thus emphasize the importance of considering heterogeneity. However, they only describe how to include a covariate which is known to cause heterogeneity. Especially in recent clinical trials there are far too many possible individual patient characteristics recorded, which might cause heterogeneity. Simply including all of them as covariates would end in over-fitting the survival model. Therefore it is of importance to not only find a way to include factors causing heterogeneity in survival analysis, but also to detect the covariates with (the most) impact on the survival time. Furthermore, sometimes not a sole covariate, but the combination of two or more covariates might be causing the underlying heterogeneity, motivating the following recommendation for the consideration of heterogeneity in survival analysis.

As mentioned in the beginning of this chapter, survival analysis is often used in the analysis of individual patient data. Consequently the analysis is highly influenced by underlying heterogeneity between patients. As patients certainly differ regarding their characteristics, it is very likely, that the sole use of averages or an average model to estimate the outcome of interest for all patients is not expedient. Keeping in mind, that nowadays many patient characteristics are collected in addition to the outcome, we recommend using them as possible predictors. In order to use them in the best possible way, we discuss the endpoint-oriented clustering in the next chapter. The approach identifies clusters of patients that are as homogeneous as possible. Besides performing a cluster-based analysis on these as described in Chapter 5, we can extend it to the newly invented cluster-based survival analysis, as described in Chapter 6, offering a new possibility to reliably evaluate medical interventions while at the same time addressing underlying heterogeneity in the patient data. Even though these new approaches are able to address heterogeneity caused by the individual patient data,

they are still explainable and therefore include a unique explanation why the specific prediction was made.

4. Geometric clustering of individual patient data

In this chapter we briefly discuss the geometric clustering approach originally introduced by Brieden and Gritzmann [27]. Furthermore, we present the transformation technique used throughout this thesis and discuss the selection of input variables for the geometric clustering approach. Thereby we use an automated selection approach developed by Brieden and Gritzmann [26] and introduce a new alternative to quantiles while classifying variables of ordinal or cardinal scale.

4.1 Geometric clustering

The geometric clustering approach is supposed to identify sufficiently large and especially homogeneous clusters of patients. It is based on methods of combinatorial optimization. For an introduction on basic combinatorial optimization please refer to [63]. The geometric clustering approach was originally developed by Brieden and Gritzmann for an agricultural economics application [27]. However, due to it being a general technique to explore high-dimensional data and especially identify structures in it, the approach is not limited to this application. In [67] it is used to reduce heterogeneity in meta-analysis, another widely used statistical method in medical data analysis. The general approach of using it in medical data analysis is presented in [26]. It has also been applied to other economic applications, such as entrepreneurship in Russia [33]. All applications of the approach resulted in generating new insight into the investigated problem, finding structures, which have been hidden before. That is why we use the geometric clustering technique to identify homogeneous clusters of patients in the cluster-based analysis presented in Chapter 5

as well as the newly invented cluster-based survival analysis introduced in Chapter 6.

In the following section we want to give an overview about the most important parts of the geometric clustering and its mathematical background. Further information as well as studies on structure, complexity and efficient approximations of the problem can be found in [19] [21] [28] [29] [25] [30] [31]. A summary is provided in [22] [20].

4.1.1 Terminology

The approach is based on the available patient data being expressed as geometric objects. By expressing them as points in a geometric space we enable the possibility to measure their distance by the Euclidean norm. Obviously some of the patients' characteristics might be random variables with a non-metric sample space making it difficult to think of them as geometric objects. We will use a transformation technique introduced in Section 4.2 to interpret each patient as a point in $\mathbb{R}^p$ with p being the number of available patients' characteristics.

Definition 4.1 (patient data set of characteristics)**.** *Let $n \in \mathbb{N}$ be the number of patients and $p \in \mathbb{N}$ the number of available patients' characteristics. The patient data set of characteristics is then defined as*

$$X := \{x_j\}_{j=1}^{n} \subset \mathbb{R}^p$$

with $x_j = \left(x_j^1, \ldots, x_j^p\right) \in \mathbb{R}^p$ for $j = 1, \ldots, n$.

Next, we need to define the patients' outcome, that we aim to predict by the explainable methods. Therefore let random variable Y denote the outcome of a patient with corresponding sampling space Ω. Depending on the outcome, Y follows different distributions. In medical data analysis the outcome is often Bernoulli distributed (e.g. remission or response) or it follows some unknown distribution which can be assumed to be discrete (e.g. survival time).

Definition 4.2 (patient data)**.** *Let X be the patient data set of characteristics and Y the corresponding outcome with sampling space Ω. Then the patient data set is defined as*

$$G := \left\{ \left(x_j, y_j \right) \right\}_{j=1}^{n} \subset \mathbb{R}^p \times \Omega$$

with $x_j \in X$ and $y_j \in \Omega$.

The aim is to split the patient data set of characteristics X into $k \in \mathbb{N}$ homogeneous clusters. Therefore, we define this kind of partition as the following k-clustering.

Definition 4.3 (k-clustering)**.** *Let $k \in \mathbb{N}$ be the number of clusters. A k-clustering $Cl = (Cl_1, \ldots, Cl_k)$ is a partition of the data set X into k nonempty sets $Cl_1, \ldots, Cl_k$. The size of cluster Cl_i thereby describes the number of patients in it and is denoted by κ_i for $i = 1, \ldots, k$.*

The identified patient data sets need to have a certain amount of patients in them in order to achieve high statistical power. Otherwise it is not possible to significantly detect differences amongst the patients' outcome. Therefore, we extend the upper definition to a (k,l,u)-clustering, restricting the size of each cluster by an upper and lower bound.

Definition 4.4 ((k,l,u)-clustering)**.** *A k-clustering $Cl = (Cl_1, \ldots, Cl_k)$ with $l = (l_1, \ldots, l_k) \in \mathbb{N}^k$, $u = (u_1, \ldots, u_k) \in \mathbb{N}^k$ fulfilling $l_i \leq \kappa_i \leq u_i$ for $i = 1, \ldots, k$ is called (k,l,u)-clustering.*

The geometric clustering approach aims to find the best clustering amongst all possible clustering. Therefore we define the set of all feasible (k,l,u)-clusterings.

Definition 4.5 (set of feasible (k,l,u)-clusterings)**.** *Let X bet the patient data set of characteristics of $n \in \mathbb{N}$ patients and $k \in \mathbb{N}$ the number of clusters. Then*

$$\mathscr{C}(X,k,l,u) := \{Cl : Cl \text{ is a } (k,l,u)\text{-clustering of } X\}$$

is the set of all feasible (k,l,u)-clusterings of X.

Note, that we have not yet defined how to distinguish between 'good' and 'bad' clusterings or as a matter of fact how to identify the 'optimal' clustering.

In the following we want to define the bounded-shape partition polytope used to derive feasible clusterings. It is possible to describe the bounded-shape partition polytope (*BSPP*) as the convex hull of so-called cluster sum vectors [28] [29]. However, we will immediately define the *BSPP* by linear constraints, which are being used in the optimization problem in the following section [75] [11].

Definition 4.6. *Let $Cl = (Cl_1, \ldots, Cl_k)$ be a k-clustering of the patient data set of characteristics $X = \{x_j\}_{j=1}^{n}$. Then for all $i = 1, \ldots, k$ and $j = 1, \ldots, n$*

$$\xi_{ij} = \begin{cases} 1 & \text{if } x_j \in Cl_i \\ 0 & \text{if } x_j \notin Cl_i \end{cases}$$

denotes the decision variable indicating whether cluster Cl_i contains patient x_j.

Definition 4.7 (bounded-shape partition polytope (*BSPP*))**.** *Let $k, l_i, u_i \in \mathbb{N}$ for $i = 1, \ldots, k$ with $\sum_{i=1}^{k} l_i \leq n \leq \sum_{i=1}^{k} u_i$. The polytope defined by*

$$\begin{aligned}
\sum_{j=1}^{n} \xi_{ij} &\leq u_i && \forall i = 1, \ldots, k \\
\sum_{j=1}^{n} \xi_{ij} &\geq l_i && \forall i = 1, \ldots, k \\
\sum_{i=1}^{k} \xi_{ij} &= 1 && \forall j = 1, \ldots, n \\
\xi_{ij} &\geq 0 && \forall i = 1, \ldots, k \quad \forall j = 1, \ldots, n
\end{aligned}$$

is called the bounded-shape partition polytope (BSPP(k, l, u)).

4.1.2 Clustering as an optimization problem

Our aim is to partition patients into clusters which are as homogeneous as possible. This corresponds to finding a feasible clustering which maximizes a specific target function f.

Definition 4.8 (bounded-shape partition problem)**.** *The following integer linear program (ILP) is called bounded-shape partition problem.*

$$
\begin{aligned}
&\max f(\xi) && && \\
&\sum_{j=1}^{n} \xi_{ij} \leq u_i && \forall i = 1,\ldots,k && \\
&\sum_{j=1}^{n} \xi_{ij} \geq l_i && \forall i = 1,\ldots,k && \\
&\sum_{i=1}^{k} \xi_{ij} = 1 && \forall j = 1,\ldots,n && \\
&\xi_{ij} \in \{0,1\} && \forall i = 1,\ldots,k && \forall j = 1,\ldots,n
\end{aligned}
\tag{4.1}
$$

The optimal (k,l,u)-clustering can be obtained by the solution $\xi^*_{ij} \in \{0,1\}^{k\times n}$ of the bounded-shape partition problem [75]. It is also shown, that due to the unimodularity of the constraints the bounded-shape partition problem can be solved in polynomial time [75]. Moreover due to every solution of the relaxed problem already being integral, it suffices to solve the following relaxation of the bounded-shape partition problem:

$$
\begin{aligned}
&\max f(\xi) && && \\
&\sum_{j=1}^{n} \xi_{ij} \leq u_i && \forall i = 1,\ldots,k && \\
&\sum_{j=1}^{n} \xi_{ij} \geq l_i && \forall i = 1,\ldots,k && \\
&\sum_{i=1}^{k} \xi_{ij} = 1 && \forall j = 1,\ldots,n && \\
&\xi_{ij} \geq 0 && \forall i = 1,\ldots,k && \forall j = 1,\ldots,n
\end{aligned}
$$

To derive an optimal clustering, two desirable properties are taken into account, the so-called strict separability and the homogeneity of clusters.
If a hyperplane separating the related sets in $\mathbb{R}^p$ exists, two clusters are called linearly separable. A clustering $Cl = (Cl_1, \ldots, Cl_k)$ allows linear separation, if any two clusters Cl_i and Cl_j are linearly separable for $i \neq j$, $i, j = 1, \ldots, k$. More details can be found in [67]. In this thesis we use the following theorem from Barnes, Hoffmann and Rothblum [11]:

Theorem 4.9. *Let v^* be a vertex of the BSPP. Then the clustering Cl^* associated with v^* allows strict linear separation.*

Proof. See [11]. □

Due to Theorem 4.9 we can automatically identify clusters of patients, that 'strictly' differ from each other both regarding their baseline characteristics. The strict separation of the clusters therefore results in unique assignment rules of patients into one of the clusters, which is crucial for the methods and results to be explainable. This especially enables the possibility to assign completely new patients into one of the clusters by looking at their respective patient characteristics.
The second property is the identified clusters being as homogeneous as possible. Hence, the target function in the bounded-shape partition problem (4.1) should describe the homogeneity inside the clusters. By maximizing the distance between the centres of each cluster we achieve exactly the desired property. Therefore the target function can be formulated as

$$\max_{Cl \in \mathscr{C}} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \left\| c_i - c_j \right\|_2^2$$

with the centre of gravity

$$c_i = \frac{1}{\kappa_i} \sum_{j=1}^{n} x_j \xi_{ij}$$

for each cluster Cl_i for $i = 1, \ldots, k$. The distance between the centres is thereby measured by the Euclidean norm, denoted by $\|.\|_2$. However, this target function leads to the optimization problem being nonlinear and therefore is $\mathbb{NP}$-hard [17].

Two alternatives exist, the Least-Square Assignment (LSA) and Cluster Sum Assignment (CSA), based on piecewise linear approximation, initially proposed by [29]. We will be using the latter in this thesis. A detailed description of both can be found in [112]. In [112] it is shown that under some conditions both approaches are equivalent. By using the linear target function of the Cluster Sum Assignment for the bounded-shape partition problem, we are able to find an optimal clustering by solving a linear program [112] [67].

The geometric clustering approach discussed in this and the previous section forms the basis of the partition of the patient data G into homogeneous cluster of patients. Based on the introduction and especially the given references, we can assume the identified cluster of patients to be as homogeneous as possible.

It is crucial for the geometric clustering approach that all available data is represented as geometric objects. In the following section we will discuss a transformation technique resulting in all patients being interpreted as points in the geometric space.

4.2 Transformation of data

The applied geometric clustering technique as described above, is conducted in a geometric space and therefore requires a transformation of all input variables into quantitative values, regardless the level of scale. In this section we will give an overview of the transformation technique introduced in [26]. The probabilistic foundations of the following transformation technique can be found in [90].
The aim is to cluster patients which are similar regarding both their patient characteristics as well as their respective outcome. Therefore the expected value of the outcome Y given all the patient characteristics is of interest. The main idea is to replace each characteristic value of a variable $X^i = x$ by the estimated conditional expected value of the outcome $E\left(Y|X^i = x\right)$. Therefore, let $X^1, \ldots, X^p$, $p \in \mathbb{N}$, be the available input variables, i.e. patient characteristics. Each data point, i.e. patient, can then be formally written as $x = \left(x^1, \ldots, x^p\right)$.

The transformation technique is motivated by the 'Naive Bayes approach', which aims to estimate the following probability:

$$P\left(Y = y | X^1 = x^1, \ldots, X^p = x^p\right)$$

By applying Bayes' rule for conditional probabilities and the 'naive' assumption about independence of the input variables $X^1, \ldots, X^p$ it holds:

$$P\left(X^1 = x^1, \ldots, X^p = x^p | Y = y\right) = \prod_{i=1}^{p} P\left(X^i = x^i | Y = y\right) \tag{4.2}$$

Therefore, the estimation of the upper probability simplifies to the estimation of the marginal probabilities $P\left(X^i = x^i | Y = y\right)$, which are much easier to estimate.

We want to use a similar approach to estimate the following conditional expected value:

$$E\left(Y | X^1 = x^1, \ldots, X^p = x^p\right)$$

Due to the definition of the expected value, we can express the upper conditional expected values as the following:

$$E\left(Y | X^1 = x^1, \ldots, X^p = x^p\right) = \sum_{y \in \Omega} y \cdot P\left(Y = y | X^1 = x^1, \ldots, X^p = x^p\right)$$

Using the 'naive' assumption (4.2) we can substitute the conditional probabilities by

$$P\left(Y = y | X^1 = x^1, \ldots, X^p = x^p\right) = \sum_{i=1}^{p} \beta^i P\left(Y = y | X^i = x^i\right)$$

with

$$\beta^i = \widehat{\beta}^i \frac{P\left(Y = y | X^1 = x^1, \ldots, X^p = x^p\right)}{P\left(Y = y | X^i = x^i\right)}$$

and $\sum_{i=1}^{p} \widehat{\beta}^i = 1$.

Therefore the desired expected value can be expressed as convex combination of the marginal one-dimensional expected values:

$$E\left(Y|X^1=x^1,\ldots,X^p=x^p\right)=\sum_{i=1}^{p}\beta^i E\left(Y=y|X^i=x^i\right)$$

with

$$\beta^i=\widehat{\beta}^i\frac{E\left(Y|X^1=x^1,\ldots,X^p=x^p\right)}{E\left(Y|X^i=x^i\right)}$$

and $\sum_{i=1}^{p}\widehat{\beta}^i=1$.

The marginal one-dimensional conditional expected values $E\left(Y|X^i=x\right)$ are the new values replacing the original value $X^i=x$ for $i=1,\ldots,p$. Each data point, i.e. each patient, $x_j\in X$, $x_j=\left(x_j^1,\ldots x_j^p\right)$ is then represented by the following vector of their conditional expected values:

$$\left(x_j^1,\ldots,x_j^p\right)\to\left(E\left(Y|X^1=x_j^1\right),\ldots,E\left(Y|X^p=x_j^p\right)\right)$$

The transformed vectors are always geometric objects in $\mathbb{R}^p$ and therefore the geometric clustering approach described in Section 4.1 can be applied.

In the following we want to describe how these conditional expected values can be properly estimated and derive special formulations for a Bernoulli distributed outcome. An introduction into the estimations used in the following can be found in [8].

Theorem 4.10. *Let $G=\left\{\left(x_j,y_j\right)\right\}_{j=1}^{n}\subset\mathbb{R}^p\times\Omega$ with $x_j=\left(x_j^1,\ldots x_j^p\right)$ be a sample of p random variables $X^1,\ldots,X^p$ and Y the discrete outcome. Then*

$$\widehat{\theta}\left(Y|X^i=x\right)=\frac{\sum_{j=1}^{n}y_j\cdot 1_{\left\{x_j^i=x\right\}}}{\sum_{j=1}^{n}1_{\left\{x_j^i=x\right\}}}$$

is an unbiased estimator for the conditional expected value $E\left(Y|X^i=x\right)$ for $i=1,\ldots,p$.

Proof. According to Bayes' rule the conditional expected value can be written as

$$E\left(Y|X^i = x\right) = \sum_{y\in\Omega} y\frac{P\left(Y = y, X^i = x\right)}{P\left(X^i = x\right)}$$

for $i = 1,\ldots,p$. Therefore, the estimators for the joint probability

$$\widehat{P}\left(Y = y, X^i = x\right) = \frac{\sum_{j=1}^{n} y \cdot 1_{\left\{x_j^i = x\right\}}}{n}$$

and the probability

$$\widehat{P}\left(X^i = x\right) = \frac{\sum_{j=1}^{n} 1_{\left\{x_j^i = x\right\}}}{n}$$

only consist of the (conditional) frequencies and lead to a mean estimation of the upper conditional expected value $E\left(Y|X^i = x\right)$. □

Remark 4.11. *Let $Y \sim Ber(p)$ be a Bernoulli distributed random variable with parameter p. The conditional expected value can then be obtained by the conditional probability of success:*

$$E\left(Y|X = x\right) = P\left(Y = 1|X = x\right) =: p_x$$

Due to Remark 4.11 Theorem 4.10 leads to the following corollary.

Corollary 4.12. *Let $G = \left\{\left(x_j, y_j\right)\right\}_{j=1}^{n} \subset \mathbb{R}^p \times \{0,1\}$ with $x_j = \left(x_j^1, \ldots x_j^p\right)$ be a sample of d random variables $X^1,\ldots,X^p$ and Y denote the binary outcome. Then*

$$\widehat{\theta}\left(Y|X^i = x\right) = \frac{\sum_{j=1}^{n} 1_{\left\{y_j=1\right\}} 1_{\left\{x_j^i = x\right\}}}{\sum_{j=1}^{n} 1_{\left\{x_j^i = x\right\}}}$$

is an unbiased estimator for the conditional expected value $E\left(Y|X^i = x\right)$ or the conditional probability $P\left(Y = 1|X^i = x\right)$ for $i = 1,\ldots,p$.

Using these estimators we can define the transformed patient data, on which the geometric clustering approach stated in Section 4.1 can be conducted.

Definition 4.13 (transformed patient data)**.** *Let X bet the patient data set of characteristics, Y the corresponding outcome with sampling space Ω and G the derived patient data. The transformed patient data set is defined by*

$$\widehat{G} := \left\{ \left(\widehat{x}_j, y_j \right) \right\}_{j=1}^{n} \subset \mathbb{R}^p \times \Omega$$

with $\widehat{x}_j = \left(\widehat{x}_j^1, \ldots, \widehat{x}_j^p \right)$ *and* $\widehat{x}_j^i = \widehat{\theta}\left(Y | X^i = x_j^i \right)$ *for* $i = 1, \ldots, p$ *and* $j = 1, \ldots, n$.

Usually not many unique characteristics x_j^i exist for each input variable X^i, $i = 1, \ldots, p$. Therefore the number of transformed values is relatively small. If however too many unique values exist, a prior classification might be necessary or reasonable. For variables of nominal scale, a classification is most likely based on expert knowledge in the respective field. Variables of ordinal or cardinal scale can be summarized in quantiles in order to eliminate a high number of unique characteristics. Based on the concept of quantiles, we will present an alternative method in Section 4.3.2 to find a suitable classification of variables of ordinal and cardinal scale.

Remark 4.14. *Often clinical trials face the problem of high dropout rates, even when this is not the inspected outcome. While regarding a binary outcome (i.e. response/remission) the outcome of dropout patients is therefore assumed to be zero. In this case the introduced transformation technique is biased by those dropout patients. In order to avoid that, dropout patients can be left out while determining the transformed values. Before performing the geometric clustering approach the dropout patients are assigned the transformed values based on their respective characteristic values.*

At this point respective algorithms can be defined to divide the patient data set G into homogeneous clusters, combining the transformation technique and the geometric clustering. In [112] and [67] the particular algorithms are stated and explained. For this thesis we conclude, that once we transform the patient data G into $\widehat{G}$ we can obtain homogeneous clusters of patients fulfilling the desired properties discussed in this section.

The success of the geometric clustering approach with the described transformation technique highly relies on the used input variables. In the next section we discuss an automatic approach to select the input variables promising the highest impact on the outcome. Additionally, we introduce a new approach to classify variables of ordinal and cardinal scale before using them for the geometric clustering.

4.3 Selection of variables for geometric clustering

Due to the collection of large amounts of data, especially in medical research, there are usually too many baseline characteristics available for the desired data analysis. Some small clinical trials collect nearly as many possible baseline characteristics as there are patients enrolled. Any analysis performed on all available baseline characteristics will be extremely over-fitted. Therefore, we use a method introduced by Brieden and Gritzmann to automatically select those baseline characteristics with the highest potential for being a predictor for the desired outcome [26].
For characteristics of ordinal or cardinal scale it is a common to classify those into quantiles before the actual data analysis. Based on the geometric clustering described in Section 4.1 we introduce a new method to more appropriately classify characteristics of ordinal or cardinal scale.

4.3.1 Automated selection approach

To determine the baseline characteristics with the highest impact on the outcome, we need to define how we intend to measure this impact. Therefore, let $X^1,\dots,X^p$ be the available baseline characteristics and Y the outcome. Each baseline characteristic X^i has m^i unique values. Therefore, let $a^i_1,\dots,a^i_{m_i}$ denote those unique values for baseline characteristic X^i. Furthermore, let

$$\widehat{a}^i_j := \widehat{\theta}\left(Y|X^i = a^i_j\right)$$

denote the respective estimation of the conditional expected value for $j = 1,\dots,m^i$.

We measure the impact on the outcome by the following outcome-variance.

Definition 4.15 (outcome-variance)**.** *With the upper notation and* $\bar{a}^{\,i} := \frac{1}{m^i}\sum_{j=1}^{m^i} \widehat{a}^{\,i}_j$ *we call*

$$OVar\left(X^{\,i}\right) := \frac{1}{m^i}\sum_{j=1}^{m^i}\left(\widehat{a}^{\,i}_j - \bar{a}^{\,i}\right)^2$$

the outcome-variance of baseline characteristic $X^{\,i}$ *for* $i = 1,\ldots,p$.

Remark 4.16. *The outcome-variance describes the empirical variance of the unique transformed values of a baseline characteristic.*

Let $M \in \mathbb{N}$ be the number of baseline characteristics supposed to perform the geometric clustering approach on. Those M characteristics can be determined by the following procedure. First of all, the outcome-variance $OVar\left(X^{\,i}\right)$ has to be calculated for each available patient characteristic $X^{\,i}$. Afterwards the patient characteristics $X^{\,1},\ldots,X^{p}$ are ranked according to their outcome-variance $OVar\left(X^{\,i}\right)$. Finally those M characteristics with the highest outcome-variance are selected for the geometric clustering. The method is summarized in Procedure 4.1.

input : all available patient characteristics $X^{\,1},\ldots,X^{\,p}$
output : M patient characteristics for the geometric clustering

begin

1. **for** $i = 1,\ldots,p$ **do**
   calculate $OVar\left(X^{\,i}\right)$
   end
2. rank $X^{\,i}$ according to $OVar\left(X^{\,i}\right)$
3. select M patient characteristics with the highest outcome-variance

end

Procedure 4.1: Selection of patient characteristics for geometric clustering

The resulting M baseline characteristics provide the input for the geometric clustering approach.

Remark 4.17. *The choice of M depends on the amount of patients and is usually influenced by the application.*

A classification of characteristics of ordinal and cardinal into quantiles might be necessary or reasonable. Otherwise the number of possible values is too high and every value rarely occurs. Thereby $x^i(q)$ denotes the q^{th} quantile of a patient characteristic X^i, if at least $100 \cdot q\%$ of all observations are less or equal to $x^i(q)$ and simultaneously at least $100 \cdot (1-q)\%$ of all observations are greater or equal to $x^i(q)$.

Remark 4.18. *We call a variable is classified into quantiles, if the borders of the classes represent the respective quantiles.*

The remaining question is, into how many quantiles the characteristic should be classified. In order to determine the most appropriate amount of quantiles we can use an adaption of Procedure 4.1.

Therefore, let X^i be a characteristic of ordinal or cardinal scale. For $j = 2, \ldots, n_{quantiles}$ with $n_{quantiles}$ denoting the maximum amount of quantiles, let X^{ij} represent characteristic X^i grouped into j quantiles. Then the number of quantiles can be determined by the following procedure. First of all, the outcome-variance $OVar\left(X^{ij}\right)$ has to be calculated for patient characteristic X^i grouped into j quantiles. Afterwards the grouped patient characteristics $X^{i1}, \ldots, X^{in_{quantiles}}$ are ranked according to their outcome-variance $OVar\left(X^{ij}\right)$. Finally, patient characteristic X^i is grouped into j^* quantiles, with j^* being the argument j maximizing the outcome-variance $OVar\left(X^{ij}\right)$:

$$j^* = \underset{j \in [2, n_{quantiles}]}{argmax} \; OVar\left(X^{ij}\right)$$

The method is summarized in Procedure 4.2.

input : cardinal or ordinal input variable X^i; maximum number of quantiles $n_{quantiles}$
output : number of quantiles for variable X^i

1. **for** $j = 2, \ldots, n_{quantiles}$ **do**
   calculate $OVar\left(X^{ij}\right)$
   end
2. rank X^{ij} according to $OVar\left(X^{ij}\right)$
3. use $\underset{j \in [2, n_{quantiles}]}{argmax} \; OVar\left(X^{ij}\right)$ quantiles for X^i

Procedure 4.2: Determine number of quantiles for characteristic X^i

With Procedure 4.2, every patient characteristic of ordinal or cardinal scale can be classified into the most auspicious amount of quantiles.

Remark 4.19. *The choice of $n_{quantiles}$ depends highly on the amount of data points, i.e. patients. It might also be set by the application.*

In the final section of this chapter, we present an alternative to quantiles.

4.3.2 Optimized classes instead of quantiles

In order to apply the transformation technique it is usually reasonable to classify patient characteristics of ordinal and cardinal scale. Even though quantiles are appropriate in many situations and easily applicable, another classification might be more suitable for the regarded patient characteristic. In the following we introduce a new method to classify patient characteristics of ordinal or cardinal scale in an optimal manner. The approach is closely related to the geometric clustering technique discussed in the beginning of this chapter. Opposed to the general clustering approach, we only look at one baseline characteristic and the outcome. We want to classify the patients, i.e. data points, such that the resulting classes are homogeneous regarding

both the characteristic and the outcome. On the other hand, the classes should be as heterogeneous as possible to each other. In other words, we want to minimize the variance of the outcome inside the classes and maximize it between them. Thereby we want to make sure, that every patient is assigned to exactly one class and that the allocation is cohesive and ascending, as it is the case for a classification based on quantiles.

To illustrate the advantage of optimized classes, we take a look at the following small examples including ten patients, where the baseline characteristic age has to be categorized and the response of every patient to an administered treatment is given. Grouping the patients into two quantiles results in a response rate of 0.2 in the first group and a response rate of 1 in the second group. However, if we were to group patients from the ages of 30 to 33 and 34 to 39 the first class shows a response rate of 0 and the second class shows a response rate of 1. The example is displayed in Table 4.1.

patient	1	2	3	4	5	6	7	8	9	10
age	30	31	32	33	34	35	36	37	38	39
response	0	0	0	0	1	1	1	1	1	1
2 quantiles	0.20					1.00				
optimized classes	0.00				1.00					

Table 4.1: Quantiles and optimized classes for a patient characteristic (1)

The advantage of optimized classes becomes even more significant, if the response to a treatment is further spread as in the following example with ten patients. Grouping the patients into optimized classes results in more homogeneous classes, that are additionally heterogeneous to each other, compared with a classification into two, three, or four quantiles. The example is displayed in Table 4.2.

patient	1	2	3	4	5	6	7	8	9	10
age	30	31	32	33	34	35	36	37	38	39
response	0	0	1	1	1	0	0	0	0	1
2 quantiles	0.60					0.20				
3 quantiles	0.50			0.33			0.33			
4 quantiles	0.33			1.00		0.00		0.33		
optimized classes	0.00		1.00			0.00			1.00	

Table 4.2: Quantiles and optimized classes for a patient characteristic (2)

This classification can be determined for any patient characteristic by the following adaption of the bounded-shape partition problem (4.1).

Let X be the patient characteristic, which has to be categorized into $k \in \mathbb{N}$ classes. Without loss of generality we assume the patient data set to be sorted according to the values of patient characteristic X.

$$C = (C_1, \ldots, C_k)$$

is a partition of the patient data set into k classes. C has the same properties as a clustering Cl, as it describes the special case with only one baseline characteristic. We give it a special notation in order to distinguish it from the original clustering approach. However, due to similar properties, the number of patients κ_i in every class C_i can also be bounded by lower and upper bounds l_i and u_i. In addition to the decision variables ξ_{ij} indicating, whether class C_i contains patient j, we need to introduce the following variables indicating the number of the class, to which patient j is assigned to.

$$c_j = i$$

indicates, that patient j is assigned to class C_i. Using these variables we can make sure, that the allocation is cohesive and ascending by adding the respective constraints to (4.1). Finally, we adapt the target function $f(\xi)$, such that it minimizes the

variance inside each class and maximizes the variance between all classes. Therefore, patient characteristic X can be categorized into k optimized classes by solving the following integer linear program:

$$
\begin{aligned}
&\max f(\xi) && \\
&\sum_{j=1}^{n} \xi_{ij} \leq u_i && \forall i = 1,\ldots,k \\
&\sum_{j=1}^{n} \xi_{ij} \geq l_i && \forall i = 1,\ldots,k \\
&\sum_{i=1}^{k} \xi_{ij} = 1 && \forall j = 1,\ldots,n \\
&c_j \leq c_{j+1} && \forall j = 1,\ldots,n-1 \\
&\xi_{ij} \in \{0,1\} && \forall i = 1,\ldots,k \quad \forall j = 1,\ldots,n \\
&c_j \in \{1,\ldots,k\} && \forall j = 1,\ldots,n
\end{aligned}
\tag{4.3}
$$

In Part IV we will classify patient characteristics of ordinal or cardinal into optimized classes instead of regular quantiles.

5. Cluster-based analysis of individual patient data

The geometric clustering approach discussed in the previous chapter identifies clusters of patients, that are as homogeneous as possible. The clusters are not only homogeneous inside, but also to differ from each other. Otherwise the division into clusters and especially different outcomes across those clusters cannot be justified. Furthermore, they are uniquely defined by baseline characteristics.
In this chapter we will discuss an explainable method to reliably predict a patient's outcome based on the clusters resulting from the geometric clustering approach. Besides serving as method itself, it is used in the newly invented cluster-based survival analysis introduced in Chapter 6 to form homogeneous clusters of patients. Due to the clusters being uniquely defined by baseline characteristics, the method does not only predict the outcome, but provides an explanation for it. We begin by defining how to determine the value of a cluster for an arbitrary outcome. We will also derive special formulations for a binary outcome. Based on the cluster value we will present confidence intervals for its prediction. Thereby we present both approximate and exact intervals. Furthermore we will discuss the usage of exact intervals as well as the usage of confidence intervals opposed to hypothesis testing in Section 5.3.2. We will also argue, why we refrain from using hypothesis tests.
Most of the derivations in this chapter are based on the work of Brieden and Gritzmann [26]. Furthermore, we will make use of several estimations, which can be found in [8].

5.1 Clustering

In order to perform a cluster-based analysis, we begin by deriving the clustering. After transforming the patient data G into $\widehat{G}$, as described in Section 4.2, selecting the input variables with possible classification into quantiles as described in Section 4.3, deciding on the number of clusters k, and applying the geometric clustering algorithm as referred to in Section 4.1 we get a (k,l,u)-clustering

$$Cl = (Cl_1, \ldots, Cl_k)$$

as partition of the transformed patient data $\widehat{G}$. Thereby each cluster represents a collective of patients with similar combinations of their patient characteristic values.

5.2 Cluster-based statistical evaluation

We begin by describing the determination of the cluster value before presenting confidence intervals for it.

5.2.1 Determination of cluster value

The main assumption is that patients belonging to different clusters differ regarding their outcome. In the following we will define the cluster value in order to justify this assumption. We will use the unbiased estimator for the expected value of the outcome of patients as cluster value. First, we define the patient data set of cluster Cl_i.

Definition 5.1 (patient data set of cluster Cl_i)**.** *Let $Cl = (Cl_1, \ldots, Cl_k)$ be a (k,l,u)-clustering of the transformed patient data $\widehat{G} = \left\{ \left(\widehat{x}_j, y_j \right) \right\}_{j=1}^{n}$ and $\xi = \left(\xi_{ij} \right) \in \{0,1\}^{k \times n}$ the corresponding cluster assignment of Cl. Then*

$$\widehat{G}_i := \left\{ \left(\widehat{x}_j, y_j \right) | \xi_{ij} = 1 \right\}_{j=1}^{n} \subseteq \widehat{G}$$

is the patient data set of cluster Cl_i with size κ_i. The corresponding patient data with the non-transformed data is denoted by G_i, for $i = 1, \ldots, k$.

Arbitrary outcome

The aim is to use the realization of the unbiased estimator for the expected value of the patients' outcome as cluster value. Thus we assume there is a true outcome denoted by random variable Y_i in cluster Cl_i following some distribution. Note, that the distribution does not necessarily have to be known. The true (unknown) expected value of Y_i is denoted by y_i with variance σ_i^2, i.e.

$$E(Y_i) = y_i \qquad\qquad V(Y_i) = \sigma_i^2$$

for $i = 1,\ldots,k$. Therefore the outcome of each patient j in cluster Cl_i can be described as random variable Y_{ij} following the very same distribution with realization y_{ij}, for $j = 1,\ldots,\kappa_i$, with the Y_{ij}'s being independent from each other. The (true) expected value of Y_{ij} is also y_i and the corresponding variance σ_i^2. With the upper notation the unknown expected outcome of patients in cluster Cl_i can be estimated via the following theorem.

Theorem 5.2. *Let $Cl = (Cl_1,\ldots,Cl_k)$ be a (k,l,u)-clustering of the transformed patient data $\widehat{G}$ and let $\widehat{G}_i := \left\{\left(\widehat{x}_j, y_j\right)\right\}_{j=1}^{\kappa_i}$ be the transformed patient data set of cluster Cl_i, for $i = 1,\ldots,k$. Then*

$$\widehat{Y}_i := \frac{1}{\kappa_i}\sum_{j=1}^{\kappa_i} Y_{ij}$$

is an unbiased estimator for the expected value of the patients' outcome with corresponding estimation

$$\widehat{y}_i := \frac{1}{\kappa_i}\sum_{j=1}^{\kappa_i} y_{ij}$$

in cluster Cl_i for $i = 1,\ldots,k$. The estimation is used as cluster value, so

$$f(Cl_i) = \widehat{y}_i$$

for cluster Cl_i, $i = 1,\ldots,k$.

Furthermore, the unknown variance σ_i^2 of the patients' outcome in cluster Cl_i can be estimated via the following theorem.

Theorem 5.3. *Let $Cl = (Cl_1, \ldots, Cl_k)$ be a (k,l,u)-clustering of the transformed patient data $\widehat{G}$ and let $\widehat{G}_i := \left\{ \left(\widehat{x}_j, y_j \right) \right\}_{j=1}^{\kappa_i}$ be the transformed patient data set of cluster Cl_i, for $i = 1, \ldots, k$. Then*

$$\widehat{\Sigma}_i^2 := \frac{1}{\kappa_i - 1} \sum_{j=1}^{\kappa_i} \left(Y_{ij} - \widehat{Y}_i \right)^2$$

is an unbiased estimator for the variance of the patients' outcome with corresponding estimation

$$\widehat{\sigma}_i^2 := \frac{1}{\kappa_i - 1} \sum_{j=1}^{\kappa_i} \left(y_{ij} - \widehat{y}_i \right)^2$$

in cluster Cl_i, for $i = 1, \ldots, k$.

The confidence intervals presented in Section 5.2.2 use both the estimator for the expected value and the estimator for the standard deviation, offering a possibility to assess the homogeneity within each cluster. Before constructing these confidence intervals, we will derive implications for a binary outcome.

Binary outcome

In many applications the investigated patients' outcome is binary. For example reaching an event of interest can directly be modelled as binary outcome, as the event can either be reached or not. Typical examples for events of interest include the response to a medication or the continuance of a treatment. Due to the estimators with corresponding estimations simplifying for a binary outcome, we will therefore present the respective representations.

We assume there is a true binary outcome in cluster Cl_i denoted by random variable $Y_i \sim Ber(p_i)$ with $p_i = P(Y_i = 1)$. The true (unknown) expected value of Y_i is

denoted by y_i with variance ${\sigma_i}^2$, i.e.

$$E(Y_i) = y_i \qquad\qquad V(Y_i) = {\sigma_i}^2$$

for $i = 1,\ldots,k$. Therefore the outcome of each patient j in cluster Cl_i can be described as random variable $Y_{ij} \sim Ber(p_i)$ with realization y_{ij}, for $j = 1,\ldots,\kappa_i$, with the Y_{ij}'s being independent from each other. The (true) expected value of Y_{ij} is also y_i and the corresponding variance ${\sigma_i}^2$.
For a Bernoulli distributed random variable Y_i it holds:

$$E(Y_i) = p_i \qquad\qquad \text{and} \qquad\qquad V(Y_i) = p_i \cdot (1 - p_i)$$

Therefore by estimating the probability of success of patients in cluster Cl_i, we estimate the expected value of the outcome of a patient in cluster Cl_i. The estimator and corresponding estimation are provided by the following corollary.

Corollary 5.4. *Let $Cl = (Cl_1,\ldots,Cl_k)$ be a (k,l,u)-clustering of the transformed patient data $\widehat{G}$ and let $\widehat{G}_i := \left\{\left(\widehat{x}_j, y_j\right)\right\}_{j=1}^{\kappa_i}$ be the transformed patient data set of cluster Cl_i with binary outcome, for $i = 1,\ldots,k$. Then*

$$\widehat{P}_i := \frac{1}{\kappa_i} \sum_{j=1}^{\kappa_i} 1_{\{Y_{ij}=1\}}$$

is an unbiased estimator for the probability of success of the patients' outcome with corresponding estimation

$$\widehat{p}_i := \frac{1}{\kappa_i} \sum_{j=1}^{\kappa_i} 1_{\{y_{ij}=1\}}$$

in cluster Cl_i for $i = 1,\ldots,k$. The estimation is used as cluster value, so

$$f(Cl_i) = \widehat{p}_i$$

for cluster Cl_i for $i = 1,\ldots,k$.

The unknown variance can be estimated by the following corollary.

Corollary 5.5. *Let $Cl = (Cl_1, \ldots, Cl_k)$ be a (k,l,u)-clustering of the transformed patient data $\widehat{G}$ and let $\widehat{G}_i := \left\{ \left(\widehat{x}_j, y_j \right) \right\}_{j=1}^{\kappa_i}$ be the transformed patient data set of cluster Cl_i with binary outcome for $i = 1, \ldots, k$. Then*

$$\widehat{\Sigma}_i^2 := \widehat{P}_i \cdot \left(1 - \widehat{P}_i \right)$$

is an unbiased estimator for the variance of the patients' outcome with corresponding estimation

$$\widehat{\sigma}_i^2 := \widehat{p}_i \cdot (1 - \widehat{p}_i)$$

in cluster Cl_i, for $i = 1, \ldots, k$.

5.2.2 Confidence intervals for the cluster value

Using these estimations we can construct $(1-\alpha)$-confidence intervals $\mathtt{I}^{\alpha}(\widehat{y}_i)$ for the expected outcome y_i of patients in cluster Cl_i, $i = 1, \ldots, k$. The intervals provide a prediction for the expected outcome of patients with those baseline characteristics defined by the respective cluster. Furthermore, they provide a measure for the homogeneity within each cluster. Narrow confidence intervals point to a homogeneous cluster, whereas wide confidence intervals point to a rather heterogeneous cluster. However, as we can see in the following, the size of the underlying patient data set influences the range of the confidence interval. Therefore the intervals sometimes do not seem to be narrow in particular due to small data sets. They still provide valuable information about the expected outcome of a patient, especially while comparing different clusters of patients with each other leading to the third application of them, which we will discuss in Section 5.3.

We will begin by describing a confidence interval for an arbitrary outcome, before deriving special formulations for a binary outcome. Furthermore we will discuss the use and advantage of exact confidence intervals.

Arbitrary outcome

Due to the central limit theorem, for $\kappa_i > 30$ the estimator $\widehat{Y}_i$ is approximately normally distributed with expected value

$$E\left(\widehat{Y}_i\right) = y_i$$

and standard deviation

$$\sigma_{\widehat{Y}_i} = \frac{\widehat{\sigma}_i}{\sqrt{\kappa_i}}$$

for cluster Cl_i, $i = 1, \ldots, k$. Therefore

$$\frac{\widehat{Y}_i - y_i}{\sigma_{\widehat{Y}_i}} \overset{approx}{\sim} \mathcal{N}(0,1)$$

and thus

$$\mathtt{I}^{\alpha}_{approx}(y_i) \coloneqq \left[\widehat{y}_i - z_{1-\frac{\alpha}{2}} \cdot \sigma_{\widehat{Y}_i}, \widehat{y}_i + z_{1-\frac{\alpha}{2}} \cdot \sigma_{\widehat{Y}_i}\right]$$

is an approximate $(1-\alpha)$-confidence interval for the expected outcome of patients in cluster Cl_i, for $i = 1, \ldots, k$, where $z_{1-\frac{\alpha}{2}}$ denotes the $\left(1-\frac{\alpha}{2}\right)$-percentile of the standard normal distribution. It holds that

$$P\left(y_i \in \mathtt{I}^{\alpha}_{approx}(y_i)\right) = 1 - \alpha$$

and therefore in $(1-\alpha) \cdot 100\%$ of all cases the true expected value of the patients' outcome y_i lies in $\mathtt{I}^{\alpha}_{approx}(y_i)$ for cluster Cl_i, $i = 1, \ldots, k$.

The confidence interval $\mathtt{I}^{\alpha}_{approx}(y_i)$ is suitable for an arbitrary outcome, as long as the cardinality in the respective cluster is high enough ($\kappa_i > 30$). However, if the cardinality is too low, we can use the following confidence interval.

Remark 5.6. *For an arbitrary outcome Y_i with expected mean value y_i and variance ${\sigma_i}^2$ it holds that*

$$\frac{\widehat{Y}_i - y_i}{\sigma_{\widehat{Y}_i}} \overset{approx}{\sim} t_{\kappa_i - 1}$$

with $\sigma_{\widehat{Y}_i} = \frac{\sigma_i}{\sqrt{\kappa_i}}$ and thus

$$\mathtt{I}^{\alpha}_{approx}(y_i) := \left[\widehat{y}_i - t_{\kappa_i-1}\left(1-\frac{\alpha}{2}\right)\cdot\sigma_{\widehat{Y}_i}, \widehat{y}_i + t_{\kappa_i-1}\left(1-\frac{\alpha}{2}\right)\cdot\sigma_{\widehat{Y}_i}\right]$$

is an approximate $(1-\alpha)$-confidence interval, even for small κ_i for the expected outcome of patients in cluster Cl_i $i = 1,\ldots,k$, where $t_{\kappa_i-1}\left(1-\frac{\alpha}{2}\right)$ denotes the $\left(1-\frac{\alpha}{2}\right)$-percentile of the Student's t-distribution with $\kappa_i - 1$ degrees of freedom.

In the following we will state the normal approximation confidence interval for a binary outcome, before discussing an exact confidence interval. Afterwards we will discuss the benefit of using this exact interval, even if it is a rather conservative one.

Binary outcome

Based on the confidence interval for an arbitrary outcome we present a similar normal approximation confidence interval for a binary outcome. The binary estimator $\widehat{P}_i$ has the expected value

$$E\left(\widehat{P}_i\right) = p_i$$

with standard deviation

$$\sigma_{\widehat{P}_i} = \frac{\widehat{\sigma}_i}{\sqrt{\kappa_i}} = \frac{\widehat{p}_i \cdot (1-\widehat{p}_i)}{\sqrt{\kappa_i}}$$

for cluster Cl_i, $i = 1,\ldots,k$. Therefore,

$$\mathtt{I}^{\alpha}_{approx}(p_i) = \left[\widehat{p}_i - z_{1-\frac{\alpha}{2}}\cdot\sigma_{\widehat{P}_i}, \widehat{p}_i + z_{1-\frac{\alpha}{2}}\cdot\sigma_{\widehat{P}_i}\right]$$

is an approximate $(1-\alpha)$-confidence interval for the expected outcome of patients in cluster Cl_i for $i = 1,\ldots,k$, where $z_{1-\frac{\alpha}{2}}$ denotes the $\left(1-\frac{\alpha}{2}\right)$-quantile of the standard

normal distribution. It holds that

$$P\left(p_i \in \mathtt{I}^{\alpha}_{approx}(p_i)\right) = 1-\alpha$$

and therefore in $(1-\alpha)\cdot 100\%$ of all cases the true expected value of the patients' outcome p_i lies in $\mathtt{I}^{\alpha}_{approx}(p_i)$ for cluster Cl_i, $i = 1,\ldots,k$.

In the following we want to examine an exact confidence interval, namely the Clopper-Pearson confidence interval. The true coverage of this interval does not fall below $(1-\alpha)$ [140]. Therefore the interval is usually wider than it has to be, making it a conservative choice. After introducing the interval, we will discuss why a conservative choice like this is appropriate, especially while evaluating medical interventions.

Let Z_i be a random variable describing the number of positive events in cluster Cl_i:

$$Z_i = \sum_{i=1}^{\kappa_i} Y_{ij}$$

The outcome of every patient Y_{ij} is binary and therefore Bernoulli distributed. Hence, Z_i follows a binomial distribution:

$$Z_i \sim Bin(\kappa_i, p)$$

Furthermore, let z_i be the number of positive events observed in cluster Cl_i:

$$z_i = \sum_{i=1}^{\kappa_i} y_{ij}$$

Finally, let $p_{i,lower}$ be the solution in p to $P_p(Z_i \geq z_i) = \frac{\alpha}{2}$ and $p_{i,upper}$ be the solution in p to $P_p(Z_i \leq z_i) = \frac{\alpha}{2}$. The Clopper-Pearson confidence interval for the expected outcome of patients in cluster Cl_i for $i = 1,\ldots,k$ is defined as:

$$\mathtt{I}_{CP}^{\alpha}(p_i) = \left[p_{i,lower}, p_{i,upper}\right]$$

Due to the connection between the binomial distribution and beta distribution $p_{i,lower}$ and $p_{i,upper}$ can also be obtained via the percentiles of the latter [140].

Remark 5.7. *Let $\beta(\alpha, p, q)$ describe the α-percentile of a beta distribution with parameters p and q. Then the Clopper-Pearson confidence interval $\mathtt{I}_{CP}^{\alpha}(p_i)$ can be obtained by*

$$p_{i,lower} = \beta\left(\frac{\alpha}{2}, z_i, \kappa_i - z_i + 1\right)$$

and

$$p_{i,upper} = \beta\left(1 - \frac{\alpha}{2}, z_i + 1, \kappa_i - z_i\right)$$

for cluster Cl_i, $i = 1, \ldots, k$.

Remark 5.8. *There is no closed-form expression for $p_{i,lower}$ and $p_{i,upper}$ for most choices of z_i. However, it can be suitably approximated using statistical software solutions.*

It holds that

$$P\left(p_i \in \mathtt{I}_{CP}^{\alpha}(p_i)\right) = 1 - \alpha$$

and therefore in $(1-\alpha)\cdot 100\%$ of all cases the true expected value of the patients' outcome p_i lies in $\mathtt{I}_{CP}^{\alpha}(p_i)$ for cluster Cl_i, $i = 1, \ldots, k$. Furthermore, the true coverage does not fall below $(1-\alpha)$.

This concludes the cluster-based analysis of individual patient data. The method helps generating new insight especially into heterogeneous patient data. It addresses the heterogeneity by dividing the patient data into homogeneous clusters. Based on them, it gives a prediction for the outcome of a patient. In addition to the prediction, it also provides a unique explanation based on baseline characteristics. Therefore, this method for explainable analytics can be used for well-founded clinical decision-making. Before introducing cluster-based survival analysis in the following chapter, we discuss how the division into clusters can be justified.

5.3 Justification of different outcome across clusters

5.3.1 Comparison of clusters

Confidence intervals as discussed in the previous section provide a prediction for the expected outcome of patients. By constructing a confidence interval for each cluster of patients, we get a prediction for patients with those baseline characteristics defined by the respective cluster. This is of vital importance, when comparing clusters with each other.

The division of the patient data into clusters is only justified, if the outcome differs across the clusters. Therefore all confidence intervals have to be considered simultaneously. If the confidence intervals of the clusters do not overlap for a reasonable choice of level of confidence α as illustrated in Figure 5.1, we can conclude that the collectives differ regarding their outcome.

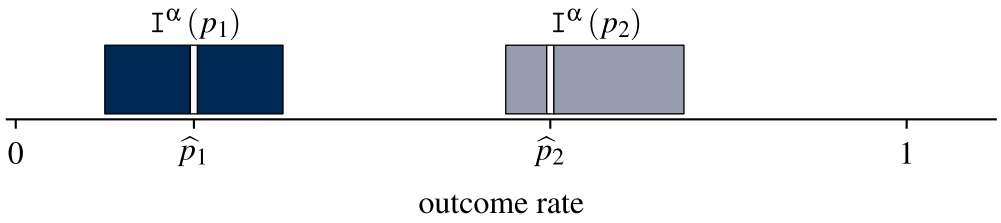


Figure 5.1: Two non-overlapping confidence intervals for a binary outcome

Every cluster is uniquely identified by a specific choice of patient characteristics (i.e. age, severity of illness) leading to a unique explanation for the respective prediction. This implies, that (new) patients can be uniquely assigned to one of the clusters yielding to a prediction of the patients' outcome. Thus, based on the patient's characteristics it is possible to evaluate whether a specific medical intervention is beneficial for the individual patient or not and there exists a unique explanation for it.

Remark 5.9. *While looking at real data sets, the intervals might not always be completely disjoint from each other. However, as in every statistical evaluation, the results have to be interpreted and set into context.*

Before introducing the newly invented cluster-based survival analysis, another explainable method for the analysis of patient individual data, we want to motivate the use of exact confidence intervals opposed to approximate intervals and the usage of confidence intervals opposed to hypothesis testing throughout this thesis.

5.3.2 Use of confidence intervals

Motivation behind the use of exact intervals

Confidence intervals always describe a range in which the parameter of interest (e.g. response rate) lies with a certain degree of confidence. However, the true coverage of those intervals might differ from the chosen level of confidence α, i.e. it might be smaller or greater than $(1-\alpha)100\%$. While using approximations as intervals, like the normal approximation interval, the true coverage might both lie above or under $(1-\alpha)100\%$. When using an exact interval, like the Clopper-Pearson interval, the true coverage is $(1-\alpha)100\%$ or even more. Hence, exact intervals might be too wide but never too narrow, whereas approximation intervals can be too narrow.
As mentioned, the exact intervals might be too wide. As a matter of fact they usually are [36]. However, even if a narrow confidence interval might seem tempting, it can strongly mislead in those cases, where their true coverage was overestimated. Taking it even further, wrong conclusions might be drawn on non reliable confidence intervals. Therefore, we decided to use exact confidence intervals whenever there is a possiblity for that. By using exact intervals, we can make sure of the results being reliable, which can then be used for further well-founded decision-making. Even more, if we can observe differences between clusters of patients while considering conservative exact intervals, the true difference might be even greater.
Many people have studied intervals for this so called binomial proportion. An overview is provided in [113]. The Clopper-Pearson interval was one of the earlier introduced exact intervals, but there exist other suggestions [36]. Even though other exact intervals exist, some possibly being more narrow, Clopper-Pearson intervals are an appropriate choice both because of their exactness and their simplicity in interpreting the results.

Before moving on to the next chapter, we want to motivate the use of confidence intervals opposed to often encountered hypothesis testing.

Motivation behind using confidence intervals instead of hypothesis testing

At this point of a statistical evaluation it is very common to test some hypotheses. In this thesis we will refrain from doing so, due to two reasons:

1. While testing hypotheses it is crucial, that those hypotheses are formulated a priori and not within the method itself. If we were to pose hypotheses, they would concern the difference in the outcome (e.g. response rates) between at least two clusters of patients. Since the specific patient characteristics defining those clusters are part of the results of our method, we can obviously not formulate the hypotheses a priori but only after conducting the clustering.

2. It has been widely discussed that the appropriate use of confidence intervals offers the same information as the testing of some hypotheses without committing any methodological mistakes as mentioned above [18] [23] [150]. Especially while keeping in mind, that parametric test statistics include the very same components as the respective confidence intervals. Moreover, the (parametric) testing of a hypothesis results in either accepting or declining the hypothesis. So the outcome of the test is (loosely speaking) a binary one. The hypothesis itself is stated for a specific parameter of interest, e.g. whether the true response rate of a cluster is greater or less than the observed response rate in the data. Compared to this, confidence intervals do not aim to estimate just the parameter itself, but offer a range in which the parameter lies with a certain confidence. Having a range for the parameter of interest (i.e. response rate), instead of just a specific estimate is much more realistic and reliable especially while drawing further conclusions from it, e.g. for the further treatment of patients.

The comparison of different clusters of patients is still possible while using confidence intervals, as described above. That is why we use confidence intervals

throughout this thesis to evaluate the outcome and especially the difference in the outcome across clusters of patients.
For further details on the usage of confidence intervals instead of hypothesis testing, please refer to [50] [62] [57] [131] [114].

6. Cluster-based survival analysis of individual patient data

In this chapter we introduce a new method for explainable analytics in medicine, the cluster-based survival analysis. It provides explainable predictions for the survival of a patient while simultaneously addressing underlying heterogeneity inside the patient data. Moreover it generates new findings by detecting subgroups, i.e. clusters, of patients, for whom a treatment or medical intervention is more beneficial both compared with other clusters of patients and other treatments. The method can be carried out in an automated way, enabling the opportunity to apply it to other data sets as well as other applications.

We will begin by describing how to derive the clusters of patients, on which the survival analysis will be performed. Afterwards we will move on to the specific cluster-based survival function and its estimation. We will present several possibilities to analyze it. Furthermore, we will introduce cluster-based survival models, with the cluster-based Cox proportional-hazards model as representative. We will discuss different alternatives to analyze it and extend it to the cluster-based stratified Cox model. Finally we will discuss how to justify the assumption of different survival times and hazard rates across clusters and treatments, before moving on to the practical application in the next part of this thesis.

In this thesis we will use the cluster-based survival analysis to compare different kinds of treatment or medication with each other. The main assumption is, that the time to discontinuation of a treatment is not only influenced by the treatment, but also by patient characteristics. Naturally, the method can be applied to compare other

groups with each other, as long as there is a unique assignment of patients or data points in general into one of the groups. For the sake of clarity and without loss of generality we will from now on speak of treatments, which we want to compare with each other.

6.1 Clustering

Survival analyses are conducted on two kinds of information about the outcome of a patient. The time until the event of interest occurs Y_j with characteristics $y_j > 0$ are recorded together with the information, whether patient j discontinued the treatment D_j with characteristics

$$d_j = \begin{cases} 0 & \text{event does not occur in observed period} \\ 1 & \text{event occurs in observed period} \end{cases}$$

for $j = 1, \ldots, n$. Patients with no event in the observed period ($d_j = 0$) do obviously not have a time until the event of interest occurs (see censoring in Chapter 3). Therefore we need a transformed outcome Z_j for the clustering. Z_j denotes the amount of incomplete time in the observed period. Patients who did not experience the event in the observed time period have no incomplete time. Patients who did experience the event missed out on $y^* - y_j$ of further time (e.g. days) with y^* being the time horizon of the observed period. Therefore the transformed outcome of incomplete time Z_j with characteristics

$$z_j = d_j \cdot (y^* - y_j) = \begin{cases} 0 & \text{if } d_j = 0 \\ y^* - y_j & \text{if } d_j = 1 \end{cases}$$

for patient $j = 1, \ldots, n$ is used for the following clustering.
Y_j and therefore also Z_j can both follow a continuous or a discrete distribution, depending on the unit in which the time until the event happens is measured. For the sake of clarity and without loss of generality, the following derivations assume Y_j and Z_j to follow a continuous distribution.

Remark 6.1. *Even though we need to transform the outcome for deriving the clustering, the original information about the outcome is used in the subsequent survival analyses on each cluster.*

In many survival analyses the primary investigation is the difference between the administered treatments. One possibility is to include the covariate describing the treatment as input variable for the clustering. However, as we explicitly want to be able to distinguish between patients receiving different treatments, we will refrain from doing so. Instead, we split the patient data according to the administered treatment prior to the subsequent analyses. Thereby we aim to identify clusters of patients, for whom a specific treatment is more beneficial than for other patients receiving the same treatment. Afterwards we will then analyze the outcome of patients with similar baseline characteristics, who received another treatment. The procedure is described in the following.

Let T be the set of treatments we want to compare with each other. Furthermore, let one of the patient characteristics included in the patient data X denote the administered treatment. In the following we assume, that the patient characteristic describing the treatment is available and unique for every patient.

Definition 6.2 (treatment patient data)**.** *Let X be the patient data set of characteristics and Z the corresponding transformed outcome with sampling space Ω. Then the treatment patient data set of treatment $t \in T$ is defined by*

$$G^t := \left\{(x_j, z_j) | x_j^{treatment} = t\right\}_{j=1}^{n} \subset \mathbb{R}^p \times \Omega$$

with $x_j \in X$ and $z_j \in \Omega$.

In the following we will only use one of the treatments in T to derive the clustering and especially the cluster-defining baseline characteristics. Based on that, we will assign patients from the other treatments into clusters in a similar way. In order to distinguish, which treatment was used for deriving the clustering, we need the following definition.

Definition 6.3 (cluster-treatment)**.** *We call treatment $t^* \in T$, which is used for the clustering, the cluster-treatment.*

Next, we need to define a (k,l,u,t)-clustering.

Definition 6.4 ((k,l,u,t)-clustering)**.** *A (k,l,u)-clustering on patients receiving treatment $t \in T$ is called (k,l,u,t)-clustering and denoted by*

$$Cl^t = \left(Cl_1^t, \ldots, Cl_k^t\right)$$

Let $t^* \in T$ be the cluster-treatment and G^{t^*} the corresponding treatment patient data. G^{t^*} is transformed into $\widehat{G}^{t^*}$ as described in Section 4.2. Based on that, the patient characteristics are selected as described in Section 4.3. A prior classification into quantiles might be necessary. By applying the geometric clustering algorithm as referred to in Section 4.1, we then get a (k,l,u,t^*)-clustering:

$$Cl^{t^*} = \left(Cl_1^{t^*}, \ldots, Cl_k^{t^*}\right)$$

Cl^{t^*} provides a unique assignment of patients receiving cluster-treatment t^* into one of the clusters solely based on their patient characteristics, formalized in the following definition. This step is crucial for the entire method being explainable.

Definition 6.5 (cluster defining characteristics)**.** *Let $t^* \in T$ be the cluster-treatment and $Cl^{t^*} = \left(Cl_1^{t^*}, \ldots, Cl_k^{t^*}\right)$ the corresponding (k,l,u,t^*)-clustering. Then the mapping*

$$\Phi_{t^*i} : \left(X^1, \ldots, X^p\right) \to \left(x^1, \ldots, x^p\right)$$

denotes the unique assignment based on the patients' characteristics $X^1, \ldots, X^p$ into cluster $Cl_{t^*i}^{t^*}$ for $i = 1, \ldots, k$, with $x^j \subseteq \mathbb{R}$ denoting the respective range of values for patient characteristic X^j in cluster $Cl_i^{t^*}$.

Those patient characteristics $X^1, \ldots, X^p$ are also available for patients receiving a different treatment. Therefore all other patients can be assigned into clusters by the very same assignment rules. For any treatment $t \in T$, not necessarily the

cluster-treatment, we can determine the corresponding (k,l,u,t)-clustering:

$$Cl^t_{t^*} = \left(Cl^t_{t^*1}, \ldots, Cl^t_{t^*k}\right)$$

with

$$Cl^t_{t^*i} = \left\{x_j \in X | x_j^{treatment} = t, x_j \text{ fulfils } \Phi_{t^*i}\right\}$$

We still want to indicate, that the clustering itself has not necessarily been derived on treatment t. Therefore we add another index for the cluster-treatment t^*, on which the clustering has been derived on.

By adding cluster-treatment t^* as additional index, the notations might seem overloaded. However, we still want to be able to tell, which of the treatments was used as cluster-treatment. Furthermore, once the actual names or abbreviations are used instead of t^* and t, the notation gets more clear, as we will see in Part IV.

Remark 6.6. *Let $t^* \in T$ be the cluster-treatment and $Cl^t_{t^*}$ be a (k,l,u,t)-clustering for every $t \in T$. Then*

$$Cl_{t^*} = (Cl_{t^*1}, \ldots, Cl_{t^*k})$$

with

$$Cl_{t^*i} = \bigcup_{t \in T} Cl^t_{t^*i}$$

is a (k,l,u)-clustering.

Procedure 6.1 summarizes the division of the patient data into homogeneous clusters, based on the cluster-treatment and the assignment of all other patients into clusters.

The remaining question is, which of the available treatments to use as cluster-treatment. We will discuss three different possibilities for that, the manual choosing, the automatic choosing, and the repeated choosing:

input : all input variables $X^1,\ldots,X^p$, cluster-treatment t^*
output : (k,l,u,t)-clustering for every treatment $t \in T$

1. transform G^{t^*} into $\widehat{G}^{t^*}$
2. select input variables for clustering using Procedure 4.1
3. derive (k,l,u,t^*)-clustering to get unique assignment rules Φ_{t^*i} for every cluster $Cl^t_{t^*i}$ $i = 1,\ldots,k$
4. **for** $t \in T$ **do**
   derive (k,l,u,t)-clustering

$$Cl^t_{t^*} = \left(Cl^t_{t^*1},\ldots,Cl^t_{t^*k}\right)$$

   with

$$Cl^t_{t^*i} = \left\{x_j \in X | x_j^{treatment} = t, x_j \text{ fulfils } \Phi_{t^*i}\right\}$$

   end

Procedure 6.1: Clustering for cluster-based survival analysis

1. **Manual choosing**

   The cluster-treatment might be clear from the application, e.g. the trial from which the data originates. There might also be some other specifications implying the cluster-treatment, e.g. choosing a 'reference' treatment. In this case, the cluster-treatment is chosen manually.

2. **Automatic choosing**

   In this case no treatment can be defined manually. One option is to derive a clustering (separately) for each treatment. Afterwards the difference between the derived clusters regarding the desired outcome (e.g. survival time) is evaluated, for example by looking at the variance of cluster values. The treatment providing the clustering with the largest variance is then used as cluster-treatment in the cluster-based survival analysis.

3. **Repeated choosing**

   Another option is to perform repeated cluster-based survival analyses, once using every treatment as cluster-treatment.

Since repeating the cluster-based survival analysis with each treatment as cluster-treatment does not differ methodologically from choosing the cluster-treatment, we will assume the cluster-treatment to be fixed in the following. It does not matter if the decision was made automatically or set by the application.

Next, we want to define the cluster-based survival function and introduce an estimation for it, related to the Kaplan-Meier estimate in a common survival analysis.

6.2 Cluster-based survival function

In order to perform a cluster-based survival analysis, we need to define the cluster-based survival function.

Definition 6.7 (cluster-based survival function)**.** *Let Y be a random variable denoting the time until the event of interest happens, $t^* \in T$ the cluster-treatment and $t \in T$ an arbitrary treatment. Furthermore, let $Cl^t_{t^*} = \left(Cl^t_{t^*1}, \ldots, Cl^t_{t^*k}\right)$ be a (k,l,u,t)-clustering. Then*

$$S^t_{t^*i}(y) \coloneqq P\left(Y > y \,|\, \textit{patient belongs to cluster } Cl^t_{t^*i}\right)$$

is called the cluster-based survival function for patients in cluster $Cl^t_{t^*i}$.

Note, that in the definition above it does not matter whether treatment $t \in T$ was the chosen cluster-treatment t^*.

The true cluster-based survival function is unknown for every treatment. Therefore we will use the Kaplan-Meier method to estimate the cluster-based survival function. Let $\widehat{S}^t_{t^*i}(y)$ denote the estimation of the true (unknown) cluster-based survival function $S^t_{t^*i}(y)$, i.e. $\widehat{S}^t_{t^*i}(y)$ estimates the probability of patients in cluster $Cl^t_{t^*i}$ surviving longer than y.

Lemma 6.8. *The cluster-based survival function can be estimated by*

$$\widehat{S}^t_{t^*i}(y) = \prod_{y_{(f)} \leq y} \frac{n_{(f)t^*i} - d_{(f)t^*i}}{n_{(f)t^*i}}$$

with $n_{(f)t^*i}$ being the number at risk at $y_{(f)}$ in cluster $Cl^t_{t^*i}$ and $d_{(f)t^*i}$ being the number of observed events at $y_{(f)}$ in cluster $Cl^t_{t^*i}$. The number of patients at risk at time $y_{(f)}$ in cluster $Cl^t_{t^*i}$ can be expressed by

$$n_{(f)t^*i} = \sum_{j=1}^{n} 1_{\{y_j \geq y_{(f)}\}} \cdot 1_{\{\textit{patient } j \textit{ in cluster } Cl^t_{t^*i}\}}$$

and the number of of patients with an event at exactly $y_{(f)}$ in cluster $Cl^t_{t^*i}$ can be expressed by

$$d_{(f)t^*i} = \sum_{j=1}^{n} 1_{\{y_j = y_{(f)}\}} \cdot 1_{\{d_j = 1\}} \cdot 1_{\{\textit{patient } j \textit{ in cluster } Cl^t_{t^*i}\}}$$

Proof. The estimation corresponds to a Kaplan-Meier estimation restricted to patients in cluster $Cl^t_{t^*i}$. □

Therefore, $\widehat{S}^t_{t^*i}(y)$ measures the fraction of patients in cluster $Cl^t_{t^*i}$ surviving at least until y. According to Lemma 6.8 it has to be re-estimated if an event occurs ($n_{(f)t^*i}$ and $d_{(f)t^*i}$ change) or the set at risk changes due to censoring ($n_{(f)t^*i}$ changes).

The variance $\widehat{\sigma}^2_{\widehat{S}^t_{t^*i}(y)}$ of the estimate $\widehat{S}^t_{t^*i}(y)$ can be estimated by Greenwood's formula [61]:

Lemma 6.9. *Let $\widehat{S}^t_{t^*i}(y)$ be the estimator for the cluster-based survival function of cluster $Cl^t_{t^*i}$, $i = 1, \ldots, k$, $t \in T$. The variance is given by Greenwood's formula [61]:*

$$\widehat{\sigma}^2_{\widehat{S}^t_{t^*i}(y)} = \left(\widehat{S}^t_{t^*}(y)\right)^2 \cdot \sum_{y_{(f)} \leq y} \frac{d_{(f)t^*i}}{n_{(f)t^*i}\left(n_{(f)t^*i} - d_{(f)t^*i}\right)}$$

6.3 Cluster-based analysis of survival function

In this section we want to discuss how to analyze and especially compare the estimated cluster-based survival functions. Besides a confidence interval for the estimator, we will take a closer look at the cluster-based median survival time and the respective confidence interval. Finally, we will introduce the cluster-based restricted mean survival time, as a possibility to quantify the entire estimated cluster-based survival function to compare it with others.
As mentioned in Chapter 3 and discussed in Section 5.3.2, we will not be using log-rank tests, due to us not being able to pose the hypotheses prior to the method.

6.3.1 Confidence interval for the cluster-based survival function

The Kaplan-Meier estimator $\widehat{S}^t_{t^*i}(y)$ is asymptotically normally distributed [97]. Using Lemma 6.9 we can therefore construct the following $(1-\alpha)$-confidence interval for the cluster-based survival function of cluster $Cl^t_{t^*i}$

$$\mathtt{I}^{\alpha}\left(S(y)^t_{t^*i}\right) := \left[\widehat{S}^t_{t^*i}(y) - z_{1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_{\widehat{S}^t_{t^*i}(y)}, \widehat{S}^t_{t^*i}(y) + z_{1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_{\widehat{S}^t_{t^*i}(y)}\right]$$

where $z_{1-\frac{\alpha}{2}}$ is the $\left(1-\frac{\alpha}{2}\right)$-percentile of the standard normal distribution.
The entire cluster-based survival function is estimated, hence for every $y > 0$ the true probability of patients in cluster $Cl^t_{t^*i}$ surviving longer than y is unknown. However, it holds that

$$P\left(S^t_{t^*i}(y) \in \mathtt{I}^{\alpha}\left(S(y)^t_{t^*i}\right)\right) = 1-\alpha$$

and therefore in $(1-\alpha)\cdot 100\%$ of all cases, the true probability of patients in cluster $Cl^t_{t^*i}$ surviving longer than y lies in $\mathtt{I}^{\alpha}\left(S(y)^t_{t^*i}\right)$.

6.3.2 Cluster-based median survival time

Similar to a common survival analysis, we will use the median survival time to compare clusters and treatments with each other. Therefore we need the following definition.

Definition 6.10 (cluster-based median survival time)**.** *Let $S^t_{t^*i}(y)$ be the survival function of treatment $t \in T$ for cluster $Cl^t_{t^*i}$ $i = 1,\dots,k$. If $S^t_{t^*i}(M^t_{t^*i}) = 0.5$, then $M^t_{t^*i}$ is called the cluster-based median survival time of $S^t_{t^*i}(y)$.*

$M^t_{t^*i}$ represents the time, after which half of all patients in cluster $Cl^t_{t^*i}$ discontinued their treatment. The true cluster-based median survival time is unknown, but it can be estimated in the following way:

Remark 6.11. *Let $\widehat{S}^t_{t^*i}(M^t_{t^*i})$ be the estimated cluster-based survival probability at the true cluster-based median survival time. We denote the estimated median survival time by $\widehat{M}^t_{t^*i}$. It can be estimated by the argument, for which $\widehat{S}^t_{t^*i}(y)$ falls below* 0.5 *for the first time.*

Using the distributional property from Lemma 3.12 from the common survival analysis we can construct the following $(1-\alpha)$-confidence interval for the cluster-based median survival time:

Lemma 6.12. *Let*

$$\left(\widehat{S}^t_{t^*i}(y) - 0.5\right)^2 < c_\alpha \cdot \widehat{\sigma}^2_{\widehat{S}^t_{t^*i}(M^t_{t^*i})} \tag{6.1}$$

where c_α denotes the $(1-\alpha)$-percentile of the χ^2_1-distribution. Then

$$\mathtt{I}^\alpha\left(M^t_{t^*i}\right) := \left[y^t_{t^*i,lower}, y^t_{t^*i,upper}\right]$$

is a $(1-\alpha)$-confidence interval for the median survival time, with $y^t_{t^*i,lower}$ being the minimum and $y^t_{t^*i,upper}$ being the maximum y fulfilling (6.1).

The cluster-based median survival time of cluster $Cl^t_{t^*i}$ has to be estimated, hence $M^t_{t^*i}$ such that $S^t_{t^*i}(M^t_{t^*i}) = 0.5$ is unknown. However, it holds that

$$P\left(M^t_{t^*i} \in \mathtt{I}^\alpha\left(M^t_{t^*i}\right)\right) = 1-\alpha$$

and therefore in $(1-\alpha)\cdot 100\%$ of all cases, the true median survival time of cluster $Cl^t_{t^*i}$ lies in $\mathtt{I}^\alpha(M^t_{t^*i})$.

6.3.3 Cluster-based restricted mean survival time

In the following, we want to introduce the cluster-based restricted mean survival time. It offers the possibility to compare the entire estimated cluster-based survival functions with each other.

Definition 6.13 (cluster-based restricted mean survival time)**.** *Let $S^t_{t^*i}(y)$ be the survival function of treatment $t \in T$ and let Y denote the survival time with limited time horizon y^* (e.g. end of the clinical trial). The cluster-based restricted mean survival time (RMST) is the mean of the survival time limited to y^* in cluster $Cl^t_{t^*i}$, i.e. $E(\min(Y, y^*))$. It equals the area under the cluster-based survival function from $y = 0$ to $y = y^*$, i.e.*

$$\widehat{RMST}^{\,t}_{t^*i} = E(\min(Y, y^*)) = \int_0^{y^*} S^t_{t^*i}(y)dy$$

for cluster $Cl^t_{t^*i}$ $i = 1, \ldots, k$.

Remark 6.14. *It holds that*

$$\widehat{RMST}^{\,t}_{t^*i} \in [0, y^*]$$

for cluster $Cl^t_{t^*i}$ $i = 1, \ldots, k$.

The cluster-based restricted mean survival time describes the average survival time of patients in cluster $Cl^t_{t^*i}$ from baseline to the limited time horizon y^*. It equals y^*, if no event occurs in this cluster until the end of the time horizon. The restricted mean survival time of a cluster converges towards zero, with the time interval also converging towards zero. This is the case, if all events in this cluster occur immediately after the starting point. Both extreme cases are visualized in Figure 3.2.

The cluster-based restricted mean survival time takes the entire estimated cluster-based survival function into account. Obviously the estimated cluster-based survival functions may not overlap for the cluster-based restricted mean survival time to be interpreted correctly.

In the following section we will introduce the cluster-based Cox proportional-hazards model. One main assumption of this model are proportional-hazards between the compared groups. This assumptions also leads to non overlapping survival functions, the above mentioned restriction.

6.4 Cluster-based survival models

Based on cluster-based survival functions, we will now introduce the newly invented cluster-based survival models enabling the possibility to take additional covariates into account. Cluster-based survival models can be used with any kind of survival model. In this thesis, we will use the Cox proportional-hazards model as representative, but the approach is not limited to it. If the true underlying model is known for certain (i.e. the survival times follow an exponential distribution), the respective parametric model can be used instead of estimating it. However, since this is often not the case, the cluster-based Cox PH model offers the opportunity to closely approximate the true model.

We will put special interest into the covariate describing the administered treatment, as we will use cluster-based survival models to compare different treatments with each other.

6.4.1 Cluster-based hazard rate

The cluster-based survival function $S^t_{t^*i}(y)$ describes the probability of not having the event of interest, given the survival up to y. In cluster-based Cox PH models, the measure of effect is the cluster-based hazard rate $h^t_{t^*i}(y)$, describing the probability for patients in the respective cluster of having an event at $y > 0$, given the survival up to y. We will use cluster-based survival models to compare two treatments $t, \tilde{t} \in T$ with each other. Therefore we need the following definition of a cluster-based hazard rate for two treatments.

Definition 6.15 (cluster-based hazard rate)**.** *Let Y describe survival times for an event of interest and $t^* \in T$ the cluster-treatment. Furthermore let $t, \tilde{t} \in T$ be any two treatments, not necessarily the cluster-treatment with respective clusterings $Cl^t_{t^*} = \left(Cl^t_{t^*1}, \ldots, Cl^t_{t^*k}\right)$ and $Cl^{\tilde{t}}_{t^*} = \left(Cl^{\tilde{t}}_{t^*1}, \ldots, Cl^{\tilde{t}}_{t^*k}\right)$. For any $y > 0$ the cluster-based hazard rate for t and $\tilde{t}$ is defined as*

$$h^{t\tilde{t}}_{t^*i}(y) := \lim_{\Delta y \to 0} \frac{P\left(y \leq Y < y + \Delta y | Y \geq y, \text{ patient in } Cl^t_{t^*i} \cup Cl^{\tilde{t}}_{t^*i}\right)}{\Delta y} \tag{6.2}$$

for cluster Cl_{t^*i}, $i = 1, \ldots, k$.

Usually the goal is to compare two groups with respect to their cluster-based hazard rates. Thereby the cluster-based hazard ratio, i.e. the ratio of their respective cluster-based hazard rates, is used as measure for the effect of the difference. For the evaluation of different values acquired by the hazard ratio please refer to Chapter 3.

The true cluster-based hazard rate for two treatments and therefore all cluster-based hazard ratios between any two groups are unknown. In the following we will show, how the cluster-based hazard rate and cluster-based hazard ratio between two treatments can be estimated by the Cox PH model.

6.4.2 Cluster-based Cox proportional-hazards model

Model formulation

Let $X = \left(X^1, \ldots, X^p\right)$ denote p independent covariates. Furthermore let $X^{t\tilde{t}}$ be a covariate describing the administered treatment with characteristics

$$x^{t\tilde{t}}_j = \begin{cases} 1 & \text{patient } j \text{ receives treatment } t \\ 0 & \text{patient } j \text{ receives treatment } \tilde{t} \end{cases}$$

for patient $j = 1, \ldots, n$.

Then for any $y > 0$ the cluster-based Cox PH model for treatments t and $\tilde{t}$ is defined as

$$h_{t^*i}^{t\tilde{t}}\left(y, X^{t\tilde{t}}, X\right) = h_{0\ t^*i}^{t\tilde{t}}(y) \cdot exp\left(\beta_{t^*i}^{t\tilde{t}} \cdot X^{t\tilde{t}} + \sum_{l=1}^{p} \beta_{t^*i}^{l} X^{l}\right) \tag{6.3}$$

for cluster Cl_{t^*i}, $i = 1, \ldots, k$. $h_{t^*i}^{t\tilde{t}}\left(y, X^{t\tilde{t}}, X\right)$ is the expected hazard of patients in cluster Cl_{t^*i} receiving treatment t or $\tilde{t}$ at time y given the covariates $X^{t\tilde{t}}, X^1, \ldots, X^p$, and $h_{0\ t^*i}^{t\tilde{t}}(y)$ is the baseline hazard when all covariates are equal to zero. The cluster-based Cox PH model therefore assumes the hazard for patients receiving treatment t or $\tilde{t}$ in cluster Cl_{t^*i} to be the product of the baseline hazard and the exponential function of the linear combination of the covariates. The baseline hazard does not depend on the covariates yielding to proportional hazards over time.

The observed survival times Y are assumed to be independent from each other. However, no assumption is made about the cluster-based baseline hazard $h_{0\ t^*i}^{t\tilde{t}}(y)$. Therefore the cluster-based Cox PH model is a semi-parametric model.

In Chapter 7, as well as many other survival analyses, the administered treatment is investigated. Hence, besides comparing different clusters with each other, we will furthermore distinguish between the administered treatments. The covariate describing the treatment is therefore discussed in the following remark.

Remark 6.16 (treatment covariate)**.** *The covariate describing the administered treatment receives a special notation in (6.3), as the primary investigation of the CATIE data in Chapter 7 is the difference between the administered treatments. Obviously, we could just consider it as one of the other covariates $X^1, \ldots, X^p$. However, as we will especially be analyzing the hazard ratio between two treatments, we already want to introduce this special notation, in order to be able to refer to it at any time. As mentioned above, groups defined by something besides treatments can be compared with each other as well using the cluster-based survival models. As a matter of fact, a cluster-based survival model can also be formulated without prior specification of groups wished to be compared with each other. In this case,*

the covariate describing the treatment - or other kind of group specification - $X^{t\tilde{t}}$ will just be left out.

During a cluster-based survival analysis covariates can be used for the clustering as well as in the subsequent survival model, which we will discuss in the following remark.

Remark 6.17 (use of covariate)**.** *Usually, while fitting a cluster-based survival model, the covariates already used for the clustering part are not included as covariates in the survival model part. However, this is not a necessity. Without loss of generality, we still speak of p covariates $X^1,\ldots,X^p$ in the cluster-based survival model. It depends on the origin of the data, whether to use a specific covariate in the survival model. Usually covariates of cardinal scale should be prioritized, as they can only be part of the prior clustering after transformation into quantiles.*
In the analysis of CATIE in Chapter 7, we will only consider those covariates originally included in the first publication to guarantee comparability.

The cluster-based survival model can be extended to time-varying covariates. However, our aim is to find a set of patient characteristics to reliably predict the time to the event of interest. These patient characteristics have to be available at baseline in order to make a prediction for new patients. That is why from now on, we restrict the covariates to be fixed over time.

Partial-likelihood estimation of β's

The cluster-based survival model (6.3) is estimated on the data representing patients receiving treatment t or $\tilde{t}$ in cluster Cl_{t^*i}. The coefficients β^l,$\beta^{t\tilde{t}}$ are estimated via a partial-likelihood function as introduced for the common Cox PH model in [43].

Definition 6.18 (cluster-based partial likelihood estimator)**.** *Let* $X = \left(X^1, \ldots, X^p\right)$ *and* $X^{t\tilde{t}}$ *denote* $p+1$ *independent covariates and (6.3) the related cluster-based Cox PH model. Then*

$$\widehat{\beta}^{\,l}_{t^*i}$$

is called the partial-likelihood estimator (PLE) of $\beta^{\,l}_{t^*i}$*,* $l = 1, \ldots, p$ *and*

$$\widehat{\beta}^{\,t\tilde{t}}_{t^*i}$$

the partial-likelihood estimator (PLE) of $\beta^{\,t\tilde{t}}_{t^*i}$*.*

The estimator $\widehat{\beta}^{\,l}_{t^*i}$ quantifies the influence of covariate X^l on the hazard rate for patients in cluster Cl_{t^*i} receiving treatment t or $\tilde{t}$, i.e. the influence on the probability of having an event for patients in cluster Cl_{t^*i} receiving treatment t or $\tilde{t}$. Similarly, $\widehat{\beta}^{\,t\tilde{t}}_{t^*i}$ quantifies the influence of the covariate describing the treatment on the hazard rate for patients in cluster Cl_{t^*i} receiving treatment t or $\tilde{t}$. Therefore $\widehat{\beta}^{\,t\tilde{t}}_{t^*i}$ can be used to measure the impact of the administered treatment on the probability of having an event, as we will see in the following section.

The estimated variances of the coefficients can be obtained as in most maximum-likelihood estimations. The interested reader is referred to [71].

Definition 6.19 (cluster-based variance of partial likelihood estimator)**.** *Let* $\widehat{\beta}^{\,l}_{t^*i}$ *denote the partial-likelihood estimation of* $\beta^{\,l}_{t^*i}$ *for* $l = 1, \ldots, p$ *and* $\beta^{\,t\tilde{t}}_{t^*i}$ *for* $X^{t\tilde{t}}$ *in (6.3). Then*

$$\widehat{\sigma}^{\,2}_{\widehat{\beta}^{\,l}_{t^*i}}$$

is called the estimated cluster-based variance of $\widehat{\beta}^{\,l}_{t^*i}$ *for* $l = 1, \ldots, p$ *and*

$$\widehat{\sigma}^{\,2}_{\widehat{\beta}^{\,t\tilde{t}}_{t^*i}}$$

the estimated cluster-based variance of $\widehat{\beta}^{\,t\tilde{t}}_{t^*i}$*.*

Remark 6.20. *The actual estimations can be calculated using any common statistical software.*

6.5 Cluster-based analysis of survival model

In the following we will discuss how to analyze the cluster-based survival model. We will discuss, how two groups can be compared with each other using the estimations of the cluster-based Cox PH model, to estimate their hazard ratio and present a confidence interval for the estimated hazard ratio. As mentioned before, we will not be testing the coefficients of the cluster-based Cox PH model. During a cluster-based survival analysis, specific clusters of patients on which the Cox PH model will be applied, are determined within the method. Hence, we are not able to define hypotheses for testing the coefficients prior to the method.

6.5.1 Estimated hazard ratio

By using the estimated coefficients of the cluster-based Cox PH model (6.3), we are able to estimate the hazard ratio between any two groups. Therefore let $X_1 = (x_1^1,\ldots,x_1^p,x_1^{t\tilde{t}})$, $X_2 = (x_2^1,\ldots,x_2^p,x_2^{t\tilde{t}})$ denote the characteristics of the covariates for group 1 and group 2 respectively. Then

$$\widehat{HR}_{t^*i} = exp\left(\widehat{\beta}_{t^*i}^{t\tilde{t}}\left(x_1^{t\tilde{t}} - x_2^{t\tilde{t}}\right) + \sum_{l=1}^{p}\widehat{\beta}_{t^*i}^{l}\left(x_1^l - x_2^l\right)\right) \tag{6.4}$$

estimates the hazard ratio between those groups for cluster Cl_{t^*i}, for $i = 1,\ldots,k$. Hence, it is a measure for the probability of an event occuring in group 1 compared with the probability of an event occuring in group 2.

If we want to estimate the effect of a single covariate X^i on the hazard rate, we assume the other covariates to have the same characteristic.

Therefore (6.4) simplifies to

$$\widehat{HR}_{t^*i}^{\,l} = exp\left(\widehat{\beta}_{t^*i}^{\,l}\left(x_1^{\,l} - x_2^{\,l}\right)\right) \tag{6.5}$$

for covariate $X^{\,l}$, $l = 1,\ldots,p$ and

$$\widehat{HR}_{t^*i}^{\,t\tilde{t}} = exp\left(\widehat{\beta}_{t^*i}^{\,t\tilde{t}}\left(x_1^{\,t\tilde{t}} - x_2^{\,t\tilde{t}}\right)\right) \tag{6.6}$$

for covariate $X^{\,t\tilde{t}}$.
The level of scale of the involved covariates has to be taken into account for the hazard ratio in (6.4), (6.5), and (6.6) to be interpreted correctly. Calculating the difference $\left(x_1^{\,l} - x_2^{\,l}\right)$ is reasonable, if the covariate $X^{\,l}$ is of cardinal or ordinal scale. It may also be of nominal scale, if this differences can be interpreted in a reasonable way. That is the case for covariate $X^{\,t\tilde{t}}$, as we will see in the following derivations.

We will use cluster-based survival models to compare different treatments with each other. Therefore, we can use the following corollary to estimate the hazard ratio between two treatments.

Corollary 6.21. *Let $\widehat{\beta}_{t^*i}^{\,t\tilde{t}}$ denote the partial-likelihood estimation of $\beta_{t^*i}^{\,t\tilde{t}}$ in (6.3). Then*

$$\widehat{HR}_{t^*i}^{\,t\tilde{t}} = exp\left(\widehat{\beta}_{t^*i}^{\,t\tilde{t}}\right)$$

is the estimated cluster-based hazard ratio between t and $\tilde{t}$ in cluster Cl_{t^*i}, $i = 1,\ldots,k$.

Proof. Due to the definition of $X^{\,t\tilde{t}}$ there are only two possible characteristic values, i.e. 0 and 1. Since t is associated with 1 and $\tilde{t}$ with 0 it holds:

$$exp\left(\widehat{\beta}_{t^*i}^{\,t\tilde{t}}\left(x_1^{\,t\tilde{t}} - x_2^{\,t\tilde{t}}\right)\right) = exp\left(\widehat{\beta}_{t^*i}^{\,t\tilde{t}}(1-0)\right) = exp\left(\widehat{\beta}_{t^*i}^{\,t\tilde{t}}\right)$$

□

The hazard ratio $\widehat{HR}_{t^*i}^{\,t\tilde{t}}$ compares the probability of an event occurring for patients receiving treatment t with the probability of an event occurring for patients receiving

treatment $\tilde{t}$ in cluster Cl_{t^*i}. Treatment $\tilde{t}$ can therefore be referred to as reference group.

In the following we derive a confidence interval for the coefficients in the cluster-based Cox PH model and the respective hazard ratio between treatments in a cluster.

6.5.2 Confidence interval for β's and hazard rates

The partial likelihood estimators $\widehat{\beta}^{\,l}_{t^*i}$ and $\widehat{\beta}^{\,t\tilde{t}}_{t^*i}$ have the same distributional properties as full maximum likelihood estimators [71]. Therefore they are also asymptotically normally distributed and we can construct the following confidence interval for coefficient $\beta^{\,l}_{t^*i}$ in (6.3)

$$\mathtt{I}^{\alpha}\left(\beta^{\,l}_{t^*i}\right) := \left[\widehat{\beta}^{\,l}_{t^*i} - z_{1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_{\widehat{\beta}^{\,l}_{t^*i}}, \widehat{\beta}^{\,l}_{t^*i} + z_{1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_{\widehat{\beta}^{\,l}_{t^*i}}\right]$$

where $z_{1-\frac{\alpha}{2}}$ is the $\left(1-\frac{\alpha}{2}\right)$-percentile of the standard normal distribution for $l = 1,\ldots,p$ and

$$\mathtt{I}^{\alpha}\left(\beta^{\,t\tilde{t}}_{t^*i}\right) := \left[\widehat{\beta}^{\,t\tilde{t}}_{t^*i} - z_{1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_{\widehat{\beta}^{\,t\tilde{t}}_{t^*i}}, \widehat{\beta}^{\,t\tilde{t}}_{t^*i} + z_{1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_{\widehat{\beta}^{\,t\tilde{t}}_{t^*i}}\right]$$

for $\widehat{\beta}^{\,t\tilde{t}}_{t^*i}$.

Due to Corollary 6.21 we get the following cluster-based $(1-\alpha)$-confidence interval for the hazard ratio between two treatments t and $\tilde{t}$ in cluster Cl_{t^*i}

$$\mathtt{I}^{\alpha}\left(HR^{\,t\tilde{t}}_{t^*i}\right) := \left[exp\left(\widehat{\beta}^{\,t\tilde{t}}_{t^*i} - z_{1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_{\widehat{\beta}^{\,t\tilde{t}}_{t^*i}}\right), exp\left(\widehat{\beta}^{\,t\tilde{t}}_{t^*i} + z_{1-\frac{\alpha}{2}} \cdot \widehat{\sigma}_{\widehat{\beta}^{\,t\tilde{t}}_{t^*i}}\right)\right]$$

where $z_{1-\frac{\alpha}{2}}$ is the $\left(1-\frac{\alpha}{2}\right)$-percentile of the standard normal distribution.

The cluster-based hazard ratio between t and $\tilde{t}$ in cluster $Cl_{t^*i}^{t\tilde{t}}$ has to be estimated, hence $HR_{t^*i}^{t\tilde{t}}$ is unknown. However, it holds that

$$P\left(HR_{t^*i}^{t\tilde{t}} \in \mathtt{I}^{\alpha}\left(HR_{t^*i}^{t\tilde{t}}\right)\right) = 1-\alpha$$

and therefore in $(1-\alpha)\cdot 100\%$ of all cases, the true hazard ratio between t and $\tilde{t}$ in cluster $Cl_{t^*i}^{t\tilde{t}}$ lies in $\mathtt{I}^{\alpha}\left(HR_{t^*i}^{t\tilde{t}}\right)$. Thus, $\mathtt{I}^{\alpha}\left(HR_{t^*i}^{t\tilde{t}}\right)$ is a $(1-\alpha)$-confidence interval for the probability of an event occuring for patients receiving treatment t compared with the probability of an event occuring for patients receiving treatment $\tilde{t}$ in cluster Cl_{t^*i}.

6.6 Extension: The cluster-based stratified Cox model

One of the underlying assumptions of the cluster-based Cox proportional-hazards model, is - as the name suggests - the proportional behaviour of the hazards (PH assumption). However, some of the covariates might not satisfy this assumption. Inspired by the stratified Cox model (SC model) we introduce the following cluster-based stratified Cox Model (cluster-based SC model) in order to include covariates violating the assumption of proportional hazards. During the analysis of the survival data in Chapter 7 one of the covariates is known to violate the PH assumption. Therefore, cluster-based SC models will be used throughout the entire analysis.

Let $X = \left(X^1,\ldots,X^p\right)$ and $X^{t\tilde{t}}$ denote $p+1$ independent covariates satisfying the PH assumption and Z a categorical stratification covariate with values $1,\ldots,K$ not satisfying the PH assumption. Then for any $y>0$, the cluster-based SC model for patients receiving treatments $t,\tilde{t}$ in cluster Cl_{t^*i} can be written as

$$h_{t^*i}^{t\tilde{t}}\left(y,X,X^{t\tilde{t}},Z\right) = h_{Z0\ t^*i}^{t\tilde{t}}(y)\cdot exp\left(\beta_{t^*i}^{t\tilde{t}}\cdot X^{t\tilde{t}} + \sum_{l=1}^{p}\beta_{t^*i}^{l}X^l\right)$$

for cluster Cl_{t^*i}, $i=1,\ldots,k$, allowing the cluster-based baseline hazard $h_{Z0\ t^*i}^{t\tilde{t}}(y)$ to depend on the stratum, but assuming the effect of the covariates being the same for

each stratum. If more covariates are violating the PH assumption, a new stratification variable can be defined by forming combinations of the categories of these covariates and assigning those combinations as categories to the newly defined variable [87]. The coefficients $\beta_{t^*i}^{l}$ and $\beta_{t^*i}^{t\tilde{t}}$ can be estimated in the same way as in the cluster-based Cox PH model by partial likelihood estimations.

The main application of the cluster-based SC model is to include variables for which the PH assumption is known to be violated. However, as the stratum-specific baseline hazard is now an essential part, the cluster-based SC model can only be reliable if enough events are observed in each stratum for each cluster. Typical examples for stratification variables include the site in clinical trials, as it is the case in CATIE. In Chapter 7 we will use this covariate as stratum, as it is known to violate the PH assumption.

To conclude this chapter, we will describe how to evaluate and justify different outcomes across clusters and treatments in survival analysis.

6.7 Justification of different outcome

One objective of the new explainable method of cluster-based survival analysis is to identify clusters of patients, which differ regarding the outcome. Confidence intervals as introduced previously provide a range in which the outcome of patients lies with a certain confidence. Additionally, they offer a prediction for the outcome of new patients, which is explainable due to the clusters being uniquely defined by baseline characteristics. Therefore they can be used to evaluate, whether the identified clusters of patients differ regarding the outcome.
Another objective is to compare the administered treatments on the identified clusters with each other.

We begin by discussing the evaluation of the difference between treatments of patients in a specific cluster. We will be comparing both the estimated survival functions and

hazard rates. Afterwards, we describe how to compare the outcome across clusters while considering one treatment individually and interpret the estimated survival functions. Additionally we discuss how to compare patients from one cluster with all patients receiving a treatment.

In the following we will assume the cluster-treatment t^* to be fixed, and the patient data to be divided into a clustering

$$Cl^t_{t^*} = \left(Cl^t_{t^*1}, \ldots, Cl^t_{t^*k}\right)$$

for every treatment $t \in T$.

6.7.1 Justification of different outcome across treatments

By the following procedure, we want to evaluate, whether the survival time of patients in a specific cluster differs between the administered treatments. Therefore let cluster Cl_{t*i} and the level of confidence α be fixed. For every treatment $t \in T$, the survival function $\widehat{S}^t_{t^*i}(y)$ and median survival time $\widehat{M}^t_{t^*i}$ with respective confidence intervals are estimated. Additionally the restricted mean survival time $RMST^t_{t^*i}$ is calculated.
Patients receiving treatment $t \in T$ have a higher survival time than patients receiving another treatment if

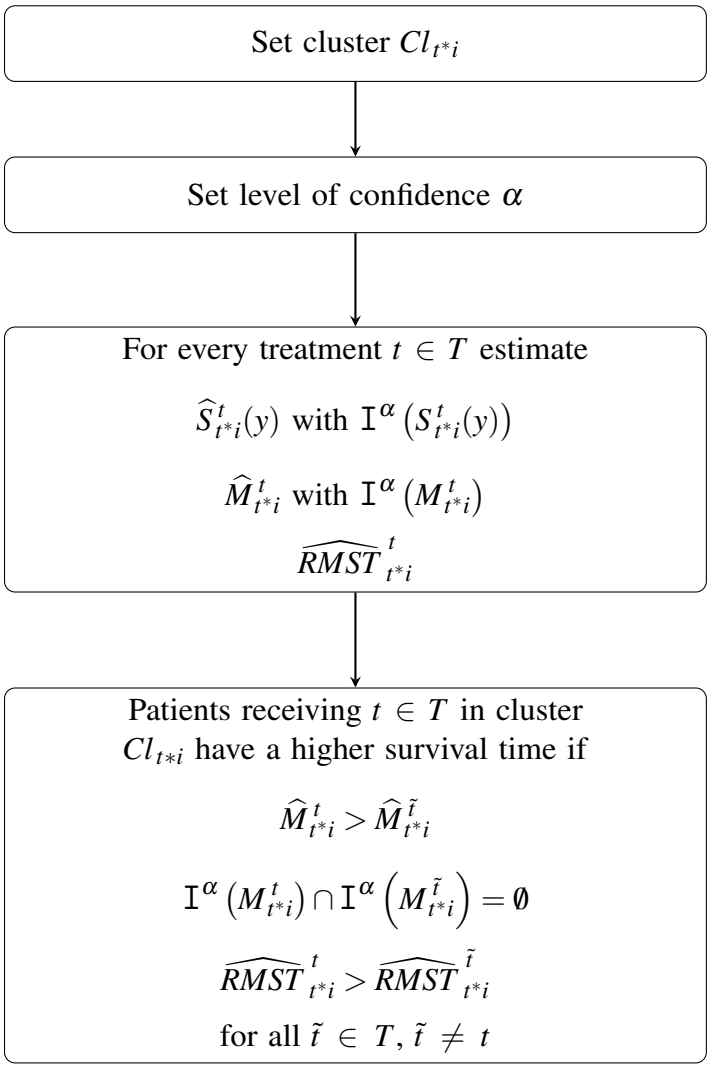
$$\widehat{M}^t_{t^*i} > \widehat{M}^{\tilde{t}}_{t^*i}$$

and the respective confidence intervals $\mathtt{I}^\alpha\left(M^t_{t^*i}\right)$ and $\mathtt{I}^\alpha\left(M^{\tilde{t}}_{t^*i}\right)$ are disjoint from each other for all $\tilde{t} \in T$ with $\tilde{t} \neq t$. Additionally, the restricted mean survival time for treatment t has to be higher than for the other treatments

$$\widehat{RMST}^{\,t}_{t^*i} > \widehat{RMST}^{\,\tilde{t}}_{t^*i}$$

for all $\tilde{t} \in T$ with $\tilde{t} \neq t$.
In this case, patients in cluster Cl_{t*i} receiving treatment t have a higher survival time than patients receiving any other treatment in this cluster. Therefore patients with the

baseline characteristics uniquely defined by cluster Cl_{t*i} are less likely to discontinue their treatment when receiving treatment t compared with any other treatment. Analogously, patients receiving a treatment might have a lower survival time, than patients receiving any other treatment. In this case, both median and restricted mean survival time have to be lower compared with all other treatments.
The procedure is summarized in Procedure 6.2.



Procedure 6.2: Evaluation of estimated survival times for patients receiving different treatments in cluster Cl_{t^*i}

In a similar way we evaluate, whether the hazard rate of patients in a specific cluster differs between the administered treatments. Therefore let cluster Cl_{t*i} and the level of confidence α be fixed. For every pair of treatments $t, \tilde{t} \in T$, the cluster-based Cox PH model is estimated. Based on that, the hazard ratio between t and $\tilde{t}$ is estimated with respective confidence interval $\mathtt{I}^{\alpha}\left(HR_{t^*i}^{t\tilde{t}}\right)$.
Patients receiving treatment $t \in T$ have a lower hazard rate than patients receiving another treatment if

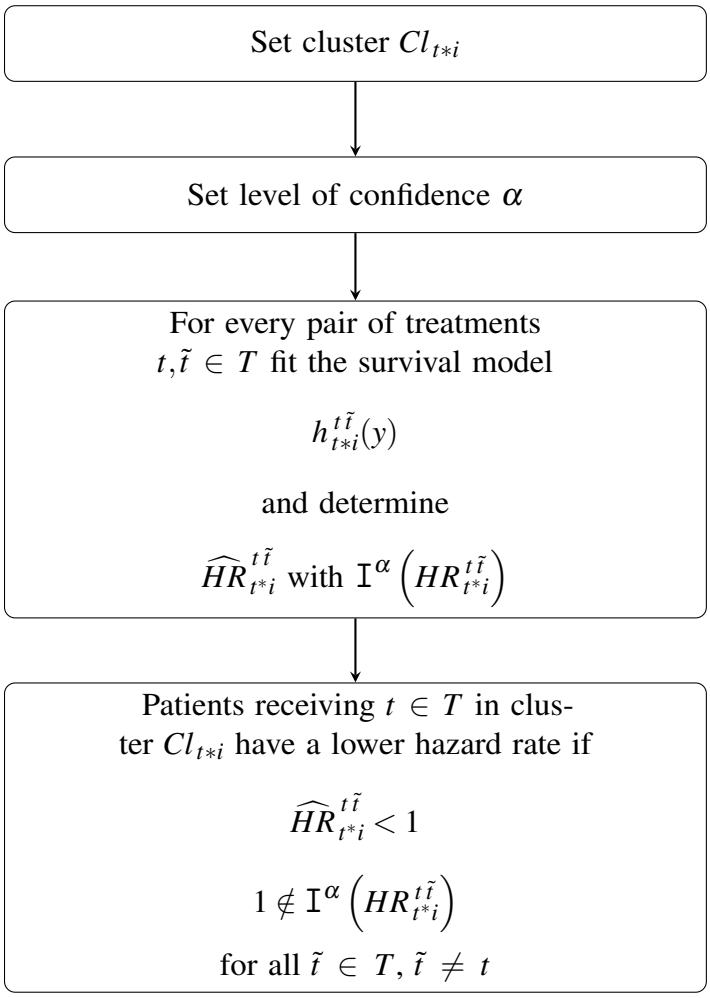
$$\widehat{HR}_{t^*i}^{t\tilde{t}} < 1$$

and 1 is not included in the respective confidence interval, i.e. $1 \notin \mathtt{I}^{\alpha}\left(HR_{t^*i}^{t\tilde{t}}\right)$ for all $\tilde{t} \in T$ with $\tilde{t} \neq t$.
In this case, patients in cluster Cl_{t*i} receiving treatment t have a lower hazard rate than patients receiving any other treatment in this cluster. Therefore patients with the baseline characteristics uniquely defined by cluster Cl_{t*i} are less likely to discontinue their treatment when receiving treatment t compared with any other treatment.
Analogously, patients receiving a treatment might have a higher hazard rate, than patients receiving any other treatment. In this case, the estimated hazard rate has to be greater than 1 for all other treatments.
The procedure is summarized in Procedure 6.3.

Remark 6.22. *While looking at real data sets, the intervals might not always be completely disjoint from each other. Eventually there are two treatments with higher survival time than the other treatments, but between those two treatments there is no difference. However, as in every statistical evaluation, the results have to be interpreted and set into context, as we will see in Part IV.*

6.7.2 Justification of different outcome across clusters

Besides comparing different treatments on one particular cluster, we also want to compare clusters with each other, while only looking at one treatment independently. We will both use it to compare two (or more) clusters with each other, as well as patients from one cluster with all patients receiving a treatment. By doing so, we can



Procedure 6.3: Evaluation of estimated hazard rates for patients receiving different treatments in cluster Cl_{t^*i}

no longer assume the survival times to be independent from each other. Therefore, we will not be using cluster-based Cox proportional-hazard models. However, we can compare their estimated survival functions with each other.

Therefore let treatment $t \in T$ and the level of confidence α be fixed. For every cluster $Cl^t_{t^*i}$, $i = 1,\ldots,k$, the survival function $\widehat{S}^t_{t^*i}(y)$ and median survival time $\widehat{M}^t_{t^*i}$ with respective confidence intervals are estimated. Additionally the restricted mean survival time $RMST^t_{t^*i}$ is calculated. Equally, the survival function without division into clusters, i.e. $\widehat{S}^t(y)$ and median survival time $\widehat{M}^t$ with respective confidence intervals are estimated as well as the restricted mean survival time $RMST^t$.

Patients in cluster $Cl^t_{t^*i}$, $i = 1,\ldots,k$ have a higher survival time than patients of the other clusters, if

$$\widehat{M}^t_{t^*i} > \widehat{M}^t_{t^*j}$$

and the respective confidence intervals $\mathtt{I}^\alpha(M^t_{t^*i})$ and $\mathtt{I}^\alpha\left(M^t_{t^*j}\right)$ are disjoint from each other for all clusters $Cl^t_{t^*j}$, $j = 1,\ldots,k$ with $j \neq i$. Additionally, the restricted mean survival time for patients in cluster $Cl^t_{t^*i}$ has to be higher compared with the other clusters, i.e.

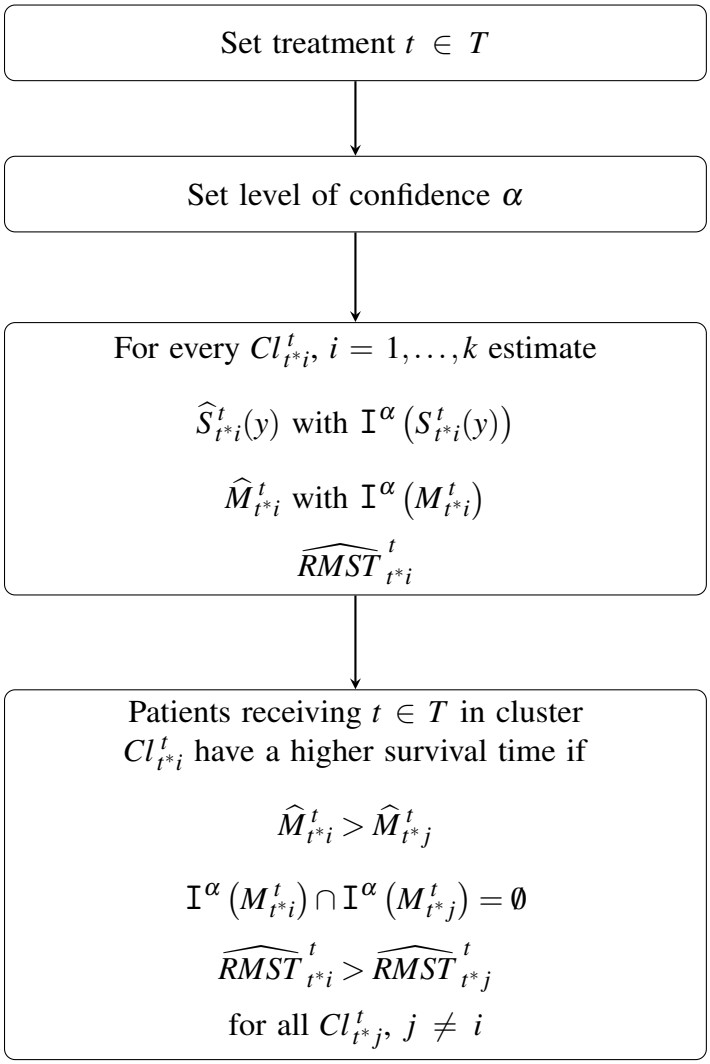
$$\widehat{RMST}^{\,t}_{t^*i} > \widehat{RMST}^{\,t}_{t^*j}$$

for all clusters $Cl^t_{t^*j}$, $j = 1,\ldots,k$ with $j \neq i$.

In this case, patients receiving treatment t in cluster Cl_{t^*i} have a higher survival time than patients in any other cluster. Therefore patients with the baseline characteristics uniquely defined by cluster Cl_{t^*i} are less likely to discontinue their treatment compared with patients with any other baseline characteristics.

Analogously, patients receiving treatment t in cluster Cl_{t^*i} might have a lower survival time than patients in any other cluster. In this case, both median and restricted mean survival time have to be less compared with all other clusters.

The procedure is summarized in Procedure 6.4.



Procedure 6.4: Evaluation of estimated survival times for patients receiving treatment t in different clusters

In a similar way, patients receiving treatment t in cluster $Cl^t_{t^*i}$, $i = 1,\ldots,k$ have a higher survival time than patients receiving treatment t in general if

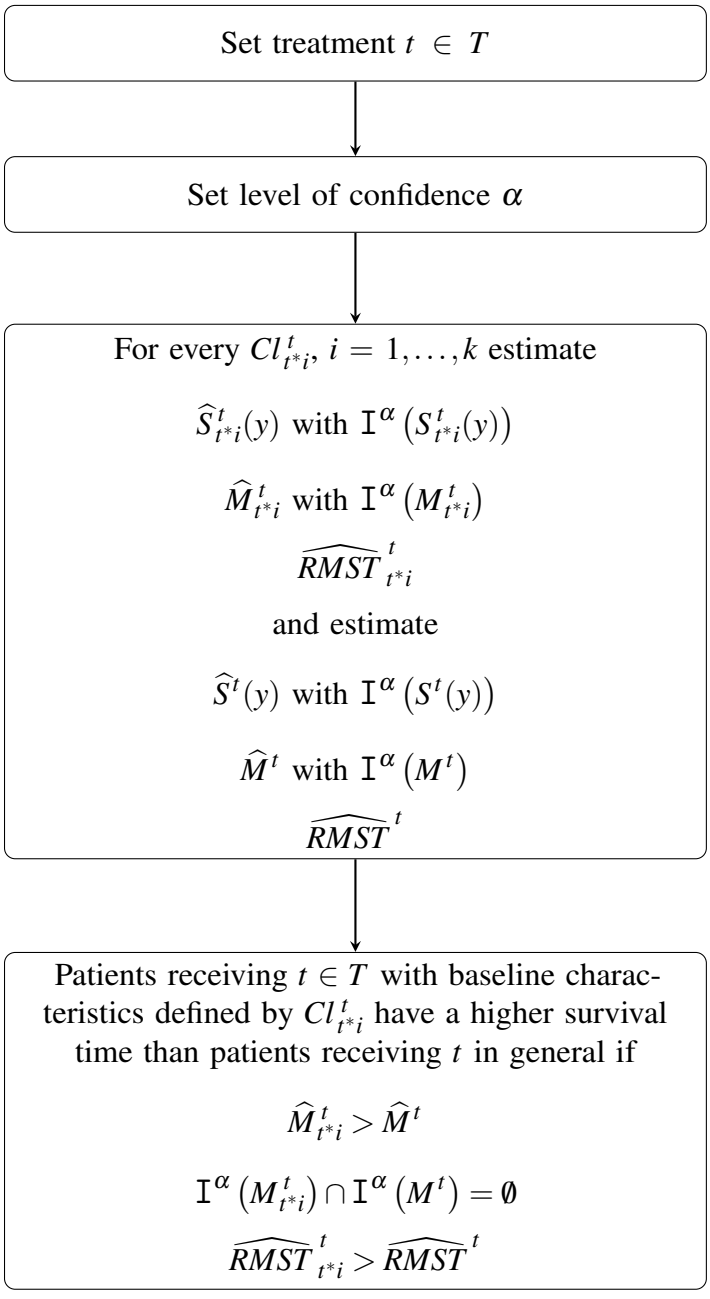
$$\widehat{M}^t_{t^*i} > \widehat{M}^t$$

and the respective confidence intervals $\mathtt{I}^\alpha\left(M^t_{t^*i}\right)$ and $\mathtt{I}^\alpha\left(M^t\right)$ are disjoint from each other. Additionally, the restricted mean survival time for patients in cluster $Cl^t_{t^*i}$ has to be higher than the restricted mean survival time of patients receiving treatment t in general:

$$\widehat{RMST}^{\,t}_{t^*i} > \widehat{RMST}^{\,t}$$

In this case, patients receiving treatment t in cluster Cl_{t^*i} have a higher survival time than patients receiving treatment t in general. Therefore patients with the baseline characteristics uniquely defined by cluster Cl_{t*i} are less likely to discontinue their treatment when receiving treatment t compared with all patients receiving treatment t.
Analogously, patients receiving treatment t in cluster Cl_{t^*i} might have a lower survival time, than patients receiving treatment t in general. In this case, both median and restricted mean survival time have to be less compared with the median and restricted mean survival time in general.
The procedure is summarized in Procedure 6.5.

This concludes Part III of this thesis. With the cluster-based survival analysis, we introduced a new method to detect differences regarding the outcome of patients between treatments as well as clusters. The method does not only detect differences, but explains them based on the respective patient characteristics, making the approach adoptable into clinical decision-making.



Procedure 6.5: Evaluation of estimated survival times for patients receiving treatment t

Part IV

Practical application

In Part IV of this thesis we apply the newly invented method for explainable analytics to a real life medical data set. The data set investigated in Chapter 7 originates from a clinical trial including patients suffering from schizophrenia. By applying the new method, we aim to gain additional knowledge for an improved treatment of patients. It is crucial, that all derived results are explainable in order to adopt them into clinical practice.

We investigate the clinical trial CATIE, a randomized clinical trial assessing the discontinuation of treatment. It begins with an overview about the clinical trial CATIE and its previous findings, on which we base our research question. Afterwards we show the results of a common survival analysis before conducting a cluster-based survival analysis. After presenting the clusters and a cluster-based analysis, we will estimate the cluster-based survival functions as well as cluster-based survival models. Finally, we will summarize the new findings by justifying different outcomes both between treatments and clusters and especially the comparison to the findings of a common survival analysis.

7. Cluster-based survival analysis: CATIE

As practical application of the cluster-based survival analysis introduced in Chapter 6, we analyze the data originating from the first phase of the U.S. Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) study. Our aim is to generate new insight into the data by applying the newly invented method for explainable analytics presented in the previous part of this thesis. We present results of a common survival analysis in comparison to the results of a cluster-based survival analysis. The results of the following analyses are summarized in the joint working paper of Brieden, Heres, Leucht, and Schiele [125].

We will begin by describing the motivation behind the clinical trial CATIE, its design, data and previous findings. Based on the original motivation behind the clinical trial, we will pose a research question we aim to answer by the cluster-based survival analysis. We will describe and interpret the derived clusters and justify the division into those, as well as compare the results of the common survival analysis with the cluster-based survival analysis.

Throughout this chapter we will be using medical terms and different scales used in the assessment of schizophrenia. A brief introduction as well as further references for the interested reader are given in Chapter 2.

The data was provided by the National Institute of Mental Health and downloaded from the NIMH Data Archive (NDA) on February 2nd 2019.

7.1 Overview

The U.S. Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) study was one of the largest independent investigations into schizophrenia treatments ever conducted [135]. Almost 1500 patients were enrolled and five antipsychotics, both first- and second-generation antipsychotics were included in the first two phases of the study. The focus of the investigation was the discontinuation of the treatment, a very common phenomenon while administering antipsychotics. Reasons for the discontinuation of treatment include inadequate therapeutic effects, unacceptable side effects as well as the patients' refusal to take the assigned antipsychotic. The study was especially designed to investigate the discrepancy between the efficacy of antipsychotics under ideal circumstances (i.e. under clinical supervision) and their effectiveness in realistic and less-than-ideal circumstances (i.e. in daily life). The main conjecture was, that second-generation antipsychotics show later discontinuation of the treatment in comparison to the older first-generation antipsychotics. A detailed overview about the history of the CATIE study is given in [135].

7.1.1 Design of CATIE

The CATIE schizophrenia trial was a multi-site, multi-phase randomized controlled trial with patients chronically suffering from schizophrenia in need of a new medication [135]. Patients were followed for at least 18 months and data was collected from 2001 to 2004.

Phase 1/1A was a randomized, controlled, double-blind comparison of five different treatments. The administered treatments were Olanzapine, Quetiapine, Risperidone, and Ziprasidone as second-generation antipsychotics and Perphenazine as first-generation antipsychotic. Ziprasidone was added to the trial after its approval in 2002. The distribution of the 1432 patients to the treatments is displayed in Table 7.1. Patients suffering from tardive dyskinesia (TD) were excluded from randomization to Perphenazine. Therefore this part of the study is referred to as phase 1A. The study design of phase 1/1A is visualized in Figure 7.1

	treatment	no. of patients included in analysis
O	Olanzapine	330
P	Perphenazine	257
Q	Quetiapine	329
R	Risperidone	333
Z	Ziprasidone	183
	total	1432

Table 7.1: Distribution of patients to treatments in CATIE phase 1/1A

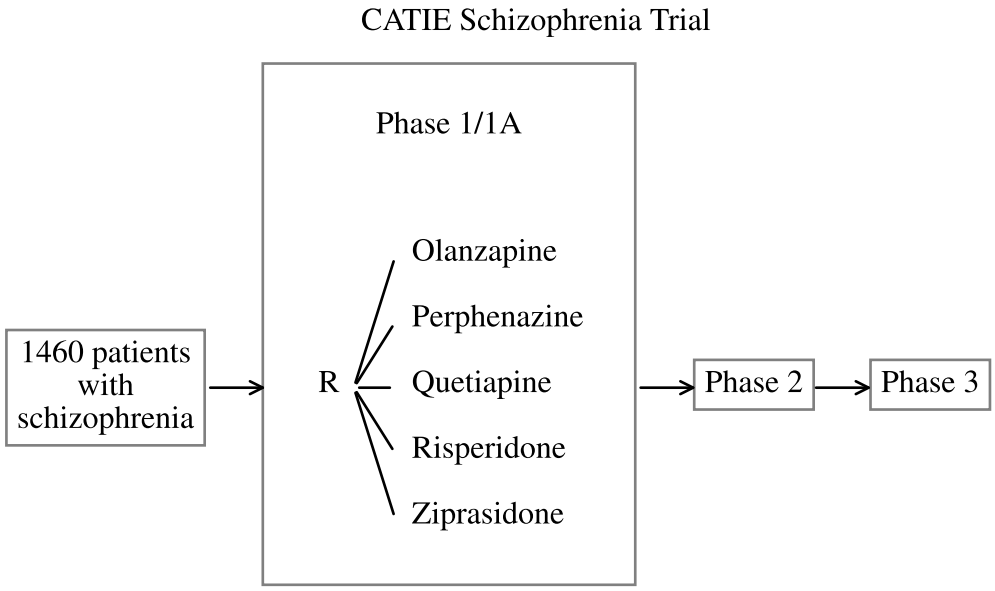


Figure 7.1: Flow chart CATIE [138]; R = randomization

Phase 2 included choosing one of two randomization pathways and phase 3 was open-label. Patients only moved on to the next phase, if they discontinued their treatment in the previous phase. Therefore patients doing well on the first prescribed treatment stayed on that treatment for the duration of the 18 month treatment period [135]. A detailed overview about the design and study protocol of the CATIE study is given in [135].

7.1.2 Data description

A total of 1460 patients were enrolled in the study. 1432 of those patients received at least one dose of the prescribed treatment. The following baseline characteristics were available at a patient individual level.

- **Characteristics at nominal level of scale**
  adjunctive or concomitant medication, gender, employment status, family interview, hospitalization in the past three months, insight & treatment attitudes questionnaire (ITAQ), living situation, marital status, other medical diagnoses (hypertension, diabetes, heart disease, pulmonary disease, HIV, hepatitis), primal diagnosis, psychiatric history (schizophrenia, major depression, obsessive–compulsive disorder, alcohol use or abuse, drug use or abuse, mental illness in family), race, tardive dyskinesia, veteran
- **Characteristics at ordinal level of scale**
  Abnormal Involuntary Movement Scale (AIMS), Barnes Akathisia Rating Scale (BARS), Calgary Depression Scale for Schizophrenia (CLGRY), Clinical Global Impression Scale (CGI-S), MacArthur Competency Assessment, Neurocognitive battery results, Positive and Negative Syndrome Scale (PANSS), PANSS subscales, PANSS Marder factors, Quality of life questionnaire, Simpson-Angus-Scale (SAS)
- **Characteristics at cardinal level of scale**
  age, age first treated for behavioural problems, age first antipsychotic was prescribed, number of previous hospitalizations (last year and lifetime), Short Form Health Survey

In phase 1/1A, the primary outcome was measured by days until discontinuation of the treatment as displayed in Table 7.2. We can assume the outcome to follow an unknown discrete distribution.

In medical data analyses it is common to assume one month to equal 28 days. Therefore we use this transformation in the following tables and figures.

phase	outcome	definition	values
1/1A	day of discontinuation	day of discontinuation phase 1/1A	$1-646$

Table 7.2: Outcome phase 1/1A in CATIE

A detailed overview about the data collected in the CATIE study is given in [135].

7.1.3 Previous findings

The first analysis of the CATIE data was published in [99]. Contrary to first conjectures, the main result was that all of the treatments were related to high and early discontinuation of the treatment. Only Olanzapine was found to be associated with significantly later discontinuation of the treatment in comparison to Quetiapine, but not to one of the other treatments. However a benefit of Olanzapine compared with all other treatments was detected, even if not all comparisons were significant according to the choice of level of significance in [99].
Other findings about the CATIE data have been published ever since for phase 1/1A as well as the other phases regarding different aspects of the study. Many of them analyzed predictors for the response to a treatment measured by the PANSS scale instead of discontinuation of treatment. [118] analyzed genetic predictors for the response of a patient. [142] analyzed predictors for the PANSS score on multiple points in time. However, they only included those characteristics available at the measured time point and they did not aim to predict response by the end of the study based on available baseline characteristics. Other publications furthermore analyzed predictors for other outcomes of the CATIE trial. [105] investigated metabolic syndrome (MS) as predictor for physical health and other outcomes. [116] analyzed negative symptoms as predictor for the general functioning of a patient. There also exist publications analyzing predictors for the outcome in other phases of the CATIE trial. [103] [137] [136] analyzed switching the medication as possible predictor for the outcome in other phases.
There are two publications forming different groups of responders. [98] formed three trajectories of treatment response measured by the PANSS scale. They provide an

overview about the average baseline characteristics in every trajectory, but the trajectories are not uniquely defined by those baseline characteristics. However, they did point out substantial heterogeneity in schizophrenia. [41] divided all patients into up to four groups of different responders (optimal, global, average, and non-responders). They assessed response to a treatment by seven different scales instead of just one measure. Even though they identify different response groups, they do not indicate, how the response groups can be uniquely defined based on baseline characteristics. Inversely they conclude, that demographics and clinical variables do not predict drug response well. Furthermore, they only considered differences regarding the response on the used scales. They did not consider different response groups on a single scale (e.g. PANSS). However, they demonstrate potential for personalized medicine.
Only [80] directly investigated predictors for the primary outcome of phase 1/1A, discontinuation of the treatment, by logistic regression. They identified low scores on neurocognitive tests, previous reported side effects, negative attitude to medication, comorbid depression and psychosocial factors (unemployment, homelessness, living alone) as the most consistent predictors. However, they only considered discontinuation at given time points instead of time to discontinuation and furthermore did not differentiate between treatments.

7.1.4 Research question

The main focus of the CATIE study was the discontinuation of the administered treatment with careful investigation of the difference between the involved treatments. However, the main analysis did not involve any further specification of subgroups who might discontinue the treatment later than other subgroups.
Our main conjecture is, that individual patient characteristics, such as the psychiatric history of the patient, influence the time to discontinuation of the treatment. By applying the new approach of cluster-based survival analysis, we aim to find clusters of patients, who discontinued the treatment later than other patients. Thereby we also aim to identify clusters of patients with substantial benefit from one of the investigated treatments opposed to the other treatments. Furthermore, we do not

only want to detect differences regarding the patient's outcome, but uniquely explain them in terms of their baseline characteristics.

7.2 Survival analysis

In the following we will present the results from a common survival analysis using notations introduced in Chapter 3. Let

$$T = \{O, Q, R, P, Z\}$$

denote the set of treatments. All 1432 patients receiving at least one dose of the treatment were included in the following analysis. We will begin by estimating and analysing the survival function for every treatment. Afterwards we will estimate and analyze the hazard ratio between pairs of treatment by estimating the respective Cox proportional-hazards model.

7.2.1 Survival function

The true survival function $S^t(y)$ of patients receiving treatment $t \in T$ is unknown. Therefore we determine the respective Kaplan-Meier estimator $\widehat{S}^t(y)$ for every treatment $t \in T$. The estimated survival functions are displayed in Figure 7.2. The separate estimated survival functions with 90%-confidence intervals $\mathtt{I}^{0.1}\left(S^t(y)\right)$ are included in Appendix A.

The estimated survival function of patients receiving Olanzapine $\widehat{S}^O(y)$ lies above all other estimated survival functions of patients receiving any other treatment. However, there is no difference between the other treatments, as their respective estimated survival functions lie fairly close to each other.

The estimated median survival times $\widehat{M}^t$ with respective confidence intervals $\mathtt{I}^{0.1}\left(M^t\right)$ are presented in Table 7.3.

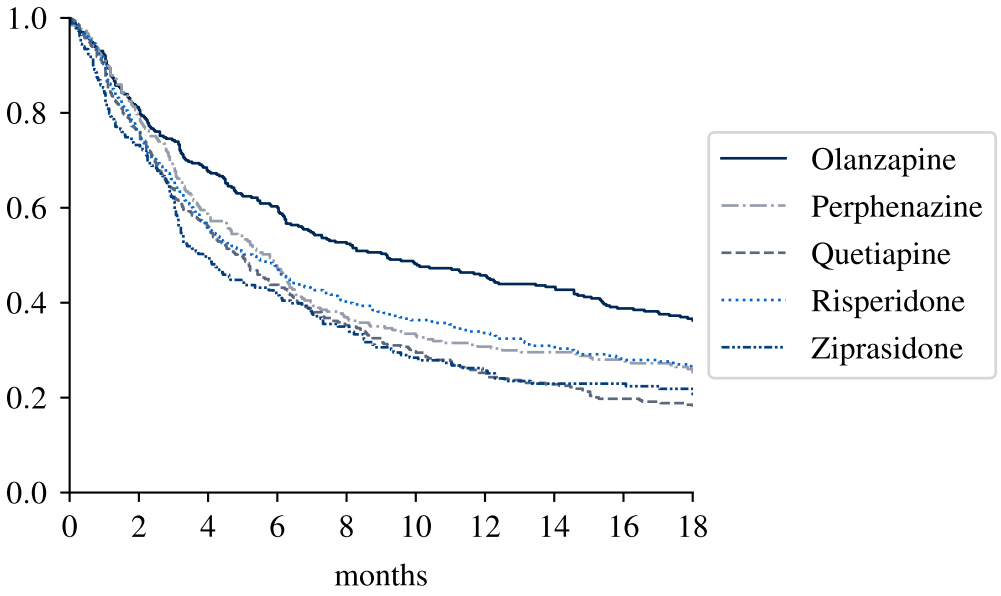


Figure 7.2: Estimated survival functions for every treatment in the common survival analysis

treatment t	estimated median survival time $\widehat{M}^t$	90%-confidence interval $\mathtt{I}^{0.1}(M^t)$
Olanzapine (O)	9.1	$[7.0, 12.0]$
Perphenazine (P)	6.2	$[4.6, 6.2]$
Quetiapine (Q)	4.9	$[4.1, 5.5]$
Risperidone (R)	5.3	$[4.2, 6.2]$
Ziprasidone (Z)	3.9	$[3.2, 5.0]$

Table 7.3: Estimated median survival times with 90%-confidence intervals for every treatment in the common survival analysis

The estimated median survival time of patients receiving Olanzapine $\widehat{M}^O$ is 9.1 months and therefore higher than the estimated median survival time of patients receiving one other of the other four treatments. The 90%-confidence interval is

$$\mathtt{I}^{0.1}\left(M^O\right) = [7.0, 12.0]$$

and not overlapping with the 90%-confidence intervals of any of the other treatments, as displayed in Figure 7.3.

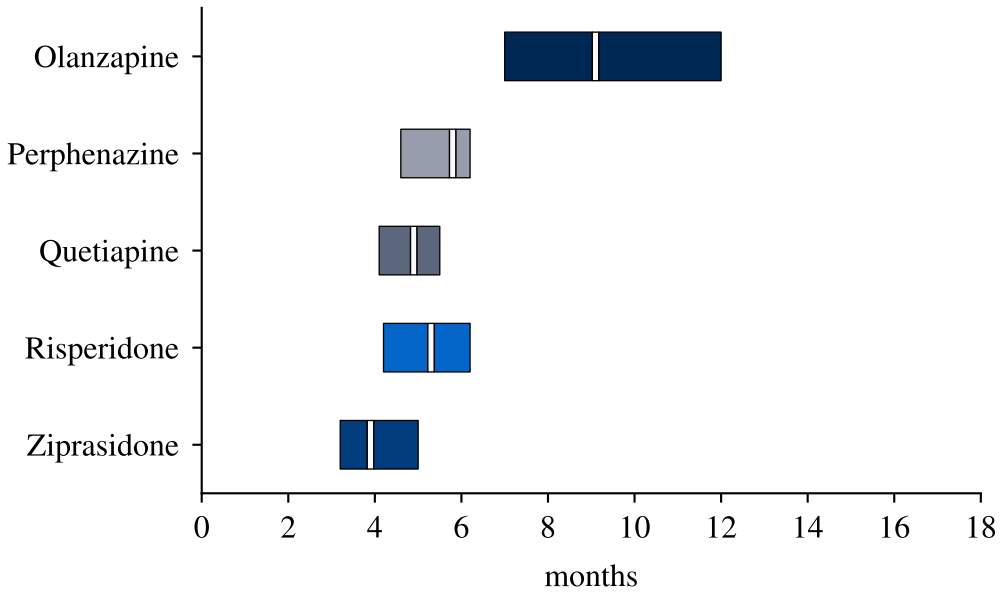


Figure 7.3: Estimated median survival times with 90%-confidence intervals common survival analysis

Therefore Olanzapine seems to be related to later and less discontinuation of the treatment. However, the gap between $\mathtt{I}^{0.1}\left(M^{O}\right)$ and $\mathtt{I}^{0.1}\left(M^{Q}\right)$, $\mathtt{I}^{0.1}\left(M^{R}\right)$,$\mathtt{I}^{0.1}\left(M^{P}\right)$, $\mathtt{I}^{0.1}\left(M^{Z}\right)$ is not very large indicating the presence of heterogeneity especially between patients receiving Olanzapine.

The estimated restricted mean survival times underline this observation, as displayed in Table 7.4. With 10.0 months, the estimated restricted mean survival time of patients receiving Olanzapine is higher than the estimated restricted mean survival time of patients receiving any of the other treatments. However, the difference is not very extensive, again implying hidden heterogeneity inside the patient data.

treatment t	estimated restricted mean survival time $\widehat{RMST}^t$
Olanzapine (O)	10.0
Perphenazine (P)	8.1
Quetiapine (Q)	7.3
Risperidone (R)	8.2
Ziprasidone (Z)	7.1

Table 7.4: Estimated restricted mean survival times for every treatment in the common survival analysis

The estimated survival functions with estimated median and restricted mean survival times indicate, that patients receiving Olanzapine discontinue the treatment later than patients receiving any of the other treatments. However, the difference is not tremendous, pointing towards the presence of heterogeneity inside the patient data. In the following section we will therefore estimate the respective pairwise hazard ratios in order to confirm the observation.

7.2.2 Survival model

In the following we will estimate the hazard ratio between two treatments by Cox proportional-hazard models. In order to guarantee comparability we will use the same covariates as in the first analysis of the CATIE data. Therefore let $X = (X^1, X^2, X^{t\tilde{t}})$ be three independent covariates, with X^1 denoting the tardive dyskinesia status and X^2 whether the patient had an exacerbation of schizophrenia in the preceding three months. Furthermore let $X^{t\tilde{t}}$ be the covariate describing the administered treatment with characteristics

$$x_j^{t\tilde{t}} = \begin{cases} 1 & \text{patient } j \text{ receives treatment } t \\ 0 & \text{patient } j \text{ receives treatment } \tilde{t} \end{cases}$$

for patient $j = 1, \ldots, n$, similar to the notation used in Chapter 6. Let Z denote the covariate describing the site, which is known to violate the proportional-hazards assumption [99].

The stratified Cox model

$$h^{t\tilde{t}}(y,X,Z) = h_{Z0}(y) \cdot exp\left(\beta^{1}X^{1} + \beta^{2}X^{2} + \beta^{t\tilde{t}}X^{t\tilde{t}}\right) \tag{7.1}$$

describes the hazard rate for patients receiving treatments $t,\tilde{t} \in T$. Using the estimated coefficient $\widehat{\beta}^{t\tilde{t}}$ in (7.1), we are able to estimate the hazard ratio

$$\widehat{HR}^{t\tilde{t}} = exp\left(\widehat{\beta}^{t\tilde{t}}\right)$$

between any two treatments $t,\tilde{t} \in T$.

The models involving patients receiving Perphenazine were limited to patients without tardive dyskinesia, as those patients were excluded from randomization to Perphenazine. Furthermore, the models involving Ziprasidone were limited to the patients, who enrolled after the inclusion of Ziprasidone into the trial.

The estimated hazard ratios with respective 90%-confidence intervals are presented in Table 7.5.

comparison $t\tilde{t}$	estimated hazard ratio $\widehat{HR}^{t\tilde{t}}$	90%-confidence interval $\mathtt{I}^{0.1}\left(HR^{t\tilde{t}}\right)$
Olanzapine - Perphenazine (O-P)	0.75	$[0.64, 0.89]$
Olanzapine - Quetiapine (O-Q)	0.60	$[0.52, 0.70]$
Olanzapine - Risperidone (O-R)	0.75	$[0.65, 0.88]$
Olanzapine - Ziprasidone (O-Z)	0.74	$[0.61, 0.89]$

Table 7.5: Estimated hazard ratios between treatments with 90%-confidence intervals in the common survival analysis

The estimated hazard ratio between Olanzapine and Quetiapine

$$\widehat{HR}^{OQ} = 0.60$$

entails that patients receiving Olanzapine only have a risk of 60% of discontinuing

their treatment compared with patients receiving Quetiapine. Vice versa patients receiving Quetiapine are $\frac{1}{0.6} \approx 1.67$ times more likely to discontinue their treatment than patients receiving Olanzapine. The corresponding 90%-confidence interval is

$$\mathtt{I}^{0.1}\left(HR^{OQ}\right) = [0.52, 0.70]$$

and does not include 1.
Analogously patients receiving Perphenazine or Risperidone are $\frac{1}{0.75} \approx 1.33$ and patients receiving Ziprasidone are $\frac{1}{0.74} \approx 1.35$ times more likely to discontinue their treatment compared with patients receiving Olanzapine.
The estimated hazard ratios with respective 90%-confidence intervals are visualized in Figure 7.4.

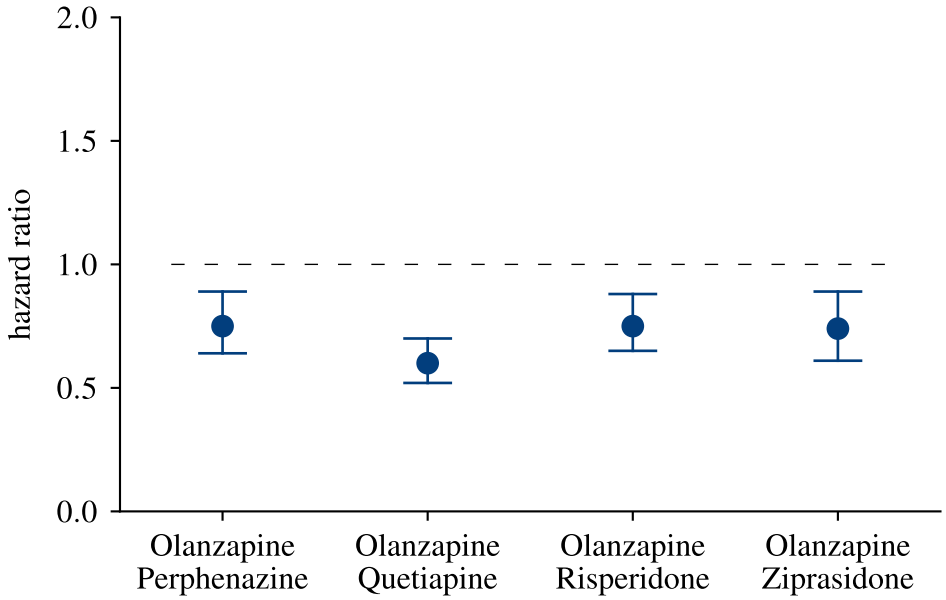


Figure 7.4: Estimated hazard ratios in common survival model with 90%-confidence intervals between Olanzapine and every other treatment

Even though all hazard ratios between Olanzapine and any other treatment are below 1, those between Olanzapine and Perphenazine, Risperidone, and Ziprasidone are not small in particular, supporting the conjecture of underlying heterogeneity.

The estimated hazard ratios between all other pairs of treatments do not indicate earlier or later discontinuation associated with any of the other treatments, as all of their confidence intervals include 1. In particular the estimated hazard ratio between Perphenazine and Risperidone is

$$\widehat{HR}^{PR} = 0.95$$

with corresponding 90%-confidence interval

$$\mathtt{I}^{0.1}\left(HR^{PR}\right) = [0.80, 1.11]$$

obviously including 1. The same results appear for any other pair of treatments. The estimated hazard ratios between pairs of treatments not listed in Table 7.5 are included in Appendix A.

Due to the conjecture of underlying heterogeneity in the patient data and the absence of further results about the discontinuation of treatment associated with medications besides Olanzapine, we perform a cluster-based survival analysis in the following section.

7.3 Cluster-based survival analysis

Based on the newly invented theory described in Chapter 6 we present the results of a cluster-based survival analysis of the CATIE data. On the following pages we will show many results. For reasons of clarity, we will therefore outline the order in which we will present the results.
To begin with, in Section 7.3.1 we will derive a clustering on one of the treatments, namely Olanzapine. Besides describing the clusters' typologies, we will discuss the clusters from a clinical point of view. At the end of Section 7.3.1 we will assign patients from all other treatments into clusters as preparation for the subsequent analyses. In Section 7.3.2 we will estimate a separate survival function for every

treatment in every cluster and present key figures to analyze the survival functions (e.g. median survival times). In Section 7.3.3 we will estimate the hazard ratio between any two pairs of treatments separately for each cluster. In both sections the results will be presented in the ordering based on the numbering of the clusters. Furthermore, there will not yet be any comparisons in both sections, as they only intend to present the results separately both for clusters and treatments. The comparison between treatments and clusters is included in Section 7.4.1 and Section 7.4.2 respectively. In Section 7.4.1 we begin with comparing treatments with each other, while only considering one of the clusters separately. The first part is concerned with comparing Olanzapine with all of the other treatments and the second part deals with comparing Perphenazine and Risperidone. Besides comparing the treatments with each other, we will also compare the cluster-based results to the results of the previously presented common survival analysis. Finally, Section 7.4.2 includes a direct comparison of the clusters. Thereby the survival functions of both clusters will be compared with each other as well as with the survival function of the common survival analysis.

Before deriving the clustering, we have to transform the outcome as described in Section 6.1. Let Y_j be the time until the event of interest occurs with characteristics $y_j > 0$ and D_j the information, whether patient j discontinued the treatment with characteristics

$$d_j = \begin{cases} 0 & \text{patient does not discontinue treatment in phase 1/1A} \\ 1 & \text{patient discontinues treatment in phase 1/1A} \end{cases}$$

for $j = 1, \ldots, n$. Patients not discontinuing their treatment ($d_j = 0$) do obviously not have a recorded time until discontinuation of the treatment. Therefore we need the following transformation to apply the clustering approach.
The transformed outcome Z_j represents the amount of incomplete days in phase 1/1A. The observed time period was 18 months, therefore we use $y^* = 548$ days as the time horizon of phase 1/1A. Patients who did not discontinue their treatment have

no incomplete days in phase 1/1A. Patients who did discontinue their treatment missed out on $548 - y_j$ days of additional treatment in phase 1/1A. The respective characteristics are obtained by

$$z_j = \begin{cases} 0 & \text{if } d_j = 0 \\ 548 - y_j & \text{if } d_j = 1 \end{cases}$$

for patient $j = 1, \ldots, n$ as displayed in Table 7.6.

	outcome	definition	values
Y_j	day of discontinuation	day of discontinuation of phase 1/1A	$1 - 646$
Z_j	incomplete days	incomplete days in phase 1/1A	$0 - 548$

Table 7.6: Original and transformed outcome in CATIE

All $n = 1432$ patients were included in the following cluster-based analysis. The treatment patient data sets are

$$G^O = \left\{ \left(x_j, z_j \right) \right\}_{j=1}^{330}$$
$$G^P = \left\{ \left(x_j, z_j \right) \right\}_{j=1}^{257}$$
$$G^Q = \left\{ \left(x_j, z_j \right) \right\}_{j=1}^{329}$$
$$G^R = \left\{ \left(x_j, z_j \right) \right\}_{j=1}^{333}$$
$$G^Z = \left\{ \left(x_j, z_j \right) \right\}_{j=1}^{183}$$

where x_j describe the available baseline characteristics and z_j the incomplete days in phase 1/1A. The aim is to investigate, which baseline characteristics indicate a high amount of incomplete days or respectively early discontinuation of treatment.

Following up on the conjecture about heterogeneity between patients receiving Olanzapine, we will manually choose Olanzapine as cluster-treatment, i.e.

$$t^* = O$$

for all further analyses.

7.3.1 Clustering

In this section we will derive the clustering on patients receiving Olanzapine. Afterwards we will describe the clusters' typologies and interpret them clinically.
The clustering was derived on all $n^O = 330$ patients receiving Olanzapine. Before deriving the clustering on G^O, we categorized some characteristics of nominal and ordinal scale into pre-defined categories. All other characteristics of ordinal scale and characteristics of cardinal scale were automatically assigned into quantiles and optimized classes as described in Section 4.3.1. All characteristics were transformed as described in Section 4.2. Afterwards we selected the variables for the geometric clustering approach as described in Section 4.3.
The selected characteristics with respective classes are displayed in Table 7.7. A complete list of all available characteristics, with respective values and classes can be found in Appendix A.

characteristic	values	class number	class
no. of previous hospitalizations (HOSP)	$0, 1, 2, 3, > 4$	1 2	0 ≥ 1
anxiety disorder in the past month (ANX)	yes, no		
drug abuse in the past 5 years (DRUG)	yes, no		

Table 7.7: Selected characteristics for clustering Olanzapine in CATIE

We applied the geometric clustering approach without specifying lower and upper bounds for each cluster using the selected variables and the described transformation to identify $k = 3$ clusters. The choice of $k = 3$ was set empirically. The transformed patients' outcome $\widehat{Z}_i$ in the patient data set $\widehat{G}_i^O$ of cluster Cl_{O-i} follows some unknown discrete distribution with possible occurrences between 0 ('no incomplete days', i.e. finished phase 1/1A without discontinuing the treatment) and 548 ('maximum incomplete days', i.e. discontinued the treatment immediately).

The estimated amount of incomplete days $\widehat{z}_{O-i}$ is used as cluster value, i.e.

$$f(Cl_{O-i}) = \hat{z}_{O-i} := \frac{1}{\kappa_{O-i}} \sum_{j=1}^{\kappa_{O-i}} z_{O-ij}$$

where κ_{O-i} is the number of patients in cluster Cl_{O-i} for $i = 1,\ldots,3$. Based on Theorem 5.3 the variance is given by

$$\widehat{\sigma}^2_{O-i} := \frac{1}{\kappa_{O-i} - 1} \sum_{j=1}^{\kappa_{O-i}} \left(z_{O-ij} - \widehat{z}_{O-i}\right)^2$$

for cluster Cl_{O-i} for $i = 1,\ldots,3$. The division of the 330 patients receiving Olanzapine into three clusters is shown in Table 7.8. 'Incomplete days' is referred to the average amount of incomplete days in phase 1/1A in the respective cluster. Therefore patients assigned to cluster Cl_{O-1} due to their individual combination of characteristic values have 199.2 days of incomplete treatment in phase 1/1A. In addition, the estimated day of discontinuation $\widehat{y}_i$ of patients in the respective cluster is provided.

cluster Cl_{O-i}	1	2	3	total
patients κ_i	154	88	88	330
incomplete days $\widehat{z}_i$	199.2	270.9	320.0	250.5
day of discontinuation $\widehat{y}_i$	334.5	268.6	221.7	286.9

Table 7.8: Clustering patients receiving Olanzapine in CATIE

The focus of the approach is the prediction of the outcome. Therefore, it is reliable to combine clusters with similar cluster values. Due to clusters Cl_{O-2} and Cl_{O-3} being similar in regards of the estimated incomplete days, we merged them into 2 resulting clusters presented in Table 7.9.

cluster Cl_{O-i}	1	2	total
patients κ_i	154	176	330
incomplete days $\widehat{z}_i$	199.2	295.5	250.5
day of discontinuation $\widehat{y}_i$	334.5	245.2	286.9

Table 7.9: Clustering patients receiving Olanzapine in CATIE with merged clusters

We allowed this automatic merging of clusters, as long as the difference from the original cluster value does not differ more than 30 days from the cluster value of the merged cluster, i.e.

$$\left| f\left(Cl^{original}\right) - f\left(Cl^{merged}\right) \right| \leq 30$$

for all $Cl^{original}$ combined to Cl^{merged}.
Cluster Cl_{O-1} consists of $\kappa_{O-1} = 154$ patients with estimated incomplete days of 199.2, whereas cluster Cl_{O-2} consists of $\kappa_{O-2} = 176$ with estimated incomplete days of 295.5.

Cluster Cl_{O-1}

The cluster is uniquely defined by three baseline characteristics. Patients who did not have an anxiety disorder in the past month (no ANX), who did not abuse drugs in the past five years (no DRUG) and have not been hospitalized in the past year (HOSP $= 0$) are assigned to Cl_{O-1}. The division is visualized in Table 7.10.
All characteristics reflect the overall situation of the patient. No anxiety disorder combined no with drug abuse and no hospitalization indicate stable conditions for the patient. Especially the finding, that non-anxious patients tend to stay on their medication was unknown before, even though the CATIE data is available to numerous researchers [125]. Patients experiencing these stable conditions only have

ANX	DRUG	HOSP	cluster Cl_{O-1} incomplete days $\widehat{z}_1$	cluster Cl_{O-1} no. of patients κ_1	cluster Cl_{O-2} incomplete days $\widehat{z}_2$	cluster Cl_{O-2} no. of patients κ_1
no	no	0	199.2	154		
		≥ 1			270.9	88
no	yes	$0, \geq 1$			311.2	52
yes	no, yes	$0, \geq 1$			332.6	36
		total	199.2	154	295.5	176

Table 7.10: Cluster typology for patients receiving Olanzapine in CATIE

estimated 199.2 days of incomplete treatment in phase 1/1A. The estimated (non transformed) day of discontinuation is 334.5. Therefore patients in Cl_{O-1} discontinue their treatment late and are overall less likely to discontinue their treatment.

Cluster Cl_{O-2}

The cluster is uniquely defined by the same three baseline characteristics. Patients who did have an anxiety disorder in the past month (ANX) and/or abused drugs in the past five years (DRUG) and/or have been hospitalized in the past year (HOSP≥ 1) are assigned to Cl_{O-2}. The division is visualized in Table 7.10.
If one component reflecting the overall situation becomes adverse, the overall situation of the patient is less stable. Especially the finding, that anxious patients tend to discontinue their medication early was unknown before, revealing the importance of further addressing these patients [125]. All in all, patients not experiencing these stable conditions have estimated 295.5 days of incomplete treatment. The estimated (non transformed) day of discontinuation is 245.5. Therefore patients in Cl_{O-2} discontinue their treatment early and are overall more likely to discontinue their treatment.

In the following, we will assign patients receiving any other treatments into clusters based on the typology derived on Olanzapine, preparing the subsequent analyses.

Using the unique baseline characteristics defining clusters Cl_{O-1} and Cl_{O-2}, we are able to assign all patients receiving any other treatment into clusters as described in Section 6.1. The division of all $n = 1432$ patients into two clusters is shown in Table 7.11.

	cluster Cl_{O-1}		cluster Cl_{O-2}	
treatment t	incomplete days $\widehat{z}^t_{O-1}$	no. of patients κ^t_{O-1}	incomplete days $\widehat{z}^t_{O-2}$	no. of patients κ^t_{O-2}
Olanzapine (O)	199.2	154	295.5	176
Perphenazine (P)	275.0	114	336.5	143
Quetiapine (Q)	315.2	143	352.1	186
Risperidone (R)	293.5	133	314.8	200
Ziprasidone (Z)	312.0	82	362.9	101
	274.4	626	329.0	806

Table 7.11: Clustering CATIE for all treatments

Patients in cluster Cl^{Q}_{O-1}, i.e. patients with those baseline characteristics defined by Cl_{O-1} receiving Quetiapine have estimated 315.2 of incomplete treatment in phase 1/1A, whereas patients in cluster Cl^{Q}_{O-2} have estimated 352.1 days of incomplete treatment in phase 1/1A. The estimated day of discontinuation $\widehat{y}^t_{O-1}$ and $\widehat{y}^t_{O-2}$ for every treatment $t \in T$ is shown in Table 7.12.

	cluster Cl_{O-1}		cluster Cl_{O-2}	
treatment t	day of discontinuation $\widehat{y}^t_{O-1}$	no. of patients κ^t_{O-1}	day of discontinuation $\widehat{y}^t_{O-2}$	no. of patients κ^t_{O-2}
Olanzapine (O)	334.5	154	245.5	176
Perphenazine (P)	261.5	114	206.2	143
Quetiapine (Q)	225.6	143	190.5	186
Risperidone (R)	244.9	133	223.7	200
Ziprasidone (Z)	228.7	82	181.2	101
	263.4	626	212.3	806

Table 7.12: Clustering CATIE for all treatments with original outcome

Patients in cluster Cl^{Q}_{O-1} discontinue their treatment with Quetiapine after estimated 225.6 days of treatment, whereas patients in cluster Cl^{Q}_{O-2} discontinue their treatment with Quetiapine already after estimated 190.5 days of treatment.

The results shown in Table 7.12 already indicate, that patients in cluster Cl_{O-1} discontinue the treatment with Olanzapine later than patients receiving any other treatment, i.e.

$$\widehat{y}^{O}_{O-1} >> \max_{t=P,Q,R,Z} \widehat{y}^{t}_{O-1}$$

whereas the difference does not seem to be as substantial for patients in cluster Cl_{O-2}. Especially the estimated day of discontinuation of patients receiving Risperidone is approximately the same estimated day of discontinuation as patients receiving Olanzapine:

$$\widehat{y}^{O}_{O-2} \approx \widehat{y}^{R}_{O-2}$$

Therefore, we estimate cluster-based survival functions for both clusters Cl_{O-1} and Cl_{O-1} in the following section.

7.3.2 Cluster-based survival function

In this section we will separately estimate a survival function in every cluster for every treatment. We begin by presenting the estimated cluster-based survival function, cluster-based median survival time and cluster-based restricted mean survival time for patients in cluster Cl_{O-1}. Afterwards we will show the respective results for patients in cluster Cl_{O-2}. The comparison of treatments and clusters is included in Section 7.4 for both clusters and all treatments. However, we will already point out some main observations while presenting the results.

The true survival functions S^{t}_{O-1} and S^{t}_{O-2} of patients receiving treatment $t \in T$ in cluster Cl_{O-1} and Cl_{O-2} are unknown. Using the approach introduced in Chapter 6 we will therefore estimate the survival functions in every cluster for every treatment.

Cluster Cl_{O-1}

Figure 7.5 shows the estimated survival functions $\widehat{S}^t_{O-1}$ for patients in cluster Cl_{O-1} receiving treatment $t \in T$. The separate estimated cluster-based survival functions with 90%-confidence intervals $\mathtt{I}^{0.1}\left(S^t_{O-1}(y)\right)$ are included in Appendix A.

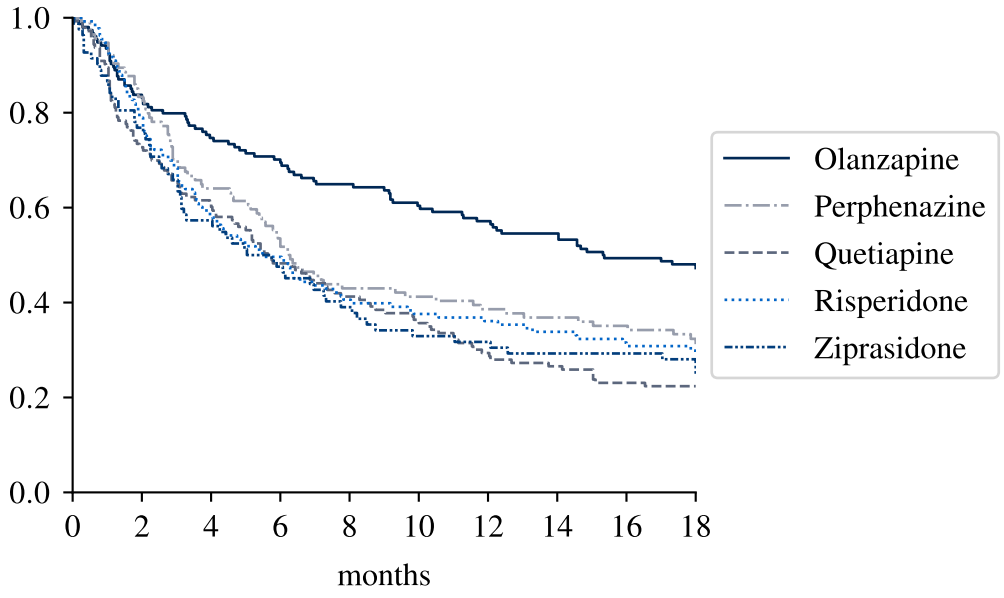


Figure 7.5: Estimated cluster-based survival functions for patients in cluster Cl_{O-1}

The estimated cluster-based survival function of patients receiving Olanzapine in cluster Cl_{O-1} lies above all other estimated survival functions, indicating that patients with baseline characteristics defined by Cl_{O-1} discontinue the treatment with Olanzapine later compared with all other treatments. Furthermore, the estimated cluster-based survival function of Perphenazine $\widehat{S}^P_{O-1}$ lies above the estimated cluster-based survival functions of Quetiapine $\widehat{S}^Q_{O-1}$, Risperidone $\widehat{S}^R_{O-1}$, and Ziprasidone $\widehat{S}^Z_{O-1}$. Hence, patients in cluster Cl_{O-1} discontinue the treatment with Perphenazine later than the treatment with Quetiapine, Risperidone, and Ziprasidone.
Based on the estimated cluster-based survival functions we estimate the cluster-based median survival times, displayed in Table 7.13 with respective 90%-confidence intervals.

treatment t	estimated median survival time $\widehat{M}^{t}_{O-1}$	90%-confidence interval $\mathtt{I}^{0.1}\left(M^{t}_{O-1}\right)$
Olanzapine (O)	15.3	$[12.1, \infty[$
Perphenazine (P)	6.2	$[5.4, 9.3]$
Quetiapine (Q)	5.6	$[4.7, 7.4]$
Risperidone (R)	5.5	$[4.1, 7.3]$
Ziprasidone (Z)	5.0	$[3.2, 7.3]$

Table 7.13: Estimated cluster-based median survival times with 90%-confidence intervals for patients in cluster Cl_{O-1}

After 15.3 months, just half of all patients in cluster Cl_{O-1} receiving Olanzapine have discontinued their treatment. On the other hand, after 6.2, 5.6, 5.5, and 5.0 months already half of all patients in cluster Cl_{O-1} receiving Perphenazine, Quetiapine, Risperidone, and Ziprasidone respectively discontinued their treatment.
We already want to note, that the estimated cluster-based median survival time of patients receiving Olanzapine is much higher than any other estimated cluster-based median survival time. Due to the upper bound of $\mathtt{I}^{0.1}\left(S^{O}_{O-1}(y)\right)$ never falling below 0.5, the upper bound of the 90%-confidence interval of the cluster-based median survival time of patients receiving Olanzapine $\mathtt{I}^{0.1}\left(M^{O}_{O-1}\right)$ can not be determined and is therefore assumed to be higher than the time horizon of 18 months.

The estimated cluster-based restricted mean survival times underline the previous observations as shown in Table 7.14.

treatment t	estimated restricted mean survival time $\widehat{RMST}^{t}_{O-1}$
Olanzapine (O)	11.7
Perphenazine (P)	9.2
Quetiapine (Q)	8.0
Risperidone (R)	8.6
Ziprasidone (Z)	8.0

Table 7.14: Estimated cluster-based restricted mean survival times for patients in cluster Cl_{O-1}

The estimated cluster-based restricted mean survival time of patients receiving Olanzapine in cluster Cl_{O-1} is 11.7 months and therefore higher than any other estimated cluster-based restricted mean survival time of patients receiving any other treatment in cluster Cl_{O-1}.

Cluster Cl_{O-2}

Figure 7.6 shows the estimated cluster-based survival functions $\widehat{S}^t_{O-2}$ for patients in cluster Cl_{O-2} receiving treatment $t \in T$. The separate estimated cluster-based survival functions with 90%-confidence intervals $\mathtt{I}^{0.1}\left(S^t_{O-2}(y)\right)$ are included in Appendix A.

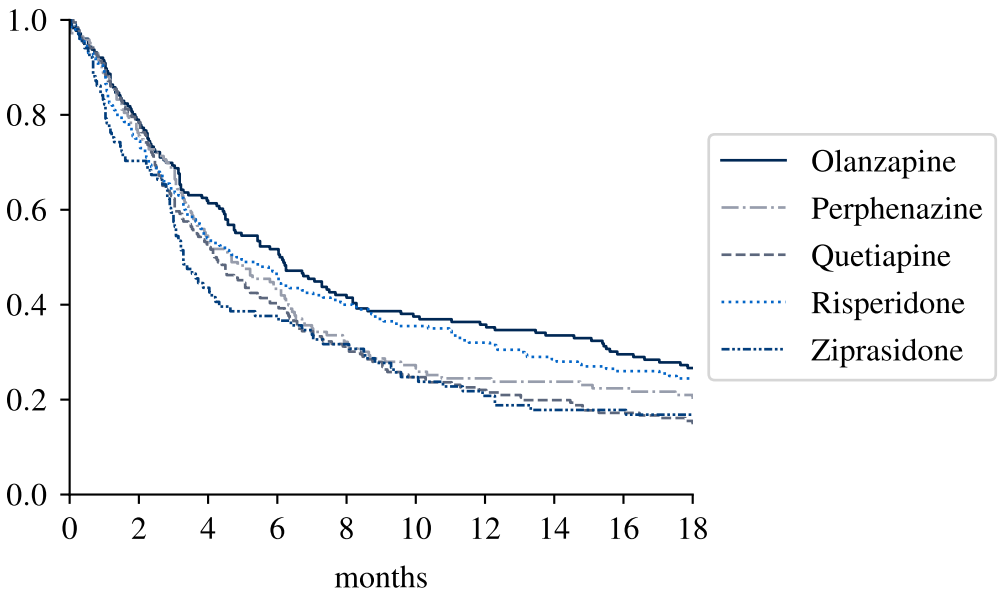


Figure 7.6: Estimated cluster-based survival functions for patients in cluster Cl_{O-2}

Opposed to the previous cluster, the estimated cluster-based survival functions of patients in cluster Cl_{O-2} lie very close together. Therefore, patients with the baseline characteristics defined by cluster Cl_{O-2} discontinue their treatment early, no matter which treatment was administered.

Based on the estimated cluster-based survival functions we estimate the cluster-based median survival times, shown in Table 7.15 with respective 90%-confidence intervals.

treatment t	estimated median survival time $\widehat{M}^{t}_{O-2}$	90%-confidence interval $\mathtt{I}^{0.1}\left(M^{t}_{O-2}\right)$
Olanzapine (O)	6.1	$[4.8, 7.3]$
Perphenazine (P)	4.7	$[3.7, 6.1]$
Quetiapine (Q)	4.2	$[3.5, 5.1]$
Risperidone (R)	5.0	$[3.8, 6.1]$
Ziprasidone (Z)	3.3	$[3.0, 4.1]$

Table 7.15: Estimated cluster-based median survival times with 90%-confidence intervals for patients in cluster Cl_{O-2}

Half of all patients in cluster Cl_{O-2} receiving Olanzapine, Perphenazine, Quetiapine, Risperidone, and Ziprasidone have discontinued their treatment after 6.1, 4.7, 4.2, 5.0, and 3.3 months respectively. There is no substantial difference between the treatments.

The cluster-based restricted mean survival times shown in Table 7.16 underline the previous observations, as they are very similar between the administered treatments.

treatment t	estimated restricted mean survival time $\widehat{RMST}^{t}_{O-2}$
Olanzapine (O)	8.6
Perphenazine (P)	7.2
Quetiapine (Q)	6.8
Risperidone (R)	8.0
Ziprasidone (Z)	6.4

Table 7.16: Estimated cluster-based restricted mean survival times for patients in cluster Cl_{O-2}

Based on the estimated cluster-based survival functions we will estimate cluster-based survival models in the following section.

7.3.3 Cluster-based survival model

In this section we will estimate the hazard ratio between any two pairs of treatments separately for both clusters. We begin by presenting the estimated cluster-based hazard ratios for patients in cluster Cl_{O-1}. Afterwards we will show the respective results for patients in cluster Cl_{O-2}. The comparison of treatments and clusters is included in Section 7.4 for both clusters and all treatments. However, we will already point out some main observations while presenting the results.

The true cluster-based hazard ratio between any two pairs of treatment $t,\tilde{t} \in T$ is unknown. We estimate the cluster-based hazard ratio between two treatments by cluster-based Cox proportional hazard models, as introduced in Chapter 6. Therefore, let $X^{t\tilde{t}}$ be the covariate describing the administered treatment with characteristics

$$x_j^{t\tilde{t}} = \begin{cases} 1 & \text{patient } j \text{ receives treatment } t \\ 0 & \text{patient } j \text{ receives treatment } \tilde{t} \end{cases}$$

for patient $j = 1,\ldots,n$, similar to the notation used in Chapter 6.

Let Z denote the covariate describing the site, which is known to violate the proportional-hazards assumption [99]. The cluster-based stratified Cox model

$$h_{O-i}^{t\tilde{t}}(y, X^{t\tilde{t}}, Z) = h_{Z0\ O-i}^{t\tilde{t}}(y) \cdot exp\left(\beta_{O-i}^{t\tilde{t}} X_{O-i}^{t\tilde{t}}\right) \tag{7.2}$$

describes the hazard rate for patients in cluster Cl_{O-i} receiving treatments $t,\tilde{t} \in T$, $i = 1,2$. Using the estimated coefficient $\widehat{\beta}_{O-i}^{t\tilde{t}}$ in (7.2), we are able to estimate the hazard ratio

$$\widehat{HR}_{O-i}^{t\tilde{t}} = exp\left(\widehat{\beta}_{O-i}^{t\tilde{t}}\right)$$

between any two treatments $t,\tilde{t} \in T$ in cluster Cl_{O-i} .

In order to guarantee comparability to the common survival model we also analyzed cluster-based survival models with the same covariates as in the first analysis of the CATIE data. Therefore let $X = (X^1, X^2)$ be two independent covariates, with X^1 denoting the tardive dyskinesia status and X^2 whether the patient had an exacerbation of schizophrenia in the preceding three months. The cluster-based stratified Cox model with additional covariates is then denoted by:

$$h^{t\tilde{t}}_{O-i}(y, X, X^{t\tilde{t}}, Z) = h^{t\tilde{t}}_{Z0\ O-i}(y) \cdot exp\left(\beta^1_{O-i} X^1_{O-i} + \beta^2_{O-i} X^2_{O-i} + \beta^{t\tilde{t}}_{O-i} X^{t\tilde{t}}_{O-i}\right) \quad (7.3)$$

The estimated cluster-based hazard rates from (7.3) only differ marginally from the estimated cluster-based hazard rates from (7.2). The results are included in Appendix A. Due to the estimated cluster-based hazard rates being similar to each other, we refrain from overloading the model and use (7.2) for all further analyses.

All cluster-based Cox PH models involving patients receiving Perphenazine were limited to patients without tardive dyskinesia, as those patients were excluded from randomization to Perphenazine. Furthermore, the models involving Ziprasidone were limited to the patients who enrolled after the inclusion of Ziprasidone into the trial.

We will begin by presenting the estimated cluster-based hazard ratios between two treatments in cluster Cl_{O-1} and afterwards present the results for cluster Cl_{O-2}. The comparison of treatments and clusters is included in Section 7.4 for both clusters and all treatments. However, we will already point out the main observations while presenting the results.

Cluster Cl_{O-1}

Table 7.17 shows the estimated cluster-based hazard ratios between Olanzapine and all other treatments for patients in cluster Cl_{O-1} with respective 90%-confidence intervals.

comparison $t\tilde{t}$	estimated hazard ratio $\widehat{HR}_{O-1}^{t\tilde{t}}$	90%-confidence interval $\mathtt{I}^{0.1}\left(HR_{O-1}^{t\tilde{t}}\right)$
Olanzapine - Perphenazine (OP)	0.76	$[0.58, 1.01]$
Olanzapine - Quetiapine (OQ)	0.48	$[0.37, 0.61]$
Olanzapine - Risperidone (OR)	0.59	$[0.45, 0.75]$
Olanzapine - Ziprasidone (OZ)	0.65	$[0.47, 0.89]$

Table 7.17: Estimated cluster-based hazard ratios between Olanzapine and all other treatments with 90%-confidence intervals for patients in cluster Cl_{O-1}

The estimated hazard ratio between Olanzapine and Quetiapine

$$\widehat{HR}_{O-1}^{OQ} = 0.48$$

entails that patients in cluster Cl_{O-1} receiving Olanzapine only have a chance of 48% of discontinuing their treatment compared with patients receiving Quetiapine. Vice versa patients receiving Quetiapine are $\frac{1}{0.48} \approx 2.08$ times more likely to discontinue their treatment than patients receiving Olanzapine in cluster Cl_{O-1}.
The corresponding 90%-confidence interval is

$$\mathtt{I}^{0.1}\left(HR_{O-1}^{OQ}\right) = [0.37, 0.61]$$

and does not include 1. Furthermore, patients in cluster Cl_{O-1} receiving Perphenazine are $\frac{1}{0.76} \approx 1.32$, patients receiving Risperidone are $\frac{1}{0.59} \approx 1.69$ and patients receiving Ziprasidone are $\frac{1}{0.65} \approx 1.54$ times more likely to discontinue their treatment than patients receiving Olanzapine. Only the 90%-confidence interval for the estimated cluster-based hazard ratio between Olanzapine and Perphenazine just includes 1, as the upper bound is 1.01. All other 90%-confidence intervals do not include 1, as visualized in Figure 7.7.

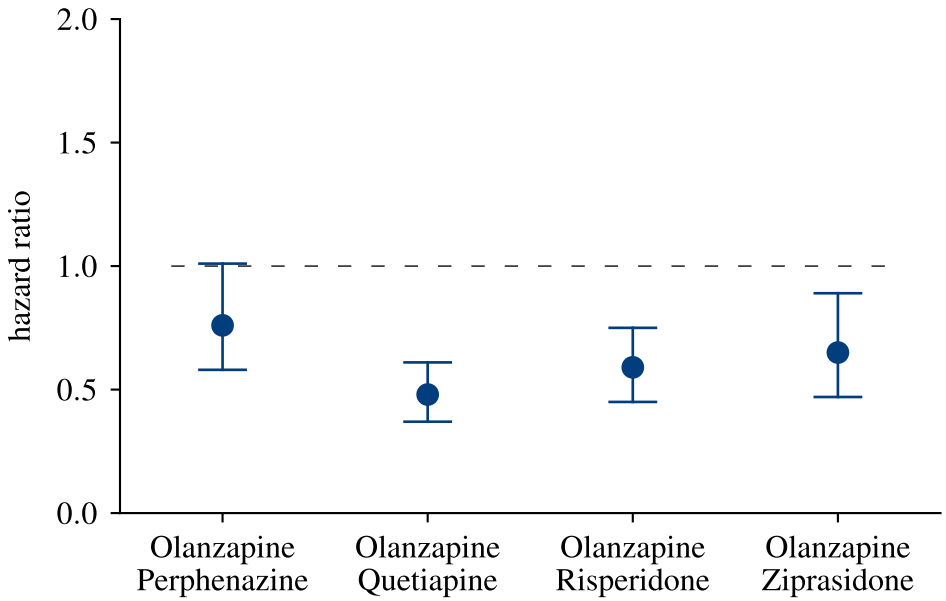


Figure 7.7: Estimated cluster-based hazard ratios between Olanzapine and all other treatments with 90%-confidence intervals for patients in cluster Cl_{O-1}

Furthermore the estimated cluster-based hazard ratio between Risperidone and Perphenazine is

$$\widehat{HR}_{O-1}^{PR} = 0.71$$

with corresponding 90%-confidence interval

$$\mathtt{I}^{0.1}\left(HR_{O-1}^{PR}\right) = [0.55, 0.94]$$

not including 1. Therefore patients receiving Risperidone in cluster Cl_{O-1} are $\frac{1}{0.71} \approx 1.4$ times more likely to discontinue their treatment than patients receiving Perphenazine. We will discuss this observation in Section 7.4.

The estimated cluster-based hazard ratios between any other pairs of treatment do not indicate earlier or later discontinuation associated with any of the other treatments, as all of their 90%-confidence intervals include 1. The results are included in Appendix A.

Cluster Cl_{O-2}

Table 7.17 shows the estimated cluster-based hazard ratios between Olanzapine and all other treatments for patients in cluster Cl_{O-2} with respective 90%-confidence intervals.

comparison $t\tilde{t}$	estimated hazard ratio $\widehat{HR}_{O-2}^{t\tilde{t}}$	90%-confidence interval $\mathtt{I}^{0.1}\left(HR_{O-2}^{t\tilde{t}}\right)$
Olanzapine - Perphenazine (OP)	0.85	$[0.69, 1.06]$
Olanzapine - Quetiapine (OQ)	0.69	$[0.57, 0.84]$
Olanzapine - Risperidone (OR)	0.96	$[0.79, 1.16]$
Olanzapine - Ziprasidone (OZ)	0.92	$[0.71, 1.19]$

Table 7.18: Estimated cluster-based hazard ratios between Olanzapine and all other treatments with 90%-confidence intervals for patients in cluster Cl_{O-2}

The estimated hazard ratio between Olanzapine and Quetiapine

$$\widehat{HR}_{O-2}^{OQ} = 0.69$$

entails that patients receiving Olanzapine only have a chance of 69% of discontinuing their treatment compared with patients receiving Quetiapine. Vice versa patients receiving Quetiapine are $\frac{1}{0.69} \approx 1.45$ times more likely to discontinue their treatment than patients receiving Olanzapine. The corresponding 90%-confidence interval is

$$\mathtt{I}^{0.1}\left(HR_{O-2}^{OQ}\right) = [0.57, 0.84]$$

and does not include 1. All other pairwise hazard ratios do not show similar results. Patients in cluster Cl_{O-2} receiving Risperidone are only $\frac{1}{0.96} \approx 1.04$ and patients

receiving Ziprasidone are only $\frac{1}{0.92} \approx 1.09$ times more likely to discontinue their treatment than patients receiving Olanzapine. Both hazard ratios are very close to one. Therefore patients in cluster Cl_{O-2} do not discontinue the treatment with Olanzapine later than the treatment with Risperidone or Ziprasidone. Additionally, patients receiving Perphenazine are just $\frac{1}{0.85} \approx 1.18$ times more likely to discontinue their treatment than patients receiving Olanzapine and the respective 90%-confidence interval for the estimated cluster-based hazard ratio between Olanzapine and Perphenazine includes 1, as visualized in Figure 7.8.

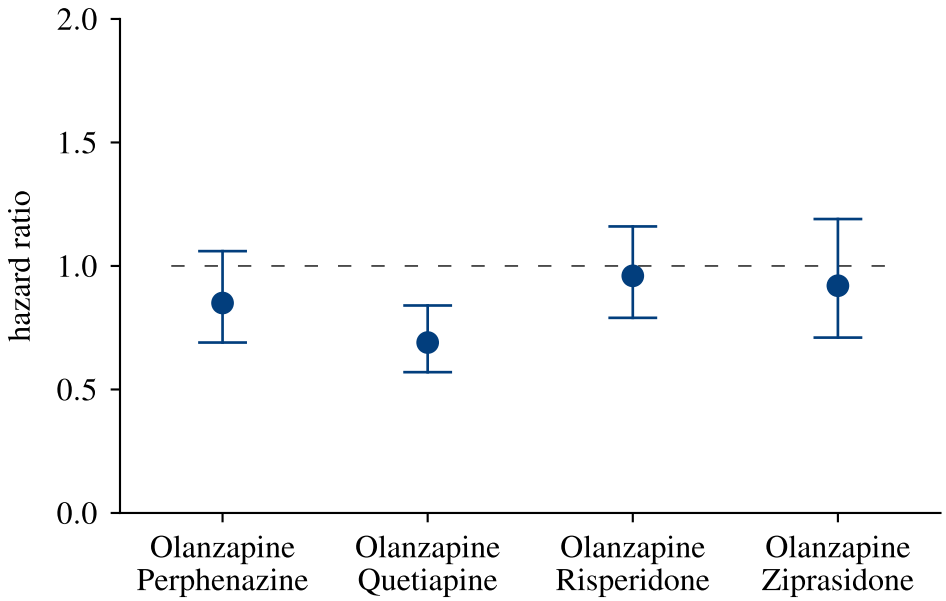


Figure 7.8: Estimated cluster-based hazard ratios between Olanzapine and all other treatments with 90%-confidence intervals for patients in cluster Cl_{O-2}

The estimated cluster-based hazard ratios between any other pairs of treatment do not indicate earlier or later discontinuation associated with any of the other treatments, as all of their 90%-confidence intervals include 1. The results are included in Appendix A.

7.4 Justification of different time to discontinuation

Up to this point, the results were presented separately regarding clusters and treatments. The comparisons are included in this section. We will begin by comparing treatments with each other while regarding one of the clusters separately, thus justifying the assumption about different outcomes across treatments. Thereby we will especially be comparing the results of the common survival analysis with the cluster-based survival analysis. Afterwards we will directly compare the clusters to each other, to justify the division into clusters by looking at the difference between the clusters regarding the outcome.

7.4.1 Justification across treatments

In this section we will compare treatments with each other while regarding both clusters separately. We begin by comparing Olanzapine with all other treatments. The results will be presented in the ordering based on the numbering of the clusters. Thereby we will both compare the survival functions, as well as the hazard ratios resulting from the respective survival models. Besides comparing the treatments with each other, we will highlight the difference to the results of the common survival analysis, emphasizing the importance of the cluster-based consideration. Afterwards we will compare Perphenazine with Risperidone and highlight different behaviours on both clusters, especially in comparison with the common survival analysis.

Olanzapine versus other treatments

In the common survival analysis Olanzapine was found to be associated with later discontinuation of the treatment compared with all other treatments (see Section 7.2). We will begin with the comparison of treatments for patients in cluster Cl_{O-1}. Afterwards respective comparisons will be made for patients in cluster Cl_{O-2}.

The cluster-based survival analysis presented in Section 7.3 indicates, that patients in cluster Cl_{O-1} discontinue the treatment with Olanzapine later compared with the treatment with Perphenazine, Quetiapine, Risperidone, and Ziprasidone.

Applying Procedure 6.2 on cluster Cl_{O-1} with level of confidence $\alpha = 0.1$ we can observe

$$\begin{aligned}
\widehat{M}^{O}_{O-1} &= 15.3 > 6.2 = \widehat{M}^{P}_{O-1} \\
\widehat{M}^{O}_{O-1} &= 15.3 > 5.6 = \widehat{M}^{Q}_{O-1} \\
\widehat{M}^{O}_{O-1} &= 15.3 > 5.5 = \widehat{M}^{R}_{O-1} \\
\widehat{M}^{O}_{O-1} &= 15.3 > 5.0 = \widehat{M}^{Z}_{O-1}
\end{aligned}$$

with disjoint 90%-confidence intervals

$$\begin{aligned}
\mathtt{I}^{0.1}\left(M^{O}_{O-1}\right) \cap \mathtt{I}^{0.1}\left(M^{P}_{O-1}\right) &= [12.1,\infty[\cap [5.4,9.3] = \emptyset \\
\mathtt{I}^{0.1}\left(M^{O}_{O-1}\right) \cap \mathtt{I}^{0.1}\left(M^{Q}_{O-1}\right) &= [12.1,\infty[\cap [4.7,7.4] = \emptyset \\
\mathtt{I}^{0.1}\left(M^{O}_{O-1}\right) \cap \mathtt{I}^{0.1}\left(M^{R}_{O-1}\right) &= [12.1,\infty[\cap [4.1,7.3] = \emptyset \\
\mathtt{I}^{0.1}\left(M^{O}_{O-1}\right) \cap \mathtt{I}^{0.1}\left(M^{Z}_{O-1}\right) &= [12.1,\infty[\cap [3.2,7.3] = \emptyset
\end{aligned}$$

for all estimated cluster-based median survival times in cluster Cl_{O-1} as displayed in Figure 7.9.

The 90%-confidence interval of the estimated cluster-based median survival time of patients receiving Olanzapine in cluster Cl_{O-1} is not just disjoint from all other 90%-confidence intervals, but there is a substantial gap, which was not observed in the common survival analysis (see Figure 7.3).

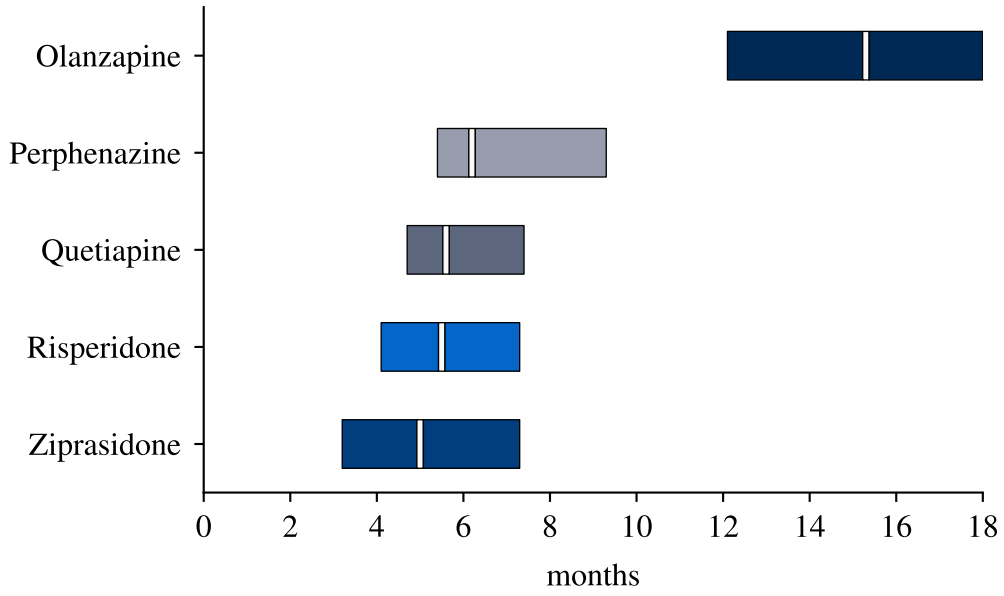


Figure 7.9: Estimated cluster-based median survival times for patients in cluster Cl_{O-1}

Furthermore, the estimated cluster-based restricted mean survival time of Olanzapine is higher than the estimated cluster-based restricted mean survival time of any other treatment:

$$\widehat{RMST}^{O}_{O-1} = 11.7 > 9.2 = \widehat{RMST}^{P}_{O-1}$$
$$\widehat{RMST}^{O}_{O-1} = 11.7 > 8.0 = \widehat{RMST}^{Q}_{O-1}$$
$$\widehat{RMST}^{O}_{O-1} = 11.7 > 8.6 = \widehat{RMST}^{R}_{O-1}$$
$$\widehat{RMST}^{O}_{O-1} = 11.7 > 8.0 = \widehat{RMST}^{Z}_{O-1}$$

Therefore patients in cluster Cl_{O-1} discontinue the treatment with Olanzapine substantially later compared with all other treatments.

The observations are strengthened by the analysis of the respective cluster-based survival models. Applying Procedure 6.3 on cluster Cl_{O-1} with level of confidence $\alpha = 0.1$ we can observe the hazard ratios between Olanzapine and all other treatments

to be less than one with all but one of the 90%-confidence intervals not including 1:

$$\widehat{HR}_{O-1}^{OP} = 0.76 \text{ with } \mathtt{I}^{0.1}\left(HR_{O-1}^{OP}\right) = [0.58, 1.01]$$
$$\widehat{HR}_{O-1}^{OQ} = 0.48 \text{ with } \mathtt{I}^{0.1}\left(HR_{O-1}^{OQ}\right) = [0.37, 0.61]$$
$$\widehat{HR}_{O-1}^{OR} = 0.59 \text{ with } \mathtt{I}^{0.1}\left(HR_{O-1}^{OR}\right) = [0.45, 0.75]$$
$$\widehat{HR}_{O-1}^{OZ} = 0.65 \text{ with } \mathtt{I}^{0.1}\left(HR_{O-1}^{OZ}\right) = [0.47, 0.89]$$

Even though $\mathtt{I}^{\alpha}\left(HR_{O-1}^{OP}\right)$ does include 1, its upper bound is just 1.01. Therefore we can still assume patients receiving Olanzapine to have a substantially lower risk of discontinuing the treatment with Olanzapine compared with Perphenazine, and obviously also Risperidone and Ziprasidone.
Furthermore, the cluster-based hazard ratios were lower for three out of four treatments compared with the common survival analysis:

$$\widehat{HR}_{O-1}^{OQ} = 0.48 < 0.60 = \widehat{HR}^{OQ}$$
$$\widehat{HR}_{O-1}^{OR} = 0.59 < 0.75 = \widehat{HR}^{OR}$$
$$\widehat{HR}_{O-1}^{OZ} = 0.65 < 0.74 = \widehat{HR}^{OZ}$$

Therefore patients in cluster Cl_{O-1} show an even lower risk than initially suggested by the common survival analysis of discontinuing the treatment with Olanzapine compared with Quetiapine, Risperidone, and Ziprasidone. The risk of discontinuing the treatment with Olanzapine compared with Perphenazine is approximately similar to the suggestion by the common survival analysis:

$$\widehat{HR}_{O-1}^{OP} = 0.76 \approx 0.75 = \widehat{HR}^{OP}$$

The estimated cluster-based hazard ratios of cluster Cl_{O-1} in comparison to the estimated hazard ratios from the common survival analysis are displayed in Figure 7.10.

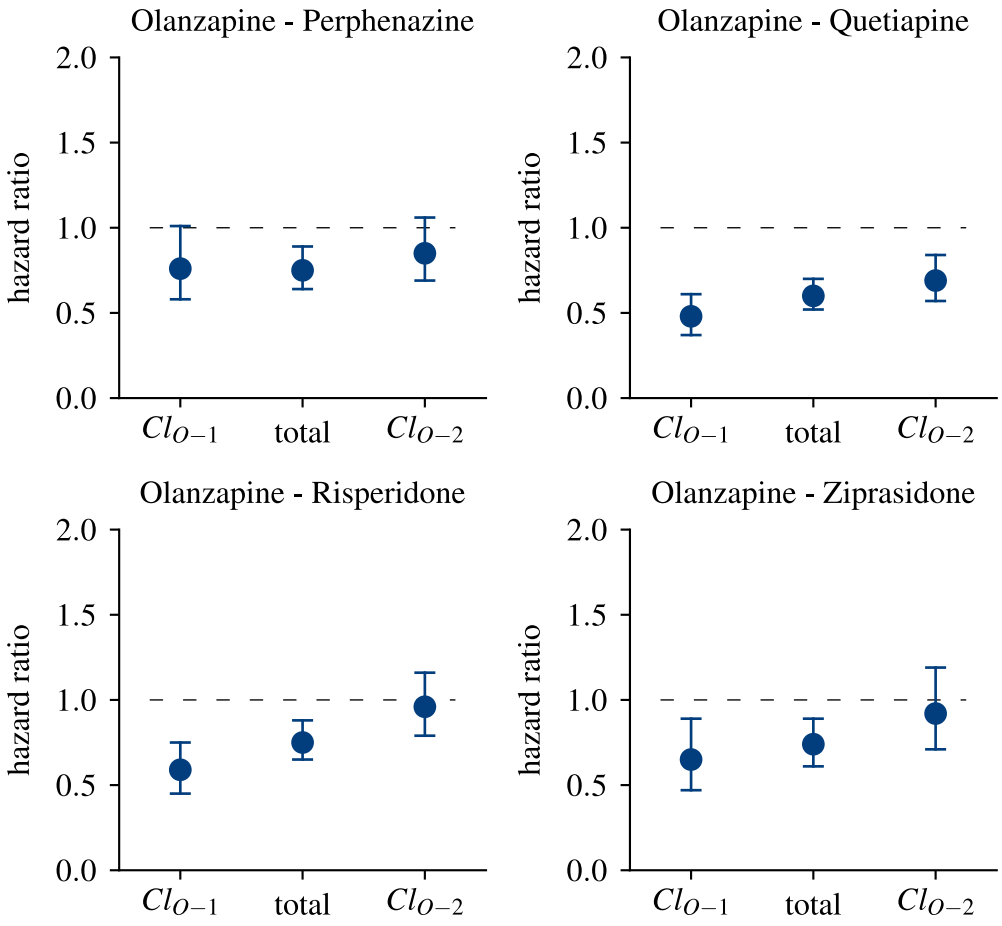


Figure 7.10: Estimated cluster-based hazard ratios between Olanzapine and all other treatments for patients in cluster Cl_{O-1} and Cl_{O-2} in comparison to the hazard ratios from the common survival models with 90%-confidence intervals

Thus, the benefit of Olanzapine compared with the other treatments was strongly underestimated for patients in cluster Cl_{O-1}.

On the other hand the results presented in Section 7.3 indicate, that there is no such difference between the treatments regarding the discontinuation of treatment for patients in cluster Cl_{O-2}. Applying Procedure 6.2 on cluster Cl_{O-2} with level of confidence $\alpha = 0.1$ we can not observe the same results. The estimated cluster-based median survival time of patients receiving Olanzapine is just marginally higher than

those of the other treatments

$$\widehat{M}^{O}_{O-2} = 6.1 > 4.7 = \widehat{M}^{P}_{O-2}$$
$$\widehat{M}^{O}_{O-2} = 6.1 > 4.2 = \widehat{M}^{Q}_{O-2}$$
$$\widehat{M}^{O}_{O-2} = 6.1 > 5.0 = \widehat{M}^{R}_{O-2}$$
$$\widehat{M}^{O}_{O-2} = 6.1 > 3.3 = \widehat{M}^{Z}_{O-2}$$

with all but one of the respective 90%-confidence intervals overlapping

$$\mathtt{I}^{0.1}\left(M^{O}_{O-1}\right) \cap \mathtt{I}^{0.1}\left(M^{P}_{O-1}\right) = [4.8, 7.3] \cap [3.7, 6.1] \neq \emptyset$$
$$\mathtt{I}^{0.1}\left(M^{O}_{O-1}\right) \cap \mathtt{I}^{0.1}\left(M^{Q}_{O-1}\right) = [4.8, 7.3] \cap [3.5, 5.1] \neq \emptyset$$
$$\mathtt{I}^{0.1}\left(M^{O}_{O-1}\right) \cap \mathtt{I}^{0.1}\left(M^{R}_{O-1}\right) = [4.8, 7.3] \cap [3.8, 6.1] \neq \emptyset$$
$$\mathtt{I}^{0.1}\left(M^{O}_{O-1}\right) \cap \mathtt{I}^{0.1}\left(M^{Z}_{O-1}\right) = [4.8, 7.3] \cap [3.0, 4.1] = \emptyset$$

as displayed in Figure 7.11.

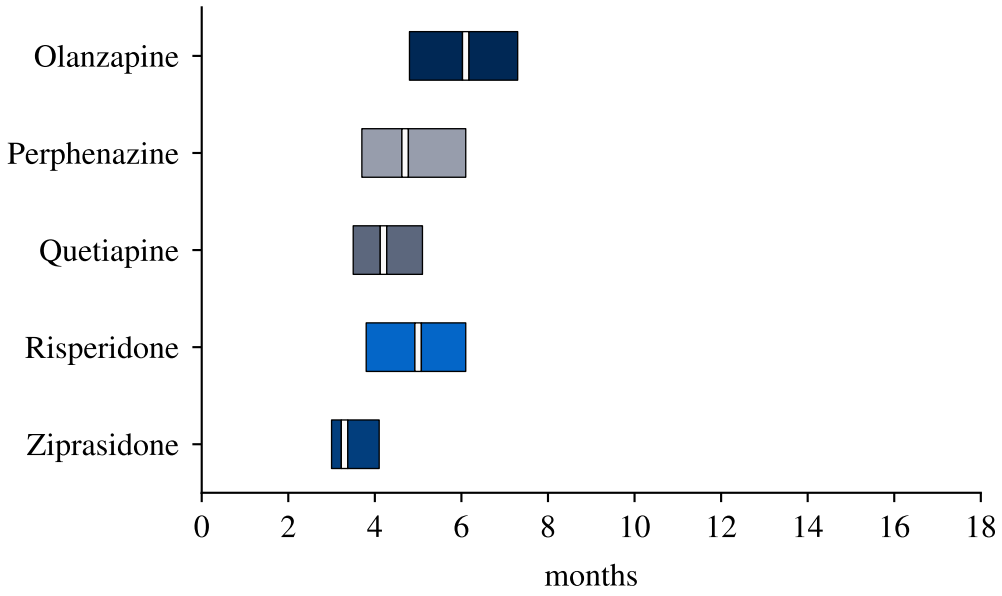


Figure 7.11: Estimated cluster-based median survival times with 90%-confidence intervals for patients in cluster Cl_{O-2}

Furthermore, the cluster-based restricted mean survival time of Olanzapine is only marginally higher compared with the other treatments:

$$\widehat{RMST}^{\,O}_{O-2} = 8.6 > 7.2 = \widehat{RMST}^{\,P}_{O-2}$$
$$\widehat{RMST}^{\,O}_{O-2} = 8.6 > 6.8 = \widehat{RMST}^{\,Q}_{O-2}$$
$$\widehat{RMST}^{\,O}_{O-2} = 8.6 > 8.0 = \widehat{RMST}^{\,R}_{O-2}$$
$$\widehat{RMST}^{\,O}_{O-2} = 8.6 > 6.4 = \widehat{RMST}^{\,Z}_{O-2}$$

Therefore patients in cluster Cl_{O-2} do not discontinue their treatment with Olanzapine substantially later compared with all other treatments.

The observations are strengthened by the analysis of the respective cluster-based survival models. Applying Procedure 6.3 on cluster Cl_{O-2} with level of confidence $\alpha = 0.1$ we can only observe the hazard ratio between Olanzapine and Quetiapine to be less than one with the respective 90%-confidence interval not including 1. Even though all other cluster-based hazard ratios are also less than one, the difference is marginal and especially all 90%-confidence intervals include 1:

$$\widehat{HR}^{\,OP}_{O-2} = 0.85 \text{ with } \mathtt{I}^{0.1}\left(HR^{\,OP}_{O-2}\right) = [0.69, 1.06]$$
$$\widehat{HR}^{\,OQ}_{O-2} = 0.69 \text{ with } \mathtt{I}^{0.1}\left(HR^{\,OQ}_{O-2}\right) = [0.57, 0.84]$$
$$\widehat{HR}^{\,OR}_{O-2} = 0.96 \text{ with } \mathtt{I}^{0.1}\left(HR^{\,OR}_{O-2}\right) = [0.79, 1.16]$$
$$\widehat{HR}^{\,OZ}_{O-2} = 0.92 \text{ with } \mathtt{I}^{0.1}\left(HR^{\,OZ}_{O-2}\right) = [0.71, 1.19]$$

Consequently in cluster Cl_{O-2} Olanzapine can not be associated with a lower hazard rate compared with all other treatments - the opposite of the observation in cluster Cl_{O-1}. Furthermore the cluster-based hazard ratios were higher for all four treatments compared with the common survival analysis:

$$\widehat{HR}^{\,OP}_{O-2} = 0.85 > 0.75 = \widehat{HR}^{\,OP}$$

$$\widehat{HR}_{O-2}^{OQ} = 0.69 > 0.60 = \widehat{HR}^{OQ}$$
$$\widehat{HR}_{O-2}^{OR} = 0.96 > 0.75 = \widehat{HR}^{OR}$$
$$\widehat{HR}_{O-2}^{OZ} = 0.92 > 0.74 = \widehat{HR}^{OZ}$$

Therefore, patients in cluster Cl_{O-2} show a much higher risk than initially suggested by the common survival analysis of discontinuing the treatment with Olanzapine. The risk of discontinuing Olanzapine is similar to the risk of discontinuing any of the other treatments.
The estimated cluster-based hazard ratios of cluster Cl_{O-2} in comparison to the estimated hazard ratios from the common survival analysis as well as the estimated cluster-based hazard ratios of cluster Cl_{O-1} are displayed in Figure 7.10.

To summarize, the results from the common survival analysis were strongly misleading due to undetected underlying heterogeneity inside the patient data. The discontinuation of treatment with Olanzapine was simultaneously over- and underestimated. Patients with those baseline characteristics defined by Cl_{O-1} have a much lower risk of discontinuing their treatment with Olanzapine compared with any of the other treatments. Olanzapine is associated with much later discontinuation of the treatment for patients living in stable conditions due to no hospitalizations in the past year, no anxiety disorder in the past month, and no drug abuse in the past five years. The risk of discontinuing the treatment is much lower than initially proposed by the common survival analysis.
On the other hand patients with those baseline characteristics defined by Cl_{O-2} do not have a lower risk of discontinuing their treatment with Olanzapine compared with any of the other treatments. Therefore, Olanzapine is not associated with later discontinuation of the treatment, if the patient does not experience stable conditions. Thus the initially proposed suggestion, that Olanzapine is related to later discontinuation of the treatment for all patients, does not hold for patients in cluster Cl_{O-2}.

Perphenazine versus Risperidone

Up to now the observed differences all included the treatment, on which the clustering was derived on (i.e. Olanzapine). Besides a thorough analysis of the results including Olanzapine, it is also important to take a look at the other medications and especially possible consequences in comparison to the common survival analysis. In the following we will therefore describe observed differences between two other treatments, namely Perphenazine and Risperidone, while maintaining the clustering derived on Olanzapine.
The common survival analysis in Section 7.2 did suggest, that Perphenazine and Risperidone are both associated with equally early discontinuation of the treatment. Their estimated survival functions and the estimated median survival times were nearly identical and the estimated hazard ratio was close to 1. We will compare Perphenazine and Risperidone with each other separately for patients in cluster Cl_{O-1} and cluster Cl_{O-2}. Besides highlighting differences, we will also compare the results to the common survival analysis.

The cluster-based survival analysis revealed, that patients with those baseline characteristics defined by cluster Cl_{O-1} discontinue the treatment with Perphenazine later compared with Risperidone. On the other hand patients with the baseline characteristics defined by cluster Cl_{O-2} tend to discontinue the treatment with Risperidone later compared with Perphenazine. The estimated survival functions for patients in cluster Cl_{O-1} and cluster Cl_{O-2} are shown in Figure 7.12 together with the estimated survival functions of the common survival analysis representing all patients without division into clusters. The estimated cluster-based survival function of patients receiving Perphenazine in cluster Cl_{O-1} lies above the estimated survival function of patients receiving Risperidone in cluster Cl_{O-1}. The estimated cluster-based survival functions of patients receiving Perphenazine and Risperidone in cluster Cl_{O-2} lie very close together in the beginning. However after 6 months of treatment, the estimated cluster-based survival function of patients receiving Risperidone lies above the estimated cluster-based survival function of patients receiving Perphenazine.

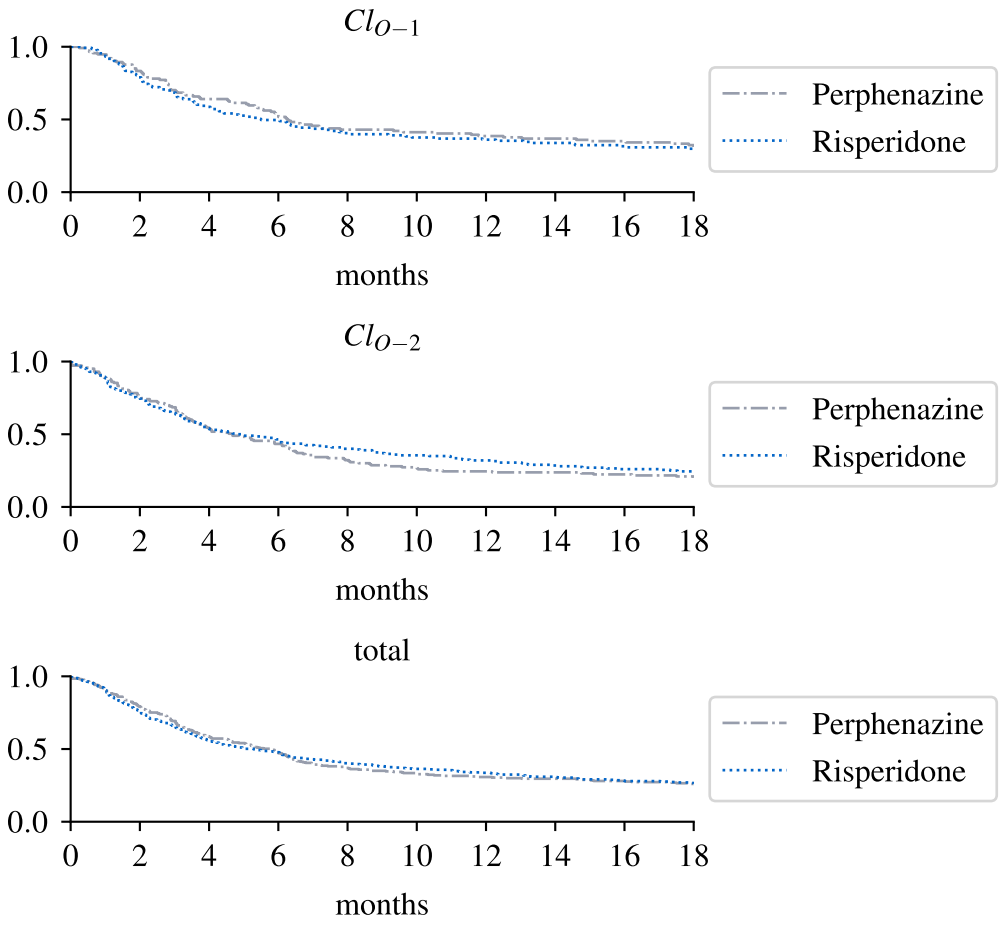


Figure 7.12: Estimated cluster-based survival functions for Risperidone and Perphenazine for patients in cluster Cl_{O-1}, cluster Cl_{O-2}, and in the common survival analysis

The cluster-based median survival times of patients receiving Perphenazine and Risperidone in cluster Cl_{O-1} and Cl_{O-2} are quite similar to each other

$$\widehat{M}^{P}_{O-1} = 6.2$$
$$\widehat{M}^{R}_{O-1} = 5.5$$

and

$$\widehat{M}^{P}_{O-2} = 4.7$$
$$\widehat{M}^{R}_{O-2} = 5.0$$

with their respective 90%-confidence intervals overlapping:

$$\mathtt{I}^{0.1}\left(M^{P}_{O-1}\right) \cap \mathtt{I}^{0.1}\left(M^{R}_{O-1}\right) = [5.4, 9.3] \cap [4.1, 7.3] \neq \emptyset$$

and

$$\mathtt{I}^{0.1}\left(M^{P}_{O-2}\right) \cap \mathtt{I}^{0.1}\left(M^{R}_{O-2}\right) = [3.7, 6.1] \cap [3.8, 6.1] \neq \emptyset$$

However, the estimated cluster-based hazard ratios reveal substantial differences between Perphenazine and Risperidone. The estimated hazard ratio between Risperidone and Perphenazine from the common survival analysis was very close to 1, i.e.

$$\widehat{HR}^{PR} = 0.95$$

with the respective 90%-confidence interval

$$\mathtt{I}^{0.1}\left(HR^{PR}\right) = [0.80, 1.11]$$

clearly including 1.

The estimated cluster-based hazard ratio exposes later discontinuation of treatment with Perphenazine in cluster Cl_{O-1} compared with the treatment with Risperidone

$$\widehat{HR}_{O-1}^{PR} = 0.71$$

with the respective 90%-confidence interval

$$\mathtt{I}^{0.1}\left(HR_{O-1}^{PR}\right) = [0.55, 0.94]$$

not including 1.
On the other hand, the estimated cluster-based hazard ratio discloses later discontinuation of treatment with Risperidone in cluster Cl_{O-2} compared with the treatment with Perphenazine

$$\widehat{HR}_{O-2}^{PR} = 1.15$$

with respective 90%-confidence interval

$$\mathtt{I}^{0.1}\left(HR_{O-2}^{PR}\right) = [0.93, 1.43]$$

including 1. Even though $\mathtt{I}^{0.1}\left(HR_{O-2}^{PR}\right)$ does include one, we can still observe different behaviours in both clusters.
The estimated cluster-based hazard ratios are shown in Figure 7.13 together with the estimated hazard ratio from the common survival analysis.

To summarize, the common survival analysis failed to detect any difference between Perphenazine and Risperidone due to underlying heterogeneity inside the patient data. However, patients with those baseline characteristics defined by cluster Cl_{O-1} discontinue the treatment with Perphenazine later than the treatment with Risperidone. Therefore, the second-generation antipsychotic Perphenazine is associated with later discontinuation of the treatment for patients living in stable conditions due to no hospitalizations in the past year, no anxiety disorder in the past month, and no drug abuse in the past five years. On the other hand, patients with those baseline characteristics defined by cluster Cl_{O-2} discontinue the treatment with Risperidone later.

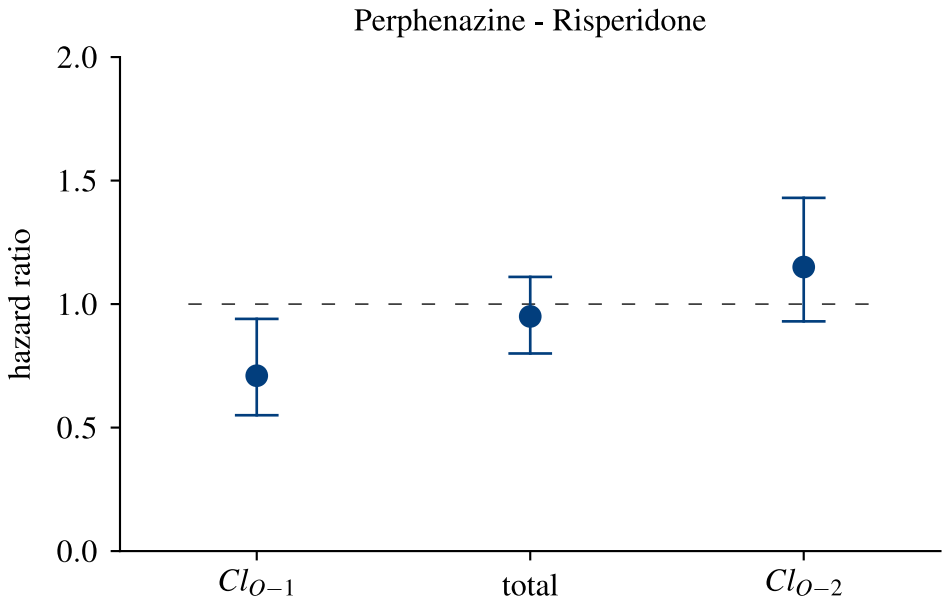


Figure 7.13: Estimated hazard ratio and cluster-based hazard ratios between Risperidone and Perphenazine for patients in cluster Cl_{O-1} and Cl_{O-2}

Therefore, the first-generation antipsychotic Risperidone is associated with later discontinuation of the treatment, if the patient does not experience stable conditions.

7.4.2 Justification across clusters

To conclude the analysis of the CATIE data, we want to justify the initial division into clusters of the patient data originating from patients receiving Olanzapine. Therefore we will directly compare cluster Cl_{O-1} and Cl_{O-2} with each other for patients receiving Olanzapine. The previous analysis already indicates differences between the two clusters. However, we want to emphasize the results by directly comparing them with each other as well as the cluster-based results with the results suggested by the common survival analysis.

The common survival analysis found Olanzapine to be associated with late discontinuation of the treatment. The cluster-based survival analysis in Section 7.3 indicates, that patients in cluster Cl_{O-1} discontinue their treatment with Olanzapine even later

than originally suggested. The estimated survival function and cluster-based survival functions are displayed in Figure 7.14.

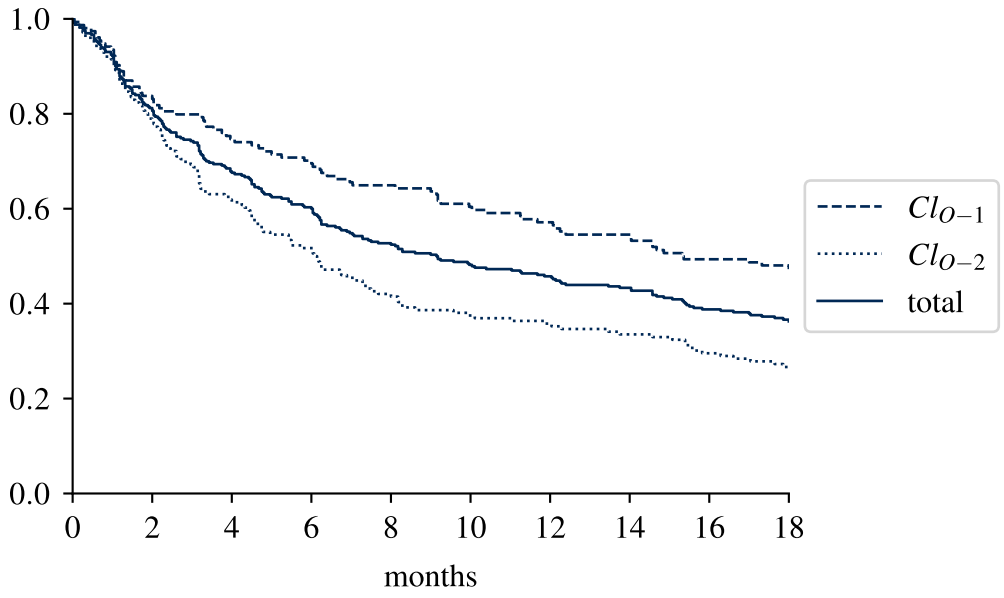


Figure 7.14: Estimated cluster-based survival functions for patients receiving Olanzapine in cluster Cl_{O-1}, Cl_{O-2}, and patients receiving Olanzapine in general

The estimated cluster-based survival function of patients receiving Olanzapine in cluster Cl_{O-1} clearly lies above the estimated cluster-based survival function of patients receiving Olanzapine in cluster Cl_{O-2} and the estimated survival function of patients receiving Olanzapine of the common survival analysis. By applying Procedure 6.4 and Procedure 6.5 on treatment Olanzapine with level of confidence $\alpha = 0.1$ we can observe the estimated cluster-based median survival time of cluster Cl_{O-1} to be larger than the cluster-based median survival time of cluster Cl_{O-1} and the estimated median survival time proposed by the common survival analysis

$$\widehat{M}^{O}_{O-1} = 15.3 > 6.1 = \widehat{M}^{O}_{O-2}$$
$$\widehat{M}^{O}_{O-1} = 15.3 > 9.1 = \widehat{M}^{O}$$

with non-overlapping 90%-confidence intervals

$$\mathtt{I}^{0.1}\left(M^{O}_{O-1}\right) \cap \mathtt{I}^{0.1}\left(M^{O}_{O-2}\right) = [12.1, \infty[\cap [4.8, 7.3] = \emptyset$$
$$\mathtt{I}^{0.1}\left(M^{O}_{O-1}\right) \cap \mathtt{I}^{0.1}\left(M^{O}\right) = [12.1, \infty[\cap [7.0, 12.0] = \emptyset$$

as displayed in Figure 7.15.

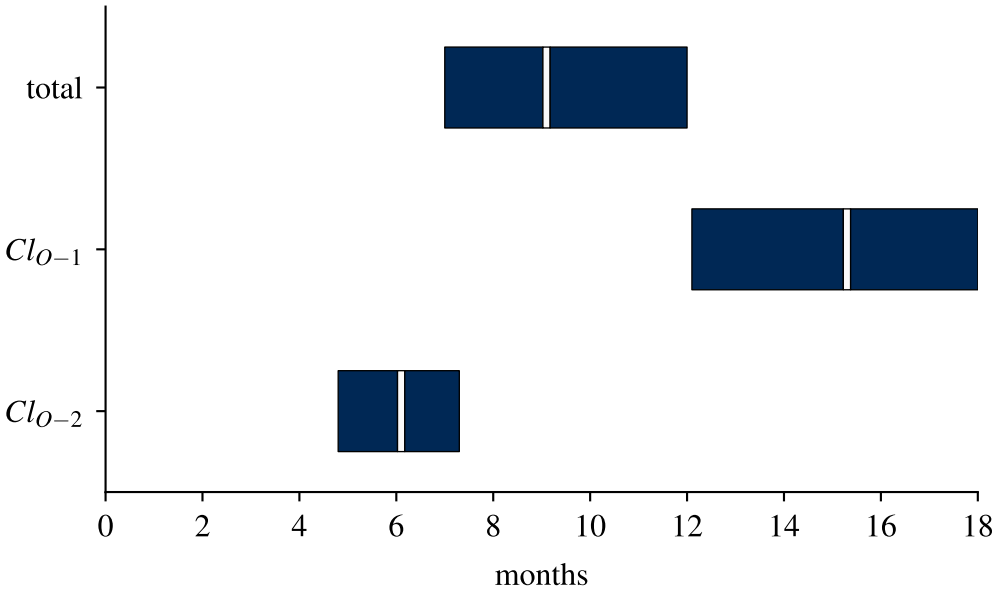


Figure 7.15: Estimated cluster-based median survival times with 90%-confidence intervals for patients receiving Olanzapine in cluster Cl_{O-1}, Cl_{O-2}, and patients receiving Olanzapine general

The cluster-based survival analysis in Section 7.3 also indicates, that patients in cluster Cl_{O-2} discontinue their treatment with Olanzapine much earlier than originally suggested. We can observe the estimated cluster-based median survival time of cluster Cl_{O-2} to be less than the estimated cluster-based median survival time of cluster Cl_{O-1} and the estimated median survival time proposed by the common

survival analysis

$$\widehat{M}^{O}_{O-2} = 6.1 < 15.3 = \widehat{M}^{O}_{O-1}$$
$$\widehat{M}^{O}_{O-2} = 6.1 < 9.1 = \widehat{M}^{O}$$

with non-overlapping 90%-confidence interval between clusters Cl_{O-1} and Cl_{O-2}

$$\mathtt{I}^{0.1}\left(M^{O}_{O-2}\right) \cap \mathtt{I}^{0.1}\left(M^{O}_{O-1}\right) = [4.8, 7.3] \cap [12.1, \infty[= \emptyset$$

and hardly overlapping confidence interval between cluster Cl_{O-2} and the common survival analysis

$$\mathtt{I}^{0.1}\left(M^{O}_{O-2}\right) \cap \mathtt{I}^{0.1}\left(M^{O}\right) = [4.8, 7.3] \cap [7.0, 12.0] \neq \emptyset$$

as displayed in Figure 7.15.

To summarize, the division of the patient data originating from patients receiving Olanzapine was justified. Patients with those baseline characteristics defined by cluster Cl_{O-1} discontinue their treatment substantially later than patients receiving Olanzapine in general and patients with those baseline characteristics defined by cluster Cl_{O-2}. The cluster-based survival function of Cl_{O-1} clearly lies above both the estimated cluster-based survival function of Cl_{O-2} and the estimated survival function of Olanzapine proposed by the common survival analysis. The confidence intervals of the estimated median survival times even show a large gap. Therefore Olanzapine is only associated with late discontinuation of the treatment for patients living in stable conditions due to no hospitalizations in the past year, no anxiety disorder in the past month and no drug abuse in the past five years. Opposed to that, patients in cluster Cl_{O-2} can be found to be associated with substantially earlier discontinuation of the treatment. Therefore Olanzapine can not be associated with late discontinuation of the treatment for patients not experiencing stable conditions. Heterogeneity inside the patient data prevented revealing these associations during a common survival analysis. The cluster-based survival analysis therefore contradicts

the initial results of the common survival analysis suggesting Olanzapine to always be related to late discontinuation of the treatment. On the other hand heterogeneity inside the patient data also prevented identifying a cluster, i.e. subgroup, of patients, who do indeed discontinue their treatment with Olanzapine much later than initially suggested.

7.5 Conclusion CATIE

By applying the new method for explainable analytics presented in Part III we were able to identify hidden structures inside the individual patient data in order to answer the posed research question. The explainable method allowed to generate new insight into the CATIE data, in order to improve the treatment for future patients suffering from schizophrenia. Furthermore, heterogeneity inside the patient data was not only detected, but overcome by the division into clusters.

According to the common survival analysis presented in Section 7.2, Olanzapine seemed to be associated with later discontinuation of the treatment compared with all other treatments. The cluster-based survival analysis detected that Olanzapine was simultaneously over- and underestimated due to underlying heterogeneity inside the patient data. Olanzapine is only associated with late discontinuation of the treatment for patients experiencing stable conditions due to no hospitalizations in the past year, no anxiety disorder in the past month, and no drug abuse in the past five years. Once these stable conditions are not met, Olanzapine is associated with equally early discontinuation of the treatment as any of the other treatments. Considering that further decisions about the care of patients are made based on clinical trials like CATIE, it is very dangerous to disregard heterogeneity inside the patient data. The sole use of average models like in the common survival can be strongly misleading. Physicians might refrain from prescribing Olanzapine, even though the patient does live in stable conditions. On the other hand patients not living in stable conditions might not expect that they have to discontinue the treatment, e.g. due to adverse events, even though there is still a high risk, even while administering Olanzapine.

Furthermore, it is crucial to explain and understand the detected differences between different groups of patients regarding the discontinuation of the treatment. By applying the new method for explainable analytics we were not only able to reveal those differences, but to explain why those predictions were made in regards of the respective patient characteristics. A non-explainable method could only detect this subgroup, but would not be able to explain the difference based on baseline characteristic of the patients.

The common survival analysis also failed to detect any difference between Perphenazine and Risperidone. Due to reduced heterogeneity during the cluster-based survival analysis, we were able to identify, that patients living in stable conditions discontinue the treatment with Perphenazine later compared with Risperidone. On the other hand, patients discontinue the treatment with Risperidone later than the treatment with Perphenazine, if those stable conditions are not met.

To conclude, while regarding large data sets from clinical trials, there is usually undetected heterogeneity distorting the results from common methods, like the common survival analysis. Patients enrolled in clinical trials (intentionally) differ regarding their characteristics. Hence, generating average results for all patients is often not reasonable. By applying the new explainable method for the analysis of individual patient data, we were able to address this heterogeneity. Furthermore, all detected differences were explainable regarding the respective patient characteristics providing the possibility to adopt them in clinical practice.

Part V

Conclusion

In this thesis we introduced a new explainable method for the analysis of individual patient data to generate new findings for an improved future patient care. Furthermore, we presented and discussed another explainable method for medical data analytics. In particular, the methods provide reliable data for the health economic evaluation of medical interventions and evidence-based medicine in general. The approaches are able to evaluate and especially predict the outcome of a patient solely based on the patient's baseline characteristics. In addition to the prediction itself, the methods provide a unique explanation for every prediction, providing the possibility of adopting them into clinical practice. Moreover, the introduced approaches address underlying heterogeneity inside the patient data by forming homogeneous clusters of patients with similar combinations of their characteristic values. The underlying assumption is, that patients inside a cluster show similar outcomes, whereas the outcome varies across clusters. Hence, besides providing explainable predictions, the methods furthermore address an often encountered problem while analyzing patient data.

The approaches are based on an endpoint-oriented clustering approach, developed by Brieden and Gritzmann. In Chapter 4, besides the mathematical background to form clusters of patients with similar combinations of their characteristic values, we discussed a transformation technique to include all baseline characteristics regardless the level of scale. Based on the transformation technique, we presented a new automated approach to select the baseline characteristics promising the most impact on the outcome of a patient. Furthermore, we introduced a new possibility to classify variables of ordinal and cardinal scale in an optimal way as alternative to a common quantile classification. The transformation technique, the automated selection of variables, and the optimized classification were used in the following methods.

First, we presented and discussed the cluster-based analysis of individual patient data. The cluster-based analysis was then used as foundation in the newly developed method of cluster-based survival analysis. We applied the new method to generate new findings for a clinical trial including patients suffering from schizophrenia.

Cluster-based analysis

In Chapter 5 we presented the cluster-based analysis of individual patient data. We discussed how to predict the outcome of a patient in the clusters identified by the endpoint-oriented clustering approach. To evaluate the reliability of the predictions, we presented confidence intervals for the outcome of patients in a cluster. We furthermore discussed, how the assumption about different outcomes across clusters can be justified by the presented confidence intervals.
The approach serves as method itself, as well as foundation for the following newly developed cluster-based survival analysis

Cluster-based survival analysis

In Chapter 6 we introduced the cluster-based survival analysis of individual patient data to estimate the time until an event of interest happens, e.g. time until discontinuation of a treatment. The main assumption is, that the time to discontinuation of a treatment is not only influenced by the treatment, but also by patient characteristics. We suggested, that patients inside a cluster derived by the endpoint-oriented clustering show similar time to discontinuation, whereas the time to discontinuation varies across clusters. Therefore we adjusted the common survival analysis presented in Chapter 3. We introduced cluster-based survival functions and the estimation for them. To evaluate their reliability, we presented confidence intervals as well as the cluster-based median and restricted mean survival time. Furthermore, we introduced cluster-based survival models in order to estimate the cluster-based hazard ratio between two treatments with respective confidence intervals. Moreover, we discussed, how the assumption about different outcomes across treatments and clusters can be justified by the presented confidence intervals.

As practical application, we analyzed the data originating from the clinical trial CATIE. The results were compared with those of the common survival analysis presented in the beginning of the respective chapter. We were able to identify hidden structures inside the patient data in order to answer the posed research question.

According to the common survival analysis, Olanzapine seemed to be associated with later discontinuation of the treatment compared with all other treatments. We were able to detect, that Olanzapine was simultaneously over- and underestimated due to underlying heterogeneity inside the patient data. The cluster-based survival analysis only found Olanzapine to be related to late discontinuation for patients experiencing stable conditions due to no recent hospitalizations, no anxiety disorder, and no drug abuse. Once these stable conditions are not met, Olanzapine is related to equally early discontinuation as the other treatments.
The common survival analysis failed to detect any differences between other pairs of treatments. By applying the cluster-based survival analysis, we were able to identify, that patients experiencing stable conditions discontinue the treatment with Perphenazine later compared with Risperidone. Once those stable conditions are not met, the effect is reversed.
Therefore, the new explainable method allowed to generate new insight into the CATIE data in order to improve the treatment for future patients suffering from schizophrenia. Thereby, heterogeneity inside the patient data was not only detected, but overcome by the division into clusters. Moreover, these new predictors were new for the CATIE data, even though it was made available for numerous researchers, as well as the treatment of patients suffering from schizophrenia in general. Especially the caution regarding anxious patients was undetected up to now and will hopefully improve future patient care.

Outlook

By applying the new explainable method for the analysis of individual patient data, we were able to generate new findings for a data set from a clinical trial. Furthermore, the method provides the possibility to generate reliable data for the health economic evaluation of medical interventions. We hope that future work regarding explainable analytics in general and in particular in medicine will result in more findings, further improving the future care of patients suffering from schizophrenia as well as other diseases. Therefore, we want to give some suggestions for possible future work.

All approaches are based on the endpoint-oriented clustering approach, consisting of the transformation of the data, the selection of variables, and the clustering itself. Regarding the transformation of the data and the related selection of variables, the concept of crossed variables shows potential for further investigations. Some variables might only show potential for predicting the outcome while regarded simultaneously. Due to the high amount of baseline characteristics available in many medical data sets, they might not be selected while regarding the one-dimensional expected values.
Furthermore, there are numerous adjustable parameters in the clustering itself, like the lower and upper bounds and the amount of clusters, allowing for further investigations.

The cluster-based analysis used exact confidence intervals to evaluate the predictions of the outcome. In particular, we used Clopper-Pearson intervals for the binomial proportion, as they presented an appropriate choice for this thesis. There are however other suggestions to be exploited, especially regarding the trade off between accuracy and reliability.
Furthermore, the effect of placebo patients in a cluster-based analysis poses an interesting topic for future work. We suggest investigating the possibility of clustering placebo patients and treatment patients separately and performing the cluster-based analysis on subsequently derived crossed treatment-placebo clusters.

We used the semi-parametric Cox proportional-hazards model as cluster-based survival model to estimate the hazard rates and hazard ratios. The examination of other models, both parametric and semi-parametric, provides potential for future investigations, especially if additional knowledge about the distribution of the survival times is available.

Last but not least, we strongly want to motivate future work into explainable analytics in general and especially in medicine. There are numerous advanced statistical methods showing potential, that have been overlooked due to the popularity of

machine learning approaches. Further investigations into those methods can lead to new findings for decision-making in general and in particular in medicine.

Appendix

A. CATIE

Additional material from the analysis of the CATIE data presented in chapter 7.

characteristic	values	class number	class
diagnosis	schizophrenia, schizophreniform disorder, schizoaffective disorder, bipolar disorder. major depression, psychosis NOS		
hospitalized in past 3 months	yes, no		
mental illness in family	yes, no		
no. of hospitalizations past year	$0, 1, 2, 3 > 4$	1 2	0 ≥ 1
no. of hospitalizations lifetime	$0, 1, 2, 3 > 4$	1 2 3	$[0, 1, 2]$ 3 ≥ 4
pre-medication	all treatments		
schizophrenia present in past month	yes, no		

Table A.1: Diagnostic and hospital-related characteristics in CATIE with respective values for patients receiving Olanzapine

characteristic	values	class number	class
PANSS	34 – 132	1	[34, 68]
		2	[69, 83]
		3	[84, 132]
PANSS-P positive subscale	7 – 35	1	[7, 16]
		2	[17, 21]
		3	[22, 35]
PANSS-N negative subscale	7 – 41	1	[7, 17]
		2	[18, 23]
		3	[24, 41]
PANSS-G general subscale	17 – 68	1	[17, 32]
		2	[33, 41]
		3	[42, 68]
PANSS Marder positive symptoms	9 – 46	1	[9, 21]
		2	[22, 28]
		3	[29, 46]
PANSS Marder negative symptoms	7 – 40	1	[7, 16]
		2	[17, 22]
		3	[23, 40]
PANSS Marder disorganized thought	10 – 34	1	[10, 14]
		2	[15, 19]
		3	[20, 34]
PANSS Marder uncontrolled hostility/excitement	4 – 16	1	[4, 5]
		2	[6, 8]
		3	[9, 16]
PANSS Marder anxiety/depression	4 – 21	1	[4, 9]
		2	[10, 12]
		3	[13, 21]

Table A.2: PANSS characteristics in CATIE with respective values for patients receiving Olanzapine

characteristic	values
chronic obstructive pulmonary disease	yes, no
diabetes (type I or II)	yes, no
hyperlipidemia	yes, no
hypertension	yes, no
osteoarthritis	yes, no
tardive dyskinesia	yes, no
major depression past month	yes, no
major depression in past 5 years	yes, no
alcohol dependence in past month	yes, no
alcohol dependence in past 5 years	yes, no
alcohol abuse in past month	yes, no
alcohol abuse in past 5 years	yes, no
drug dependence in past month	yes, no
drug dependence in past 5 years	yes, no
OCD in past month	yes, no
OCD in past 5 years	yes, no
anxiety disorder in past month	yes, no
anxiety disorder in past 5 years	yes, no
drugs in hair (cocaine, opiates, meth, THC)	yes, no

Table A.3: Characteristics related to other diseases in CATIE with respective values for patients receiving Olanzapine

characteristic	values	class number	class
Abnormal Involuntary Movement Scale AIMS, item 8	$0-5$	1 2	0 ≥ 1
Barnes Akathisia Rating Scale BARS global	$0-6$	1 2 3	0 1 ≥ 2
Calgary Depression Scale for Schizophrenia CLGRY	$0-22$	1 2 3	$[0,2]$ $[3,6]$ $[7,22]$
Insight & Treatment Attitudes Questionnaire ITAQ	$1-22$	1 2 3	$[1,18]$ $[19,21]$ 22
quality of life global score	$1-7$		
Simpson-Angus-Scale SAS	$0-14$	1 2	0 ≥ 1

Table A.4: Characteristics from other scales in CATIE with respective values for patients receiving Olanzapine

characteristic	values	class number	class
age	$18-65$	1 2 3	$[18,36]$ $[37,46]$ $[47,65]$
employment status	unemployed, full-time, part-time, retired		
gender	female, male		
living with significant other	yes, no		
married	yes, no		
race	American Indian/ Alaska Native, Asian, Black or African American, Hawaiian or Pacific Islander White	1 2	not white white

Table A.5: Characteristics from demographics in CATIE with respective values for patients receiving Olanzapine

characteristic	values	class number	class
BMI	$16.68 - 52.64$	1	$[16.68, 25.82]$
		2	$[25.83, 31.52]$
		3	$[31.53, 52.64]$
heart rate	$50 - 132$	1	$[50, 72]$
		2	$[73, 82]$
		3	$[83, 132]$
vital diastolic	$50 - 118$	1	$[50, 80]$
		2	$[81, 118]$
vital systolic	$81 - 188$	1	$[82, 118]$
		2	$[119, 130]$
		3	$[131, 188]$

Table A.6: Characteristics from vital signs in CATIE with respective values for patients receiving Olanzapine

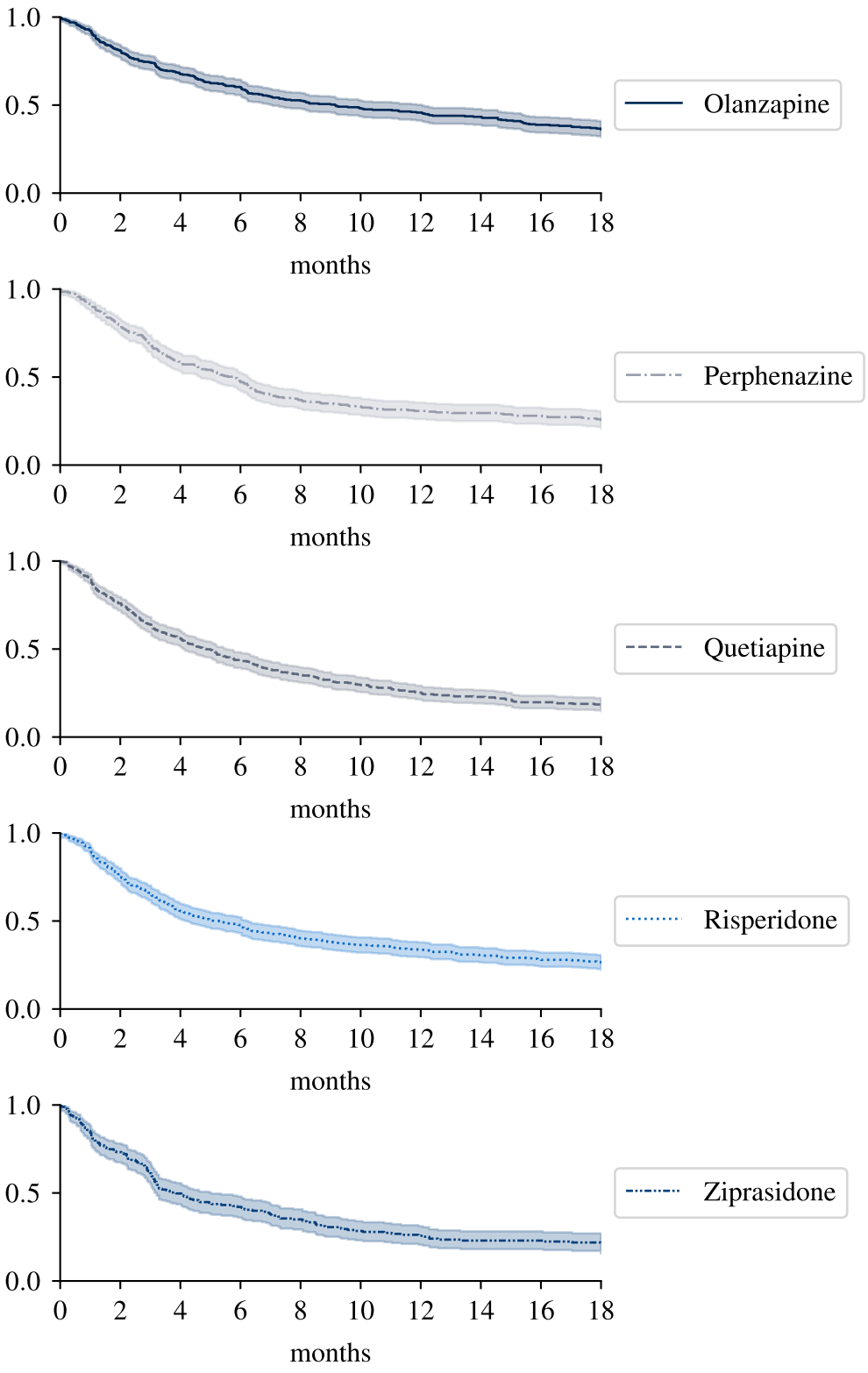


Figure A.1: Separate estimated survival functions for every treatment in the common survival analysis with 90%-confidence intervals

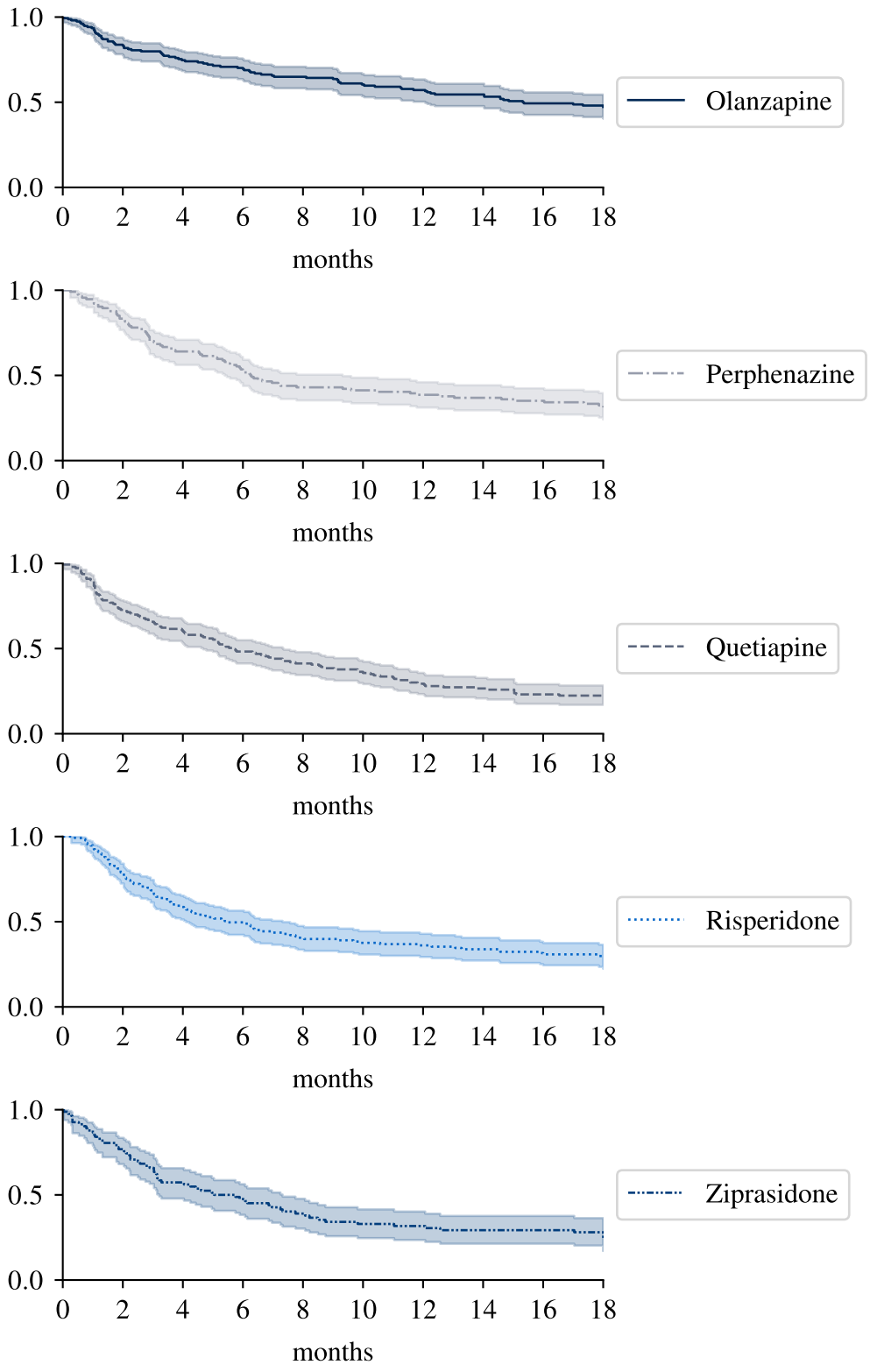


Figure A.2: Separate estimated cluster-based survival functions for patients in cluster Cl_{O-1} with 90%-confidence intervals

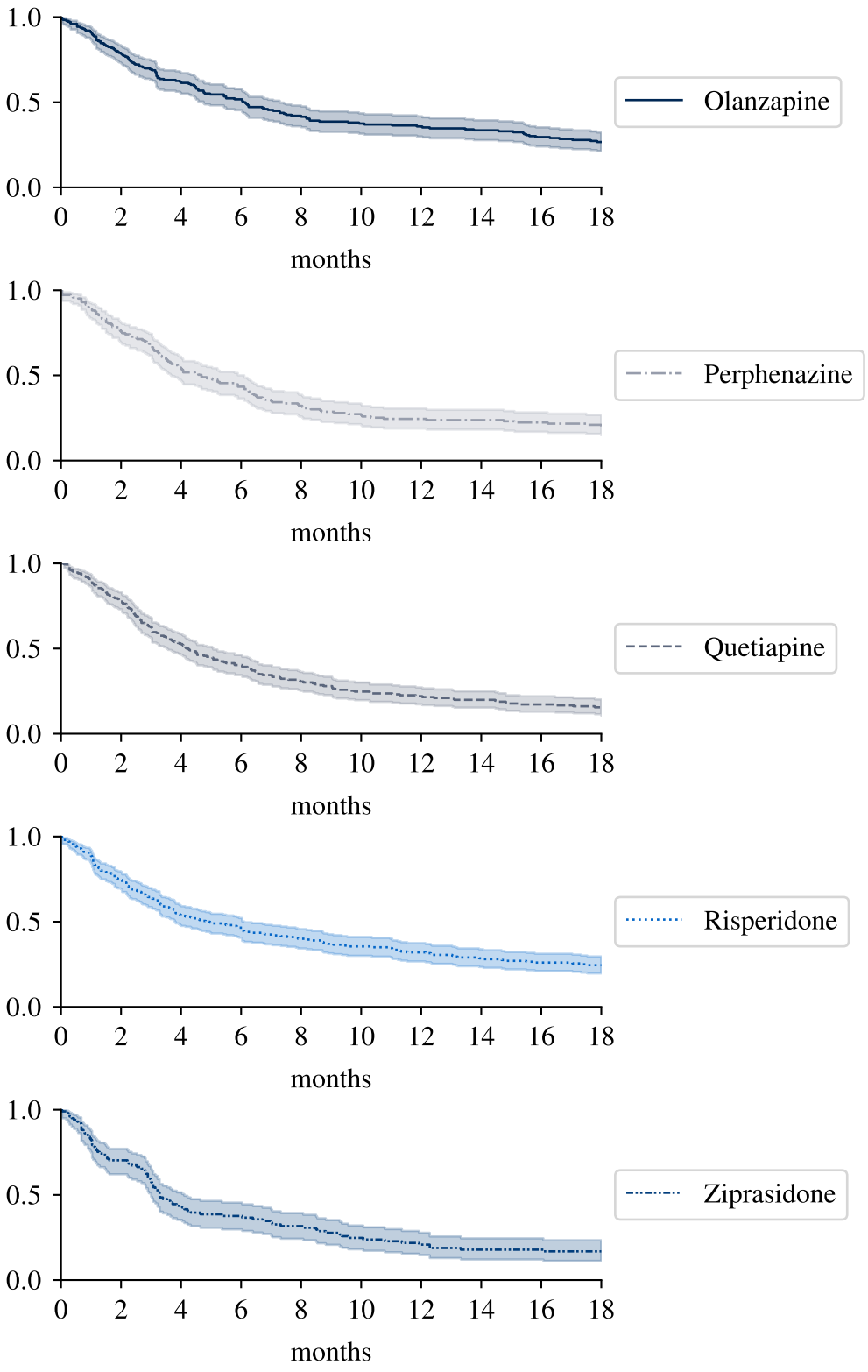


Figure A.3: Separate estimated cluster-based survival functions for patients in cluster Cl_{O-2} with 90%-confidence intervals

comparison $t\,\tilde{t}$	estimated hazard ratio $\widehat{HR}^{t\,\tilde{t}}$	90%-confidence interval $\mathtt{I}^{0.1}\left(\widehat{HR}^{t\,\tilde{t}}\right)$
Olanzapine - Perphenazine	0.75	$[0.64, 0.89]$
Olanzapine - Quetiapine	0.60	$[0.52, 0.70]$
Olanzapine - Risperidone	0.75	$[0.65, 0.88]$
Olanzapine - Ziprasidone	0.74	$[0.61, 0.89]$
Perphenazine - Quetiapine	0.89	$[0.76, 1.05]$
Perphenazine - Risperidone	0.95	$[0.80, 1, 11]$
Perphenazine - Ziprasidone	0.87	[0.70.1.07]
Quetiapine - Risperidone	1.2	$[1.04, 1.38]$
Quetiapine - Ziprasidone	0.98	$[0.81, 1.17]$
Risperidone - Ziprasidone	0.84	$[0.69, 1.02]$

Table A.7: Estimated hazard ratios between all pairs of treatments in the common survival analysis

comparison $t\,\tilde{t}$	estimated hazard ratio $\widehat{HR}_{O-1}^{t\,\tilde{t}}$	90%-confidence interval $\mathtt{I}^{0.1}\left(\widehat{HR}_{O-1}^{t\,\tilde{t}}\right)$
Olanzapine - Perphenazine	0.76	$[0.58, 1.01]$
Olanzapine - Quetiapine	0.48	$[0.37, 0.61]$
Olanzapine - Risperidone	0.59	$[0.45, 0.75]$
Olanzapine - Ziprasidone	0.65	$[0.47, 0.89]$
Perphenazine - Quetiapine	0.84	$[0.64, 1.09]$
Perphenazine - Risperidone	0.71	$[0.55, 0.94]$
Perphenazine - Ziprasidone	0.84	$[0.60, 1.20]$
Quetiapine - Risperidone	1.14	$[0.91, 1.43]$
Quetiapine - Ziprasidone	0.99	$[0.73, 1.35]$
Risperidone - Ziprasidone	0.85	$[0.60, 1.22]$

Table A.8: Estimated hazard ratios between all pairs of treatments in cluster Cl_{O-1}

comparison $t\,\tilde{t}$	estimated hazard ratio $\widehat{HR}_{O-2}^{t\,\tilde{t}}$	90%-confidence interval $\mathtt{I}^{0.1}\left(\widehat{HR}_{O-2}^{t\,\tilde{t}}\right)$
Olanzapine - Perphenazine	0.85	$[0.69, 1.06]$
Olanzapine - Quetiapine	0.69	$[0.57, 0.84]$
Olanzapine - Risperidone	0.96	$[0.79, 1.16]$
Olanzapine - Ziprasidone	0.92	$[0.71, 1.19]$
Perphenazine - Quetiapine	0.88	$[0.71, 1.09]$
Perphenazine - Risperidone	1.15	$[0.93, 1.43]$
Perphenazine - Ziprasidone	0.78	$[0.59, 1.04]$
Quetiapine - Risperidone	1.27	$[1.05, 1.54]$
Quetiapine - Ziprasidone	0.87	$[0.67, 1.13]$
Risperidone - Ziprasidone	0.83	$[0.63, 1.07]$

Table A.9: Estimated hazard ratios between all pairs of treatments in cluster Cl_{O-2}

comparison $t\tilde{t}$	estimated hazard ratio in model (7.2) $\widehat{HR}_{O-1}^{t\tilde{t}}$	estimated hazard ratio in model (7.3) $\widehat{HR}_{O-1}^{t\tilde{t}}$
Olanzapine - Perphenazine	0.76	0.78
Olanzapine - Quetiapine	0.48	0.48
Olanzapine - Risperidone	0.59	0.59
Olanzapine - Ziprasidone	0.65	0.65
Perphenazine - Quetiapine	0.84	0.83
Perphenazine - Risperidone	0.71	0.72
Perphenazine - Ziprasidone	0.84	0.85
Quetiapine - Risperidone	1.14	1.14
Quetiapine - Ziprasidone	0.99	1.00
Risperidone - Ziprasidone	0.85	0.88

Table A.10: Estimated hazard ratios in cluster-based survival model with (7.3) and without (7.2) additional covariates in cluster Cl_{O-1}

comparison $t\tilde{t}$	estimated hazard ratio in model (7.2) $\widehat{HR}_{O-2}^{t\tilde{t}}$	estimated hazard ratio in model (7.3) $\widehat{HR}_{O-2}^{t\tilde{t}}$
Olanzapine - Perphenazine	0.85	0.85
Olanzapine - Quetiapine	0.69	0.69
Olanzapine - Risperidone	0.96	0.95
Olanzapine - Ziprasidone	0.92	0.93
Perphenazine - Quetiapine	0.88	0.88
Perphenazine - Risperidone	1.15	1.15
Perphenazine - Ziprasidone	0.78	0.79
Quetiapine - Risperidone	1.27	1.26
Quetiapine - Ziprasidone	0.87	0.87
Risperidone - Ziprasidone	0.83	0.83

Table A.11: Estimated hazard ratios in cluster-based survival model with (7.3) and without (7.2) additional covariates in cluster Cl_{O-2}

Bibliography

[1] AALEN, O. O. Two examples of modelling heterogeneity in survival analysis. *Scandinavian Journal of Statistics 14*, 1 (1987), 19–25.

[2] AALEN, O. O. Heterogeneity in survival analysis. *Statistics in Medicine 7*, 11 (1988), 1121–1137.

[3] ABOU-SETTA, A. M., MOUSAVI, S. S., SPOONER, C., SCHOUTEN, J. R., PASICHNYK, D., ARMIJO-OLIVO, S., BEAITH, A., SEIDA, J. C., DURSUN, S., NEWTON, A. S., AND HARTLING, L. *First-Generation Versus Second-Generation Antipsychotics in Adults: Comparative Effectiveness*. Agency for Healthcare Research and Quality (US), Rockville (MD), 2012.

[4] ADDINGTON, D., ADDINGTON, J., AND SCHISSEL, B. A depression rating scale for schizophrenics. *Schizophrenia Research 3*, 4 (1990), 247–251.

[5] AGRAWAL, D., BERNSTEIN, P., BERTINO, E., DAVIDSON, S., DAYAL, U., FRANKLIN, M., GEHRKE, J., HAAS, L., HALEVY, A., HAN, J., JAGADISH, H. V., LABRINIDIS, A., MADDEN, S., PAPAKONSTANTINOU, Y., PATEL, J., RAMAKRISHNAN, R., ROSS, K., SHAHABI, C., SUCIU, D., VAITHYANATHAN, S., AND WIDOM, J. Challenges and opportunities with big data 2011-1, 2011. A community white paper developed by leading researchers across the United States; Available at: https://docs.lib.purdue.edu/cctech/1 (Accessed 21 July 2020).

[6] ALPAYDIN, E. *Introduction to Machine Learning*, 4th ed. MIT Press, Cambridge, Mass., 2014.

[7] AMERICAN PSYCHIATRIC ASSOCIATION. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*, 5th ed. American Psychiatric Publishing, Washington, DC, 2013.

[8] ANDERSON, T. W., AND SCLOVE, S. L. *The Statistical Analysis of Data*, 2nd ed. The Scientific Pr, Palo Alto, CA, 1986.

[9] ANDREASEN, N. C., CARPENTER, W. T., KANE, J. M., LASSER, R. A., MARDER, AND WEINBERGER. Remission in schizophrenia: Proposed criteria and rationale for consensus. *The American Journal of Psychiatry 162*, 3 (2005).

[10] ATZMUELLER, M. Onto explicative data mining: Exploratory, interpretable and explainable analysis. *Proceedings Dutch-Belgian Database Day* (2017).

[11] BARNES, E., HOFFMAN, A., AND ROTHBLUM, U. Optimal partitions having disjoint convex and conic hulls. *Mathematical Programming 54* (1992), 69–86.

[12] BARNES, T. R. A rating scale for drug-induced akathisia. *The British Journal of Psychiatry: The Journal of Mental Science 154*, 5 (1989), 672–676.

[13] BAROCAS, S., BRADLEY, E., HONAVAR, V., AND PROVOST, F. Big data, data science, and civil rights, 2017. A Computing Community Consortium (CCC) white paper; Available at: https://arxiv.org/pdf/1706.03102 (Accessed 21 July 2017).

[14] BISHOP, C. M. *Pattern Recognition and Machine Learning*, softcover reprint of the original 1st ed. Information science and statistics. Springer New York, New York, NY, 2016.

[15] BLEULER, E. *Dementia praecox: oder Gruppe der Schizophrenien*. Frank Deuticke, Leipzig/Wien, 1911.

[16] BLEULER, E. *Dementia Praecox: Or the Group of Schizophrenias*. International Univsersities Press, New York, 1950.

[17] BODLAENDER, H. L., GRITZMANN, P., KLEE, V., AND VAN LEEUWEN, J. Computational complexity of norm-maximization. *Combinatorica 10*, 2 (1990), 203–225.

[18] BORENSTEIN, M. The case for confidence intervals in controlled clinical trials. *Controlled Clinical Trials 15*, 5 (1994), 411–428.

[19] BORGWARDT, S., BRIEDEN, A., AND GRITZMANN, P. Constrained minimum-k-star clustering and its application to the consolidation of farmland. *Operational Research 11* (2009), 1–17.

[20] BORGWARDT, S., BRIEDEN, A., AND GRITZMANN, P. Geometric clustering for the consolidation of farmland and woodland. *The Mathematical Intelligencer 36* (2014), 37–44.

[21] BORGWARDT, S., BRIEDEN, A., AND GRITZMANN, P. An LP-based k-means algorithm for balancing weighted point sets. *European Journal of Operational Research 263*, 2 (2017), 349–355.

[22] BORGWARDT, S., A. BRIEDEN, AND P. GRITZMANN. Mathematics in agriculture and forestry: Geometric clustering for land consolidation. *IFORS News 7*, 4 (2013), 10–11.

[23] BRANDSTÄTTER, E. Confidence intervals as an alternative to significance testing. *Methods of Psychological Research Online 4*, 2 (1999), 33–46.

[24] BREIMAN, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science 16*, 3 (2001), 199–231.

[25] BRIEDEN, A. Geometric optimization problems likely not contained in apx. *Discrete & Computational Geometry 28* (2002), 201–209.

[26] BRIEDEN, A., AND GRITZMANN, P. Response prediction: Gaining reliable and interpretable insight from small study data. working paper.

[27] BRIEDEN, A., AND GRITZMANN, P. A quadratic optimization model for the consolidation of farmland by means of lend-lease agreements. *Proceedings of Operations Research 2004* (2004), 324–331.

[28] BRIEDEN, A., AND GRITZMANN, P. On clustering bodies: Geometry and polyhedral approximation. *Discrete & Computational Geometry 44* (2010), 508–534.

[29] BRIEDEN, A., AND GRITZMANN, P. On optimal weighted balanced clusterings: Gravity bodies and power diagrams. *SIAM Journal on Discrete Mathematics 26* (2012).

[30] BRIEDEN, A., GRITZMANN, P., KANNAN, R., KLEE, V., LOVÁSZ, L., AND SIMONOVITS, M. Approximation of diameters: randomization doesn't help. In *Proceedings of the 39th Annual Symposium of the FOCS*, IEEE, Ed. IEEE, 1998, pp. 244–251.

[31] BRIEDEN, A., GRITZMANN, P., KANNAN, R., KLEE, V., LOVÁSZ, L., AND SIMONOVITS, M. Deterministic and randomized polynomial-time approximation of radii. *Mathematika 48*, 1-2 (2001), 63–105.

[32] BRIEDEN, A., AND SCHIELE, S. Survival analysis based on constrained confidence partitioning. working paper.

[33] BRIEDEN, A., SCHIELE, S., AND SCHLATTAU, M. Cluster-based analysis of entrepeneurship in russia. working paper.

[34] BRKAN, M. Do algorithms rule the world? algorithmic decision-making and data protection in the framework of the GDPR and beyond. *International Journal of Law and Information Technology 27*, 2 (2019), 91–121.

[35] BROOKMEYER, R., AND CROWLEY, J. A confidence interval for the median survival time. *Biometrics 38*, 1 (1982), 29.

[36] BROWN, L. D., CAI, T. T., AND DASGUPTA, A. Interval estimation for a binomial proportion. *Statistical Science 16*, 2 (2001), 101–117.

[37] BUNDESREPUBLIK DEUTSCHLAND. Das Fünfte Buch Sozialgesetzbuch – Gesetzliche Krankenversicherung (Artikel 1 des Gesetzes vom 20.Dezember 1988, BGBl. I S. 2477, 2482), das zuletzt durch Artikel 311 der Verordnung vom 19. juni 2020 (BGBl. I S. 1328) geändert worden ist: SGB V, 1988.

[38] BZDOK, D. Classical statistics and statistical learning in imaging neuroscience. *Frontiers in Neuroscience 11*, 543 (2017), 1–23.

[39] BZDOK, D., KRZYWINSKI, M., AND ALTMAN, N. Points of significance: Machine learning: a primer. *Nature Methods 14*, 12 (2017), 1119–1120.

[40] CHONG, H. Y., CHAIYAKUNAPRUK, N., D B C, W., K K C, L., AND CHIOU, C. F. Global economic burden of schizophrenia: A systematic review. *Neuropsychiatric Disease and Treatment 12* (2016), 357–373.

[41] CLARK, S. L., ADKINS, D. E., AND VAN DEN OORD, EDWIN J C G. Analysis of efficacy and side effects in CATIE demonstrates drug response subgroups and potential for personalized medicine. *Schizophrenia Research 132*, 2-3 (2011), 114–120.

[42] CLOUTIER, M., AIGBOGUN, M. S., GUERIN, A., NITULESCU, R., RAMANAKUMAR, A. V., KAMAT, S. A., DELUCIA, M., DUFFY, R., LEGACY, S. N., HENDERSON, C., FRANCOIS, C., AND WU, E. The economic burden of schizophrenia in the United States in 2013. *The Journal of Clinical Psychiatry 77*, 6 (2016), 764–771.

[43] COLLETT, D. *Modelling Survival Data in Medical Research*, 3rd ed. Texts in statistical science. Chapman & Hall/CRC, Boca Raton, Fla., 2015.

[44] COX, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34*, 2 (1972), 187–220.

[45] COX, D. R., AND OAKES, D. O. *Analysis of Survival Data*, vol. 21 of *Chapman and Hall/CRC Monographs on Statistics and Applied Probability*. Routledge, Boca Raton, 2018.

[46] DAM, H. K., TRAN, T., AND GHOSE, A. Explainable software analytics. In *2018 ACM/IEEE 40th International Conference on Software Engineering: New ideas and emerging results* (Piscataway, NJ, 2018), A. Zisman and S. Apel, Eds., IEEE, pp. 53–56.

[47] DAVENPORT, T., HARRIS, J., AND ABNEY, D. *Competing on Analytics: Updated, with a New Introduction: The New Science of Winning*. Harvard Business Review Press, Brighton, MA, 2017.

[48] DAWES, R. M. The robust beauty of improper linear models in decision making. *American Psychologist 34*, 7 (1979), 571–582.

[49] DOSHI-VELEZ, F., AND KIM, B. Towards a rigorous science of interpretable machine learning, 2017. Avaiable at: https://arxiv.org/pdf/1702.08608 (Accessed 21 July 2020).

[50] DUNKLER, D., HALLER, M., OBERBAUER, R., AND HEINZE, G. To test or to estimate? P-values versus effect sizes. *Transplant International: Official Journal of the European Society for Organ Transplantation 33*, 1 (2020), 50–55.

[51] EFRON, B., AND HASTIE, T. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*, vol. 5 of *Institute of Mathematical Statistics monographs*. Cambridge University Press, New York, 2016.

[52] ESCALANTE, H. J., ESCALERA, S., GUYON, I., BARÓ, X., GÜÇLÜTÜRK, Y., GÜÇLÜ, U., AND VAN GERVEN, M., Eds. *Explainable and Interpretable Models in Computer Vision and Machine Learning*. The Springer Series on Challenges in Machine Learning. Springer International Publishing, Cham, 2018.

[53] EUROPEAN UNION. General Data Protection Regulation: GDPR, 15.05.2018.

[54] FEDERICO CABITZA, RAFFAELE RASOINI, AND GIAN FRANCO GENSINI. Unintended consequences of machine learning in medicine. *JAMA 318*, 6 (2017), 517–518.

[55] FRANCESCA ROSSI. Artificial intelligence: Potential benefits and ethical considerations: Pe 571.380. European Parliament: Briefing Legal Affairs.

[56] FREY, S. The economic burden of schizophrenia in germany: A population-based retrospective cohort study using genetic matching. *European Psychiatry 29*, 8 (2014), 479–489.

[57] GARDNER, M. J., AND ALTMAN, D. G. Confidence intervals rather than p values: estimation rather than hypothesis testing. *British Medical Journal (Clinical Research ed.) 292 (6522)* (1986), 746–750.

[58] GBD 2015 DALYS AND HALE COLLABORATORS. Global, regional, and national disability-adjusted life-years (DALYs) for 315 diseases and injuries

and healthy life expectancy (HALE), 1990–2015: a systematic analysis for the global burden of disease study 2015. *The Lancet 388* (2016), 1603–1658.

[59] GBD 2015 DISEASE AND INJURY INCIDENCE AND PREVALENCE COLLABORATORS. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the global burden of disease study 2015. *The Lancet 388* (2016), 1545–1602.

[60] GOODMAN, B., AND FLAXMAN, S. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine 38*, 3 (2017), 50–57.

[61] GREENWOOD, M. *A Report on the Natural Duration of Cancer*. Reports on public health and medical subjects. Her Majesty's Stationery Office (HMSO), United Kingdom, 1926.

[62] GRIFFITHS, P., AND NEEDLEMAN, J. Statistical significance testing and p-values: Defending the indefensible? a discussion paper and position statement. *International Journal of Nursing Studies 99* (2019), 103384.

[63] GRITZMANN, P. *Grundlagen der Mathematischen Optimierung: Diskrete Strukturen, Komplexitätstheorie, Konvexitätstheorie, Lineare Optimierung, Simplex-Algorithmus, Dualität*. Aufbaukurs Mathematik. Springer Fachmedien Wiesbaden, 2013.

[64] GUY, W. *ECDEU Assessment Manual for Psychopharmacology*. U.S. Dept. of Health, Education, and Welfare, Public Health Service, Alcohol, Drug Abuse, and Mental Health Administration, National Institute of Mental Health, Psychopharmacology Research Branch, Division of Extramural Research Programs, 1976.

[65] HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 10th ed. Springer Series in Statistics. Springer, New York, NY, 2008.

[66] HERES, S., CIRJALIU, D. M., DEHELEAN, L., MATEI, V. P., PODEA, D. M., SIMA, D., STECHER, L., AND LEUCHT, S. The SWITCH study: rationale and design of the trial. *European Archives of Psychiatry and Clinical Neuroscience 266*, 6 (2016), 513–521.

[67] HINNENTHAL, M. *Endpoint-oriented clustering of individual patient data : A new approach for the health economic evaluation of medical interventions*. Dissertation, Universität der Bundeswehr München, Fakultät für Wirtschafts- und Organisationswissenschaften, Neubiberg, 2017.

[68] HOLTE, R. C. Very simple classification rules perform well on most commonly used datasets. *Machine Learning 11*, 1 (1993), 63–90.

[69] HOLTE, R. C. Elaboration on two points raised in "classifier technology and the illusion of progress". *Statistical Science 21*, 1 (2006), 24–26.

[70] HOLZINGER, A., BIEMANN, C., PATTICHIS, C. S., AND KELL, D. B. What do we need to build explainable ai systems for the medical domain?, 2017. Available at: https://arxiv.org/pdf/1712.09923 (Accessed 21 July 2020).

[71] HOSMER, D. W., LEMESHOW, S., AND MAY, S. *Applied Survival Analysis*. John Wiley & Sons, Inc, Hoboken, NJ, USA, 2008.

[72] HOUGAARD, P. Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika 71*, 1 (1984), 75.

[73] HOUGAARD, P. Modelling heterogeneity in survival data. *Journal of Applied Probability 28*, 3 (1991), 695–701.

[74] HUAJIE JIN, AND IRIS MOSWEU. The societal cost of schizophrenia: A systematic review. *PharmacoEconomics 35*, 1 (2017), 25–42.

[75] HWANG, F., ONN, S., AND ROTHBLUM, U. Representations and characterizations of vertices of bounded-shape partition polytopes. *Linear Algebra and its Applications 278*, 1-3 (1998), 263–284.

[76] IAN GOODFELLOW, YOSHUA BENGIO, AND AARON COURVILLE. *Deep Learning*. MIT Press, Cambridge, 2016.

[77] INTERNATIONAL DATA CORPORATION. IDC's Global DataSphere Forecast shows continued steady growth in the creation and consumption of data, 8.5.2020. Available at: https://www.idc.com (About IDC - IDC Media Center); (Accessed 21 July 2020).

[78] IQWIG. General methods, 2017.

[79] JAIN, V., AND CHATTERJEE, J. M., Eds. *Machine Learning with Health Care Perspective: Machine Learning and Healthcare*, vol. 13 of *Learning and analytics in intelligent systems*. Springer, Cham, 2020.

[80] JAKUBOVSKI, E., CARLSON, J. P., AND BLOCH, M. H. Prognostic subgroups for remission, response, and treatment continuation in the clinical antipsychotic trials of intervention effectiveness (CATIE) trial. *The Journal of Clinical Psychiatry 76*, 11 (2015), 1535–1545.

[81] JOSHI, P. *Python: Real World Machine Learning*. Packt Publishing, Birmingham, 2016.

[82] K, L., KLEMPERER, D., K, K., C, B., OLLENSCHLÄGER, G., AND LUDWIG, W.-D. Interessenkonflikte in der Medizin: Mit Transparenz Vertrauen stärken. *Deutsches Ärzteblatt 108*, 6 (2011), A256 – A261.

[83] KALBFLEISCH, J. D., AND PRENTICE, R. L. *The Statistical Analysis of Failure Time Data*, 2nd ed., vol. 360 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, Hoboken, 2011.

[84] KAPLAN, E. L., AND MEIER, P. T. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53*, 282 (1958), 457–481.

[85] KAY, S. R., FISZBEIN, A., AND OPLER, L. A. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin 13*, 2 (1987), 261–276.

[86] KLEIN, J. P., VAN HOUWELINGEN, H. C., IBRAHIM, J. G., AND SCHEIKE, T. H. *Handbook of Survival Analysis*. Chapman & Hall / CRC Handbooks of Modern Statistical Methods. Taylor and Francis, Hoboken, 2014.

[87] KLEINBAUM, D. G., AND KLEIN, M. *Survival Analysis: A Self-Learning Text*, 3 ed. Statistics for Biology and Health. Springer Science+Business Media LLC, New York, NY, 2012.

[88] KOBELT, G. *Health economics: An introduction to economic evaluation*, 2. ed. ed. Office of Health Economics, London, 2002.

[89] KOLETSI, D., AND PANDIS, N. Survival analysis, part 2: Kaplan-Meier method and the log-rank test. *American Journal of Orthodontics and Dentofacial Orthopedics 152*, 4 (2017), 569–571.

[90] KOLMOGOROV, A. N., AND MORRISON, N. *Foundations of the Theory of Probability*, 2 ed. Dover Books on Mathematics. Dover Publications, Newburyport, 2018.

[91] KONNOPKA, A., KLINGBERG, S., WITTORF, A., AND KÖNIG, H.-H. Die Kosten der Schizophrenie in Deutschland: Ein systematischer Literaturüberblick. *Psychiatrische Praxis 36*, 5 (2009), 211–218.

[92] KRAEPELIN, E. *Psychiatrie: Ein Lehrbuch für Studierende und Ärzte*, 5th ed. Johann Ambrosius Barth, Leipzig, 1896.

[93] KRAEPELIN, E. *Dementia Praecox and Paraphrenia*. Chicago Medical Book Co, Chicago, 1919.

[94] KROLL, J. A. The fallacy of inscrutability. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences 376*, 2133 (2018).

[95] KRUMHOLZ, H. M. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Affairs (Project Hope) 33*, 7 (2014), 1163–1170.

[96] KULKARNI, P. *Reinforcement and Systemic Machine Learning for Decision Making*. IEEE Press series on systems science and engineering. Wiley, Hoboken, NJ, 2012.

[97] LEE, E. T., AND WANG, J. W. *Statistical Methods for Survival Data Analysis*. John Wiley & Sons, Inc, Hoboken, NJ, USA, 2003.

[98] LEVINE, S. Z., RABINOWITZ, J., FARIES, D., LAWSON, A. H., AND ASCHER-SVANUM, H. Treatment response trajectories and antipsychotic medications: Examination of up to 18 months of treatment in the CATIE chronic schizophrenia trial. *Schizophrenia Research 137*, 1-3 (2012), 141–146.

[99] LIEBERMAN, J. A., STROUP, T. S., MCEVOY, J. P., SWARTZ, M. S., ROSENHECK, R. A., PERKINS, D. O., KEEFE, R. S. E., DAVIS, S. M., DAVIS, C. E., LEBOWITZ, B. D., SEVERE, J., AND HSIAO, J. K. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *The New England Journal of Medicine 353*, 12 (2005), 1209–1223.

[100] LIPTON, Z. C. The mythos of model interpretability, 2016. Available at: https://arxiv.org/pdf/1606.03490 (Accessed 21 July 2020).

[101] MANTON, K. G., STALLARD, E., AND VAUPEL, J. W. Methods for comparing the mortality experience of heterogeneous populations. *Demography 18*, 3 (1981), 389–410.

[102] MARDER, S. R., DAVIS, J. M., AND CHOUINARD, G. The effects of risperidone on the five dimensions of schizophrenia derived by factor analysis: Combined results of the north american trials. *The Journal of Clinical Psychiatry 58*, 12 (1997), 538–546.

[103] MCEVOY, J. P., LIEBERMAN, J. A., STROUP, T. S., DAVIS, S. M., MELTZER, H. Y., ROSENHECK, R. A., SWARTZ, M. S., PERKINS, D. O., KEEFE, R. S., DAVIS, C. E., SEVERE, J., AND HSIAO, J. K. Effectiveness of clozapine versus olanzapine, quetiapine, and risperidone in patients with

chronic schizophrenia who did not respond to prior atypical antipsychotic treatment. *The American Journal of Psychiatry 163*, 4 (2006), 600–610.

[104] MENCHER, A. G. On the social deployment of science. *Bulletin of the Atomic Scientists 27*, 10 (1971), 34–38.

[105] MEYER, J. M., NASRALLAH, H. A., MCEVOY, J. P., GOFF, D. C., DAVIS, S. M., CHAKOS, M., PATEL, J. K., KEEFE, R. S., STROUP, T. S., AND LIEBERMAN, J. A. The clinical antipsychotic trials of intervention effectiveness (CATIE) schizophrenia trial: Clinical comparison of subgroups with and without the metabolic syndrome. *Schizophrenia Research 80*, 1 (2005), 9–18.

[106] MOLNAR, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub, 2019. Available at: https://christophm.github.io/interpretable-ml-book/ (Accessed 21 July 2020).

[107] MOROSINI, P. L., MAGLIANO, L., BRAMBILLA, L., UGOLINI, S., AND PIOLI, R. Development, reliability and acceptability of a new version of the DSM-IV social and occupational functioning assessment scale (SOFAS) to assess routine social functioning. *Acta Psychiatrica Scandinavica 101*, 4 (2000), 323–329.

[108] MÜLLER, A. C. *Introduction to Machine Learning with Python*, 1st ed. O'Reilly, Bejing and Boston and Farnham and Sebastopol and Tokyo, 2016.

[109] MURPHY, K. P. *Machine Learning: A Probabilistic Perspective*. Adaptive computation and machine learning series. MIT Press, Cambridge, Mass., 2012.

[110] NATIONAL INSTITUTE OF MENTAL HEALTH. Schizophrenia: 19-mh-8082, 2019.

[111] NGUYEN, A., YOSINSKI. J., AND CLUNE, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 427–436.

[112] ÖLLINGER, M. J. *Geometric clustering and its applications in binary classification problems*. Dissertation, Universität der Bundeswehr München, Fakultät für Informatik, Neubiberg, 2014.

[113] PIRES, A. M. Interval estimators for a binomial proportion: a comparison of twenty methods. *REVSTAT - Statistical Journal 6*, 2 (2008), 165–197.

[114] POCOCK, S. J., AND HUGHES, M. D. Estimation issues in clinical trials and overviews. *Statistics in Medicine 9*, 6 (1990), 657–671.

[115] POOLE, D. L., MACKWORTH, A. K., AND GOEBEL, R. *Computational Intelligence: A Logical Approach*. Oxford Univ. Press, New York, NY, 1998.

[116] RABINOWITZ, J., LEVINE, S. Z., GARIBALDI, G., BUGARSKI-KIROLA, D., BERARDO, C. G., AND KAPUR, S. Negative symptoms have greater impact on functioning than positive symptoms in schizophrenia: Analysis of CATIE data. *Schizophrenia Research 137*, 1 (2012), 147–150.

[117] RAGHUPATHI, W., AND RAGHUPATHI, V. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems 2*, 3 (2014), 1–10.

[118] RAMSEY, T. L., LIU, Q., MASSEY, B. W., AND BRENNAN, M. D. Genotypic variation in the SV2C gene impacts response to atypical antipsychotics the CATIE study. *Schizophrenia Research 149*, 1 (2013), 21–25.

[119] RASCHKA, S. *Python Machine Learning*, 3rd ed. Expert insight. Packt, Birmingham and Mumbai, 2019.

[120] RICE, D. P. The economic impact of schizophrenia. *The Journal of Clinical Psychiatry 60 Suppl 1* (1999).

[121] RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence 1*, 5 (2019), 206–215.

[122] RUDIN, C., WANG, C., AND COKER, B. The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review 2(1)* (2020).

[123] SACKETT, D. L., ROSENBERG, W. M., GRAY, J. A., HAYNES, R. B., AND RICHARDSON, W. S. Evidence based medicine: what it is and what it isn't. *BMJ (Clinical research ed.) 312*, 7023 (1996).

[124] SAMEK, W., MONTAVON, G., AND VEDALDI, A. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Lecture notes in computer series: Lecture notes in artificial intelligence. Springer, Cham, 2019.

[125] SCHIELE, S., BRIEDEN, A., LEUCHT, S., AND HERES, S. New insight into the CATIE study by constrained confidence partitioning. An innovative technique towards personalized antipsychotic drug therapy in schizophrenia treatment. working paper.

[126] SCHNEIDER, K. *Clinical Psychopathology*. Grune & Stratton, New York, 1959.

[127] SCHÖFFSKI, O. *Gesundheitsökonomische Evaluationen*. Springer, Dordrecht, 2012.

[128] SCHUMACHER, M., OLSCHEWSKI, M., AND SCHMOOR, C. The impact of heterogeneity on the comparison of survival times. *Statistics in Medicine 6*, 7 (1987), 773–784.

[129] SHALEV-SHWARTZ, S., AND BEN-DAVID, S. *Understanding Machine Learning: From theory to algorithms*. Cambridge University Press, Cambridge, 2014.

[130] SIMON, P. *Analytics: The Agile Way*. Wiley & SAS Business Series. John Wiley & Sons Inc, Hoboken, New Jersey, 2017.

[131] SIMON, R. Confidence intervals for reporting results of clinical trials. *Annals of Internal Medicine 105*, 3 (1986), 429–435.

[132] SIMPSON, G. M., AND ANGUS, J. W. A rating scale for extrapyramidal side effects. *Acta Psychiatrica Scandinavica 45*, 212 (1970), 11–19.

[133] SIMPSON, J. *Oxford Dictionary of English*, 2nd ed. Oxford Univ. Press, 1989.

[134] STATISTISCHES BUNDESAMT. Total cost of illness in millions of euro for Germany, 2015. Available at: www.gbe-bund.de (Recherche nach Themen: Ausgaben, Kosten, Finanzierung - Kosten einzelner Krankheiten - Dokumentart Tabellen) (Accessed 7 July 2020).

[135] STROUP, T. S., AND LIEBERMAN, J. A., Eds. *Antipsychotic Trials in Schizophrenia: The CATIE Project*. Cambridge University Press, Cambridge, 2010.

[136] STROUP, T. S., LIEBERMAN, J. A., MCEVOY, J. P., SWARTZ, M. S., DAVIS, S. M., CAPUANO, G. A., ROSENHECK, R. A., KEEFE, R. S. E., MILLER, A. L., BELZ, I., AND HSIAO, J. K. Effectiveness of olanzapine, quetiapine, and risperidone in patients with chronic schizophrenia after discontinuing perphenazine: a CATIE study. *The American Journal of Psychiatry 164*, 3 (2007), 415–427.

[137] STROUP, T. S., LIEBERMAN, J. A., MCEVOY, J. P., SWARTZ, M. S., DAVIS, S. M., ROSENHECK, R. A., PERKINS, D. O., KEEFE, R. S., DAVIS, C. E., SEVERE, J., AND HSIAO, J. K. Effectiveness of olanzapine, quetiapine, risperidone, and ziprasidone in patients with chronic schizophrenia following discontinuation of a previous atypical antipsychotic. *The American Journal of Psychiatry 163*, 4 (2006), 611–622.

[138] STROUP, T. S., MCEVOY, J. P., SWARTZ, M. S., BYERLY, M. J., GLICK, I. D., CANIVE, J. M., MCGEE, M. F., SIMPSON, G. M., STEVENS, M. C., AND LIEBERMAN, J. A. The national institute of mental health clinical antipsychotic trials of intervention effectiveness (CATIE) project: schizophrenia trial design and protocol development. *Schizophrenia Bulletin 29*, 1 (2003), 15–31.

[139] SU, J., VARGAS, D. V., AND KOUICHI, S. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation), Vol.23 , Issue.5 , pp. 828–841. Publisher: IEEE. 23* (2019).

[140] THULIN, M. The cost of using exact confidence intervals for a binomial proportion. *Electronic Journal of Statistics 8*, 1 (2014), 817–840.

[141] TOLLENAAR, N., AND VAN DER HEIJDEN, P. G. M. Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. *Journal of the Royal Statistical Society. Series A (Statistics in Society) 176*, 2 (2013), 565–584.

[142] VAN DEN OORD, EDWIN J.C.G., ADKINS, D. E., MCCLAY, J., LIEBERMAN, J., AND SULLIVAN, P. F. A systematic method for estimating individual responses to treatment with antipsychotics in CATIE. *Schizophrenia Research 107*, 1 (2009), 13–21.

[143] VAN DER BEEK, K., AND VAN DER BEEK, G. *Gesundheitsökonomik: Einführung*. VWL, 10-2012. Oldenbourg, München, 2011.

[144] VELLIDO, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications* (2019), 1–15.

[145] WACHTER, S., MITTELSTADT, B., AND FLORIDI, L. Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law 7*, 2 (2017), 76–99.

[146] WEINBERGER, D. R., AND HARRISON, P. J. *Schizophrenia*, 3rd ed. Wiley-Blackwell, Chichester, 2011.

[147] WELLS, A. *Statistics: An Introduction Using R*. ED-Tech Press, Waltham Abbey Essex, 2019.

[148] WEXLER, R. When a computer program keeps you in jail. *The New York Times* (13 June 2017). Available at: https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html (Accessed 21 July 2020).

[149] WU, E. Q., BIRNBAUM, H. G., SHI, L., BALL, D. E., KESSLER, R. C., MOULIS, M., AND AGGARWAL, J. The economic burden of schizophrenia in the United States in 2002. *The Journal of Clinical Psychiatry 66*, 9 (2005).

[150] YOUNG, K. D., AND LEWIS, R. J. What is confidence? part 1: The use and interpretation of confidence intervals. *Annals of Emergency Medicine 30*, 3 (1997), 307–310.

[151] ZENG, J., USTUN, B., AND RUDIN, C. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 180*, 3 (2017), 689–722.