Master's thesis
Master's Programme in Data Science

A survival analysis of benign prostatic hyperplasia procedure complications

Uula Ulkuniemi

December 9, 2022

Supervisor(s): Antti Rannikko
Esa Pitkänen

Examiner(s): Antti Honkela
Esa Pitkänen

University of Helsinki
Faculty of Science

P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Degree programme
Faculty of Science	Master's Programme in Data Science

Tekijä — Författare — Author

Uula Ulkuniemi

Työn nimi — Arbetets titel — Title

A survival analysis of benign prostatic hyperplasia procedure complications

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidantal — Number of pages
Master's thesis	December 9, 2022	87

Tiivistelmä — Referat — Abstract

This thesis presents a complication risk comparison of the most used surgical interventions for benign prostatic hyperplasia (*BPH*). The investigated complications are the development of either a post-surgery BPH recurrence (*reoperation*), an urethral stricture or stress incontinence severe enough to require a surgical procedure for their treatment. The analysis is conducted with survival analysis methods on a data set of urological patients sourced from the Finnish Institute for Health and Welfare. The complication risk development is estimated with the Aalen-Johansen estimator and the effects of certain covariates on the complication risks is estimated with the Cox PH regression model. One of the regression covariates is the Charlson Comorbidity Index score, which attempts to quantify a disease load of a patient at a certain point in time as a single number. A novel Spark algorithm was designed to facilitate the efficient calculation of the Charlson Comorbidity Index score on a data set of the same size as the one used in the analyses here. The algorithm achieved at least similar performance to the previously available ones and scaled better on larger data sets and with stricter computing resource constraints.

Both the urethral stricture and urinary incontinence endpoints suffered from a lower number of samples, which made the associated results less accurate. The estimated complication probabilities in both endpoint types were also so low that the BPH procedures couldn't be reliably differentiated. In contrast, BPH reoperation risk analyses yielded noticeable differences among the initial BPH procedures. Regression analysis results suggested that the Charlson Comoborbidity Index score isn't a particularly good predictor in any of the endpoints. However, certain cancer types that are included in the Charlson Comorbidity Index score did predict the endpoints well when used as separate covariates. An increase in the patient's age was associated with a higher complication risk, but less so than expected. In the urethral stricture and urinary incontinence endpoints the number of preceding BPH operations was usually associated with a notable complication risk increase.

ACM Computing Classification System (CCS):
Applied computing → Life and medical sciences → Health care information systems
Mathematics of computing → Probability and statistics → Statistical paradigms → Survival analysis

Avainsanat — Nyckelord — Keywords

survival analysis, benign prostatic hyperplasia

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

1. Introduction

Benign prostatic hyperplasia (*BPH*) is the most common benign tumor found in the male population. While many patients can make do without treatment, the mere amount of cases makes even the small subsection of the patients that do need interventions rather sizable. This also means that an optimal treatment choice is important for the well-being of a large patient contingent. Surgical procedures are the most effective BPH treatments. While efficacious, they are unfortunately also associated with risks of multiple types of complications. The complication risks are dependent on the chosen BPH procedure and the attributes of the treated patient, which means that quantifying the risk discrepancies between the intervention options and the effects of certain clinical characteristics of the treated patients on the risk development is important for the clinician planning the treatment of a particular patient. In addition to having the best chance of a successful treatment, a correct procedure choice can save the patient from additional procedures and troublesome complications.

The BPH procedure complications chosen to be investigated in this thesis were a BPH recurrence necessitating a new BPH procedure (*reoperation*), stress incontinence (*SI*) and urethral stricture (*US*). A BPH recurrence refers to a case where the initial BPH procedure was successful but the patient redevelops the symptoms at a later time. Stress incontinence is a condition where physical stress can cause the patient to lose their ability to control their urination. While not a life-threatening condition by itself, stress incontinence can still be an extreme nuisance in the patient's daily life. An urethral stricture refers to scar tissue that narrows the patient's urethra. It often occurs as an unwanted side effect of surgical procedures that affect the area around the urethra. Common symptoms of strictures include pain and either a reduction or a total blockage of the flow of urine, the latter being an especially severe complication which requires urgent care.

The study presented in this thesis is a retrospective analysis of patient health records done in collaboration with a group of urologists with a clinical interest in the results, which are also planned to be submitted for a release in a medical journal. The choice of the retrospective analysis was partly motivated by the fact that Finland has a long history of collecting medical data in an electronic form. As medical studies are often plagued with

small sample sizes, a well-sized source of good-quality data offers a possibility of important insight. Finnish Institute for Health and Welfare (*THL*) manages two separate registries for Finnish inpatient and outpatient care data, *Hilmo* and *Avohilmo*, respectively. Health care providers operating in Finland are obligated by law to provide the information about patient visits to the registries. The research data used in this study was formed by requesting the available inpatient and outpatient data for a cohort of patients who had received any of the certain urological diagnoses or procedures detailed later in Chapter 3. Some of the key data points included in the research data are the time periods the patients have spent in the health care facilities and both the diagnoses and the procedures that the patient received during a particular visit.

A few other authors have investigated the BPH procedure complication risks before, but their works have been limited in scope due to their less comprehensive research data sources with limited sample numbers and follow-up times. These issues are common in medical data. One reason for this is that medical data is often protected by legislation which causes getting access to it to be challenging. Another factor is that while medical facilities nowadays commonly store health records in an electronic form, the data formats are often incompatible between the facilities even inside a single country. This means that researchers aiming to obtain data from several sources need to perform the necessary harmonization work in order to merge the data sets to a usable form. Each additional data source also adds another possible source of either incomplete or incorrect data. These problems were minimized in the analyses of this thesis by mostly relying on the harmonization and validation work done by the THL personnel.

Survival analysis is a branch of statistics for the analysis of the distribution of amount of time before some event happens, i.e. the survival distribution. Rephrased in more practical terms, survival analysis attempts to answer questions such as “what is the probability of surviving for 1 year after a prostate cancer diagnosis” and “how does age affect the risk of developing a complication after a surgery”. As the thesis analyses the period of time between the first BPH procedure of a patient and the first following procedure treating an associated complication, survival analysis methods are the natural tools of choice for the task. The analysis results presented in Chapter 4 use survival analysis methods to estimate the survival distributions in the different scenarios, to test whether the survival distributions from the different initial BPH procedure options to a certain complication procedure are significantly different and to perform regression analysis to estimate the effect of the clinical characteristics of the patients on the risk development.

One of the investigated clinical characteristics of the patients is the so-called Charlson Comorbidity Index (*CCI*) score. The CCI score attempts to quantify the patient’s diagnoses into a single number to enable the stratification of the patients per the severity

of their total diagnosis load. CCI score has also proven to be a functional statistical indicator for the mortality rate of a disease. The basic idea of scoring diseases is simple, but there are certain specific characteristics which make the CCI surprisingly problematic from an algorithmic perspective. The difficulty comes from the temporal component caused by the facts that the CCI scores are assigned to groups of diseases instead of individual ones and that a patient can receive the score of a certain group only once. There are also some pairs of groups comprised of a less severe and a more severe form of the same disease. The temporal component means that the ordering of the diagnoses matters, which complicates the performance considerations. The available implementations of the CCI score calculation algorithm failed to finish due to the size of the research data. To remedy the issue a new algorithm for calculating the CCI scores was designed on the Apache Spark big data processing platform. The algorithm was designed to scale to support the sizes of the ever increasing medical data sources.

To help to familiarize the reader with the necessary medical science domain knowledge, Chapter 2 will give an overview of the background of BPH and the investigated complications in addition to introducing the reader to the medical coding systems used in Finland. This chapter is mandatory reading to understand the rest of the text. Chapter 3 describes the methodology used to perform the analyses. The methodology is comprised of the data preparation process, the novel Charlson Comorbidity Index algorithm and a brief overview of the theory behind the used survival analysis methods. Chapter 4 details the results of the performed analyses. Chapter 5 provides a discussion of the results in addition to the associated relevance and limitations. The thesis then ends with a conclusion in Chapter 6.

1.1 Research questions

- Some possible complications that can develop after a patient's first BPH procedure include recurrent BPH, stress incontinence and an urethral stricture. How do the risks for a severe enough form of these complications to require another procedure develop with time for a subset of the more common initial BPH procedures?
- How do the complication risks compare among the initial BPH procedures?
- What are the effects of certain clinical factors on the risk development?
- How do the effects of the clinical factors differ among the initial BPH procedures?
- Do the results of the previous short-term BPH procedure complication studies match ones from a cohort with longer follow-up times?
- How can the Charlson Comorbidity Index scores be efficiently computed for a very large cohort?
- Does the Charlson Comorbidity Index score predict any of the investigated compli-

cations?

- A survival analysis course meant for practitioners often consists of a set of simple and rigid methods whereas the bleeding edge research can be inaccessible without sufficient statistical training. What are the common pitfalls in applying survival analysis models and how can they be remedied?

2. Medical background

The analysis of medical data presents several challenges. Understanding the content of medical data necessitates the analyst to possess the medical domain knowledge and to understand the often Latin-based specialized vocabulary. Medical data entities such as diagnoses, procedures, laboratory tests and medicines are also often denoted in the data by their representation in an appropriate standardized medical coding system. Accessing and using medical data is also often limited by legislation and requires special considerations due to the inclusion of personally identifiable information. As an example, only the structure of the analyzed data can be presented in this thesis as the content of the research data includes personally identifiable patient information which is protected by the law [3].

The main aim of this chapter is to familiarize the reader with the prerequisite medical knowledge that is necessary to understand the research questions and the clinical motivations behind them. The main topics covered are the diagnosis and procedure coding systems that are currently used in Finland and the brief medical backgrounds of BPH, urethral strictures and stress incontinence. The investigated procedures are mentioned by name but not described in detail as further discussion would be unnecessary for the purposes of this thesis.

2.1 Electronic medical records

While the definition of an electronic medical record (*EMR*) varies in the literature, the underlying idea is to store the information about the interactions between a patient and a health care provider on a computer. EMR systems have greatly increased the efficacy and reduced the costs of care in the health care facilities which have accommodated their use [63,65]. One practical issue of EMR systems is the format that the different interactions are stored in. Medical data includes entities such as diagnoses, procedures and medicines that are not specific to a particular health care provider, so they should be encoded in a standardized way in order to enable interoperability. Medical coding systems offer an answer to the problem by providing a way to map the entities to a standardized representation in the data. They are also convenient from the perspective of data analysis as the standardized codes provide the medical data analysts a common

Table 2.1: A list of the chapters of the ICD-10 system

Chapter	Code range	Name
1	A00-B99	Certain infectious and parasitic diseases
2	C00-D49	Neoplasms
3	D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
4	E00-E89	Endocrine, nutritional and metabolic diseases
5	F01-F99	Mental, Behavioral and Neurodevelopmental disorders
6	G00-G99	Diseases of the nervous system
7	H00-H59	Diseases of the eye and adnexa
8	H60-H95	Diseases of the ear and mastoid process
9	I00-I99	Diseases of the circulatory system
10	J00-J99	Diseases of the respiratory system
11	K00-K95	Diseases of the digestive system
12	L00-L99	Diseases of the skin and subcutaneous tissue
13	M00-M99	Diseases of the musculoskeletal system and connective tissue
14	N00-N99	Diseases of the genitourinary system
15	O00-O9A	Pregnancy, childbirth and the puerperium
16	P00-P96	Certain conditions originating in the perinatal period
17	Q00-Q99	Congenital malformations, deformations and chromosomal abnormalities
18	R00-R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
19	S00-T88	Injury, poisoning and certain other consequences of external causes
20	V00-Y99	External causes of morbidity
21	Z00-Z99	Factors influencing health status and contact with health services

vocabulary for interactions with the domain expert clinicians. The later chapters of this thesis will refer to the discussed procedures and diagnoses only with their codes in the appropriate medical coding systems.

2.1.1 ICD-10

ICD-10 is a diagnosis classification system developed by the World Health Organization (*WHO*) [47]. The tenth version of the system was finalized in 1990, but it has received several smaller amendments after the initial creation. Finland's health care sector is currently storing diagnosis information using the ICD-10 system with some additional country-specific modifications. ICD-10 encodes the diagnoses as alphanumeric codes. In addition to describing the diagnosis itself, the codes are hierarchically structured to also contain additional information about the diagnosis. As an example of a such structure, the ICD-10 code for benign prostatic hyperplasia without urinary tract issues is *N40.0*. If the patient has lower urinary tract issues, then the code *N40.1* is used instead [39]. An ICD-10 code includes diagnostic criteria which must be reached in order for a clinician to be able to diagnose the patient with a disease represented by that particular code [47]. The following details of the ICD-10 system are mainly based on a Finnish ICD-10 manual by Komulainen et al. because of the country-specific codes contained in research data [39].

ICD-10 has a hierarchical structure consisting of chapters, sections, classes and subclasses. Chapters represent a high-level categorization of the diagnoses. In practice the chapters consist of ranges of diagnostic codes. The current version of the ICD-10 contains 21 chapters. These are listed in Table 2.1. ICD-10 codes always start with an alphabetic character. In most cases the first character denotes the chapter of the code. Exceptions to this are letters H, used in both sections 7 and 8, and letter D, used in both chapters 2 and 3. Usually the initial character of the code is followed by a numeric character. Exceptions to this are the chapters 1, 2, 19 and 20, which include codes that begin with a longer sequence of alphabetical characters [39].

Each chapter is further divided into sections. Each section is a range of 3-character codes. A section represents some logic for further categorizing the diseases of the chapter. As an example, chapter 1 is grouped according to the disease location, the way of infecting of the disease and the organism that causes the disease. A chapter can also have multiple levels of sections. For instance, chapter 2 has two levels of sections based alternatively on either tumor severity or the tumor location [39].

All of the sections contain one or more classes. A class is a single 3-character code. Classes can either define a disease on their own or represent an even more specialized category. To represent an actual diagnosis in the latter case the code needs to define additional one or two levels of subclasses. Subclasses are formed by adding an alphabetical character to the code as a suffix, which causes the length of an ICD-10 code to range from 3 to 5 characters, which provides a helpful limitation while processing ICD-10 codes. Each subclass further clarifies a detail about the disease, e.g. the location. A dot is usually used as a delimiter between the third and fourth characters [39]. Some examples of real ICD-10 codes are listed in Table 2.2, which shows the 20 most common ICD-10 codes found in the research data.

ICD-10 has three special code types. They are all used to encode a causal relationship between two codes. The three possible causal relationship types are the disease and its symptom, a tumor and its symptom or a harmful effect caused by a chemical/biological agent and the associated Anatomical Therapeutic Chemical (*ATC*) code of the agent. ATC is the coding system currently used in Finland to store information about medicines. For the purposes of this thesis the details of the ATC system are not necessary and are therefore omitted. Table 2.3 summarizes the different types of combination ICD-10 codes and their structures [39].

The coding system has drawn criticism from psychiatrists. Some of the claimed issues include that the system does not distinguish between problem behaviors caused by a psychiatric illness and those that are merely symptoms of another medical problem. Other criticism is mainly directed towards the diagnostic criteria for certain diagnostic codes. Some psychiatrists feel that the criteria doesn't always match the clinical presentations

Table 2.2: A list of 20 most common diagnoses in the research data.

Diagnosis	Count	Name
C61	2345077	Malignant neoplasm of prostate
I10	1577908	Essential (primary) hypertension
I48	1297654	Atrial fibrillation and flutter
N40	978068	Hyperplasia of prostate
Z491	724170	Extracorporeal dialysis
I251	580188	Atherosclerotic heart disease
K021	563659	Caries of dentine
N189	547090	Chronic renal failure, unspecified
E119	529111	Non-insulin-dependent diabetes mellitus - Without complications
G473	447118	Sleep apnoea
N0839	433867	Glomerular disorders in diabetes mellitus microalbuminuria
I702	384798	Atherosclerosis of arteries of extremities
J448	370840	Other specified chronic obstructive pulmonary disease
E11	361192	Non-insulin-dependent diabetes mellitus
E112	348224	Non-insulin-dependent diabetes mellitus - With renal complications
J189	339660	Pneumonia, unspecified
I509	303914	Heart failure, unspecified
G20	279107	Parkinson's disease
H903	263872	Sensorineural hearing loss, bilateral
R104	251966	Other and unspecified abdominal pain

Table 2.3: A summary of the combination code types of the ICD-10 system.

Code 1	Delimiter	Code 2
Causal code	*	Symptom code
Tumor code	&	Tumor symptom code
Overdose/poisoning code	#	ATC code

of the disease and that the criteria is overly complex in general [18]. Additionally, the author holds the opinion that the need for country-specific versions of the coding system can be considered a severe flaw.

2.1.2 NCSP

Finland is currently using a modified version of the Nordic Classification of Surgical Procedures (*NCSP*) as the coding system for medical procedures. NCSP was released in 1996 by Nordic Medico-Statistical Committee (*NOMESCO*) [1]. NCSP procedure codes are alphanumeric sequences of 5 characters. It is also a hierarchical system, as codes with fewer characters can form categories of connected procedure codes, but the categories are not procedures by themselves. This is in contrast with the ICD-10 system, which uses the optional suffixes for encoding additional information. The limitation to codes of length 5 means that the NCSP system is not necessary to be described in detail here, as the static structure will not require any particular concern while processing the research data. A separate procedure coding system called *PTHAVO* or *SPAT* is concurrently used in outpatient care. While SPAT codes were used in defining the research cohort, the analysed procedures themselves are not done in outpatient care so a detailed description of the SPAT system is also omitted [41].

2.2 Prostate

The prostate is a gland (a substance-producing organ) that is a part of the male reproductive system. Figure 2.1 illustrates the structure of the system. As can be seen from the picture, the prostate resides directly under the bladder. The location in the body is also the etymology of the term, as the word *prostate* originates from the Greek expression for "*one who stands before*". When viewing the male reproductive system from below, the prostate appears to be in front of the bladder. The prostate weighs about 30 grams and is about the size of a chestnut. The main function of the prostate is to produce fluid that is a key ingredient of semen. It also transforms testosterone to a more potent form called dihydrotestosterone (*DHT*). The prostate is surrounded by connective tissue and muscle fibers. These muscles also enable ejaculation by expelling the semen through the urethra [2].

The structure of the prostate comprises of three different parts. These can thought of as layers, starting from the around the center of the gland. The transition zone is located there surrounding the upper part of the urethra. The central zone resides around the transition zone. The seminal duct and the seminal vesicles (*the ejaculatory duct*) are located in the transition zone. The peripheral zone is the last layer. The tissue mass of

the prostate is divided to approximately 10% for the transition zone, about 20% for the central zone and around 70% for the peripheral zone. The transition zone tissue is the area which benign prostatic hyperplasia affects. Malignant growth, i.e. cancers, tend to affect the peripheral zone [2].

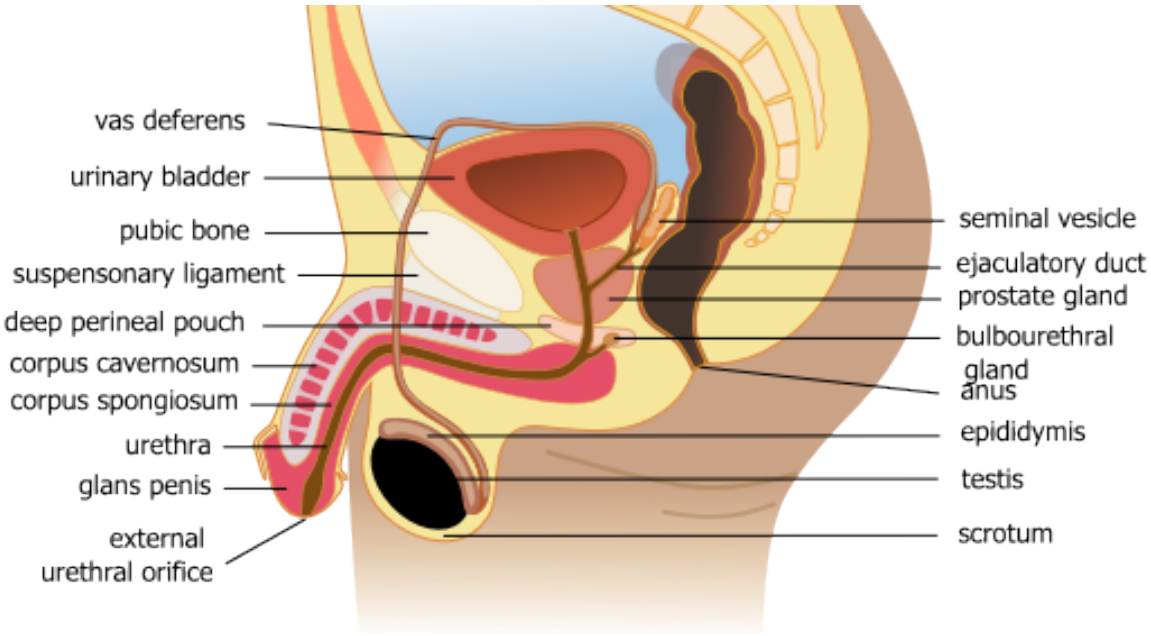

Figure 2.1: Human male reproductive system © 2021 by Wumingbai is licensed under CC BY-SA 4.0. To view a copy of this license, visit http://creativecommons.org/licenses/by-sa/4.0/ [67].

2.3 Benign prostatic hyperplasia (BPH)

Benign prostatic hyperplasia (*BPH*) is a common disease among the male population. As the medical term suggests, BPH refers to increased non-cancerous cell production in the prostate and the associated connective tissue. Worldwide autopsy studies have shown that BPH affects around 10% of the men in their 30s, 20% of the men in their 40s, 50% of the men in their 60s, and between 80% and 90% of the men in their 70s and 80s. It is postulated that most men would develop a some degree of BPH if they would live long enough. Age and normal androgen production are the only known risk factors in BPH. Increased physical activity and moderate alcohol consumption might have a preventative effect. The role of genetics is possible but remains unconfirmed [51,55].

Table 2.4: A list of the investigated BPH procedures.

NCSP code	Name	Abbreviation
KED00/KED10	Transvesical prostatectomy / Retropubic prostatectomy	OSP
KED22	Transurethral resection of prostate	TURP
KED33	Transurethral incision of prostate	TUIP
KED52	Laser resection of prostate	VLAP

BPH is not always a clinical condition. It transforms into one only if the patient presents with any of the associated symptoms. Three categories of symptoms are especially prevalent. Lower urinary tract symptoms (*LUTS*) can negatively affect the flow of urine and the need to urinate. Benign prostatic enlargement (*BEE*) refers to an increase in the size of the prostate. Bladder outlet obstruction (*BOO*) is a blockage in the bladder caused by a pressure gradient at the bladder neck and the prostatic urethra. It is important to note that BPH, LUTS, BEE and BOO can all also occur independently of each other [51,55].

Disease development varies greatly among patients. With a follow-up time of 3-5 years, two thirds of the patients experience either no change or even an improvement in their symptoms, but a longer follow-up time generally makes them worse. Patients with mild or moderate symptoms and no complications often require no treatment as long as the progression is monitored annually and the symptoms do not largely affect the quality of life. The predicted outcome in these cases is often determined by the size of the prostate. A clear increase in the size tends to result in increased severity later. In the cases where treatment is necessary, medication is initially considered if it is enough to either remove or substantially reduce the symptoms. However, the efficacy of the BPH medications is weaker than the surgical intervention options [51,55]. The NCSP codes of the BPH procedures investigated in this paper are listed in Table 2.4. Table 2.5 lists the procedures that are considered as reoperations after the initial BPH procedure. The ICD-10 code for BPH is *N40.1*.

Table 2.5: A list of the investigated BPH reoperations.

NCSP code	Name	Abbreviation
KED00/KED10	Transvesical prostatectomy / Retropubic prostatectomy	OSP
KED22	Transurethral resection of prostate	TURP
KED33	Transurethral incision of prostate	TUIP
KED52	Laser resection of prostate	VLAP
KED62	Transurethral needle ablation of prostate	TUNA
KED72	Transurethral microwave therapy of prostate	TUMT
KED76	Transurethral electrovaporization of the prostate	TUVP

2.4 Urethral stricture

Urethral stricture is a narrowing of the urethra that is caused by scar tissue. The causes of urethral strictures can be divided into 4 categories. These are unknown causes (*idiopathic*), side effects from other procedures (*iatrogenic*), inflammatory diseases and physical injuries (*traumatic*). Patients of all ages can develop an urethral stricture, but it is more common in the older population. Urethral stricture is rare in women [6,62]. Urethral stricture incidence rate for men in the UK has been estimated to range from 10 to 100 per 100 000 men, depending on the age bracket [46].

Symptoms of an urethral stricture are similar to the lower urinary tract symptoms of BPH. If left untreated, they can in some cases also develop into a life-threatening renal failure. Treatment choice depends on the symptom profile. If the symptoms are not troublesome, a treatment might not be necessary. If they are, a procedure is usually required [6,62]. The NSCP codes of the urethral stricture procedures that are investigated in this paper are listed in Table 2.6. The ICD-10 code for urethral stricture is *N35.9*.

Table 2.6: A list of the investigated urethral stricture procedures.

NCSP code	Name	Abbreviation
TKD00	Dilatation of urethra	Urethral dilation
KDV10/KDV12	Internal urethrotomy / Urethroscopic internal urethrotomy	Urethrotomy
KDH70	Plastic repair of stricture of urethra	Urethroplasty

2.5 Stress incontinence

Stress incontinence has been defined in several ways. The key symptom is an involuntary loss of urination control. The different definitions revolve around the clinical criteria according to which the diagnosis should be given. The differing criteria alternatives center around the involuntariness of the incontinence, whether it causes a social or a hygienic problem and the frequency of the incontinence [32].

Stress incontinence is a specific form if incontinence where an activity causes a sudden and involuntary loss of urine for the patient. Activities that can cause stress incontinence include effort, physical exertion, sneezing and coughing. This form incontinence is notably more prevalent in women than men. Stress incontinence can be caused by anything that either directly injures or reduces the capacity of maintaining suitable resistance of the urinary sphincter. Such causes be be categorized to neurological causes, side effects of BPH and prostate cancer treatments and physical injuries (*traumas*). Aging is a risk factor in stress incontinence for both sexes, but more steeply for women [20].

Stress incontinence is often treated with lifestyle changes. If these are not sufficient to control the condition, physical therapy and medications that improve the bladder control can be tried. Severe cases can require a surgical procedure [20]. The NCSP codes for the stress incontinence procedures investigated in this paper are listed in Table 2.7. The ICD-10 code of stress incontinence is *N39.3*.

Table 2.7: A list of the investigated stress incontinence procedures.

NCSP code	Name	Abbreviation
KDK00/KDK10	Implantation of artificial urinary sphincter around bladder neck / Implantation of artificial urinary sphincter around bulbar urethra	AUS
KDG43	Transobturatorial sling urethrocystopexy	Sling

3. Methods

The aim of this chapter is to explain and motivate the chosen methodology of the thesis. The methodology is composed of three distinct parts. These are the the Charlson Comorbidity Index and its computation, the survival analysis methods and the data preparation process. Each part will be detailed in their respective sections. The structure and the content of the research data will be described along with the data preparation process.

3.1 Charlson Comorbidity Index

Treatment efficacy studies are burdened by the fact that the treated patients are not limited to having only the disease that the treatment is meant for. These coexisting health problems are called *comorbidities*. The problem they cause stems from the fact that having multiple concurrent illnesses can make it difficult to accurately point out the cause of each of them. This can present a severe problem as the comorbidities can mask the effect of the studied treatment. Higher number of comorbidities, i.e. a high disease load, can also increase the risk of the patient to drop out of the study before completion. As the treatments are often meant for severe diseases, accounting for comorbidities to prevent incorrect conclusions is of utmost importance [15].

A traditional way to do this has been to exclude any patients with a certain number of comorbidities to reduce the probability of the confounding influence. This approach is practical but has the downside of limiting the generality of the results. It can also make the recruitment of a sufficient number of patients problematic if too many possible candidates are excluded. Another way to solve the problem is to use statistical methods to control for the patient-specific comorbidities. The vast amount of possible comorbidities makes classifying them difficult. Stratifying (*partitioning*) the patients according to their disease load requires a quantitative measure to determine it and to compare it with others [15]. The Charlson Comorbidity Index provides such a measure, which has also shown to predict various clinical outcomes well [14,16].

The CCI is not the only way to quantify the disease load. While the underlying basic idea of quantifying the disease load is shared among the alternatives, their predictive performance differs depending on the task. Austin et al. provide a general proof for

the mathematical validity of comorbidity measures. Ou et al. provide a performance comparison of several different comorbidity measurement schemes in various predictive tasks for an American diabetic patient cohort. The comparison includes the Charlson Comorbidity Index (*CCI*), Elixhauser Index (*EI*), Chronic Disease Score (*CDS*), Health-related Quality of Life Comorbidity Index (*HRQL-CI*) [48].

The viability of the CCI score for the prediction of the BPH procedure complications was chosen to be investigated because it would offer a convenient tool for clinicians if found to be effective. Additionally, to the author's knowledge, no previous work has investigated the predictive power of the CCI scores for the prediction of BPH complications.

Only the CCI is detailed in this chapter as the other mentioned measures aren't considered in the analyses of this thesis, but an interested reader is directed to Charlson et. al for the CCI, Elixhauser et al. for the EI, von Korff et al. for the CDS and Muhkerjee et al. for the HRQL-CI [15,23,45,64].

3.1.1 Definition

Charlson Comorbidity Index (*CCI*) is a weighted index that was developed to quantify the severity of a patients comorbidities to a single number. CCI was empirically developed by following a cohort of all the patients admitted in a 1-month period to the New York Hospital-Cornell Medical Center for a year and analyzing the results of their outcomes afterwards. The authors later validated the index as a good predictor of death due to a comorbidity on a cohort of breast cancer patients [15].

CCI consists of 17 different comorbidity categories. These are essentially groups of diseases that share characteristics and a similar risk of mortality. Each of the categories is assigned a score representing that risk. The scores range from 0 to 6 and increase with the severity of the risk. A total score is formed by summing the scores of all of the categories that the patient has received diagnoses from. Each category is only counted once even if the patient has multiple diagnoses from it. Some pairs of categories are also exclusionary due to representing a milder and a severe form of the same disease. For instance, if a patient has diagnoses of both diabetes without and with chronic complications, the one without chronic complications will not contribute to the patients total score [15].

As the original CCI was developed by assessing the outcomes of real patients in 1987, the scoring system was dependent on the state of the medical science at that particular time. As advances in medicine can have an effect on the treatability of different comorbidities, Quan et al. published a revision to the scoring system in 2010 that attempted to improve it to better fit the changed realities [49]. Both the original and the revised scoring systems are listed in Table 3.2. The exclusionary categories are shared by both versions and can be found in Table 3.1.

Table 3.1: Exclusionary categories of the CCI

Mild form	Severe form
Diabetes without chronic complication	Diabetes with chronic complication
Renal disease	Moderate or severe liver disease
Any malignancy, including lymphoma and leukemia, except malignant neoplasm of skin	Metastatic solid tumor

As the comorbidity index was originally developed for empirical use, it made sense to describe the diseases with their names. Retrospective studies are often using administrative health care data which contains the diagnosis information encoded with some coding system. The medical community has developed several coding systems for storing various types of health care data. Their general idea is to assign a short alphanumeric code for each recognized entity [50].

Quan et al. developed a mapping between the disease categories defined by the original CCI publication and ICD-10 codes [50]. The mapping can be found in Table 3.3. An important observation about the mapping is that it is defined only for codes with a length of 4 or less characters. Any country-specific alterations of the ICD-10 system were not considered which means that the mapping should be valid for international source data. This was also confirmed by Sundararajan et al., who found the Quan mapping to perform better in predicting hospital mortality than some of the alternatives for an Australian patient cohort [54].

The hierarchical structure of the ICD-10 system serves as a good practical example to motivate the use of a comorbidity classification instead of the individual diagnoses. ICD-10 spans tens of thousands of diagnostic codes and most of them are rarely if ever used. Figure 3.1 depicts the skewness of the distribution of the unique diagnosis codes in the research data. As similar diagnoses can have a huge number of separate codes, grouping them together is important to prevent the less used ones from appearing statistically irrelevant [66]. Figure 3.2 displays the occurrence distribution of the Charlson Comorbidity Index comorbidities in the research data.

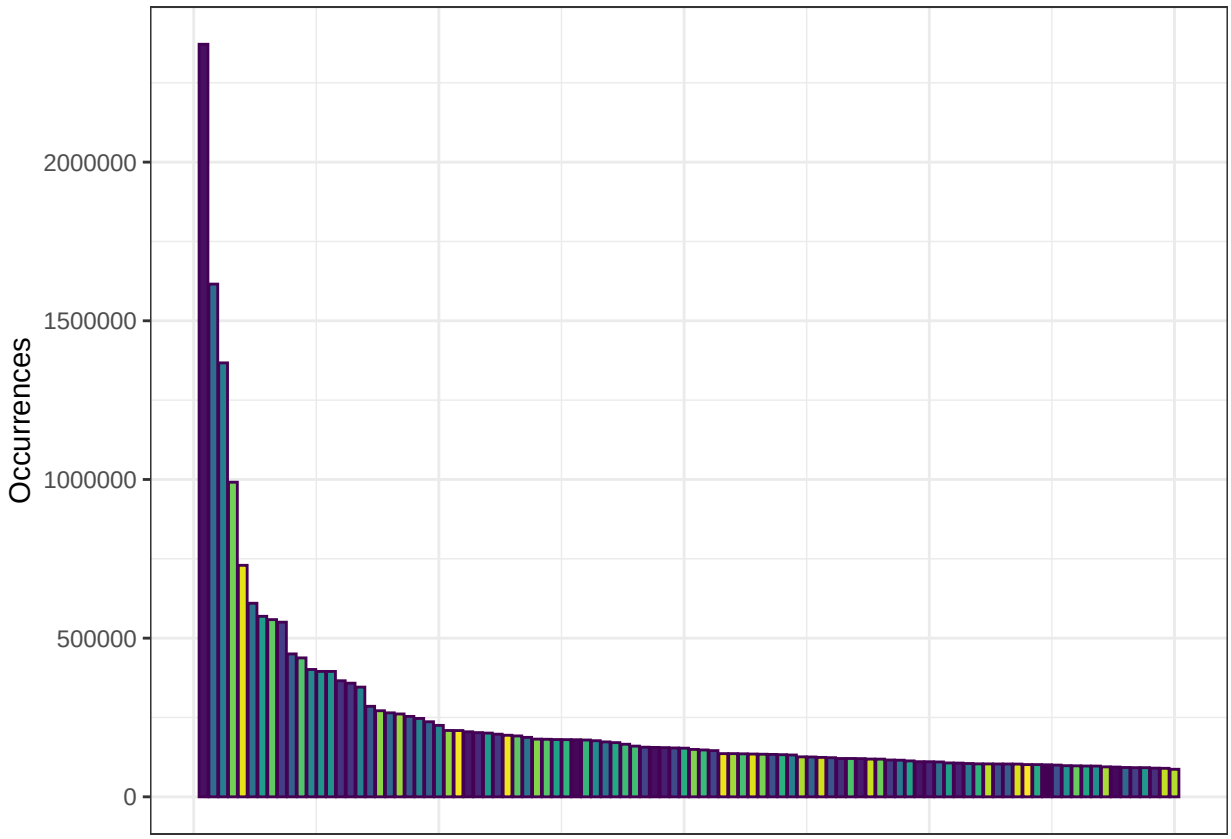

Figure 3.1: Occurrences of the 100 most common ICD-10 codes sorted by occurrence count in a decreasing order.

Table 3.2: Diseases of the CCI with their corresponding weights

Disease	Charlson weight	Quan weight
Myocardial infarction	1	0
Congestive heart failure	1	2
Peripheral vascular disease	1	0
Cerebrovascular disease	1	0
Dementia	1	2
Chronic pulmonary disease	1	1
Rheumatologic disease	1	1
Peptic ulcer disease	1	0
Mild liver disease	1	2
Diabetes without chronic complication	1	0
Diabetes with chronic complication	2	1
Hemiplegia or paraplegia	2	2
Renal disease	2	1
Any malignancy, including lymphoma and leukemia, except malignant neoplasm of skin	2	2
Moderate or severe liver disease	3	4
Metastatic solid tumor	6	6
AIDS/HIV	6	4

Table 3.3: Diseases of the Charlson Comorbidity Index with their corresponding ICD-10 codes

Disease	ICD-10 codes
Myocardial infarction	I21.x, I22.x, I25.2
Congestive heart failure	I09.9, I11.0, I13.0, I13.2, I25.5, I42.0, I42.5-I42.9, I43.x, I50.x, P29.0
Peripheral vascular disease	I70.x, I71.x, I73.1, I73.8, I73.9, I77.1, I79.0, I79.2, K55.1, K55.8, K55.9, Z95.8, Z95.9
Cerebrovascular disease	G45.x, G46.x, H34.0, I60.x-I69.x
Dementia	F00.x-F03.x, F05.1, G30.x, G31.1
Chronic pulmonary disease	I27.8, I27.9, J40.x-J47.x, J60.x-J67.x, J68.4, J70.1, J70.3
Rheumatologic disease	M05.x, M06.x, M31.5, M32.x-M34.x, M35.1, M35.3, M36.0
Peptic ulcer disease	K25.x-K28.x
Mild liver disease	B18.x, K70.0-K70.3, K70.9, K71.3-K71.5, K71.7, K73.x, K74.x, K76.0, K76.2-K76.4, K76.8, K76.9, Z94.4
Diabetes without chronic complication	E10.0, E10.1, E10.6, E10.8, E10.9, E11.0, E11.1, E11.6, E11.8, E11.9, E12.0, E12.1, E12.6, E12.8, E12.9, E13.0, E13.1, E13.6, E13.8, E13.9, E14.0, E14.1, E14.6, E14.8, E14.9
Diabetes with chronic complication	E10.2-E10.5, E10.7, E11.2-E11.5, E11.7, E12.2-E12.5, E12.7, E13.2-E13.5, E13.7, E14.2-E14.5, E14.7
Hemiplegia or paraplegia	G04.1, G11.4, G80.1, G80.2, G81.x, G82.x, G83.0-G83.4, G83.9
Renal disease	I12.0, I13.1, N03.2-N03.7, N05.2-N05.7, N18.x, N19.x, N25.0, Z49.0-Z49.2, Z94.0, Z99.2
Any malignancy, including lymphoma and leukemia, except malignant neoplasm of skin	C00.x-C26.x, C30.x-C34.x, C37.x-C41.x, C43.x, C45.x-C58.x, C60.x-C76.x, C81.x-C85.x, C88.x, C90.x-C97.x
Moderate or severe liver disease	I85.0, I85.9, I86.4, I98.2, K70.4, K71.1, K72.1, K72.9, K76.5, K76.6, K76.7
Metastatic solid tumor	C77.x-C80.x
AIDS/HIV	B20.x-B22.x, B24.x

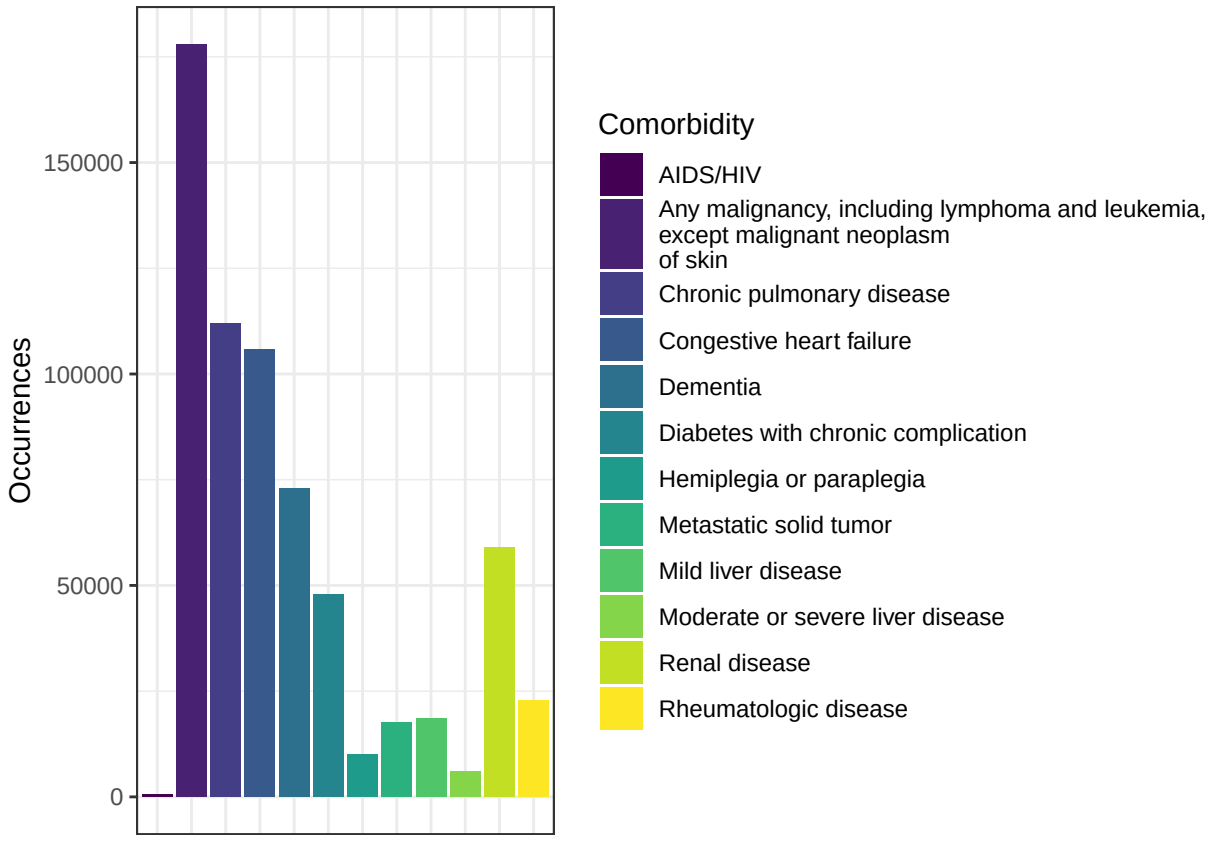

Figure 3.2: Occurrences of comorbidities in the research data.

3.2 Charlson Comorbidity Index on Spark

Current state-of-the art algorithm for computing the comorbidity scores is likely the R library **icd** published by Wasey et al. [66]. Another good alternative is the *comorbidity* R library by Gasparini et al [25]. However, both of these libraries ignore the temporal factor of the data by focusing on the computation of the CCI scores at a certain point in time. In other words, neither library will give you the history of the patient's CCI scores without running the algorithm several times on different data sets. This is understandable as both algorithms are designed to be ran on a single computer. Limiting the calculation to the total CCI score of the data set enables the algorithm designer to employ several clever design choices to reduce the resource usage. See Appendix A for a description of the *icd* algorithm. The next section details an alternative CCI score algorithm, which builds a timeline of the patient's CCI scores instead of just the total score. The usage of the big data processing platform Apache Spark enables the algorithm to perform well despite of the higher amount of work necessary.

3.2.1 Apache Spark

Apache Spark is a software platform for large-scale data processing. Spark aims to provide the infrastructure and the necessary functionality to enable users to easily write software code that scales from limited computers to huge computing clusters. At the core of Spark's computation model is the resilient distributed data set (*RDD*), which is a fault-tolerant immutable collection partitioned for parallel processing. The partitions can be stored in different *nodes* of a computing cluster. A Spark application (*driver program*) is built from operations that create and process RDDs. Before a Spark application is executed it is split into one or more *jobs* depending on which RDD operations are used. A directed acyclic graph (*DAG*) is calculated for each job representing the steps required for the computation of the job. In addition to assisting in the distribution of the tasks, the DAGs are used to optimize the execution plan. Jobs comprise of ordered *stages*. A stage is a set of *tasks*, which are the smallest unit of computation in Spark. Each task operates on a single partition. See Figure 3.3 for a graph of the computation model. After the whole execution plan is ready, the driver program requests *executors* from a *cluster manager*. Executors are processes which execute the tasks. They can reside in the same or a different computer than the driver manager. The executors will send the results back to the driver program after they are finished. Figure 3.4 summarizes the described structure of a Spark application [29,68].

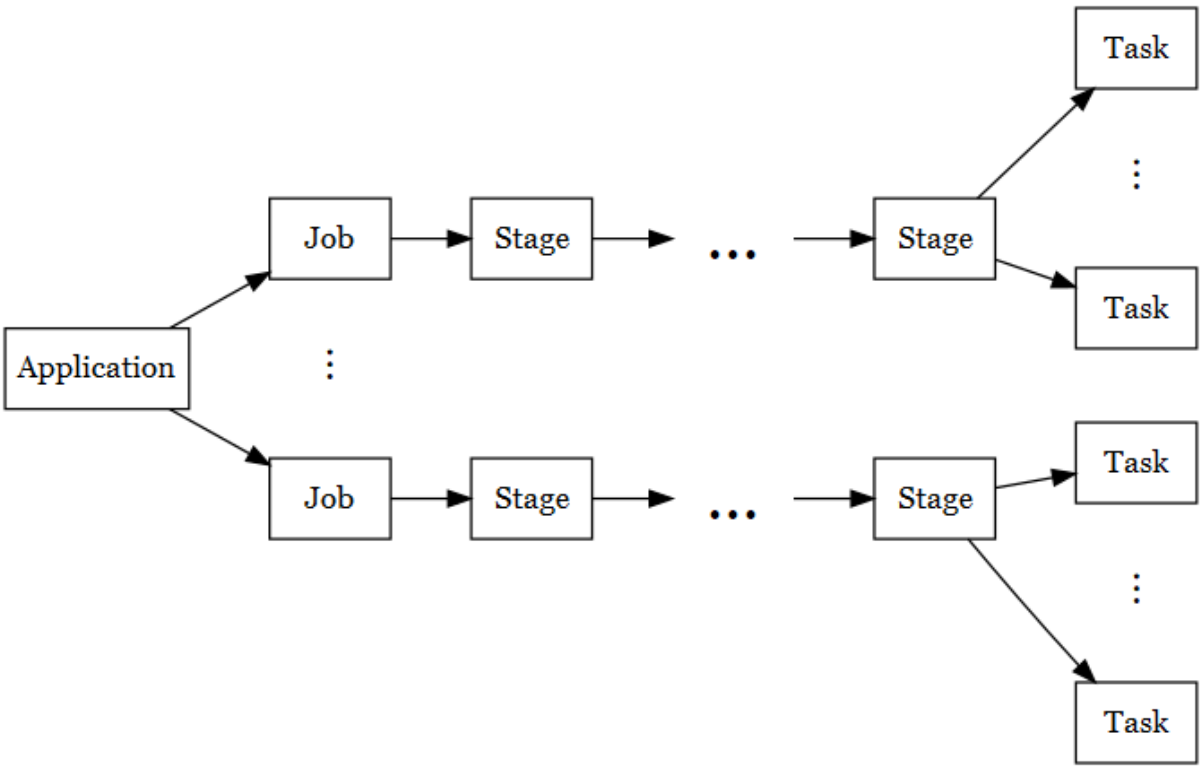

Figure 3.3: An outline of the computation model of a Spark program.

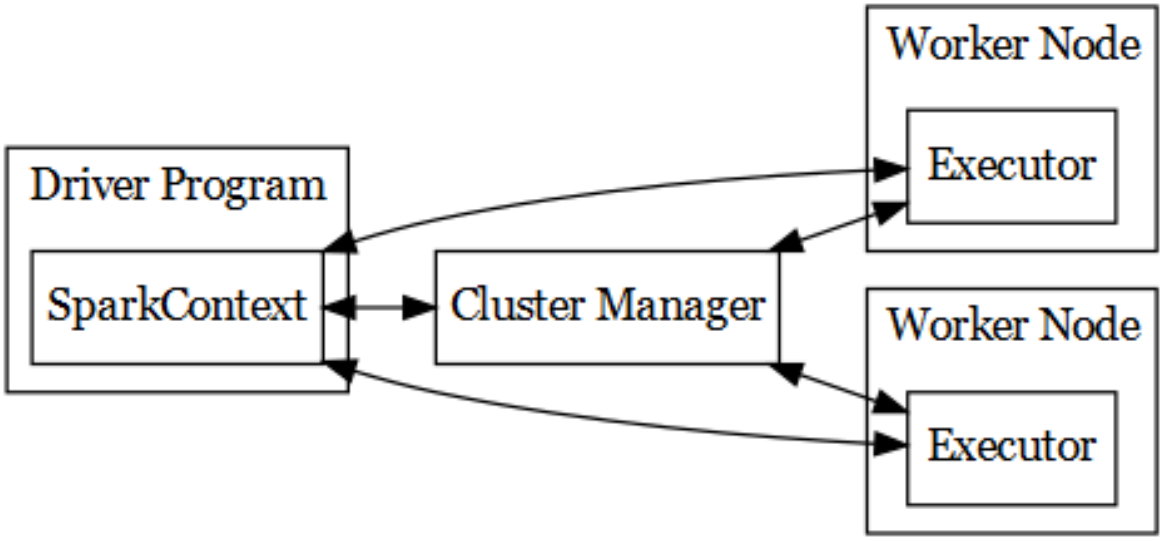

Figure 3.4: An outline of the structure of a Spark program.

Spark exposes an application programming interface (*API*) for Python, Scala, Java and R. The core programming model of Spark can be extended with modules that serve as a replacement for specialized tools meant for different workloads [68]. Spark SQL is the module meant for processing structured data. In addition to the typical tabular data analysis tools, Spark SQL can function as a Structured Query Language (*SQL*) engine. SQL is a standardized domain-specific programming language used to operate relational databases [38]. SQL is declarative by nature, which means that the user expresses their outcome intent instead of the procedural alternative of explicit instructions on how to achieve that outcome. Spark SQL translates the SQL query to Spark's native functions while also providing superior performance compared to the RDD API as knowing the data structure allows better optimization [5]. Another upside of writing an algorithm in SQL is that assuming that the standard is followed, the used SQL engine is usually not locked to Spark SQL and can be chosen as desired [68]. A concrete benefit of the declarative SQL algorithm was realized when an update to Spark increased the performance of the algorithm on the same hardware by roughly a fifth with no changes.

3.2.2 Algorithm

When designing a new algorithm for calculating the CCI scores, it should be noted that while in an optimal case a patient would have all of their valid diagnoses stored in an electronic health record system at the time of their visit, in this practice this does not necessarily hold. A physician can just as well only store only the newly given diagnoses for a particular patient visit, which means that using the whole diagnosis history of a patient is the only way to ensure the accuracy of the CCI score calculation. Computing

the CCI scores for a set of patients can then be divided to following stages:

1. Parse the either the Charlson or the Quan version of the comorbidity mapping between the CCI diseases and the ICD-10 codes
2. Match the diagnoses to the corresponding CCI categories
3. Sum the scores of each category with one or more diagnoses
4. If a patient has diagnoses of both forms of any exclusionary diseases, reduce the score of the mild form from the total score

Computational problems of the algorithm are mainly string matching, taking into account the logic of the ICD-10 coding system and handling the temporal components of stages 3 and 4. Assuming that the data has no illogical entries, stage 2 of the above can be represented by the following naive pseudocode:

```
for each patient:
  for each visit of the patient:
    for diagnosis of the visit:
      for each possible comorbidity:
        for each code of the comorbidity:
          if diagnosis is comorbidity code:
            if no preceding occurrences of the same diagnosis
               category:
              if preceding occurrences of milder form of the
                 disease:
                adjust score
              add comorbidity score of the category to the
                 total
```

As is evident, a naive implementation of the algorithm would be extremely slow. Electronic health record systems can contain hundreds of millions rows of data, so nested loops without any special considerations would not scale well. In addition to the performance considerations, stage 1 of the algorithm also requires some thought. As mentioned, Quan et al. defined their mapping between the ICD-10 codes and the CCI diseases for ICD-10 codes up to length of 4 characters despite the fact that valid ICD-10 codes can be longer than that. The mapping is also not defined in terms of exact codes, but as ranges with following forms (the delimiter symbol . is not counted as a character) [50]:

- I13.0
- I42.5-I42.9
- I21.x
- C77.x-C80.x

Table 3.4: The schema of the 'diagnosis' table.

Column	Type	Purpose
id	string	patient identifier
date	datetime	date of diagnosis
diagnosis	string	ICD-10 code

The Finnish Institute for Health and Welfare (THL) publishes a list of currently valid ICD-10 codes [60]. The list is defined as codes ranging between 3 and 5 characters in length. Because the latter characters only add details to the diagnosis, Quan's ranges can be converted to codes of the aforementioned lengths as follows:

- I13.0 -> I130, I1300-I1399
- I42.5-I42.9 -> I425-I429, I4250-I4299
- I21.x -> I21, I210-I219, I2100-I2199
- C77.x-C80.x -> C77-C80, C770-C809, C7700-C8099

Each of the generated codes must then be checked against the THL's list to ensure that it is actually a valid diagnosis code in addition to just fulfilling the syntax requirements of an ICD-10 code. When this process is done, the generated mapping can be used with naive string matching against the patient diagnosis codes truncated to the length of 4 characters. All of the ranges include codes of length 4, so such code must either exist in the generated mapping or not be a comorbidity code.

The new CCI algorithm was written in SQL as a Spark SQL application. The design of the algorithm requires two Spark SQL tables to be setup prior to execution. Because the data files can't be distributed alongside of the thesis, only the contents of the tables are described here. *diagnosis* holds the information about the patients and their diagnoses. The schema is summarized in Table 3.4. *cci* stores the information about the ICD-10 codes that describe a disease in the CCI and the associated CCI score. The schema can be found in Table 3.5. It should be noted that the disease categories of the CCI are mapped to integers to improve performance in the *disease* column of the *cci_table* table.

Given these tables, the following query adds information about the temporal ordering of the diagnoses, whether the diagnosis is a CCI disease and whether the diagnosis is either a mild form or a severe form of a pair of exclusive diagnoses to the diagnosis table (see Spark documentation for an explanation about the broadcast join hint [4]):

```
SELECT
/*+ BROADCAST(cci_table) */
diagnoses.*,
```

Table 3.5: The schema of the 'cci' table.

Column	Type	Purpose
disease	integer	disease
diagnosis	string	ICD-10 code
points	integer	CCI score
mild	integer	mild form of the disease (null if doesn't exist)
severe	integer	severe form of the disease (null if doesn't exist)
change	integer	-1 * (severe - mild) (null if doesn't exist)

```
ROW_NUMBER() OVER (PARTITION BY id, disease ORDER BY diagnoses.
    date) AS dg_ix,
cci_table.points,
cci_table.disease,
cci_table.disease_name,
cci_table.mild,
cci_table.severe,
cci_table.change
FROM diagnoses
LEFT JOIN cci_table
    ON (substring(diagnoses.diagnosis, 1, 4) = cci_table.
        diagnosis)
```

The following query selects the rows where the diagnosis is not a comorbidity, the diagnosis is not the first diagnosis of that disease and the ones where the diagnosis is a comorbidity but has no exclusionary diseases (in other words, the rows that require no further changes) while setting the CCI score to 0 for the first two row types and keeping the score assigned in the query above for the last one:

```
SELECT
ordered_diagnoses.id,
ordered_diagnoses.date,
ordered_diagnoses.age,
ordered_diagnoses.visit_duration,
ordered_diagnoses.diagnosis,
ordered_diagnoses.disease_name,
CASE
WHEN (ordered_diagnoses.dg_ix==1) AND (ordered_diagnoses.points
    IS NOT NULL) THEN ordered_diagnoses.points
```

```
ELSE 0
END as points
FROM ordered_diagnoses
WHERE ((ordered_diagnoses.disease IS NULL) OR (
    ordered_diagnoses.dg_ix != 1) OR (ordered_diagnoses.change
    IS NULL))
```

The following query selects the rows which do need to be checked for whether they need to be modified due to preceding or succeeding exclusionary diseases:

```
SELECT
ordered_diagnoses.*,
ROW_NUMBER() OVER w AS override_ix,
COUNT(1) OVER w AS count
FROM ordered_diagnoses
WHERE ((ordered_diagnoses.change IS NOT NULL) AND (
    ordered_diagnoses.dg_ix = 1))
WINDOW w AS (PARTITION BY id, (ordered_diagnoses.mild *
    ordered_diagnoses.severe) ORDER BY date, points)
```

The final query applies combines the previous query results and corrects the score in the cases where an appropriate exclusionary disease diagnosis exists:

```
SELECT * FROM diagnoses_no_changes
UNION
SELECT
diagnoses_changes.id,
diagnoses_changes.date,
diagnoses_changes.age,
diagnoses_changes.visit_duration,
diagnoses_changes.diagnosis,
diagnoses_changes.disease_name,
CASE
WHEN (diagnoses_changes.disease=diagnoses_changes.mild) AND (
    diagnoses_changes.override_ix==2) THEN 0
WHEN (diagnoses_changes.disease=diagnoses_changes.severe) AND (
    diagnoses_changes.override_ix==2) THEN (diagnoses_changes.
    points + diagnoses_changes.change)
ELSE diagnoses_changes.points END as {points_col_name}
FROM diagnoses_changes
```

The resulting algorithm calculates the CCI scores for 54 158 895 diagnoses in under five minutes on a modest laptop with an *Intel(R) Core(TM) i5-8350U* CPU and 16 GB of RAM while Spark was limited to using 10 gigabytes of the available RAM. An accurate benchmarking between the algorithms on a smaller data was not done as Spark's lazy execution makes it difficult to compute the full result of the algorithm without some follow-up I/O operations such as writing the result to a file. While similar I/O burden could be added to the *icd* algorithm, the differences in the I/O libraries would make accurate comparison impossible. It is also of questionable value to compare the running times as the goals of the algorithms are not directly comparable. *icd* algorithm has to be ran repeatedly to build a history of the CCI scores whereas the Spark algorithm does it in a one go. Despite the aforementioned limitations, the algorithm performance was still measured to be in roughly same ballpark as the *icd* implementation. The correctness of the Spark algorithm was verified to match the results of another R implementation in the *comorbidity*.

3.3 Survival analysis

The theory of survival analysis is rooted in advanced probability theory concepts, which necessitates a limitation in depth for the exposition presented here due to the high number of necessary topics. The most notable omission is the generalization of survival analysis methods to the counting process and martingale theory. The methods are detailed in the depth necessary to have a high-level understanding of the methods used to attain the results of the next chapter. The reader is assumed to be familiar with elementary statistics, probability theory and calculus concepts.

Survival analysis is the branch of statistics that provides methods for analyzing the distribution of the amount of time taken before a subject experiences an *event of interest* [36]. An equivalent way to define survival analysis is as the analysis of the distribution of the *survival time* before experiencing a *failure*. The latter definition is a historical artifact stemming from the survival analysis' roots in being used in the analysis of mortality rates [13]. The former definition is likely less confusing as it emphasizes the fact that the event of interest can be freely chosen and is not limited to death. In addition to death, other typical events of interest studied in cancer research include disease recurrence and treatment response [17].

A typical survival analysis scenario defines an *entry event* that the subjects must experience to enter the studied cohort and what is the event of interest being studied. The maximum *observation period* that the subject will be followed for after the entry event is usually limited to a some fixed amount of time. The collected data is referred to as *time-event-data*. Sometimes an event of interest can be precluded by another event. For

example, a total thyroidectomy would prevent a subject from dying from thyroid cancer. These scenarios are collectively called *competing risks* and have their own specialized analysis methods. Survival analysis also includes regression analysis methods to determine the effect of covariates on the survival time [33].

An important characteristic of time-to-event data is incomplete data, which can happen for many different reasons. For example, a subject might drop out of the study before experiencing the event or they might experience the event after the observation period has ended. Survival analysis deals with subjects with missing event observations by *censoring* the corresponding subjects. Subjects can be either right censored, left censored or interval censored. Subjects are *right censored* if they have not experienced the event of interest during their observation period. Conversely, the subjects who have experienced the event of interest before their entry event are *left censored*. If a subject is known to have experienced the event of interest during a certain interval of the observation period but the exact time is unknown, the subject is *interval censored* [36].

Censoring is important because censored subjects are handled differently to the uncensored ones by the survival analysis methods. It is usually assumed that censoring is non-informative, which means that the censoring time should be independent of the event of interest. An illustrative counterexample where the independence assumption does not hold is a scenario of a sick patient leaving a treatment efficacy study because their disease symptoms get worse and they are no longer able to participate in the study. Worsening symptoms are clearly related to whichever suitable event of interest is being studied [36].

Time-to-event can also be *truncated*. Truncation sets either a lower or an upper limit for the observation period length. Left truncation happens when a subject starts being observed some time after their entry event to the research cohort. This can happen for instance if patients who have been diagnosed with the studied disease prior to study enrollment are recruited to the study. Right truncation happens when the lengths of the observation periods have an upper limit. The situation can occur e.g. when the study has a fixed end date. If the subjects have not started their observation periods at the same time, the latter ones will not be observed for the full duration [36].

3.3.1 Survival distribution

A *procedure* is called random if it can be repeated infinitely many times and it has a well-defined set of multiple possible outcomes. Probability theory is the mathematical framework for studying random procedures. The sample set Ω is the collection of all the possible outcomes ω of a random procedure. The subsets of Ω are called *events*. Events are realized by outcomes. For example, when throwing a single die, the sample set is $\Omega = \{1, 2, 3, 4, 5, 6\}$ and the event of getting an odd number is realized by any of the

outcomes in the outcome subset $A = \{1, 3, 5\}$. F is a σ-algebra on Ω that defines the observable events, i.e. those that are assigned a probability value. A σ-algebra is defined so that [12]:

1. the empty set $\emptyset$ (*impossible event*) and the Ω (*certain event*) belong to it
2. if a subset (event) belongs to it, so does the complement of the subset
3. if the subsets A_1, A_2... belong to it, so does $\cup_{k=1}^{\infty} A_k$

A random variable is a function $X : \Omega \to R$ for which $\{\omega; X(\omega) \leq a\} \in F$, $a \in R$ holds. Let the random variable T represent the time from the start of an observation period to the event of interest. The probability distribution of T is called the *survival distribution*. The survival distribution can be characterized with several different representations. These include the *survival function* S, the failure function F, the probability density function f, the *hazard function* h and the *cumulative hazard function* H. The survival function represents the probability of survival at least up to the point in time t [33]:

$$S(t) = P(t < T)$$

for which holds

$$0 \leq S(t) \leq 1$$
$$S(0) = 1$$
$$\lim_{t \to \infty} S(t) = 0$$
$$\frac{\partial S(t)}{\partial t} \leq 0$$
$$\frac{\partial^2 S(t)}{\partial t^2} \neq 0.$$

The mean of the survival distribution is defined in terms of the survival function [33]:

$$\mu = \int_0^{\infty} S(t) dt.$$

If some of the subjects do not experience the event of interest, the integral will be infinite and so the mean can't be calculated. A solution in such cases is to define a maximum survival time in order to make the integral finite. The median of the survival distribution is defined as the smallest time t at which $S(t) \leq 0.5$. It is undefined if the survival curve never reaches 0.5 [44].

Another way to convey the same information is to use the complement of the survival function, which is referred to as either the failure function or the *cumulative risk function*. In statistical sense, the failure function is the cumulative distribution function of the

random variable T. It represents the probability of failure before the point in time t and is defined as [33]:

$$\begin{aligned} F &= P(T \leq t) \\ &= 1 - S(t). \end{aligned}$$

The probability density function $f(t)$ is a positive function depicting the rate of change (or the slope) of the cumulative distribution function, which in this case is the failure function F. As the failure function is also the complement of the survival function, the probability density function and the associated relations can be stated as [33]:

$$\begin{aligned} f(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t} \\ &= \frac{\partial F(t)}{\partial t} \\ &= -\frac{\partial S(t)}{\partial t}. \end{aligned}$$

The hazard function (*hazard rate*) $\theta(t)$ is the non-negative instantaneous failure rate at the point in time t [33]:

$$\begin{aligned} h(t) &= \frac{f(t)}{1 - F(t)} \\ &= \frac{f(t)}{S(t)}. \end{aligned}$$

The cumulative hazard function $H(t)$ is the integral of the hazard function. It represents the area under the curve of the hazard function up to the point in time t [33]:

$$H(t) = \int_0^t h(u)du.$$

Given one of the survival distribution representations, all of the others can be derived. The following states all of the others in terms of the survival function [33]:

$$\begin{aligned} S(t) &= 1 - F(t) \\ S(t) &= \exp\left(-\int_0^t h(u)du\right) \\ S(t) &= \exp\left(-H(t)\right) \\ S(t) &= 1 - \int_0^t f(u)du. \end{aligned}$$

3.3.2 Survival distribution estimation

Medical studies are often characterized by the competing risks model, where a subject can exclusively experience either the main event of interest or one of the alternative events. For example, if one is studying disease recurrence, a patient might die before a disease can recur preventing the patient from experiencing the event of interest. Competing risks scenarios are often avoided in epidemiological studies by either censoring the patients when they experience one of the competing risks or forming a composite event of interest which includes the competing risks events. Both of the approaches have downsides. A composite event of interest makes the model difficult to interpret and can easily hide important relationships. This issue is compounded if the competing risks events and the event of interest are not similar. Censoring the competing risks overestimates the risk of the event of interest, because the patients are still considered to be in risk set for the main event of interest as censoring is used to indicate that the patient experiences the event some time after being censored [22]. A better solution is to use a statistical method which can handle the estimation of the survival function while accounting for the presence of the competing risks. The Fine-Gray estimator is a method designed for dealing with competing risks scenarios [24]. While the estimator is extremely popular due to its interpretability, it has some severe statistical issues [8]. The Aalen-Johansen estimator will be introduced next instead as a better behaving estimator for competing risks scenarios.

A discrete-time stochastic process with a state space E is a sequence of n random variables X_n with values in a set E. The notation $X_k = i$ means that the process is at state i at time k. A discrete-time stochastic process is called a Markov process if it fulfills the Markov property of the next state only depending on the current state [12]:

$$P(X_{n+1} = j | X_n = i, X_0 = i_0, ..., X_{n-1} = i_{n-1}) = P(X_{n+1} = j | X_n = i).$$

A Markov process is called non-homogeneous if the quantity $P(X_{n+1} = j | X_n = i)$ is dependent on the time n. The state space is defined hereafter as $E = \{0, 1, ..., k-1, k\}$. Transition intensity $\alpha_{gh}(t)$ denotes the instantaneous transition rate from the state $g \in E$ to the state $h \in E$, $g \neq h$. $P_{gh}(s, t)$ is the probability that a subject who is in state g at time s will later be in state g at time t, $s < t$. $P(s, t)$ is a $(k + 1) \times (k + 1)$ transition matrix containing the aforementioned probabilities for each state. The competing risks scenario can be modeled as a non-homogeneous Markov process by considering subjects to start at the initial state 0 (*transient state*) and defining states 1...k as the competing risks (*absorbing states*). The transition matrix $P(s, t)$ can be estimated using the Aalen-Johansen estimator $\widehat{P}(s, t)$ [11]:

$$\widehat{P}(s,t) = \prod_{s<t_j\leq t} (\mathbb{I} + \hat{\alpha}_j)$$

where

$$g, h \in E, g \neq h$$
$$t_j = \text{time with a transition}$$
$$d_{ghj} = \text{transitions from } g \text{ to } h \text{ at } t_j$$
$$d_{gj} = \sum_{g\neq h} d_{ghj} = \text{transitions out of } g \text{ at } t_j$$
$$r_{gj} = \text{subjects in } g \text{ just prior to } t_j$$
$$\mathbb{I} = (k+1) \times (k+1) \text{ identity matrix}$$
$$\hat{a}_j = (k+1) \times (k+1) \text{ transition intensity matrix}$$
$$\hat{a}_{ghj} = \frac{d_{ghj}}{r_{gj}}, g \neq h$$
$$\hat{a}_{ggj} = \frac{-d_{gj}}{r_{gj}}.$$

The covariance estimates for data with either no tied event times or where ties have been broken at random can be computed as follows (see Borgan et al. for the case with ties) [11]:

$$g, h, m, r \in E$$
$$\widehat{\text{cov}}(\widehat{P}_{gh}(s,t), \widehat{P}_{mr}(s,t)) = \sum_{i=0}^{k} \sum_{l\neq i} \sum_{s<t_j\leq t} \widehat{P}_{gi}(s,t_{j-1})\widehat{P}_{mi}(s,t_{j-1})(\widehat{P}_{lh}(t_j,t) - \widehat{P}_{ih}(t_j,t))(\widehat{P}_{lr}(t_j,t) - \widehat{P}_{ir}(t_j,t))(r_{ij}-1)r_{ij}^{-3}d_{ilj}.$$

The transition probabilities $P_{gh}(s,t)$ are often called the cumulative incidence functions (*CIF*) in the survival analysis literature. They can be thought of as risk-specific failure functions. Assuming time-to-event data without truncation (every subject starts at time 0), the CIF for the competing risk h is [11,59]:

$$
\begin{aligned}
CI_h(t) &= P_{0h}(t) \\
&= \int_0^t \alpha_{0h}(u)S(u)du \\
&\text{where} \\
\alpha_{0h}(u) &= \text{cause-specific hazard (transition intensity)} \\
S(u) &= \text{survival function to any event} \\
&= \exp\left(-\int_0^t \sum_{h=1}^{k} \alpha_{0h}(u)du\right).
\end{aligned}
$$

The Aalen-Johansen estimator $\widehat{P}_{0h}(t)$ for the CIF is [11]:

$$
\begin{aligned}
\widehat{P}_{00}(t) &= \text{the Kaplan-Meier estimator} \\
&= \widehat{S}(t) \\
&= \prod_{0<t_j\leq t} \frac{1-d_{0j}}{r_{0j}} \\
\widehat{\text{var}}(\widehat{P}_{00}(t)) &= \text{Greenwood's formula} \\
&= \widehat{S}(t)^2 \sum_{t_j\leq t} \frac{d_j}{r_j(r_j-d_j)} \\
\widehat{P}_{0h}(t) &= \sum_{0<t_j\leq t} \widehat{P}_{00}(t_{j-1})(\frac{d_{0hj}}{r_{0j}}) \\
\widehat{\text{var}}(\widehat{P}_{0h}) &= \sum_{s<t_j\leq t} \left(\widehat{P}_{00}(s,t_{j-1})\widehat{P}_{0h}(t_j,t)\right)^2 (r_{0j}-1)r_{0j}^{-3}d_{0j} \\
&+ \sum_{s<t_j\leq t} \widehat{P}_{00}(s,t_{j-1})^2(1-2\widehat{P}_{0h}(t_j,t))(r_{0j}-1)r_{0j}^{-3}d_{0j}.
\end{aligned}
$$

Confidence bounds of the Aalen-Johansen and the Kaplan-Meier estimators are based on their asymptotic multinormality and normality, respectively. The variance formulas above are again assuming untied data. See the original paper by Johansen et al. for further details on the Aalen-Johansen estimator [34]. Survival function estimation in cases with only one (absorbing) state can be done with the Kaplan-Meier estimator mentioned above. The details of the Kaplan-Meier estimator can be found in the original paper by Kaplan et. al [35].

A common research question is to determine whether there is a statistically significant difference between two different survival function estimates. The log-rank test is a statistical test derived for that purpose [43]. Given estimates S_1 and S_2, the log-rank test can be used to test the null hypothesis that $S_1(t) \neq S_2(t)$. First, assume two groups of subjects and define the following [44]:

$$
\begin{aligned}
d_{1i} &= \text{the number of events at time i for group 1} \\
d_{2i} &= \text{the number of events at time i for group 2} \\
d_i &= d_{1i} + d_{2i} \\
n_{1i} &= \text{the number of subjects at risk before time i for group 1} \\
n_{2i} &= \text{the number of subjects at risk before time i for group 2} \\
n_i &= n_{1i} + n_{2i}.
\end{aligned}
$$

Assuming that the events in each group are independent and there is no difference between the survival functions, a conditional probability for the number of events in the group 1 at the failure time i follows a hypergeometric distribution with the following expected value and variance [44]:

$$
\begin{aligned}
P(d_{1i}|n_{1i}, n_{2i}, n_i, d_i) &= \frac{\binom{n_{1i}}{d_{1i}}\binom{n_{2i}}{d_{2i}}}{\binom{n_i}{d_i}} \\
E(d_{1i}) &= \frac{n_{1i}d_i}{n_i} \\
\operatorname{var}(d_{1i}) &= \frac{n_{1i}n_{2i}(n_i - d_i)}{n_i^2(n_i - 1)}.
\end{aligned}
$$

A Cochran-Mantel-Haenszel test statistic following a large-sample chi-squared distribution with one degree of freedom can then be constructed [7]:

$$
\frac{\sum_{i=1}^{N}\left(d_{1i} - E(d_{1i})\right)^2}{\sum_{i=1}^{N}\operatorname{var}(d_{1i})} \sim \chi_1^2.
$$

When multiple tests are performed on the same data set, the results are generally recommended to be corrected for the multiplicative false positive (type I) error rate. One method to do the correction is the Benjamini–Hochberg (BH) procedure. The description of the method is omitted here as it is rather elementary and only one of the possible options, but the underlying idea is to scale the p-values upwards to control for the increased amount of false positives. See Benjamini et al. for the details method [10]. It should be noted that some researchers disagree with the idea of doing multiple test correction at all. Rothman et al. provide several compelling arguments [52].

3.3.3 Cox proportional hazard model

Regression analysis attempts to estimate the relationships between a set of covariates and an outcome variable. In the context of regression analysis, the previous definitions for the various representations of the survival distribution are amended to include a vector

of covariates X as an additional input. As an example, the hazard function would then be defined as $h(t, X)$ [33]. The Cox proportional hazard model (*CPH model*) is likely the most popular method for the regression analysis of time-to-event data. The covariates X_k are modeled as a linear combination of a weight vector β and a subject-specific covariate vector X: [19,33]:

$$\beta^T X = \beta_0 + \beta X_1 + ... + \beta X_k.$$

The CPH model belongs to a family of statistical models requiring that the data follows the so-called proportional hazards assumption (*PH assumption*), which states that the hazard function can be separated into a product [33]:

$$h(t, X) = h_0(t)\lambda(\beta X)$$

where

$h_0 =$ a non-negative baseline hazard function

$\lambda =$ a non-negative function scaling the baseline hazard for each subject.

The baseline hazard function depends only on the time t. λ incorporates the subject-specific variability in the hazard rate with exp being the most common function used as the λ. Assuming that the PH assumption holds and using exp as λ, the quotient of the hazard functions of subjects i and j at the point in time t_k results in [33]:

$$\begin{aligned} \frac{h(t_k, X_i)}{h(t_k, X_j)} &= \exp\left(\beta X_i - \beta X_j\right) \\ &= \exp\left(\beta[X_i - X_j]\right) \\ \text{if } X_{iy} \neq X_{jy} \text{ is the only difference, then} & \\ &= \exp\left(\beta_k[X_{iy} - X_{jy}]\right) \\ \text{if } X_{iy} - X_{jy} = 1, \text{ then} & \\ &= \exp\left(\beta_k\right). \end{aligned}$$

The last quantity is known as the *hazard ratio*. It corresponds to the proportionate change in the hazard rate when the covariate at the index y increases by one, assuming that all of the other covariates remain unchanged. The coefficient at the index y depicts the proportional change in the hazard from the absolute changes in the covariates as follows [33]:

$$\beta_y = \frac{\partial \log h(t, X)}{\partial X_y}$$

When the PH assumption holds, any differences between the subjects can be explained by scaling the baseline survival function [33]:

$$
\begin{aligned}
S(t, X) &= \exp\left(-\int_0^t h(u)du\right) \\
&= \exp\left(-\lambda \int_0^t h_0(u)du\right) \\
&= \exp\left(-\int_0^t h_0(u)du\right)^{\lambda}
\end{aligned}
$$

S_0 = the baseline survival function

$$
= (S_0(t))^{\lambda}.
$$

The key insight of the CPH model is that while the PH assumption holds, the estimation of the weight vector B does not depend on the baseline hazard function h_0. Cox proved this by defining the model in terms of a *partial likelihood* (*PL*) function of the failure times [31,33]. The intuition for the partial likelihood function can be explained without delving deep into the underlying theory. Assume a sample of n subjects of which k experience the event of interest. Assume further that the survival times of the subjects are in the vector t which is in sorted in increasing order and contains no ties. Define R_i as the set of the j subjects whose survival time Y_j is equal or greater than t_i. Now the conditional probability that the subject i experiences the event of interest at t_i given that the subject is in the risk set R_i and that a single subject experiences the event at t_i, i.e. the partial likelihood function for t_i, is defined as [31]:

$$
\begin{aligned}
L_i(\beta) &= P(A|B, C) \\
&= \frac{P(A, C|B)}{P(C|B)} \\
&= \frac{h_0(t_i)\exp(\beta^T X_i)}{\sum_{j \in R_i} h_0(t_i)\exp(\beta^T X_j)} \\
&= \frac{\exp(\beta^T X_i)}{\sum_{j \in R_i} \exp(\beta^T X_j)} \\
&= \frac{\exp(\beta^T X_i)}{\sum_{t_i \leq Y_j} \exp(\beta^T X_j)}
\end{aligned}
$$

where

A = subject i fails at t_i

B = subject $i \in R_j$

C = one subject fails at t_i.

By arguing that each of the subjects is independent of each other, a partial likelihood function for the joint set of failures can be formed [31]:

$$L(\beta) = \prod_{t_i \text{ with event}} \frac{\exp(\beta^T X_i)}{\sum_{t_i \leq Y_j} \exp(\beta^T X_j)}$$

$$\log L(\beta) = \sum_{t_i \text{ with event}} \left(\beta^T X_i - \log \left(\sum_{Y_i \leq Y_j} \exp\left(\beta^T X_j\right) \right) \right)$$

$$= l(\beta).$$

The log partial likelihood $\log L(\beta)$ has been shown to behave like a regular log likelihood function and can therefore be used to perform maximum likelihood estimation (*MLE*) [31] for the weight vector β. The first derivative of the log partial likelihood is called the *score* U [44]:

$$U(\beta) = \sum_{t_i \text{ with event}} \left(X_i - \sum_{Y_i \leq Y_j} X_j p(\beta, X_j) \right)$$

where

$$p(\beta, X_k) = \frac{\exp(\beta^T X_j)}{\sum_{Y_k \leq Y_j} \exp(\beta^T X_k)}.$$

The MLE of the Cox PH model is computed with an optimization algorithm from the partial likelihood equation [56]:

$$U(\hat{\beta}) = 0.$$

The confidence bounds of $\hat{\beta}$ are based on its asymptotic normality [56]. The variance of the score is called the *information* and it is equivalent to the negative second derivative of the log partial likelihood [44]:

$$I(\beta) = -\frac{\partial U(\beta)}{\partial \beta}.$$

The *Therneau-Grambsch test* is a diagnostic test developed to investigate the fulfillment of the proportional hazards assumption. The test is derived from the score of the log partial likelihood function. The individual terms of sum of the score function for each failure time i are called the *Schoenfeld residuals* r_i:

$$r_i = X_i - \sum_{Y_i \leq Y_j} X_j p(\beta, X_j).$$

They can be interpreted to roughly correspond to the difference of the observed covariate values and their expected value. The way the residuals help in testing the PH assumption is through another result, which states that adding the residuals r_i scaled with the coefficient variance to the corresponding coefficient estimate from a Cox model approximates time-varying coefficients [56]:

$$E(r_i^*) + \hat{\beta} \approx \beta(t)$$

where

$$r_i^* = \mathrm{Var}^{-1}(\hat{\beta}, t_i) r_i.$$

If the assumption of proportional hazards holds, a plot of the sum of the scaled Schoenfeld residuals and the coefficient estimate vector against time should be horizontal, i.e. the coefficients should not be time-dependent. The corresponding test statistic follows a chi-squared distribution with number of covariates n degrees of freedom. The test has separate forms for testing either the whole model's (*global*) or the individual covariate's compliance with the proportional hazards assumption. Because the Grambsch-Therneau test has the null hypothesis of proportionality, an insignificant result indicates that the assumption of proportional hazards is not violated [28].

Interpreting the results of the Grambsch-Therneau tests is highly dependent on the number of events experienced by the subjects of that particular cohort and any possible extreme outlier values. Very low event numbers will not yield trustworthy results and very high event numbers result in the test being too sensitive. This is a common problem in diagnostic tests, as the underlying statistical assumptions are often not met in the data depicting real world. In this case the assumption of proportional hazards will likely never be exactly true, so as more data is used the more likely the test is to detect the true non-proportionality [56].

The cases where the test shows a significant result of non-proportionality should therefore be confirmed by visually inspecting the corresponding plot and using domain knowledge to determine whether the actual amount of non-proportionality has a practical effect [56]. Therneau et al. and Grambsch et al. discuss the properties of the test in detail [28,56]. Keele et al. provide a treatment on the common mistakes made in the interpretation of the Grambsch-Therneau test and the steps one should follow to remedy any issues [37].

The Cox model's partial likelihood framework supports the use of the *Wald test*, the *score test* and the *likelihood ratio test*, which are the three standard asymptotic likelihood inference tests. All of them have $\beta = 0$ as the null hypothesis with a chi-squared distribution having number of covariates $k - 1$ degrees of freedom. The test statistics are defined as follows [44]:

$$\begin{aligned} \text{Wald test statistic} &= \hat{\beta}^T I(\hat{\beta}) \hat{\beta} \\ \text{score test statistic} &= U(\beta = 0)^T I(\beta = 0)^{-1} U(\beta = 0) \\ \text{likelihood ratio test statistic} &= 2\left(l(\beta = \hat{\beta}) - l(\beta = 0)\right). \end{aligned}$$

A common need is for the covariate values to be able to change with time. For example, if one wants to assess the effect of changing laboratory measurement values dur-

ing a patient's observation period, the Cox model must be extended in order to allow for time-varying covariates. A typical error in such scenarios is to use a static covariate value even though it was not known at the start of the observation period. A historical example of the error is a study which found that heart transplants lead to significantly longer lives. The problem was a correlation with time, the patients who had received a transplant during the study were by definition the ones who had already lived long enough to receive the transplant [44]. The previous considerations assume that the covariates X_j remain fixed with respect to the time t, but the underlying proportional hazards assumption can incorporate time-varying covariates by generalizing the definition to proportionality at a point in time t [33]:

$$h(t, X_t) = h_0(t) \exp(\beta X_t).$$

The definition of the survival function is done under the assumption that the covariates are constant during a certain interval. For example, illustrating the idea with a single covariate that takes on different values before and after a point in time s [33]:

$$\begin{aligned} X &= X_1 \text{ when } t < s \\ &= X_2 \text{ when } t \geq s. \end{aligned}$$

Now the survival function is [33]:

$$\begin{aligned} S(t, X) &= \exp\left(-\int_0^t h(u)du\right) \\ &= \exp\left(-\int_0^t h_0(u) \exp\left(\beta X_u\right)du\right) \\ &= \exp\left(-\int_0^s h_0(u) \exp\left(\beta X_1\right)du - \int_s^t h_0(u) \exp\left(\beta X_2\right)du\right) \\ &= \exp\left(-\lambda_1 \int_0^s h_0(u)du - \lambda_2 \int_s^t h_0(u)du\right) \\ &= \exp\left(-\lambda_1 \int_0^s h_0(u)du\right) \exp\left(-\lambda_2 \int_s^t h_0(u)du\right) \\ &= S_0(s)^{\lambda_1} \frac{S_0(t)^{\lambda_2}}{S_0(s)^{\lambda_2}}. \end{aligned}$$

The other representations of the survival distribution can be derived similarly [33]. The assumption of proportionality at a point in time brings an additional burden for the data preparation, as the patient data must be split into intervals which end in the point in time where a covariate value changes. The use of time-varying covariates requires no other special consideration from the analysts perspective, as even though the likelihood

equations do not cancel out to equally clean forms, the additional bookkeeping is handled by the analysis routines [44].

The aforementioned significance tests and the Grambsch-Therneau test also are valid for models with time-varying covariates. The interpretation of the results with time-varying covariates is also similar, as while the relative hazard depends on time, the coefficient estimates $\hat{\beta}$ still summarize the relative impact of the different covariate values. One unfortunate downside of the time-varying covariates is that they limit the available methods for adjusting a model which contains covariates that don't fulfill the the proportional hazards assumption [56].

3.4 Data preparation

The main source for the research data is a national Finnish system called Hoitoilmoitusjärjestelmä (*Hilmo*), which is used for information-gathering and reporting by the social and health care services. It comprises of three sub-registries, which are the social service registry (*sosiaali-Hilmo*), the health care registry (*terveys-Hilmo*, *Hilmo*) and the outpatient care registry (*Avohilmo*). The outpatient care registry also includes occupational health care and domiciliary care. Health care registry includes inpatient care, outpatient surgeries and publicly funded outpatient care. Social service registry includes institutional care and housing services. Hilmo is managed by the Finnish Institute for Health and Welfare (*THL*). THL publishes a manual which details Hilmo's data sources and the details of the collected data [30].

After the research plan was approved by the HUS (*Hospital District of Helsinki and Uusimaa*) ethics committee, the research data was requested from the health care registry (*Hilmo*) and the outpatient care registry (*Avohilmo*). The requested data was specified as all of the diagnoses and all of the procedures received by a cohort of patients defined separately for each of the registries. The cohort of the Hilmo data included patients that had ever received any of the procedures listed in Table 3.9 or any of the diagnoses listed in Table 3.10. The cohort of the Avohilmo data included patients that have had any of the diagnoses listed in Table 3.10 and any of the procedures listed in Table 3.11. It should be noted that the Avohilmo procedure table uses the outpatient care SPAT procedure codes. Statistics Finland, the national statistical institution of Finland, was used as a secondary data source to enrich the Hilmo data with the time of death for all the patients of the both cohorts for whom the information was available. The details of the death data registry can be found on the Statistics Finland website [61]. All of the aforementioned data sources only include data up to the year 2018.

The Hilmo data has some major limitations. The biggest one of these is caused by the usage of visits as the base data entity. The data includes the start and the dates of

the treatment provider visitation period, but not the exact dates of the diagnoses and procedures received by the patient during their visit. For example, if a patient stayed at a treatment facility for four days and received a different diagnosis each day, the latter diagnoses can't be dated correctly. One practical negative effect of this issue was that some patients had to be excluded from the cohorts as they had received two or more BPH procedures during the visit which contained their first BPH procedure. As the order of the procedures couldn't be determined, the patients couldn't be appropriately categorized according to their initial BPH procedure. Table 3.12 displays the amount of excluded patients.

If one would follow the definitions presented in the survival analysis section to the letter, Hilmo data would be considered interval censored as the events are only known to have happened inside an interval (during a visit). Unfortunately the methods for the analysis of interval censored data are notably more complex and less developed, so the data was chosen to be considered right censored [42]. This was achieved by dating each of the procedures and diagnoses to the starting date of the visit during which they were received by the patient. While this does introduce some inaccuracies to the data, the negative effect is minimized by the fact that the vast majority of the visits lasted only a day [58]. This is displayed in Figure 3.5, which shows the distribution of the visit durations between 0 and 10 days. See Lindsey et al. for an overview on the methods for dealing with inverval censored data [42].

The data received from Hilmo was also missing the ages of the patients. While a crucial omission otherwise, the situation here was saved by the inclusion of the personal identification codes of the patients in the data. A Finnish personal identification codel (*PIC*) is an official identifier given by the Finnish government, which is partly formed by the date of birth of the person in question. The official format of a PIC is defined in the Finnish legislation. A PIC consists of the person's birth date, an identification number and a checksum symbol used for checking the validity of the PIC. The structure is summarized in table 3.8. Table 3.6 describes how the seventh character is used to depict the century of the birth date [21].

A PIC can also be used to determine the gender of the person, because the identification number formed by the combination of the characters at the indices 8, 9 and 10 is odd for males and even for females. The last character is calculated from the date of birth and the identification number by concatenating them together to form an integer with a length of 9. The result is then divided by 31 and the remainder is used to find the matching the checksum character as shown in table 3.7. If the checksum does not match, then the PIC can be determined to be invalid [21]. The PICs were used to calculate the birth dates and as an additional validation tool. 290 patients were discarded because their PIC failed the validation algorithm.

Table 3.6: Century symbols of Finnish personal identification codes.

Century	Symbol
twenty-first (2000s)	A
twentieth (1900s)	-
nineteenth (1800s)	+

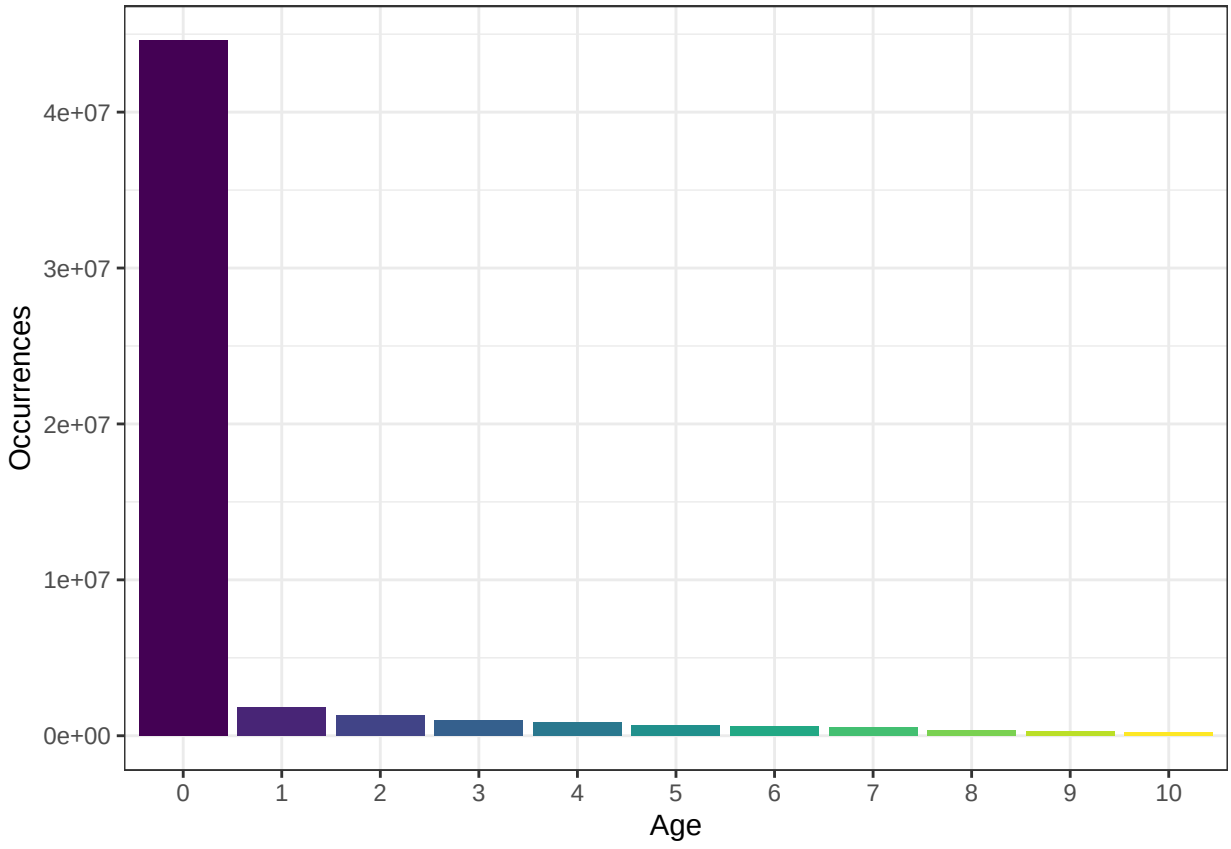

Figure 3.5: The distribution of the visitation period lengths for the first 10 days.

Table 3.7: A conversion table from the remainders to the checksum symbols.

Remainder	Checksum symbol
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	A
11	B
12	C
13	D
14	E
15	F
16	H
17	J
18	K
19	L
20	M
21	N
22	P
23	R
24	S
25	T
26	U
27	V
28	W
29	X
30	Y

Table 3.8: The structure of a Finnish personal identification code.

Position	Type	Value range	Meaning
1	integer	0-3	day of birth
2	integer	0-9	day of birth
3	integer	0-1	month of birth
4	integer	0-9	month of birth
5	integer	0-9	decade of birth
6	integer	0-9	year of birth
7	character	A or - or +	century of birth
8	integer	0-9	identification number
9	integer	0-9	identification number
10	integer	0-9	identification number
11	integer/character	0-9/A-Y	checksum symbol

The details of the data cleaning process are omitted for brevity, but the general workflow was the following:

1. Separate the diagnoses and procedures to distinct data sets
2. Transform the data to a form where a row represents either a single diagnosis or a single procedure
3. Discard patients with invalid PICs

- also discard the cases where the PIC contains an invalid date of birth

4. Discard rows with invalid visitation dates
5. Discard rows with invalid procedure or diagnosis codes (not found in the THL diagnosis/procedure list)

- remove non-alphanumeric characters first
- for diagnoses, match substrings of length 3-5
- for procedures, match the whole string (always length of 5)

6. Discard rows that have a date indicating occurrence after the date of death
7. Discard the rows that are either duplicated diagnosis codes or procedure codes inside a single date
8. Discard patients without any procedures from Table 2.4

The resulting data set constituted the full research data set which contained 54 158 895 diagnoses, 17 965 650 procedures and 649 382 unique patients. Death dates were

Table 3.9: A list of operations used for cohort selection for the Hilmo data set

NCSP range	Name
KE**	prostate operations
KC**	bladder operations
KAS**	kidney transplants
KAE**	kidney stone operations
KAT00	Extracorporeal Shock Wave Lithotripsy (ESWL)
KAT*4	dialysis
TK8**	dialysis
KBE**	ureteral stone operations
KDK**	artificial urinary sphincter operations
KDH**	reconstructive urethral operations
KDV**	urethral operations such as stents
KH1**	imaging of the urinary tract
KW**	operations due to complications of the urinary tract
TKC**	catheterization
TKD00	expansion of the urethra
TKE00	biopsia of the prostate
UKC02	cystocopy
UKD02	urethroscopy
XKC00/XKC03	pressure measurements of the bladder

Table 3.10: A list of diagnoses used as an inclusion criteria for the Hilmo and the Avohilmo cohort.

NCSP code range	Name
N17-N19	chronic kidney diseases
N20-N23	bladder stone diseases
N30	infections and dysfunctions of the bladder
N40-N42	hyperplasia of the prostate and other prostatic diseases
R30-R39	urinary problems

Table 3.11: A list of procedures used as an inclusion criteria for the Avohilmo cohort.

SPAT code	Name
SPAT1167	urinary catheter
SPAT1168	catheter education
SPAT11761	bladder puncture

available for 194 601 patients. See Figure 3.6 for an overview of the incidences of the investigated procedures in the full research data set. A data set for the analysis of the BPH reoperation data set was formed from the full research data set as follows:

1. Discard patients with two or more procedures from Table 2.4 during the visitation containing the initial BPH procedure
2. Discard patients with procedures from Table 2.5 before or during the visitation containing the initial BPH procedure
3. Discard patients with diagnoses from Table 3.13 before or during the visitation containing the initial BPH procedure

- these diagnoses represent a malignancy, i.e. not **benign** prostatic hyperplasia

The resulting data set contained 5 797 416 diagnoses, 2 029 497 procedures and 66 617 unique patients. See Figure 3.7 for an overview of the clinical characteristics of the patients of this cohort at the time of their first BPH procedure. The analysis of the urethral stricture procedures and the stress incontinence procedures required separating a new data set for them. The additional steps were as follows:

1. Discard patients with either the diagnosis for urethral stricture (ICD-10 code N35.9) or a procedure from Table 2.7 before or during the visitation containing the initial BPH procedure
2. Discard patients with either the diagnosis for stress incontinence (ICD-10 code N39.3) before or during the visitation containing the initial BPH procedure

The resulting data set contained 5 168 281 diagnoses, 1 832 774 procedures and 59 599 unique patients. See Figure 3.8 for an overview of the clinical characteristics of the patients of this cohort at the time of their first BPH procedure. The structure of the diagnosis and procedure data was the same for both the reoperation and the US/SI data sets. The diagnosis data included the following fields:

1. Patient identifier
2. Date of diagnosis

Table 3.12: Occurrences of two or more BPH operations during the visitation period that contains the first BPH operation.

Operation 1	Operation 2	Operation 3	count
KED00/KED10	KED00/KED10		1
KED00/KED10	KED22	KED33	2
KED00/KED10	KED22		37
KED00/KED10	KED33		4
KED00/KED10	KED52		1
KED00/KED10	KED62		1
KED00/KED10	KED76		1
KED22	KED22	KED33	1
KED22	KED22		28
KED22	KED33	KED76	18
KED22	KED33		1725
KED22	KED52	KED76	4
KED22	KED52		136
KED22	KED62	KED62	1
KED22	KED62		6
KED22	KED72		5
KED22	KED76		1086
KED33	KED33		8
KED33	KED52		9
KED33	KED62		1
KED33	KED76	KED76	1
KED33	KED76		46
KED52	KED52		14
KED52	KED76		28
KED62	KED76		1
KED72	KED72		2
KED76	KED76		2

Table 3.13: A list of diagnoses used as an exclusionary criteria for both data sets.

code	name
C61	Malignant neoplasm of prostate
C67	Malignant neoplasm of bladder
N31	Neuromuscular dysfunction of bladder, not elsewhere classified

3. Treatment provider visit duration in days
4. ICD-10 code
5. Patient age at the date of diagnosis

The procedure data included the following fields:

1. Patient identifier
2. Date of procedure
3. Treatment provider visit duration in days
4. NCSP code
5. Patient age at the date of procedure

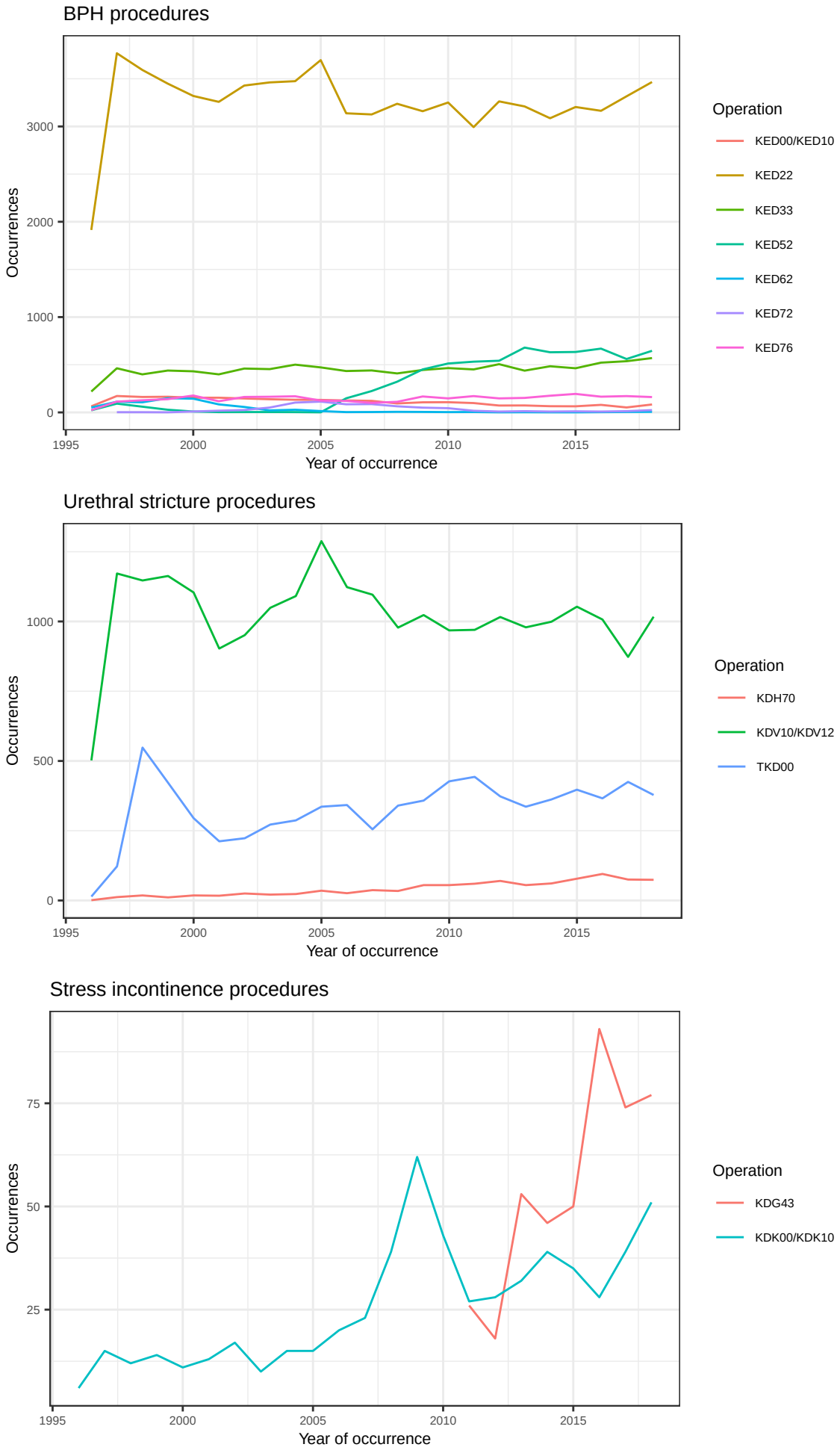

Figure 3.6: A summary of the incidence rates of the investigated procedures.

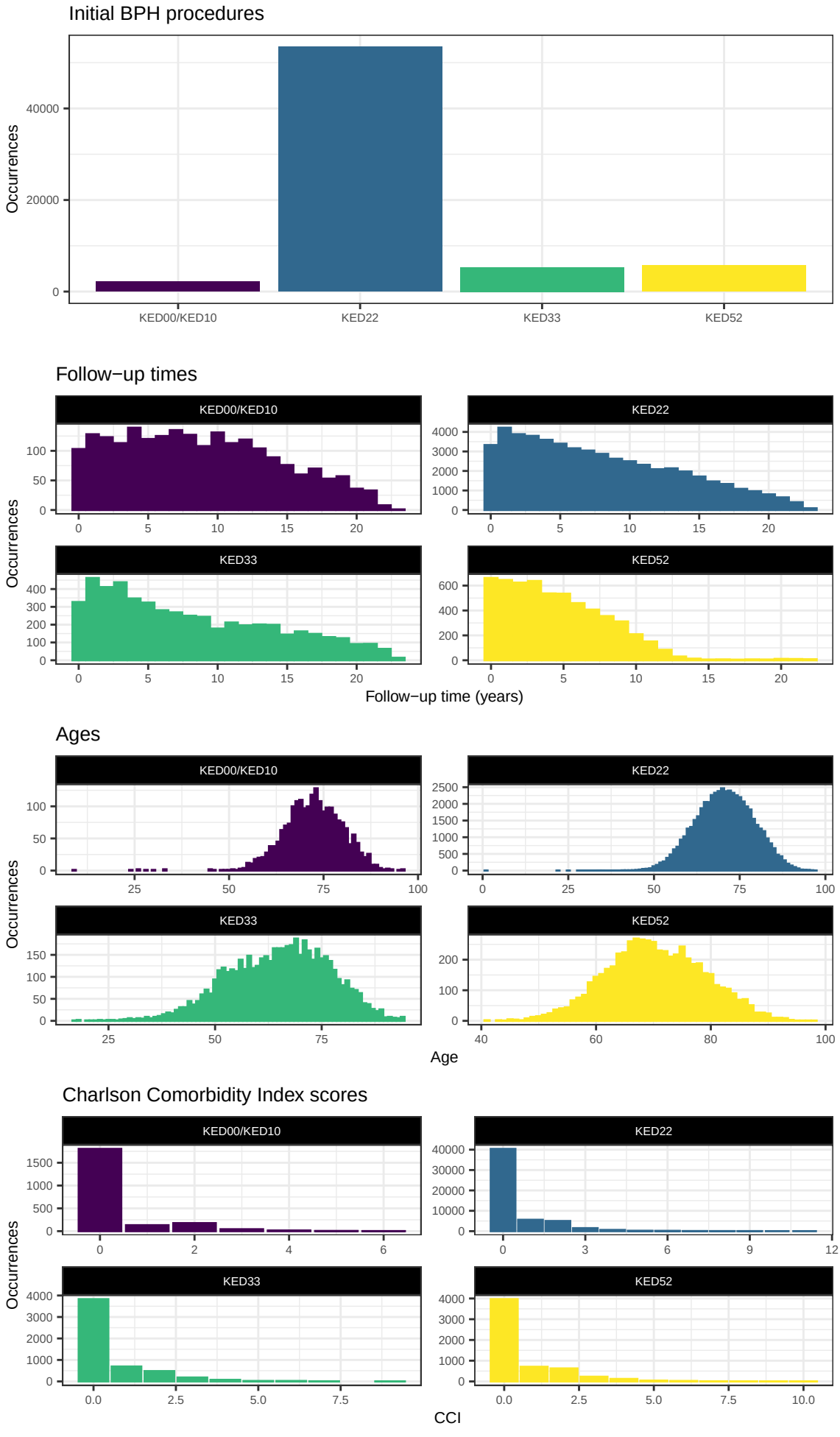

Figure 3.7: A summary of the initial BPH procedures in the reoperation cohort.

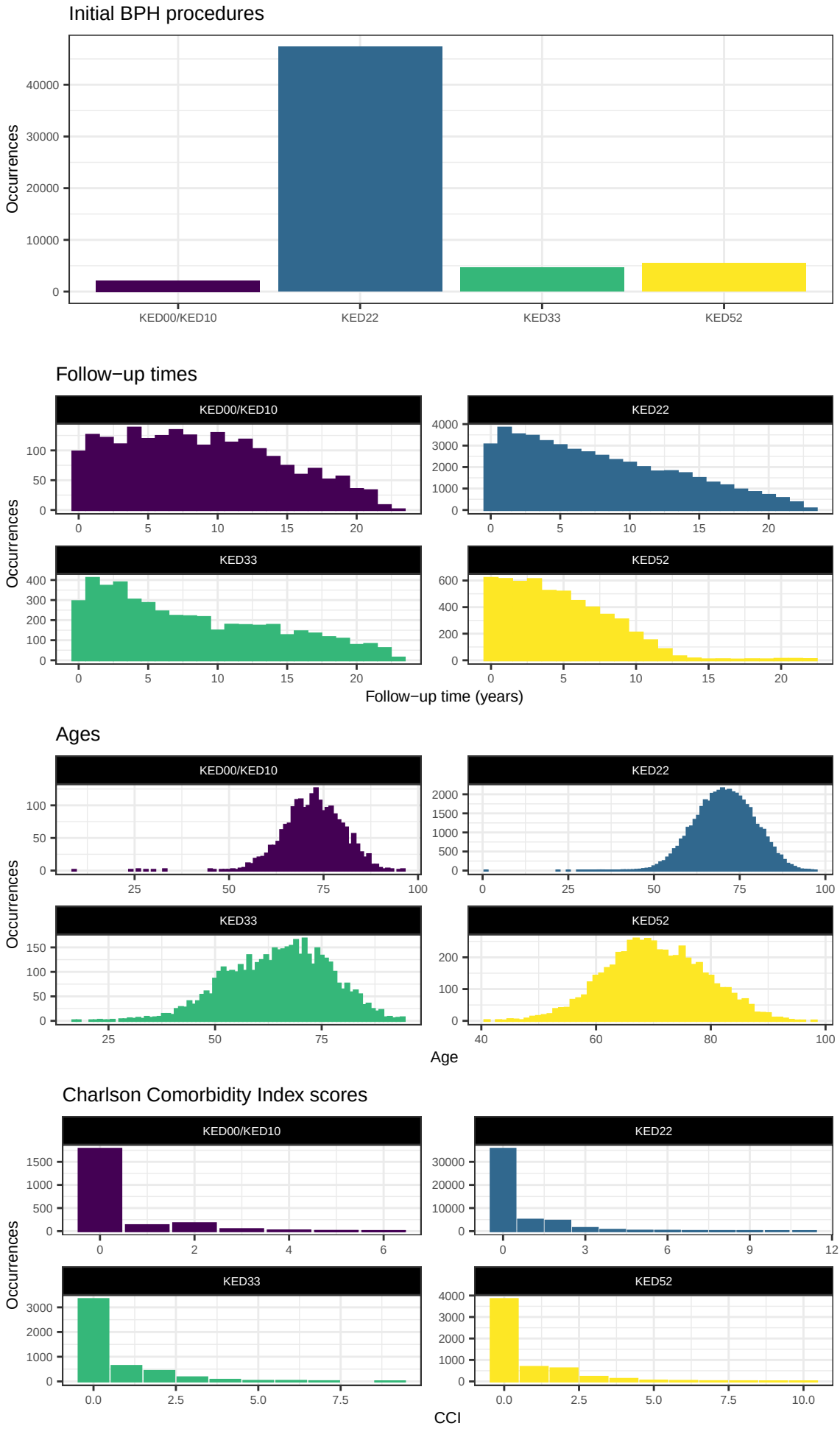

Figure 3.8: A summary of the initial BPH procedures in the US/SI cohort.

4. Results

The aim of this chapter is to detail the performed analyses and to present the associated outcomes. The results are also briefly commented with regards to how they explain the research questions, but a more thorough consideration is left to the collaborating clinicians as inferring the clinical relevance of the results requires extremely significant domain expertise.

The complication risk development after the initial BPH procedure is presented with Kaplan-Meier failure function estimates and Aalen-Johansen cumulative incidence function estimates. Despite the overlap of the methods, the results of both the Kaplan-Meier and the Aalen-Johansen estimators are presented to showcase the effect of the common error of not accounting for the competing risks. Due to the non-exclusionary nature of the investigated endpoints, the Aalen-Johansen estimates are modeled with death as the sole competing risk.

Pairwise log-rank tests were used to determine any statistically significant differences between the Kaplan-Meier survival function estimates. The results of the tests were adjusted with the Benjamini-Hochberg procedure. The differences were quantified with pairwise Cox univariate regression analyses. The results use KED22 as a baseline, which means that the presented hazard ratios are relative risks compared to the complication risks of KED22. The effect of the clinical factors was determined with multivariate Cox regression analyses. The significance of the fitted models was tested with likelihood ratio tests and the significance of the individual covariates was tested with Wald tests. The models and their covariates were also tested with Grambsch-Therneau tests to determine any violations of the proportional hazards assumption. All of the aforementioned significance test results are interpreted with the standard p-value cutoff point of 5%.

As was described in Chapter 3, the different cohort selection criterion necessitated using separate data sets for the analysis of the BPH reoperations and urethral stricture/stress incontinence procedures. All of the analyses were done separately for subcohorts formed by dividing the patients of the whole cohort according to their first BPH procedure (either *KED00/KED10*, *KED22*, *KED33* or *KED52*). The analyses were performed using the R library *survival* by Therneau et al. *survival* is a general-purpose survival analysis library that offers robust implementations for most of the traditional

survival methods. Some notable *survival* features that were not covered in the last chapter include both multi-state and basic frailty models [57].

4.1 Survival distribution analysis

4.1.1 Urethral stricture procedures

Figure 4.1 presents the Kaplan-Meier and the Aalen-Johansen risk estimates for an urethral stricture procedure after each initial BPH procedure. All of the estimates are very low and do not differ much from each other. The Aalen-Johansen estimates are expectedly somewhat lower than the corresponding Kaplan-Meier estimates. Table 4.1 lists the pairwise results of log-rank tests for the Kaplan-Meier estimates adjusted with the Benjamini-Hochberg procedure. Using a significance level of 5%, only the difference between KED22 and KED33 survival functions reaches statistical significance. A reason for this is apparent on Figure 4.1, as the other survival function estimates suffer from extremely wide confidence intervals due to low sample numbers.

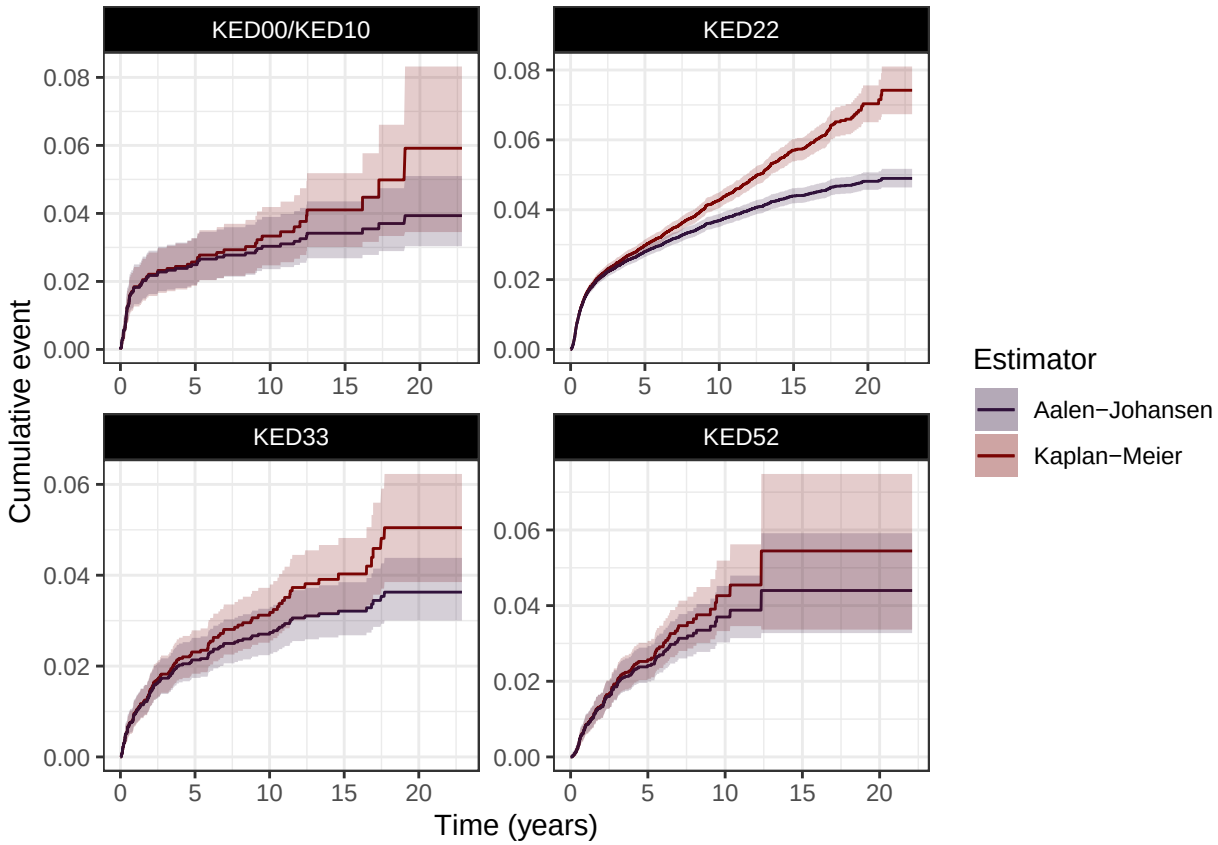

Figure 4.1: Failure function estimates for urethral stricture procedure endpoints.

Table 4.1: A pairwise comparison of log-rank test of the stricture procedures corrected with the BH procedure.

	KED00/KED10	KED22	KED33
KED22	0.238		
KED33	0.585	0.005	
KED52	0.684	0.238	0.308

4.1.2 Stress incontinence procedures

Figure 4.2 presents the Kaplan-Meier and the Aalen-Johansen risk estimates for a stress incontinence procedure after each initial BPH procedure. All of the estimates are extremely low and do not differ much from each other. The Aalen-Johansen estimates are expectedly somewhat lower than the corresponding Kaplan-Meier estimates. Table 4.2 lists the pairwise results of log-rank tests for the Kaplan-Meier estimates adjusted with the Benjamini-Hochberg procedure. Using the significance level of 5%, none of the comparisons reach statistical significance. All of the survival function estimates in Figure 4.2 suffer from low sample numbers and very wide confidence intervals.

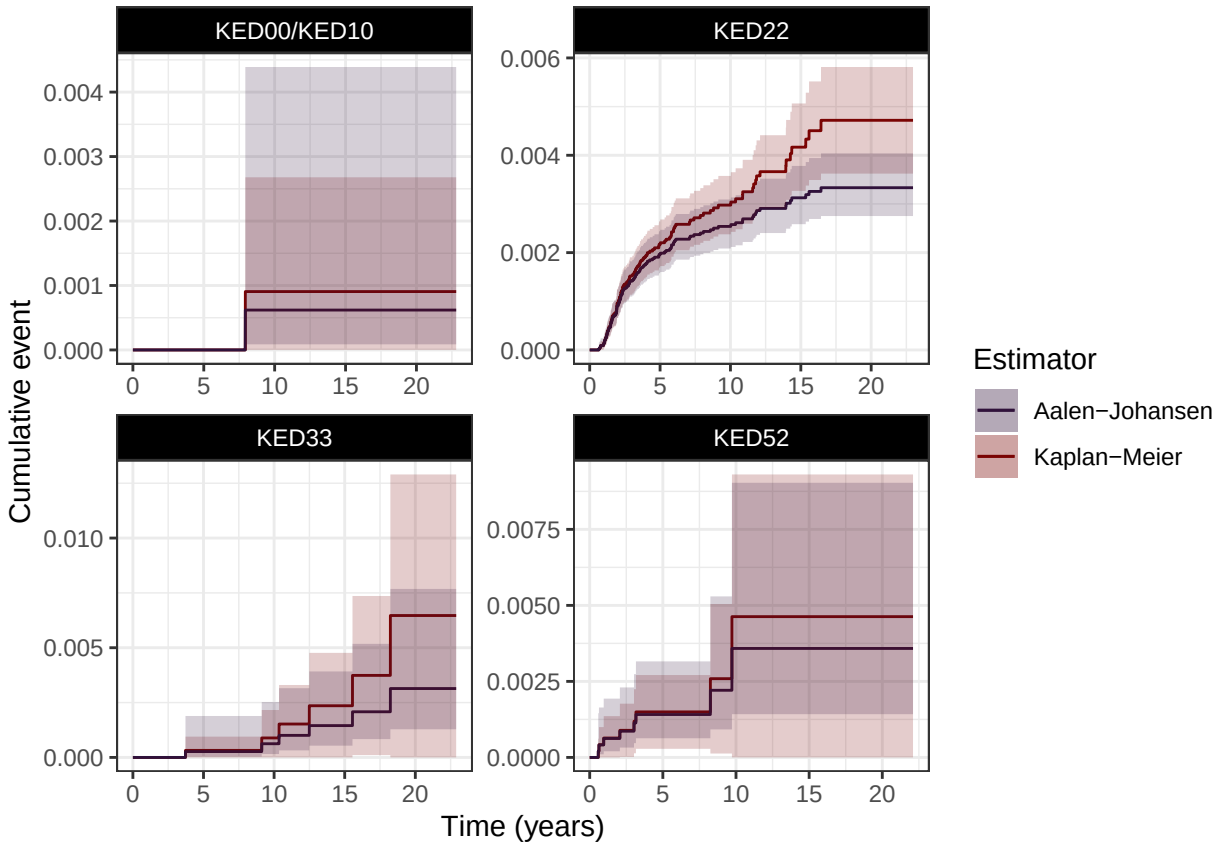

Figure 4.2: Failure function estimates for urinary incontinence procedure endpoints.

Table 4.2: A pairwise comparison of log-rank test of the stress incontinence procedures corrected with the BH procedure.

	KED00/KED10	KED22	KED33
KED22	0.133		
KED33	0.351	0.166	
KED52	0.133	0.649	0.133

4.1.3 BPH reoperations

Figure 4.3 presents the Kaplan-Meier and the Aalen-Johansen risk estimates for a BPH reoperation after each initial BPH procedure. The risk estimates present clear differences between the BPH procedures. The ordering based on the estimated risks can be read from the Figure 4.3. The Aalen-Johansen estimates are expectedly somewhat lower than the corresponding Kaplan-Meier estimates. Table 4.3 lists the pairwise results of log-rank tests for the Kaplan-Meier estimates corrected with the Benjamini-Hochberg procedure. Using the significance level of 5%, only the difference between KED33 and KED52 fails to reach statistical significance.

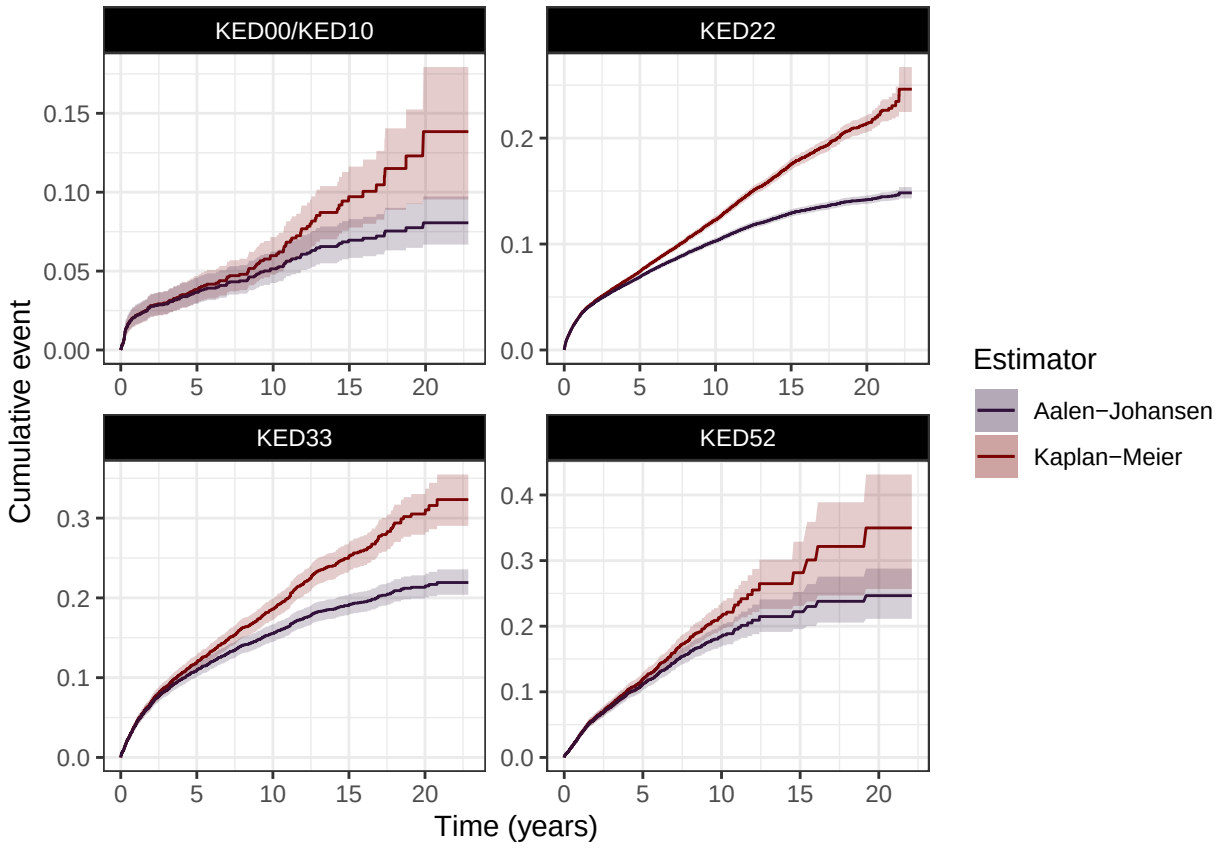

Figure 4.3: Failure function estimates for BPH reoperation endpoints.

Table 4.3: A pairwise comparison of log-rank test of the BPH reoperations corrected with the BH procedure.

	KED00/KED10	KED22	KED33
KED22	0		
KED33	0	0	
KED52	0	0	0.334

4.2 Cox PH regression

Covariates for the Cox regression analyses were chosen by the collaborating urologists according to the ones they were interested in for their clinical practice. Most of the covariates had to be encoded as time-varying covariates due to their nature. Table 4.4 lists the used covariates along with their descriptions. It contains two covariates which have not been discussed before. *clot* refers to the number of blood clot removals from the bladder done within 14 days of the initial BPH procedure. The NCSP operation code that was considered to be a clot removal was *KCV22*. *stone* is a dummy variable representing whether the subject had bladder stones removed during their initial BPH procedure. The NCSP operation codes considered to be bladder stone removals were *KCE00* and *KCE02*. These covariates were only analyzed for the urethral stricture and stress incontinence endpoints.

Table 4.4: A list of the covariates used the in the Cox regression analyses.

Covariate	Meaning	Type
age	Time-varying patient age	continuous
points	Time-varying patient CCI scores	ordered
bph_count	Time-varying number of preceding BPH procedures	ordered
stone	Time-varying number of preceding bladder stone procedures	ordered
clot	Time-varying number of preceding clotting procedures	ordered
N31	Time-varying indicator for a N31 diagnosis	dummy
C61	Time varying indicator for a C61 diagnosis	dummy
C67	Time-varying indicator for a C67 diagnosis	dummy

4.2.1 Urethral stricture procedures

Figure 4.5 summarizes the Cox regression results from each initial BPH procedure to the first urethral stricture procedure. KED00/KED10 is the only model which fails to achieve statistical significance in the global likelihood ratio test, which suggests that the chosen covariates weren't appropriate predictors for the model. This is also visible in the individual Wald tests of the KED00/KED10 model covariates. The Grambsch-Therneau plots for the individual covariates can be found in the Appendix B.1. All of the models pass the global Grambsch-Therneau test. The effect of age and the Charlson Comorbidity Index score seems to be rather small in all cases. This makes sense, as the diagnoses N31, C61 and C67 (colloquially referred to as *cancers* henceforth) display a great risk increase and they are included in the Charlson Comorbidity Index. Bladder stone and blood clot removal procedures generally have too few samples for any reasonable interpretation. Repeated BPH procedures seem to increase risk of an urethral stricture, which is to be expected. Operations affecting that region of the body are known to be a risk factor of urethral strictures [62]. Figure 4.4 presents the univariate Cox regression hazard ratio estimates of the relative urethral stricture risks using KED22 as the baseline.

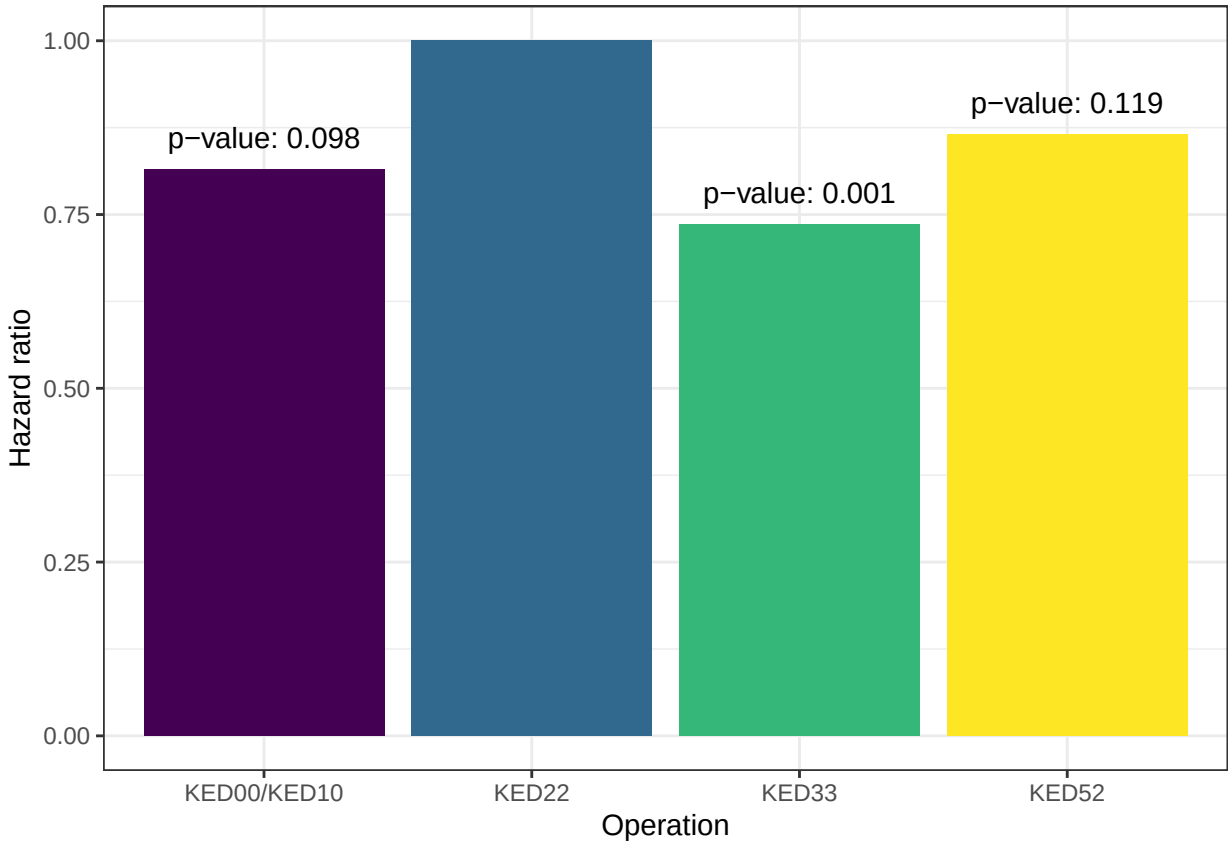

Figure 4.4: A summary of the univariate hazard ratios using KED22 as a baseline.

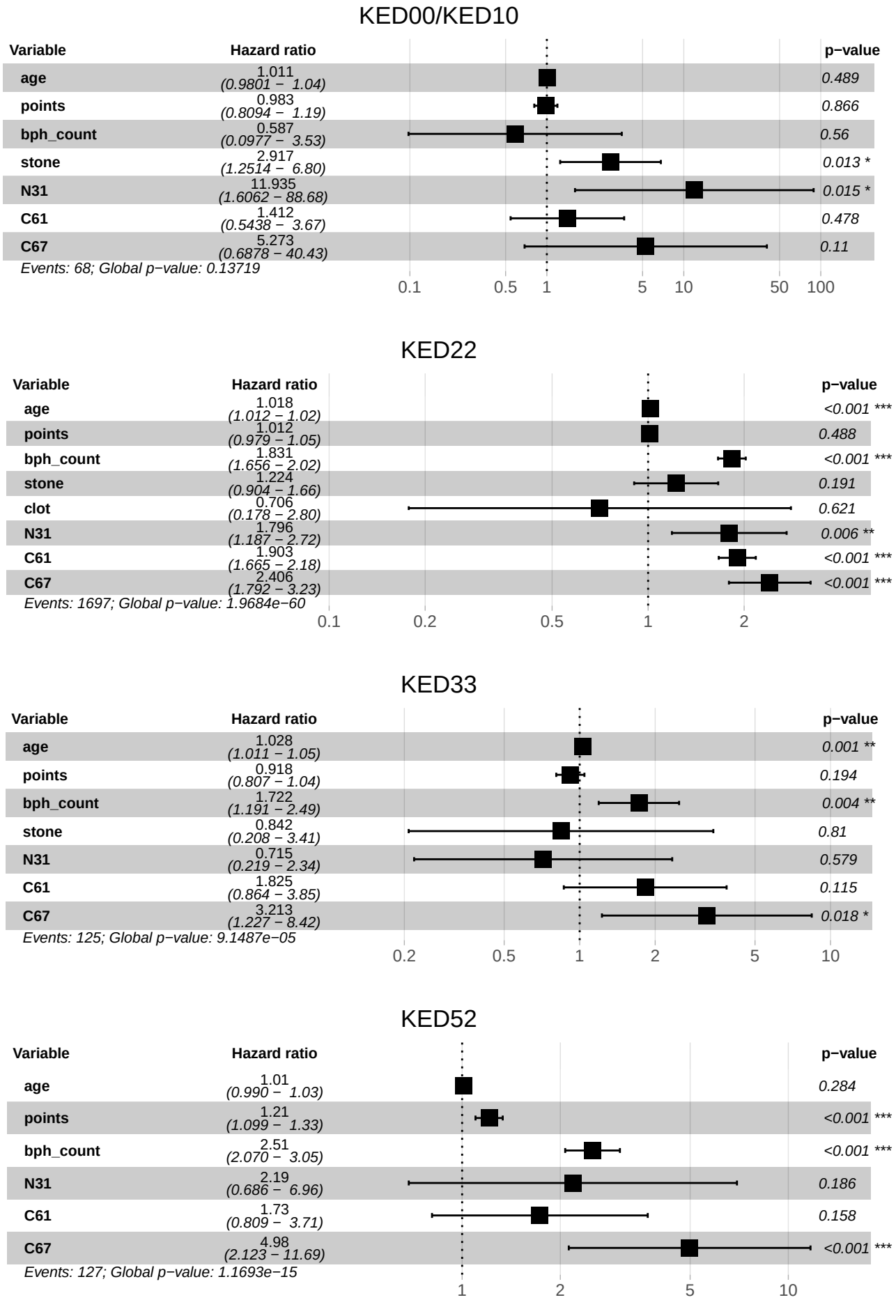

Figure 4.5: A summary of the urethral stricture Cox PH regression results.

4.2.2 Stress incontinence procedures

Figure 4.7 summarizes the Cox regression results from each initial BPH procedure to the first stress incontinence procedure. The KED00/KED10 model could not be estimated due to too few samples. KED22 and KED33 achieved statistical significance in the likelihood ratio test. KED56 model failed the likelihood ratio test, which means that a null model with no covariates fit the data better. This could be due to either the low sample number or inappropriate covariates. The Grambsch-Therneau plots for the individual covariates can be found in the Appendix B.2. All of the models pass the global Grambsch-Therneau test. A clear problem with the stress incontinence results is that there are too few samples for trustworthy inference. Where the bar of significance is passed, the results seem similar as with the urethral strictures. Cancers and repeated BPH procedures increase the risk greatly and the effect of age and the CCI index score is slight. Figure 4.6 presents the univariate Cox regression hazard ratio estimates of the relative stress incontinence risks using KED22 as the baseline.

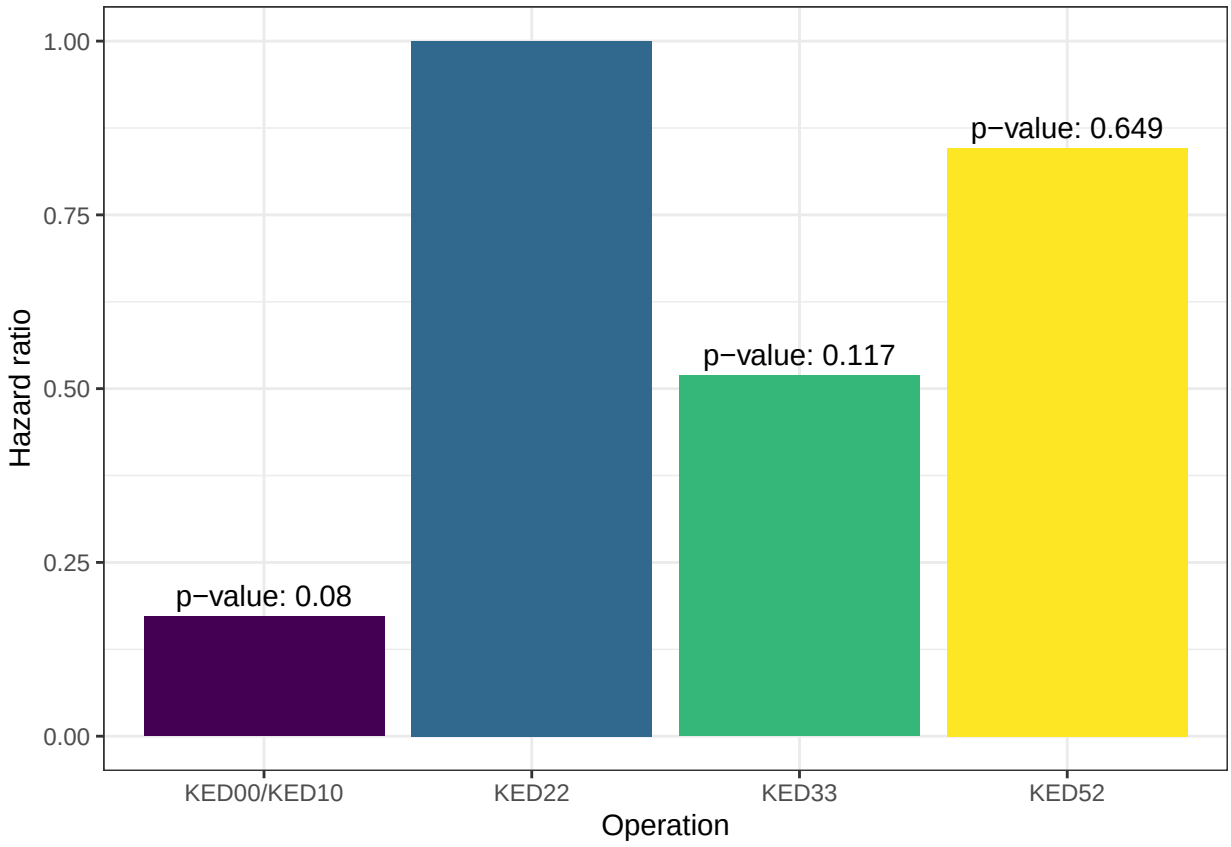

Figure 4.6: A summary of the univariate hazard ratios using KED22 as a baseline.

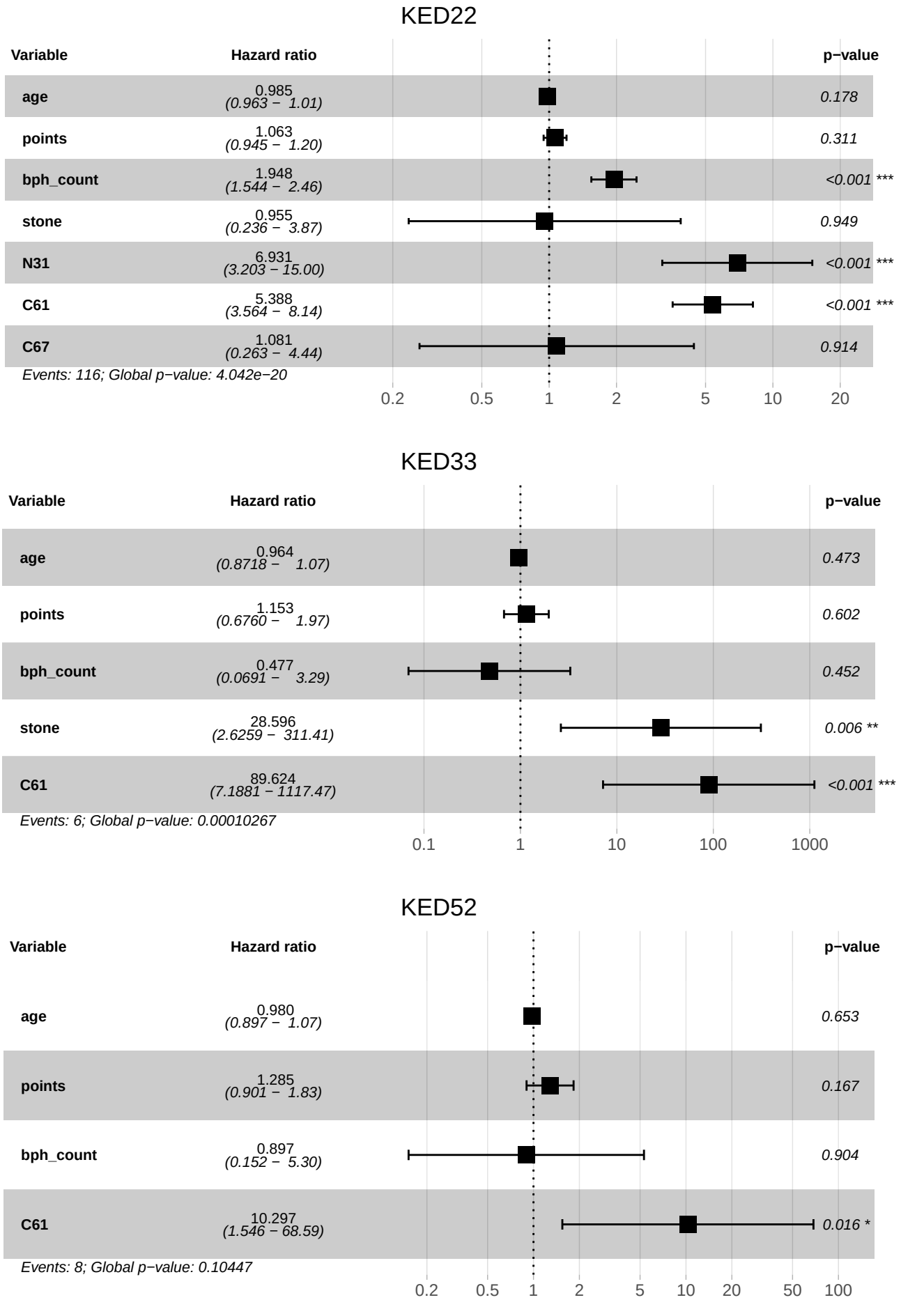

Figure 4.7: A summary of the stress incontinence Cox PH regression results.

4.2.3 BPH reoperations

Figure 4.9 summarizes the Cox regression results from each initial BPH procedure to the first BPH reoperation. The sample sizes in the reoperation models are a lot larger than in both the urethral stricture and urinary incontinence models. All of the models achieve statistical significance in the likelihood ratio test, which suggests that the models with the chosen covariates fit the data better than null models with no covariates. The Grambsch-Therneau plots for the individual covariates can be found in the Appendix B.3. KED22 and KED52 models fail the global Grambsch-Therneau test. Despite of the tests implying non-proportionality, the plots of the scaled residuals remain almost perfectly horizontal and near zero. It is therefore assumed that the test failures are caused by either the high number of observations or extreme outlier values. The regression results are similar to the results of the urethral stricture and urinary incontinence models. The effect of age and the CCI score is slight. Cancers remain the dominant factor, even though their confidence intervals narrow with the increased amount of samples. Figure 4.8 presents the univariate Cox regression hazard ratio estimates of the relative reoperation risks using KED22 as the baseline.

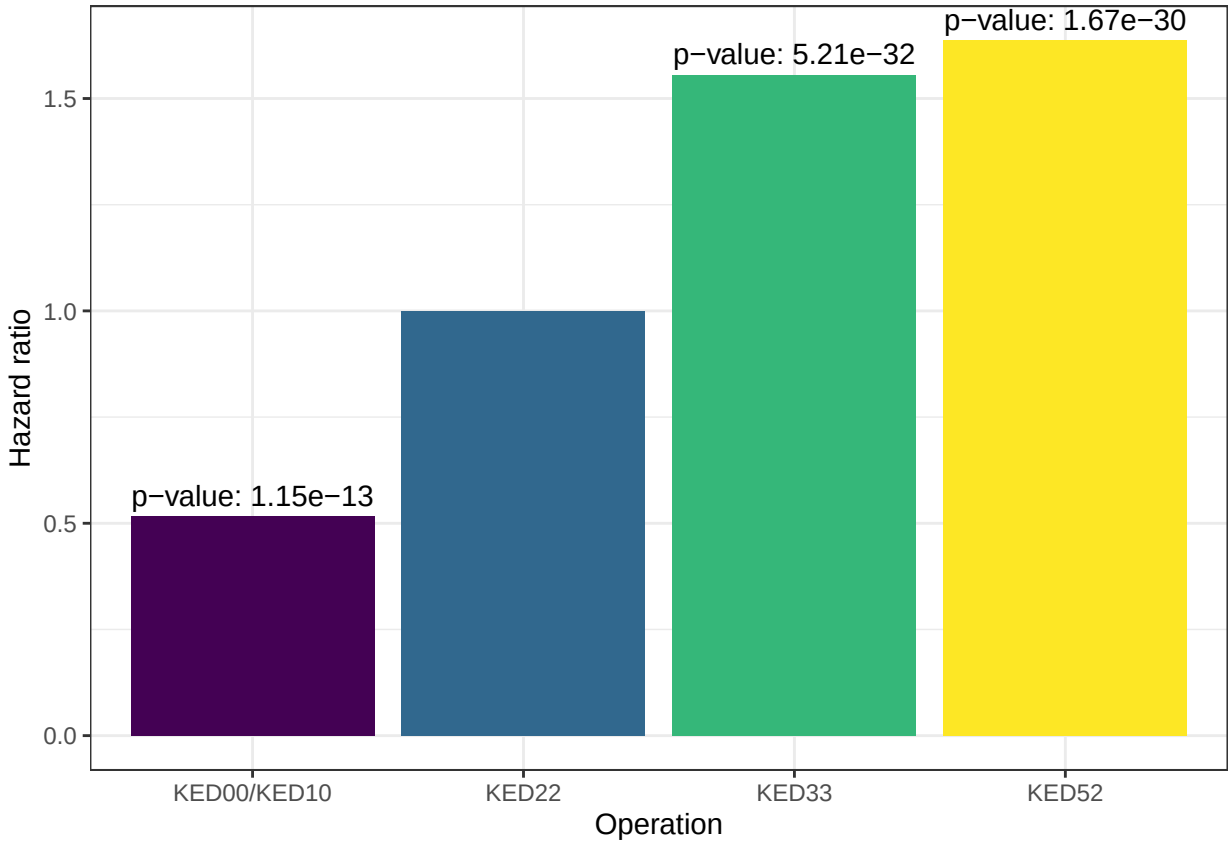

Figure 4.8: A summary of the univariate hazard ratios using KED22 as a baseline.

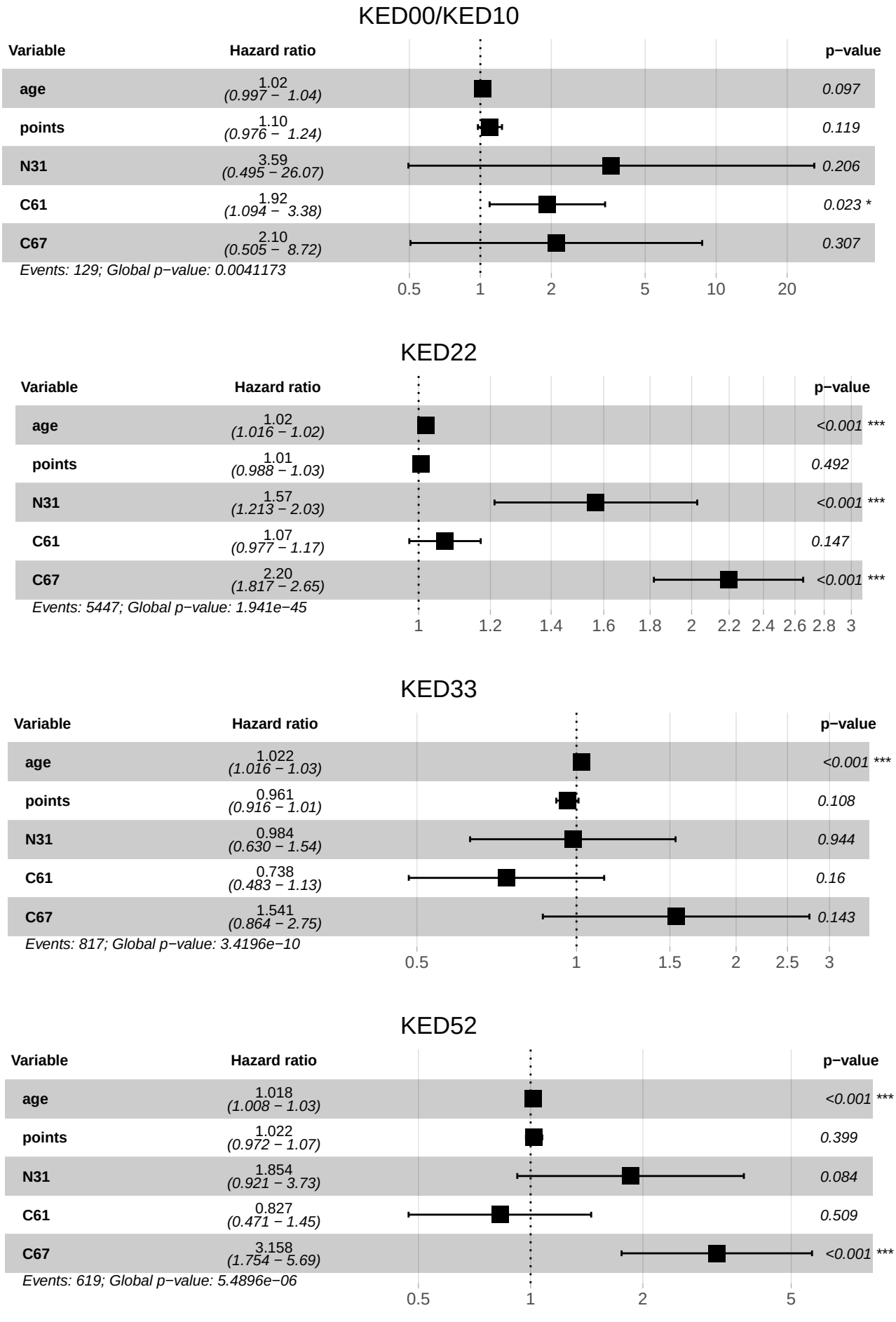

Figure 4.9: A summary of the BPH reoperation Cox PH regression results.

5. Discussion

The main aim of this thesis was to analyse and compare the complication risks of the most common BPH surgical intervention options. While this was achieved to a limited extent, the results of the analyses were limited by the surprisingly low sample numbers. Out of the three analysed complication endpoints (BPH reoperations, stress incontinence procedures and urethral stricture procedures), only the BPH reoperations contained enough samples for relatively trustworthy inference. One possible cause for the apparent lack of data is that only complications which required a new procedure were considered. While unfortunate from the perspective of the analysis results, the seemingly low urethral stricture and stress incontinence risks are of course a positive fact from the patients' point of view.

The survival function estimation provided interesting results for the BPH reoperations. Both the Kaplan-Meier and the Aalen-Johansen estimators showed notable differences in the complication risks for the initial procedure options. However, as the Kaplan-Meier estimator doesn't account for competing risks, the difference between the estimates was striking. The results of the analyses highlight the importance of accounting for any of the possibly existing competing risks. While the choice of the estimator didn't matter for the determination of the initial procedure with the highest complication risk, the absolute differences in the pointwise estimates were very large. One additional benefit of the Aalen-Johansen estimator was that it provided narrower confidence intervals.

The Cox PH regression analysis results were of varying quality. The effect of age tended to be lower than was assumed. As the patients undergoing their initial BPH procedure were already quite old at the time of the procedure, one would have expected to see a bigger risk increase as they aged further. The CCI index scores seemed to have either a low impact or none at all on the complication risks. This is likely due to the fact that most of the diseases of the CCI aren't related to the prostate and the most serious ones of those which were included in the analyses as separate covariates. The cases of blood clots and bladder stones were so infrequent that their effect could only be estimated in a few of the models. The urethral stricture and the stress incontinence analyses suggest that recurring BPH procedures might be a significant risk factor for those complication types.

It is important to keep in mind that the statistical interpretation of the regression

results can be different from the clinical interpretation. The interpretation of hypothesis test results under the selected significance level is that there was sufficient evidence to reject the null hypothesis, which conversely means that the a non-significant result only means that the evidence was not sufficient. A clinician can still find the evidence informative in some respects, be it the practical work at the clinic or as a basis for a new research question. Gelman et al. provide a good discussion about the problems of using arbitrary p-value cutoff points as the significance indicator [26].

Aside from the aforementioned low sample sizes, the results were additionally limited by the quality of the source data. As was discussed in Chapter 3, the Hilmo data only provided an interval (starting and ending dates of a visit at a treatment provider) during which either a diagnosis or a procedure was received by the patient. Even though this wasn't a problem in most of the cases due to the distribution of the visit durations, a sizable amount of diagnoses and procedures had to be arbitrarily dated to the starting date of the visit. In addition to the introduced inaccuracies, the lack of exact dates lead to excluding all of the patients who had received more than one BPH procedures during the visit of their initial BPH procedure, as the first one couldn't be determined in those cases.

Because the Cox PH model covariates were defined by the collaborating clinicians and their clinical interests without considering the the statistical validity, the models generally suffer from overly many covariates for the amount of samples available. Many of the covariates were also defined to be time-varying, which can be problematic. While the theoretical aspects of the time-varying covariates are sound and the software libraries do support their use, time-varying covariates still complicate the interpretation of the results. Notably the fact that the temporal effect is summarized by a single number can lead to inaccurate estimates if the covariate values tend to have great shifts. Time-varying covariates also limit the available means of model adjustment for fixing any violations of the PH assumption. The need for adjustments was avoided in the analyses of the thesis as the Grambsch-Therneau diagnostics didn't present any severe violations.

The Charlson Comorbidity Score algorithm lent itself well to a Spark implementation. The previous work and the associated algorithms have focused on computing CCI scores at a certain point in time. The presented Spark SQL algorithm implementation goes further by building a total timeline of the CCI scores while still remaining simple to understand and performant. The algorithm is also written in standard SQL, which should enable it to also be used in other SQL engines than just Spark SQL. One important downside of the algorithm is that it's currently limited to only the computation the CCI scores, albeit still giving a choice to use either the original Charlson or the newer Quan scoring system.

The final contribution of the thesis was the discussion of some practical issues that

can be found in survival analysis studies performed by clinical practitioners. Survival analysis is a complex topic based on advanced mathematical theory, which can make the necessary concepts difficult to grasp. However, modern software libraries make the practical analysis straightforward as long as the interpretation of the results is done correctly. A related goal of the thesis was to showcase some alternative methods for clinicians as the material meant for them is often locked into a select few. For example, the use of the Fine-Gray estimator is common in medical literature for the analysis of competitive risks scenarios. The Aalen-Johansen estimator used in the analyses of this thesis possesses better statistical properties with less constraints on data but is not as commonly taught in applied courses. Therneau et al. give convincing arguments on the superiority of the Aalen-Johansen estimator [59]. Austin et al. go further in explaining the problematic aspects of the Fine-Gray estimator [8].

The results presented here generally agree with the preceding works. Stoddard et al. found that TURP (*KED22*) has a lower risk of reoperations but a higher risk of urethral strictures than procedures based on a laser (*KED52*) [53]. Gilfrich et al. found that procedures based on a laser (*KED52*) have a similar reoperation risk with TURP (*KED22*) and that OSP (*KED00/KED10*) has a lower reoperation risk than TURP (*KED22*) [27]. The only findings differing in the presented results are the stricture rates of TURP and laser procedures as their difference did not reach statistical significance. It should be noted that none of the aforementioned papers had an exactly equal methodology, which is likely the cause of the discrepancy. The preceding studies also used notably different study setting definitions with a narrower scope and shorter observation periods compared to the analysis of this thesis.

The BPH complications remain a good candidate for future studies. The research data used here was limited in the extent of the available clinical covariates. A simple extension of the analyses presented here would be to include additional data sources with additional samples and covariates to investigate their effect. Another alternative source of future work is the choice of the used regression model. The Cox proportional hazards model was chosen to be used here mainly due to its prevalence in the medical literature. As the results are meant to be published, the use of the standard analysis methods of the field was important. Additionally, the Cox PH model remains a good choice even among more modern alternatives. However, the alternative models can be useful e.g. in scenarios where the data does not fulfill the assumptions of the Cox model and none of the available remedies work. Lee et al. provide a good overview of the alternatives, which include machine learning approaches [40].

6. Conclusions

Benign prostatic hyperplasia is one of the most common male diseases, which makes an optimal treatment choice for it important for a large number of patients. Surgical procedures are the most effective treatment options for BPH, but the choice between them depends on the individual patient profile and the complications associated with the particular procedure. Complications present a problem for the well-being of the patients which makes quantifying the complication risks important for the treating clinicians deciding on the treatment plans. Recurring BPH, urethral strictures and stress incontinence were the complications chosen to be investigated and only the cases where the complication required another procedure were included in the analysis.

The Charlson Comorbidity Index score is a widely used measure of the disease load of a patient at a particular point in time. It has proven to be an useful predictor of a multitude of negative event types. While the intuition of the score calculation procedure is easy to grasp, it presents some challenges for an algorithmic implementation. The novel Spark algorithm described in this thesis is easy to understand, has good performance characteristics and should scale well to any data sizes within reason. The CCI score was also one of the patient profile attributes used in the regression analyses.

The results presented in this thesis showed distinct differences in the reoperation risks between the choices for the initial BPH procedure. Urethral stricture and stress incontinence complication risks did not show statistically significant differences, which is attributed to the low sample numbers in both cases. Cox regression analysis results followed a similar trend in almost all of the cases. Age and the disease load as measured with the Charlson Comorbidity Index scores were associated with at most slight risk increases. The various cancers presented both the greatest risk increase and the greatest variability between the initial BPH procedures. An accurate comparison between the initial BPH procedures is challenging because the urethral stricture and stress incontinence procedures had low sample numbers resulting in extremely wide confidence intervals for the parameter estimates. One important takeaway from the Cox analyses is that the individual cancers, which are also included in the CCI score, seem to predict the complications better than the total CCI score. A higher number of preceding BPH procedures seemed to increase the risk of urethral strictures and stress incontinence, which was expected.

References

[1] 2010. *NOMESCO classification of surgical procedures* (Version 1.15 ed.). Nordic Medico-Statistical Committee, Copenhagen.

[2] 2016. *How does the prostate work?* Institute for Quality and Efficiency in Health Care (IQWiG). Retrieved July 11, 2022 from https://www.ncbi.nlm.nih.gov/books/NBK279291/

[3] FINLEX ® - Säädökset alkuperäisinä: Laki sosiaali- ja terveystietojen toissijaisesta… 552/2019. Retrieved December 9, 2022 from https://www.finlex.fi/fi/laki/alkup/2019/20190552

[4] Hints - Spark 3.3.1 Documentation. Retrieved December 8, 2022 from https://spark.apache.org/docs/latest/sql-ref-syntax-qry-select-hints.html

[5] Spark SQL and DataFrames - Spark 3.3.0 Documentation. Retrieved August 4, 2022 from https://spark.apache.org/docs/latest/sql-programming-guide.html

[6] Bashar M. Abdeen and Almostafa M. Badreldin. 2022. Urethral Strictures. In *StatPearls*. StatPearls Publishing, Treasure Island (FL). Retrieved July 11, 2022 from http://www.ncbi.nlm.nih.gov/books/NBK564297/

[7] Alan Agresti. 2013. *Categorical data analysis* (3rd ed ed.). Wiley, Hoboken, NJ.

[8] Peter C. Austin, Ewout W. Steyerberg, and Hein Putter. 2021. Fine-Gray subdistribution hazard models to simultaneously estimate the absolute risk of different event types: Cumulative total failure probability may exceed 1. *Statistics in Medicine* 40, 19 (2021), 4200–4212. DOI:https://doi.org/10.1002/sim.9023

[9] Douglas Bates and Dirk Eddelbuettel. 2013. Fast and Elegant Numerical Linear Algebra Using the RcppEigen Package. *Journal of Statistical Software* 52, (February 2013), 1–24. DOI:https://doi.org/10.18637/jss.v052.i05

[10] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (1995), 289–300. DOI:https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[11] Ørnulf Borgan. 1997. *Three Contributions to the Encyclopedia of Biostatistics: The Nelson-Aalen, Kaplan-Meier, and Aaalen-Johansen Estimators*. Department of Mathematics, University of Oslo.

[12] Pierre Brémaud. 2020. *Markov Chains: Gibbs Fields, Monte Carlo Simulation and Queues*. Springer International Publishing, Cham. DOI:https://doi.org/10.1007/978-3-030-45982-6

[13] Liberato Camilleri. 2019. History of survival analysis. Retrieved July 12, 2022 from https://www.um.edu.mt/library/oar/handle/123456789/55748

[14] Mary E. Charlson, Danilo Carrozzino, Jenny Guidi, and Chiara Patierno. 2022. Charlson Comorbidity Index: A Critical Review of Clinimetric Properties. *Psychother Psychosom* 91, 1 (2022), 8–35. DOI:https://doi.org/10.1159/000521288

[15] Mary E. Charlson, Peter Pompei, Kathy L. Ales, and C.Ronald MacKenzie. 1987. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases* 40, 5 (January 1987), 373–383. DOI:https://doi.org/10.1016/0021-9681(87)90171-8

[16] Mary Charlson, Ted P. Szatrowski, Janey Peterson, and Jeffrey Gold. 1994. Validation of a combined comorbidity index. *Journal of Clinical Epidemiology* 47, 11 (November 1994), 1245–1251. DOI:https://doi.org/10.1016/0895-4356(94)90129-5

[17] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman. 2003. Survival Analysis Part I: Basic concepts and first analyses. *Br J Cancer* 89, 2, 2 (July 2003), 232–238. DOI:https://doi.org/10.1038/sj.bjc.6601118

[18] Anna Cooper and Angela Hassiotis. 2009. Appendix A: Critique of the ICD 10 and DSM IV based Classification of Mental Disorders in Intellectual Disability. In *Intellectual Disability Psychiatry*, Angela Hassiotis, Diana Andrea Barron and Ian Hall (eds.). John Wiley & Sons, Ltd, Chichester, UK, 253–256. DOI:https://doi.org/10.1002/9780470682968.app1

[19] D. R. Cox. 1972. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 34, 2 (1972), 187–220.

[20] Giulio Del Popolo, Donatella Pistolesi, and Vincenzo Li Marzi (Eds.). 2015. *Male Stress Urinary Incontinence*. Springer International Publishing, Cham. DOI:https://doi.org/10.1007/978-3-319-19252-9

[21] Edita Publishing Oy. *Valtioneuvoston asetus väestötietojärjestelmästä*. Oikeusministeriö, Edita Publishing Oy. Retrieved July 11, 2022 from https://www.finlex.fi/fi/laki/ajantasa/2010/20100128

[22] Jessie K. Edwards, Laura L. Hester, Mugdha Gokhale, and Catherine R. Lesko. 2016. Methodologic Issues When Estimating Risks in Pharmacoepidemiology. *Curr Epidemiol Rep* 3, 4 (December 2016), 285–296. DOI:https://doi.org/10.1007/s40471-016-0089-1

[23] Anne Elixhauser, Claudia Steiner, D. Robert Harris, and Rosanna M. Coffey. 1998. Comorbidity Measures for Use with Administrative Data: *Medical Care* 36, 1 (January 1998), 8–27. DOI:https://doi.org/10.1097/00005650-199801000-00004

[24] Jason P. Fine and Robert J. Gray. 1999. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association* 94, 446 (June 1999), 496–509. DOI:https://doi.org/10.1080/01621459.1999.10474144

[25] Alessandro Gasparini. 2018. Comorbidity: An R package for computing comorbidity scores. *JOSS* 3, 23 (March 2018), 648. DOI:https://doi.org/10.21105/joss.00648

[26] Andrew Gelman and Hal Stern. 2006. The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. *The American Statistician* 60, 4 (November 2006), 328–331. DOI:https://doi.org/10.1198/000313006X152649

[27] Christian Gilfrich, Matthias May, Claus Fahlenbrach, Christian Günster, Elke Jeschke, Gralf Popken, Jens-Uwe Stolzenburg, Lothar Weissbach, Christoph von Zastrow, and Hanna Leicht. 2021. Surgical Reintervention Rates after Invasive Treatment for Lower Urinary Tract Symptoms due to Benign Prostatic Syndrome: A Comparative Study of More than 43,000 Patients with Long-Term Followup. *Journal of Urology* 205, 3 (March 2021), 855–863. DOI:https://doi.org/10.1097/JU.0000000000001463

[28] Patricia M. Grambsch and Terry M. Therneau. 1994. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 81, 3 (September 1994), 515–526. DOI:https://doi.org/10.1093/biomet/81.3.515

[29] Runxin Guo, Yi Zhao, Quan Zou, Xiaodong Fang, and Shaoliang Peng. 2018. Bioinformatics applications on Apache Spark. *GigaScience* (August 2018). DOI:https://doi.org/10.1093/gigascience/giy098

[30] Pirjo Häkkinen, Kaisa Mölläri, Sanna-Mari Saukkonen, and Riikka Väyrynen. Hilmo Sosiaali- ja terveydenhuollon hoitoilmoitus 2020. 176. Retrieved July 11, 2022 from http://urn.fi/URN:ISBN:978-952-343-346-5

[31] Frank E. Harrell. 2015. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer International Publishing, Cham. DOI:https://doi.org/10.1007/978-3-319-19425-7

[32] S. Hunskaar, E. P. Arnold, K. Burgio, A. C. Diokno, A. R. Herzog, and V. T. Mallett. 2000. Epidemiology and natural history of urinary incontinence. *Int Urogynecol J Pelvic Floor Dysfunct* 11, 5 (2000), 301–319. DOI:https://doi.org/10.1007/s001920070021

[33] Stephen P Jenkins. 2005. Survival analysis. Retrieved July 11, 2022 from https://www.iser.essex.ac.uk/files/teaching/stephenj/ec968/pdfs/ec968lnotesv6.pdf

[34] Ren Johansen. 2022. An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations. *Scand J Statist* (2022), 11.

[35] E. L. Kaplan and Paul Meier. 1958. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* 53, 282 (June 1958), 457–481. DOI:https://doi.org/10.1080/01621459.1958.10501452

[36] Christiana Kartsonaki. 2016. Survival analysis. *Diagnostic Histopathology* 22, 7 (July 2016), 263–270. DOI:https://doi.org/10.1016/j.mpdhp.2016.06.005

[37] Luke Keele. 2010. Proportionally Difficult: Testing for Nonproportional Hazards in Cox Models. *Polit. anal.* 18, 2 (2010), 189–205. DOI:https://doi.org/10.1093/pan/mpp044

[38] Brad Kelechava. 2018. The SQL Standard - ISO/IEC 9075:2016 (ANSI X3.135). Retrieved July 11, 2022 from https://blog.ansi.org/2018/10/sql-standard-iso-iec-9075-2016-ansi-x3-135/

[39] Jorma Komulainen. 2012. *Suomalainen tautien kirjaamisen ohjekirja*. THL. Retrieved July 11, 2022 from http://urn.fi/URN:ISBN:978-952-245-511-6

[40] Seungyeoun Lee and Heeju Lim. 2019. Review of statistical methods for survival analysis using genomic data. *Genomics Inform* 17, 4 (December 2019), e41. DOI:https://doi.org/10.5808/GI.2019.17.4.e41

[41] Jari Lehtonen, Jukka Lehtovirta, and Päivi Mäkelä-Bengs. 2013. *THL-Toimenpideluokitus*. THL. Retrieved from http://urn.fi/URN:ISBN:978-952-245-858-2

[42] Jane C. Lindsey and Louise M. Ryan. 1998. Methods for interval-censored data. *Statist. Med.* 17, 2 (January 1998), 219–238. DOI:https://doi.org/10.1002/(SICI)1097-0258(19980130)17:2<219::AID-SIM735>3.0.CO;2-O

[43] N. Mantel. 1966. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep* 50, 3 (March 1966), 163–170.

[44] Dirk F. Moore. 2016. *Applied Survival Analysis Using R*. Springer International Publishing, Cham. DOI:https://doi.org/10.1007/978-3-319-31245-3

[45] Bhramar Mukherjee, Huang-Tz Ou, Fei Wang, and Steven R. Erickson. 2011. A new comorbidity index: The health-related quality of life comorbidity index. *Journal of Clinical Epidemiology* 64, 3 (March 2011), 309–319. DOI:https://doi.org/10.1016/j.jclinepi.2010.01.025

[46] Anthony R. Mundy and Daniela E. Andrich. 2011. Urethral strictures: URETHRAL STRICTURES. *BJU International* 107, 1 (January 2011), 6–26. DOI:https://doi.org/10.1111/j.1464-410X.2010.09800.x

[47] World Health Organization (Ed.). 1992. *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines*. World Health Organization, Geneva.

[48] Huang-Tz Ou, Bhramar Mukherjee, Steven R. Erickson, John D. Piette, Richard P. Bagozzi, and Rajesh Balkrishnan. 2012. Comparative Performance of Comorbidity Indices in Predicting Health Care-Related Behaviors and Outcomes among Medicaid Enrollees with Type 2 Diabetes. *Population Health Management* 15, 4 (August 2012), 220–229. DOI:https://doi.org/10.1089/pop.2011.0037

[49] H. Quan, B. Li, C. M. Couris, K. Fushimi, P. Graham, P. Hider, J.-M. Januel, and V. Sundararajan. 2011. Updating and Validating the Charlson Comorbidity Index and Score for Risk Adjustment in Hospital Discharge Abstracts Using Data From 6 Countries. *American Journal of Epidemiology* 173, 6 (March 2011), 676–682. DOI:https://doi.org/10.1093/aje/kwq433

[50] Hude Quan, Vijaya Sundararajan, Patricia Halfon, Andrew Fong, Bernard Burnand, Jean-Christophe Luthi, L Duncan Saunders, Cynthia A. Beck, Thomas E. Feasby, and William A. Ghali. 2005. Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data: *Medical Care* 43, 11 (November 2005), 1130–1139. DOI:https://doi.org/10.1097/01.mlr.0000182534.19832.83

[51] Claus G Roehrborn. 2005. Benign Prostatic Hyperplasia: An Overview. *Rev Urol* 7, (2005), S3–S14. Retrieved July 11, 2022 from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1477638/

[52] Kenneth J. Rothman. 1990. No Adjustments Are Needed for Multiple Comparisons: *Epidemiology* 1, 1 (January 1990), 43–46. DOI:https://doi.org/10.1097/00001648-199001000-00010

[53] Michelina D. Stoddard, Xinyan Zheng, Jialin Mao, Alexis Te, Art Sedrakyan, and Bilal Chughtai. 2021. Safety and Efficacy of Outpatient Surgical Procedures for the Treatment of Benign Prostatic Enlargement in New York State and California (2005-2016). *Journal of Urology* 205, 3 (March 2021), 848–854. DOI:https://doi.org/10.1097/JU.0000000000001401

[54] Vijaya Sundararajan, Hude Quan, Patricia Halfon, Kiyohide Fushimi, Jean-Christophe Luthi, Bernard Burnand, and William A. Ghali. 2007. Cross-National Comparative Performance of Three Versions of the ICD-10 Charlson Index. *Medical Care* 45, 12 (December 2007), 1210–1215. DOI:https://doi.org/10.1097/MLR.0b013e3181484347

[55] Suomalaisen Lääkäriseuran Duodecimin ja Suomen Urologiyhdistyksen asettama työryhmä. 2020. Eturauhasen hyvänlaatuinen liikakasvu. Retrieved July 11, 2022 from https://www.kaypahoito.fi/hoi11010

[56] Terry M. Therneau and Patricia M. Grambsch. 2000. *Modeling Survival Data: Extending the Cox Model.* Springer New York, New York, NY. DOI:https://doi.org/10.1007/978-1-4757-3294-8

[57] Terry M. Therneau, Thomas Lumley (original S.->R port and R. maintainer until 2009), Atkinson Elizabeth, and Crowson Cynthia. 2022. *Survival: Survival Analysis.* Retrieved August 3, 2022 from https://CRAN.R-project.org/package=survival

[58] Terry Therneau. 2022. A package for survival analysis in R. (August 2022). Retrieved from https://cran.r-project.org/web/packages/survival/survival.pdf

[59] Terry Therneau, Cynthia Crowson, and Elizabeth Atkinson. Multi-state models and competing risks. Retrieved July 11, 2022 from https://cran.r-project.org/web/packages/survival/vignettes/compete.pdf

[60] THL. 2011. *Tautiluokitus ICD-10.* THL. Retrieved July 11, 2022 from http://urn.fi/URN:NBN:fi-fe201205085423

[61] Tilastokeskus. Archive of death certificates. Retrieved December 5, 2022 from https://www.stat.fi/tup/kuolintodistusarkisto/index_en.html

[62] Stefan Tritschler, Alexander Roosen, Claudius Füllhase, Christian G Stief, and Herbert Rübben. 2013. Urethral Stricture: Etiology, Investigation and Treatments. *Dtsch Arztebl Int* 110, 13 (March 2013), 220–226. DOI:https://doi.org/10.3238/arztebl.2013.0220

[63] Aykut Uslu and Jürgen Stausberg. 2021. Value of the Electronic Medical Record for Hospital Care: Update From the Literature. *Journal of Medical Internet Research* 23, 12 (December 2021), e26323. DOI:https://doi.org/10.2196/26323

[64] Michael Von Korff, Edward H. Wagner, and Kathleen Saunders. 1992. A chronic disease score from automated pharmacy data. *Journal of Clinical Epidemiology* 45, 2 (February 1992), 197–203. DOI:https://doi.org/10.1016/0895-4356(92)90016-G

[65] Samuel J. Wang, Blackford Middleton, Lisa A. Prosser, Christiana G. Bardon, Cynthia D. Spurr, Patricia J. Carchidi, Anne F. Kittler, Robert C. Goldszer, David G. Fairchild, Andrew J. Sussman, Gilad J. Kuperman, and David W. Bates. 2003. A cost-benefit analysis of electronic medical records in primary care. *The American Journal of Medicine* 114, 5 (April 2003), 397–403. DOI:https://doi.org/10.1016/S0002-9343(03)00057-3

[66] Jack O Wasey, Steven M Frank, and Mohamed A Rehman. Icd: Efficient Computation of Comorbidities from ICD Codes Using Sparse Matrix Multiplication in R. *Journal of Statistical Software*, 25.

[67] Wumingbai. 2021. Human male reproductive system. Retrieved July 11, 2022 from https://upload.wikimedia.org/wikipedia/commons/e/ea/Human_male_reproductive_system_en.svg

[68] Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. 2016. Apache Spark: A unified engine for big data processing. *Commun. ACM* 59, 11 (October 2016), 56–65. DOI:https://doi.org/10.1145/2934664

A. Algorithm of the *icd* library

Given a comorbidity mapping between the ICD-10 codes and the CCI disease categories, the algorithm for ICD-10 codes has the following stages [66]:

1. Reduce the problem by using partial matching to compute the intersection of the patient diagnoses and the comorbidity mapping codes
2. Reduce the computational cost of the algorithm by converting the ICD-10 codes of the intersection to consecutive integers
3. Create a sparse matrix with a row for each visit of a patient and a column for each computed ICD-10 code (*patient-visit matrix*)
4. Create a dense matrix with a row for each computed ICD-10 code and a column for each comorbidity (*comorbidity-matrix*)
5. Multiply the matrices from the stages 3 and 4 together

Stage 1 is especially important to reduce the computational complexity because of the skewness of the distribution of the ICD-10 codes. As comorbidities are typically

Table A.1: An example of a comorbidity mapping.

Disease	Codes
Rheumatic Hearth Disease	I098
Hypertension	I10, I11
Heart failure	I150, I110

Table A.2: An imaginary patient with four visits and the associated diagnoses.

Visit	Diagnosis 1	Diagnosis 2	Diagnosis 3
1	K401		
2	I0981	C450	
3	M352	I10	
4	I110		

uncommon diagnoses, removing the diagnosis codes which do not belong a comorbidity category reduces the needed computations by a substantial amount [66].

Following the example in the original paper, using the comorbidity mapping listed in Table A.1 and the artificial patient data for a single patient listed in Table A.2, the algorithm performs as follows:

1. The computed intersection is [I098, I11, I10]
2. Assigned integers: I0981=0, I10=1, I110=2
3. The patient-visit matrix A is defined as:

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

4. The comorbidity matrix B is defined as:

$$B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

5. The matrix multiplication results in:

$$AB = C = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \end{bmatrix}$$

6. The code *I11* belongs to two different comorbidities, so the resulting incorrect indicator must be corrected:

$$[C \neq 0] = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

The result of the algorithm can be interpreted as Table A.3. Columns of the result matrix represent the different comorbidities, the rows are the patient visits and the cell

Table A.3: The resulting matrix of the Wasey et al. algorithm

Visit	Rheumatic Disease	Hypertension	Congestive Heart Failure
1	no	no	no
2	yes	no	no
3	no	yes	no
4	no	yes	yes

values correspond to whether a particular comorbidity diagnosis was observed during a certain visit [66].

The algorithm implementation makes use of the heavily optimized *RcppEigen* linear algebra library to improve the performance of the matrix calculations [9]. The paper does not include the details of how the algorithm handles exclusionary categories, but as the categories with non-zero scores can be read from the result matrix, the computational cost of the score correction should be asymptotically irrelevant [66].

B. Grambsch-Therneau plots

B.1 Urethral strictures

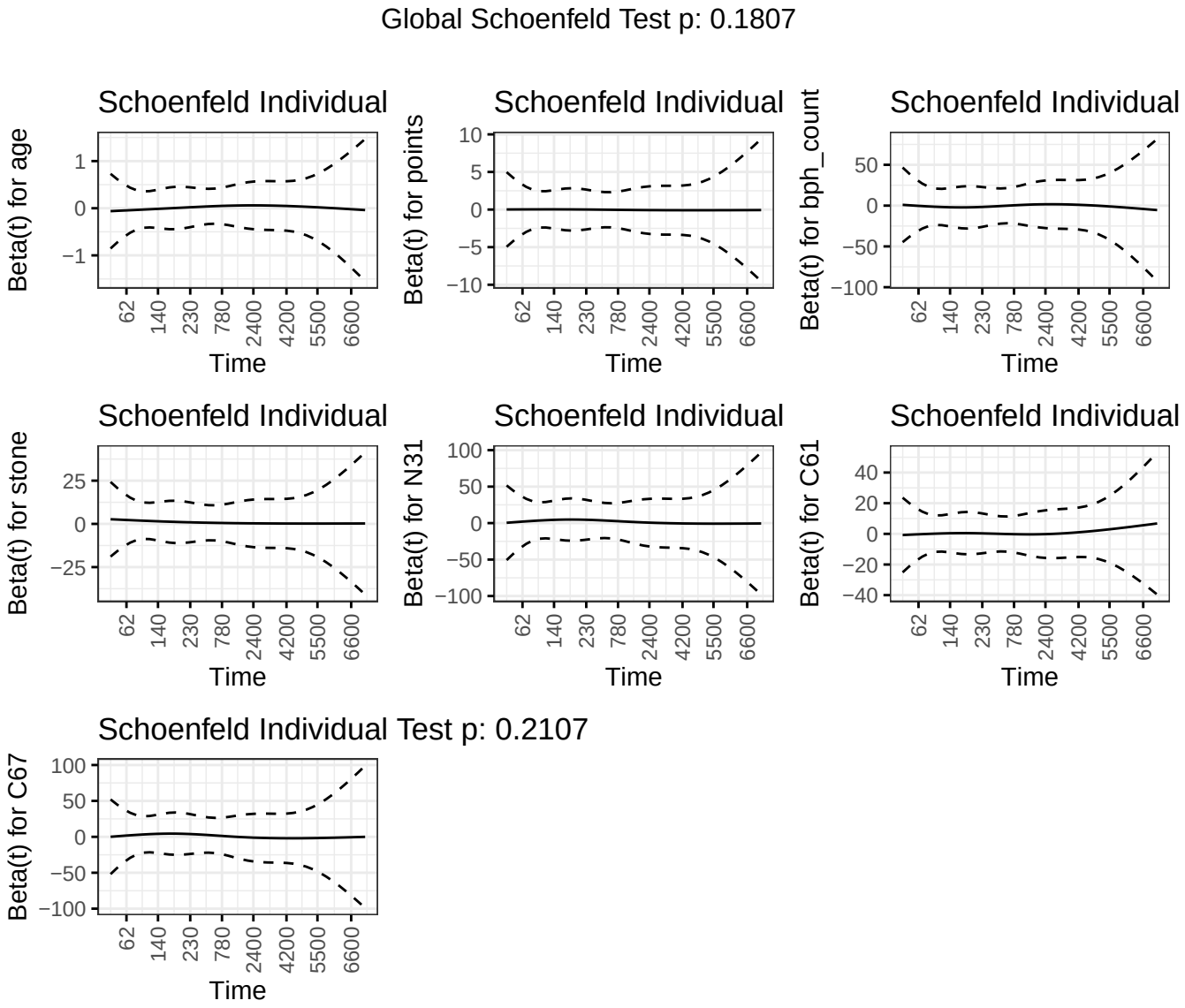

Figure B.1: KED00/10

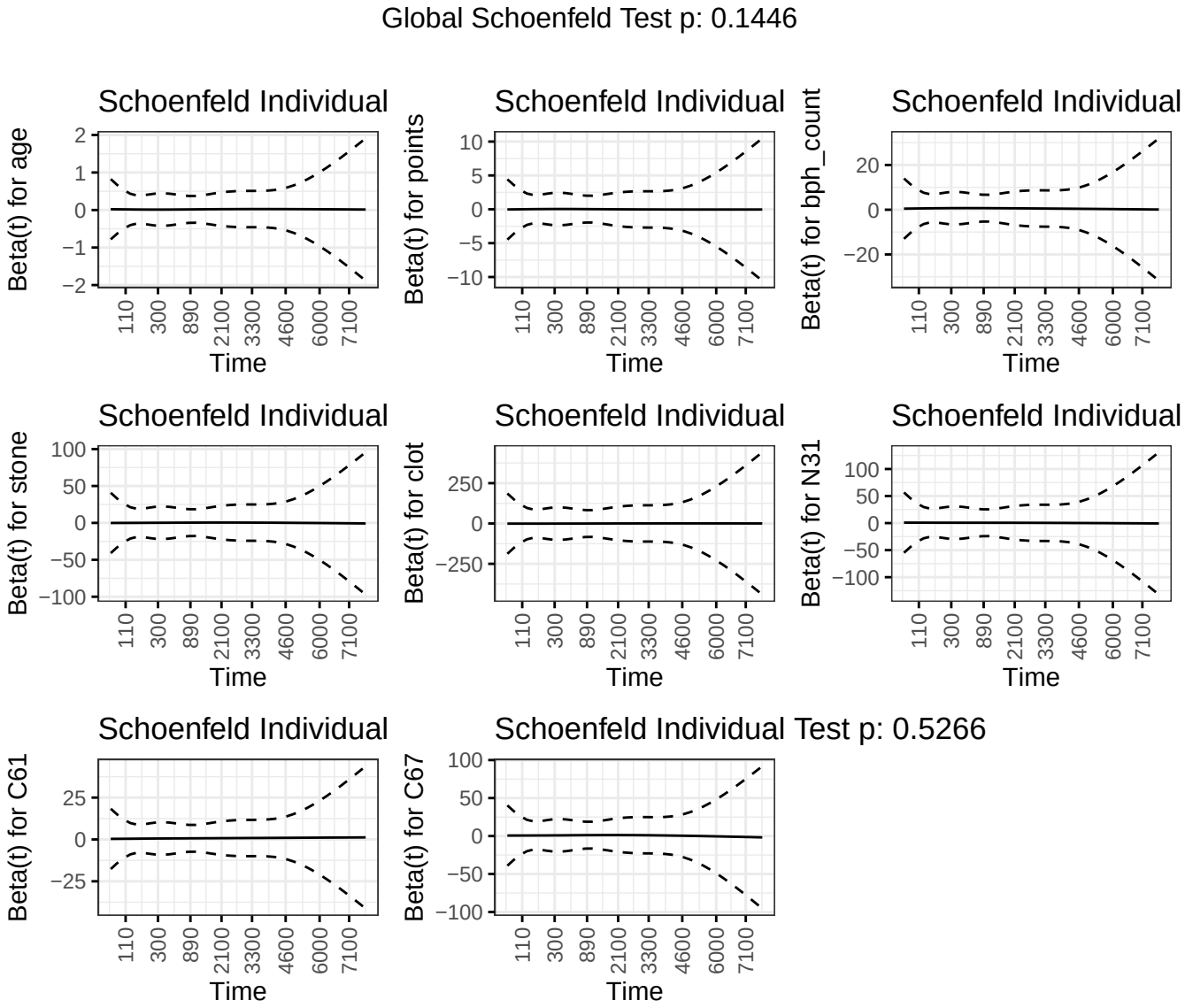

Figure B.2: KED22

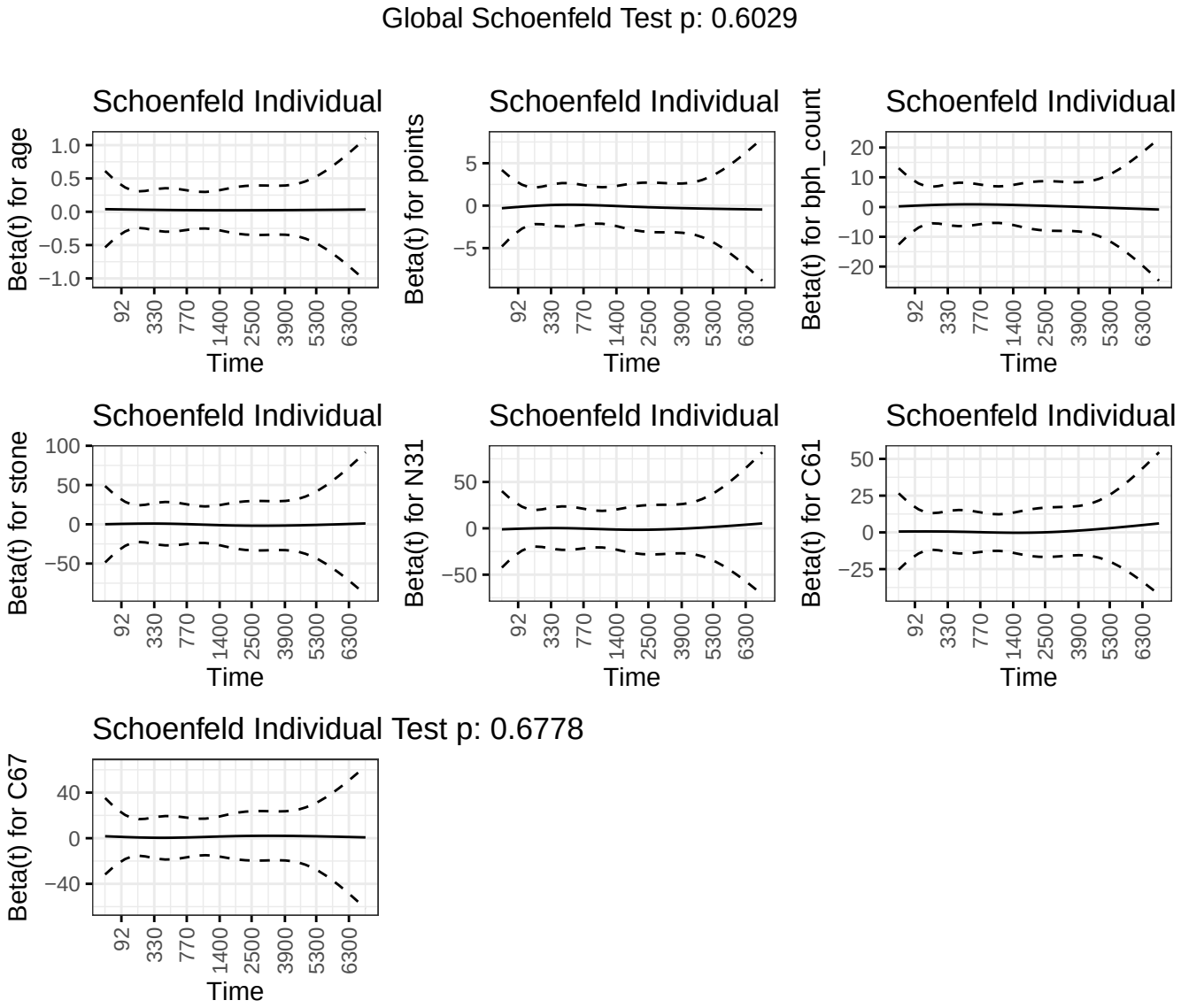

Figure B.3: KED33

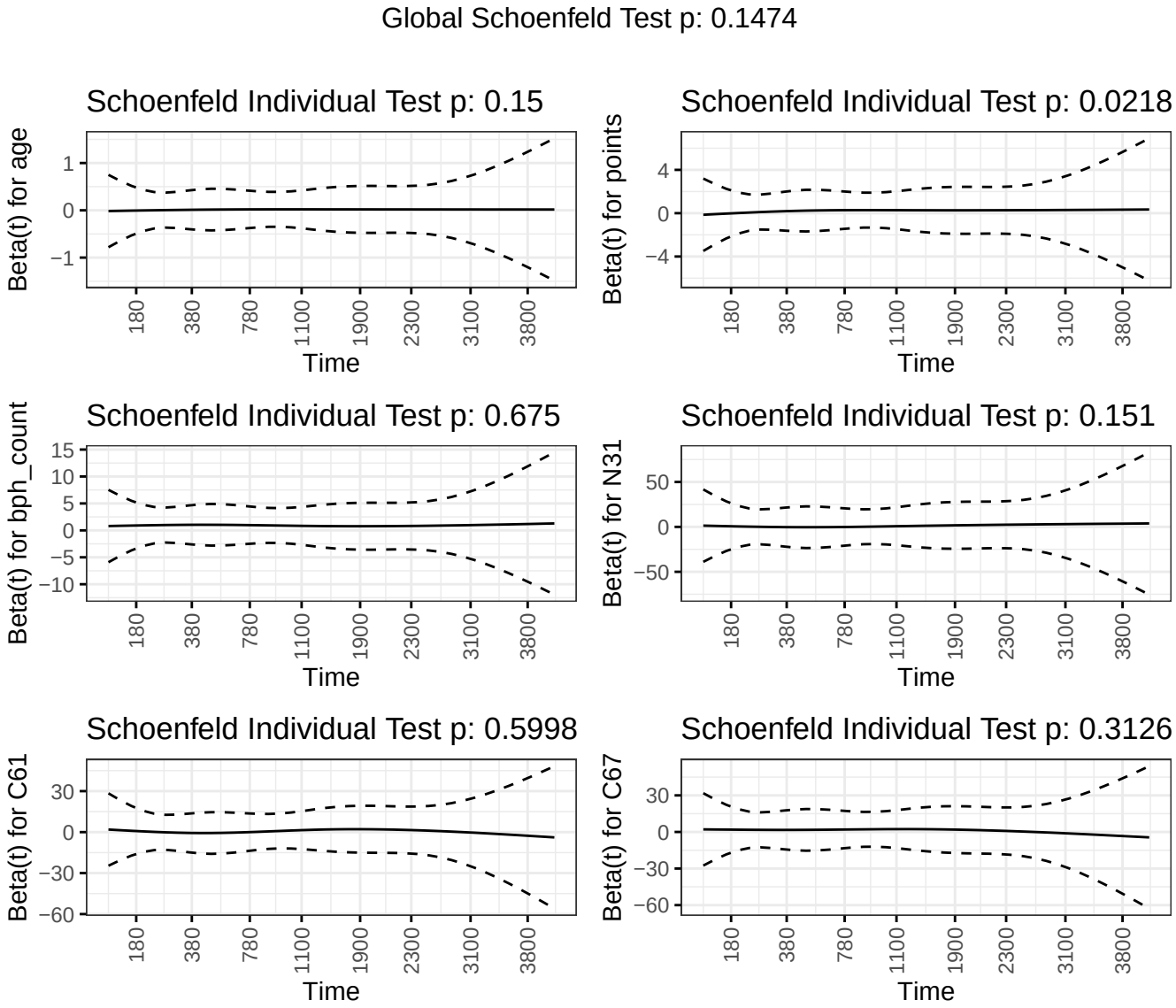

Figure B.4: KED52

B.2 Stress incontinences

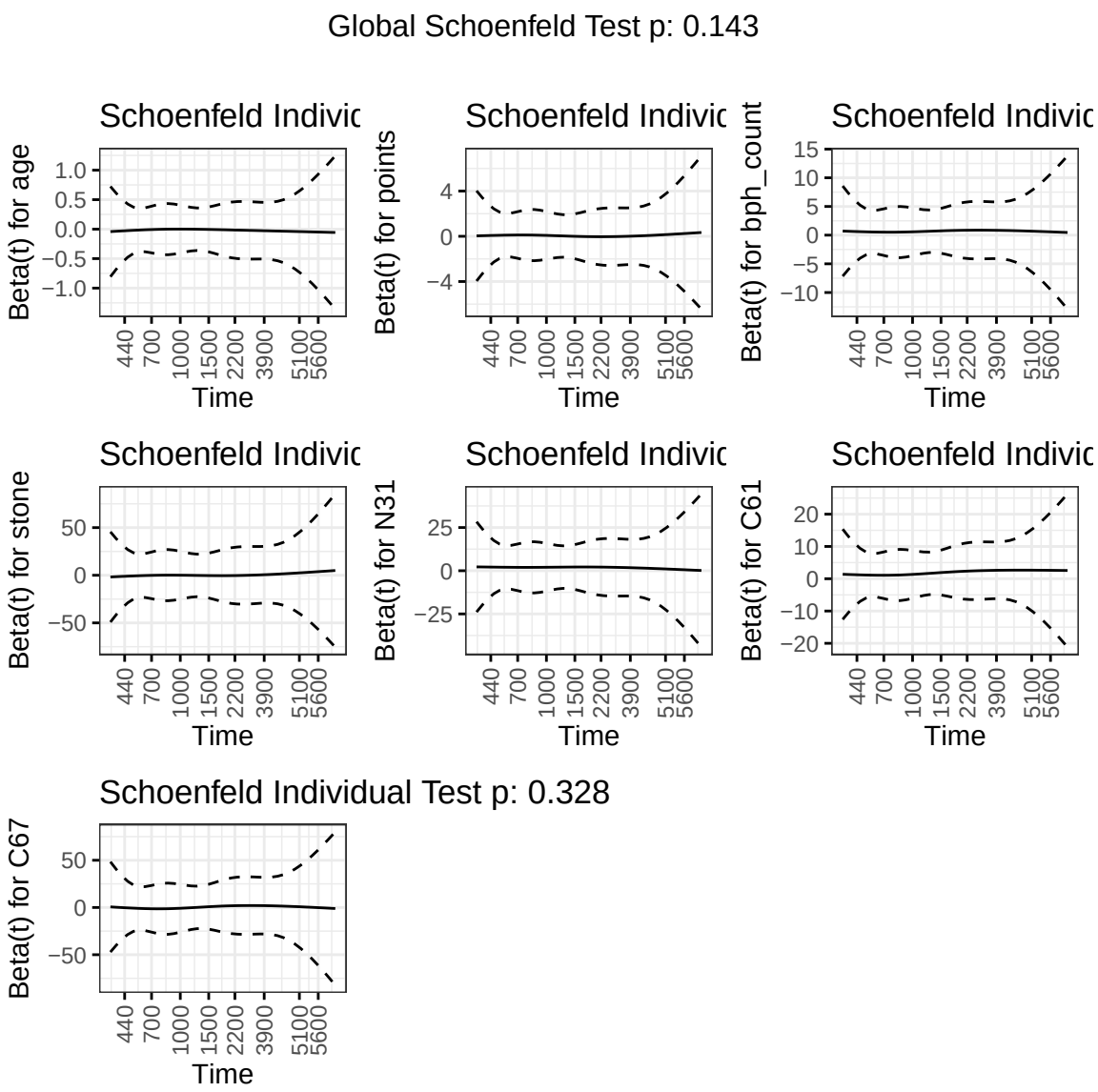

Figure B.5: KED22

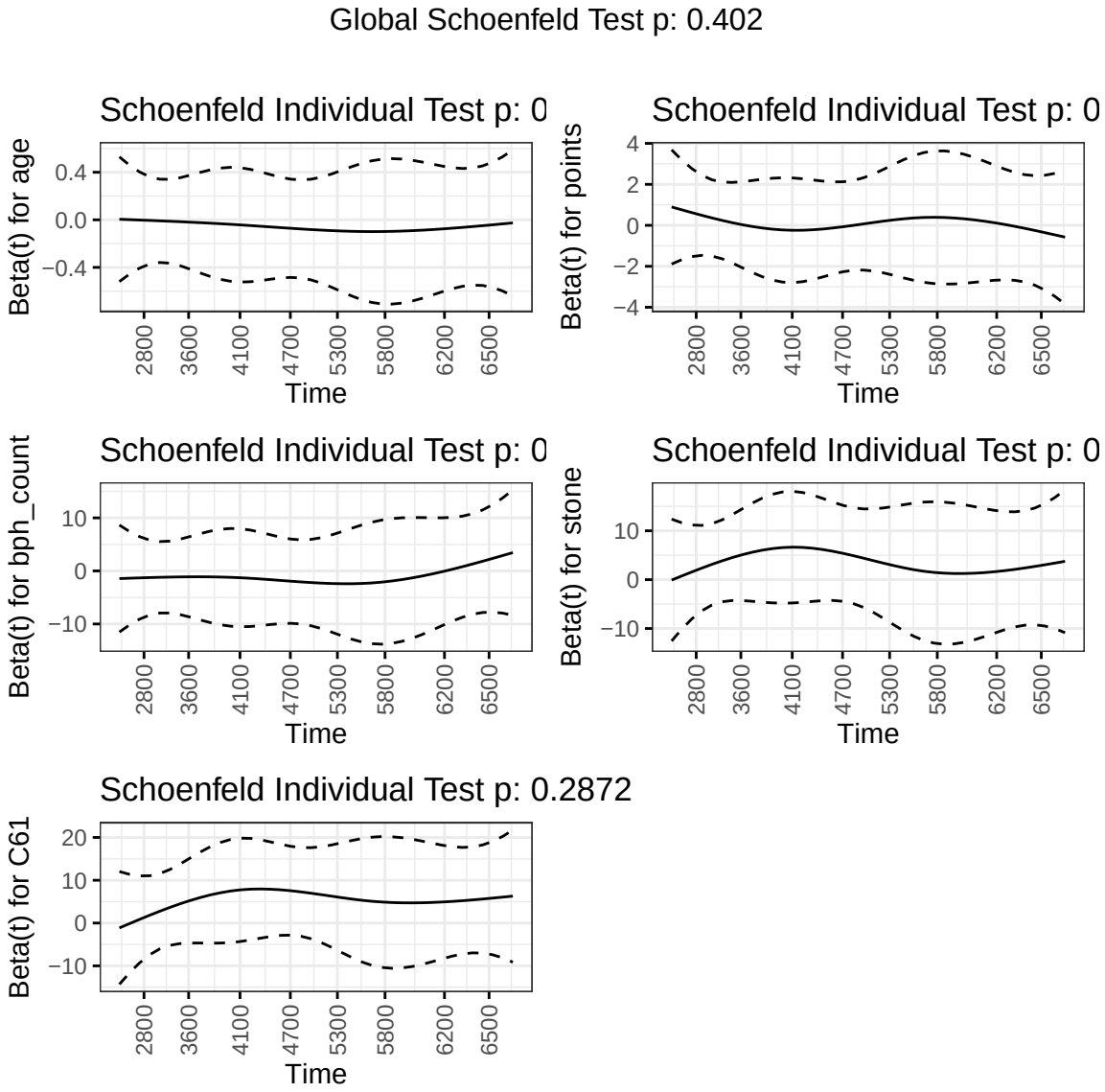

Figure B.6: KED33

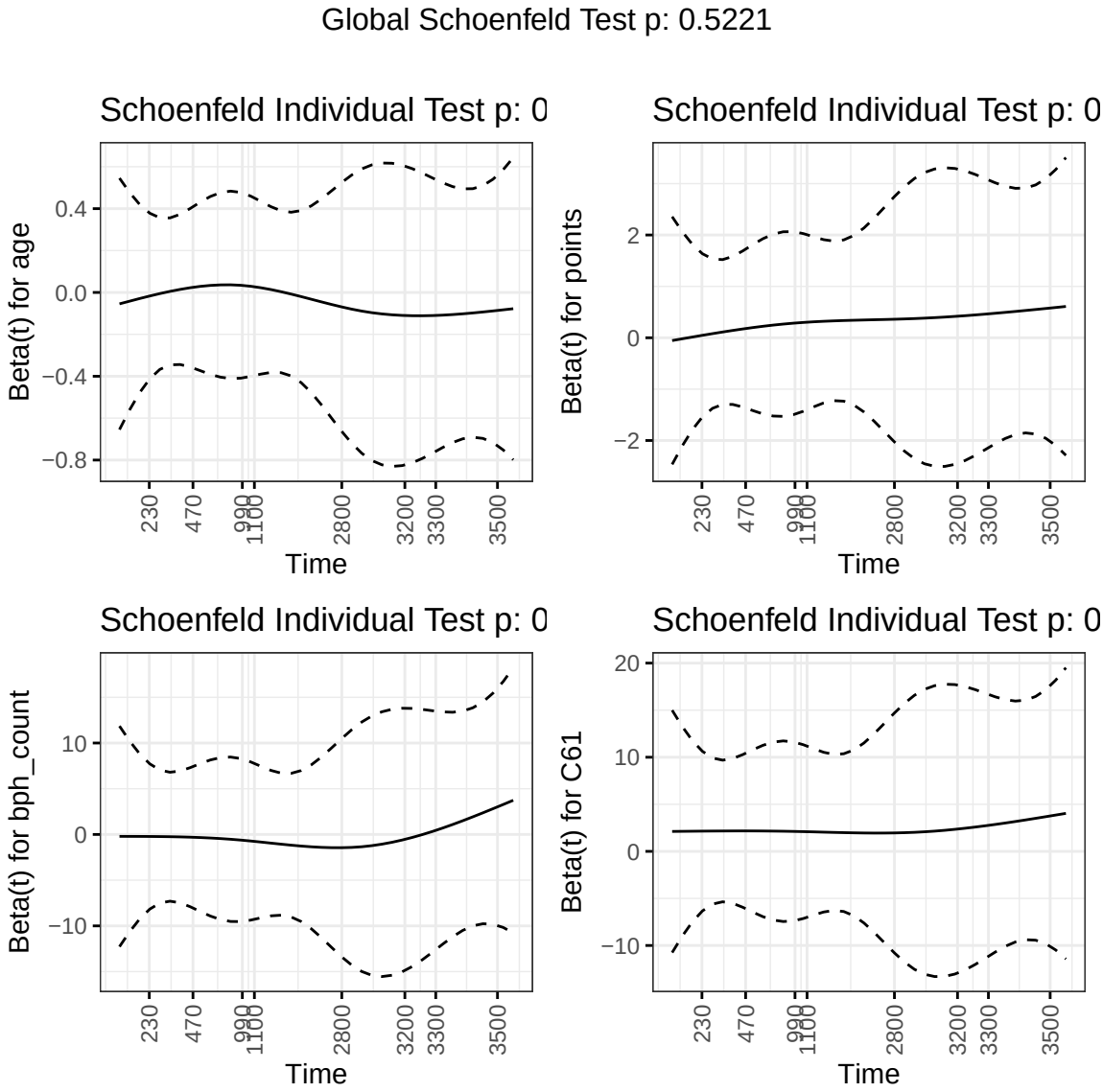

Figure B.7: KED52

B.3 BPH reoperations

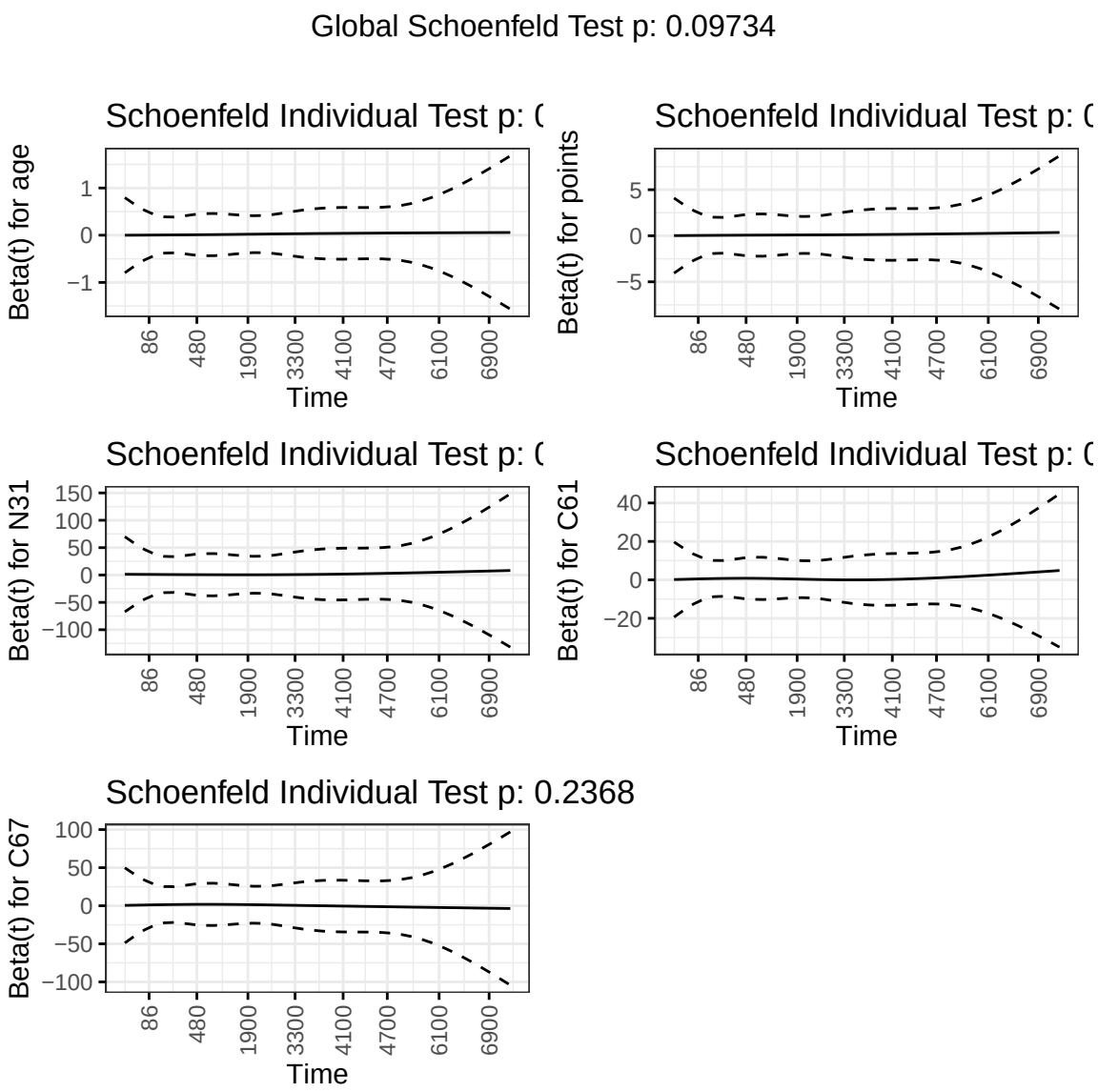

Figure B.8: KED00/10

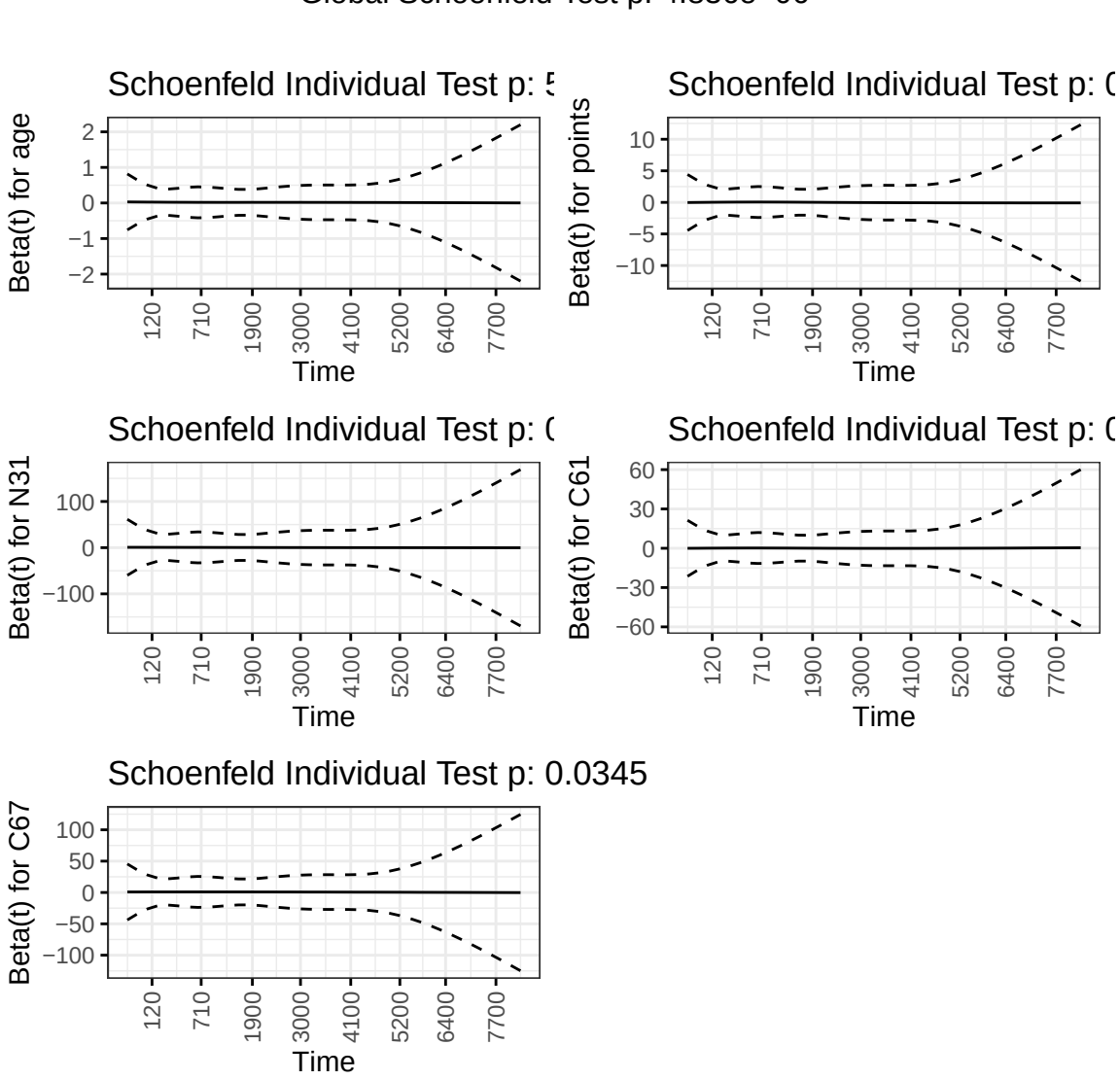

Figure B.9: KED22

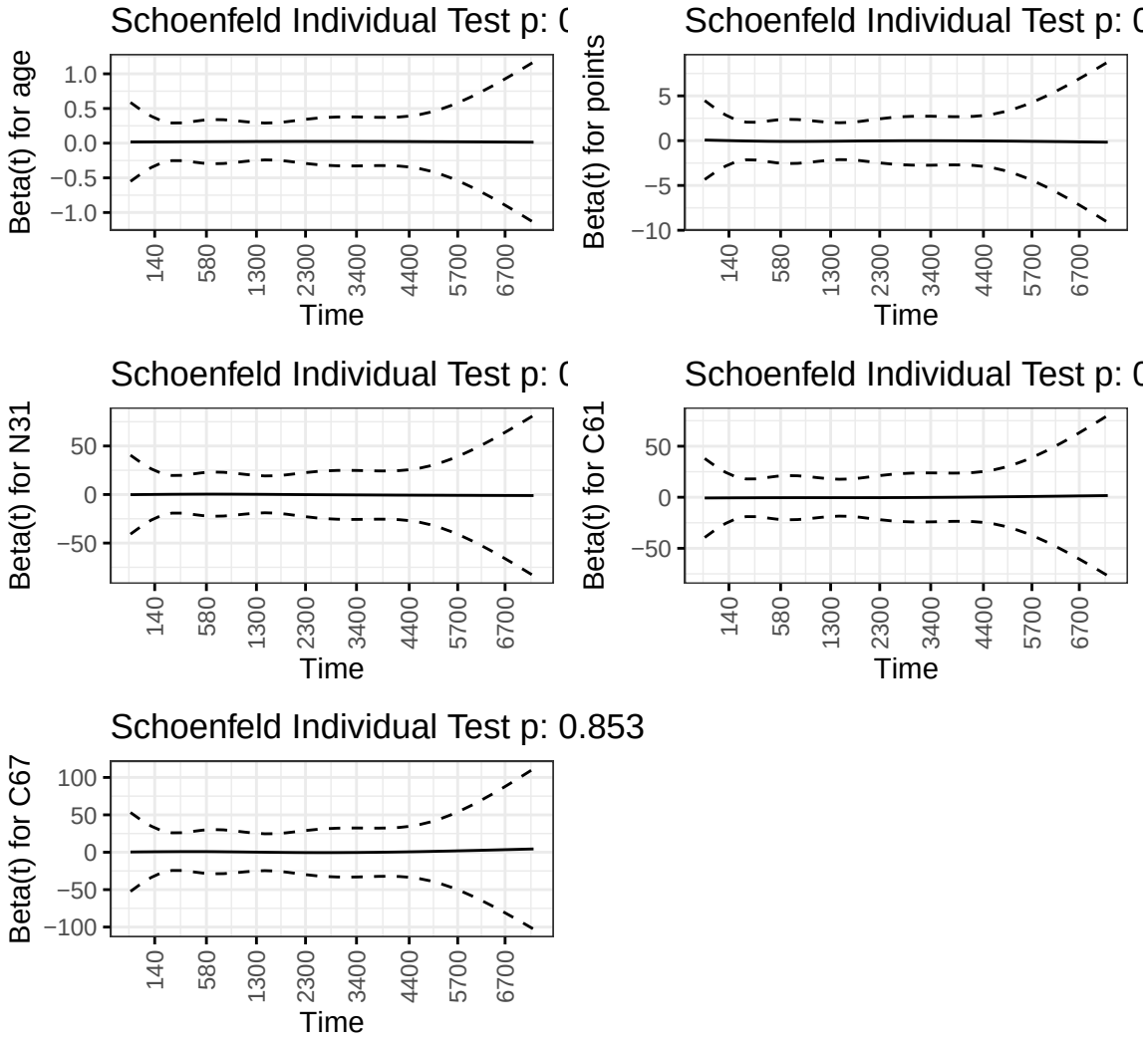

Figure B.10: KED33

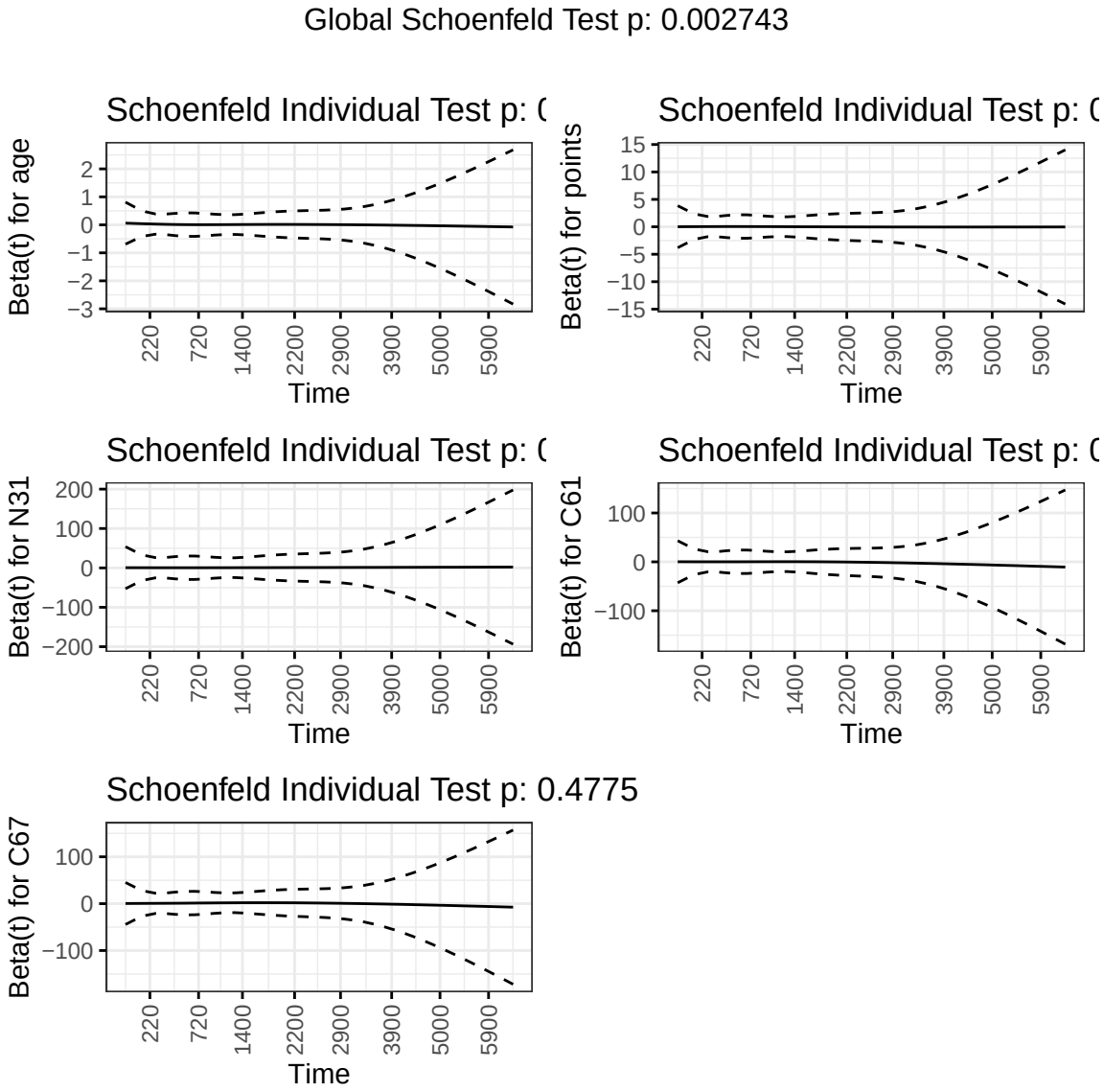

Figure B.11: KED52