UNIVERSITY OF LIVERPOOL

Exploring the potential of spinal registry data for the use in clinical trials

---

Thesis submitted in accordance with the requirements

of the University of Liverpool for the degree of

Doctor in Philosophy

by

Lukas Staudt

May 2023

Acknowledgements

I would like to express my sincere gratitude to my supervisor team Dr. Girvan Burnside, Dr. Martin Wilby, Prof. Anthony Marson and Dr. Maria Sudell for their invaluable guidance and support throughout my PhD journey. Their expertise, patience and encouragement have been a constant source of motivation and inspiration to me. A special thanks goes to Dr. Girvan Burnside who has always been my first contact whenever I needed advice and mentorship and without whom I could not have made it this far.

Additionally, I would like to thank Dr. Emin Aghayev and the Spine Tango Committee for granting me access to their data which was crucial to the completion of my thesis. Their generosity and support have made this research possible and I am truly grateful for the opportunity to use their valuable resources.

I would also like to thank the University of Liverpool's JIC Davies studentship for research in Orthopaedic Science and Practice for funding my studies and providing a very positive and supportive work environment.

My sincerest gratitude goes to my family and friends who have supported me throughout this journey. Their love, encouragement and inspiration have been a source of strength and have sustained me during the most challenging times.

A special thanks goes to my lifelong friends David and Thomas, as well as my girlfriend Nicole, whom I could always count on for support and advice and who always had time for me.

Abstract

Title: Exploring the potential of the use of spinal registry data for the use in clinical trials

Author: Lukas Staudt

Background: Sciatica describes the symptoms of low-back and leg pain most commonly due to a herniated disc that presses on the sciatic nerve. If persistent, invasive methods such as surgical microdiscectomy are required. Although being a surgery with relatively small incisions, it bears some risks of adverse events (AEs), e.g. durotomy, wound infection or in rare cases even nerve root damage. Observational registries allow for continuous data collection over indefinite time for numerous patients. One can therefore gain additional insight in subgroup demographics of the patient population and rare events. Furthermore, large numbers of patients and observations can improve the performance of prediction models.

Purpose: The aims of this study are to: (1) provide a comprehensive overview of the collected dataset from the Spine Tango registry; (2) determine the best method for imputing missing data in this routinely collected registry data set; (3) assess the predictive values of patient characteristics on patient-reported outcome measures (PROMs) and complications during surgery; and (4) examine the utilization of registries in both clinical trials and observational studies and identify strategies to increase their impact on clinical trials.

Methods: To understand the patient population and potential relationships among collected variables, thorough descriptive statistics were performed. Simulation studies were conducted to determine the best approach for imputing PROM items and scores, including the examination of missingness percentages, mechanisms, and cut-off point score calculations. The focus of prediction modeling was the routinely collected Core Outcome Measurement Index (COMI) and complications. Patients with sciatica were identified in collaboration with the Spine Tango committee, and various model approaches were compared for goodness of fit and prediction accuracy, including regression and mixed models. A literature review of both randomised controlled trials and observational studies was conducted, comparing differences in missing data, collected outcomes, study length, number of patients, and registry use. Case studies of successful registry utilization in other clinical areas were analyzed to identify potential for implementation in the present clinical focus.

Results: The international nature of the Spine Tango registry led to variability in documentation and data collection across countries. The simulation studies showed that item-based imputation was superior to score-based imputation in most scenarios. Mixed models with random intercepts and

slopes, as well as non-linear time terms, performed best in terms of model fit. Logistic regression models that defined complications as outcome were able to identify risk factors, such as prior surgery, level of spine of physical status. The utilization of registries in the field of this clinical population is underutilized, and studies from other areas demonstrate that registry use can reduce trial costs by facilitating patient identification, data collection, and event detection, as well as reducing trial-specific patient visits and improving patient retention.

Conclusion: The potential of routinely collected registry data remains under-utilized within the sciatica-affected patient population. The noteworthy resemblances observed between observational data and randomized controlled trial data, both in descriptive statistics and prognostic factors, underscore the comparability of these sources and advocate for the integration of registry data in this domain. While the integration of a registry into a trial presents complexities, successful endeavors in related fields point to an innovative trial design that harmonizes these two research approaches.

List of Abbreviations (alphabetically)

AE	Adverse Event
AIC	Akaike Information Criterion
ASA	American Society of Anaesthesiologists
AUC	Area Under Curve
BMI	Body Mass Index
BSR	British Spine Registry
CCA	Complete Case Analysis
CI	Confidence Interval
COMI	Core Outcome Measures Index
ESI	Epidural Steroid Injection
KS	Kolmogorov Smirnov
MAR	Missing At Random
MCAR	Missing Completely At Random
MCID	Minimal Clinically Important Difference
MI	Multiple Imputation
MNAR	Missing Not At Random
MRM	Modified Roland Morris
NERVES	NErve Root block VErsus Surgery
NHS	National Health Service
NRS	Numeric Rating Scale
ODI	Oswestry Disability Index
PCI	Pain Coping Inventory
PMM	Predictive Mean Matching
PROM	Patient-Reported Outcome Measure
QoL	Quality of Life
RCT	Randomised Controlled Trial
RMSE	Root-Mean-Square Error
ROC	Receiver Operating Characteristic
s.d.	Standard deviation
SAP	Statistical Analysis Plan
SBI	Sciatica Bothersome Index

SF-36	Short Form 36
SPORT	Spine Patient Outcomes Research Trial
ST	Spine Tango
STEMI	ST-segment Elevation Myocardial Infarction
VAS	Visual Analog Scale

Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Chapter outline

In this opening chapter, an overview of the content and structure of the thesis is provided. The chapter begins by delving into the medical background of sciatica, explaining its characteristics, causes, prevalent treatment options and methods of measuring pain and quality of life. Various clinical study designs are discussed, with an emphasis on the importance of randomized controlled trials (RCTs). A literature review of impactful studies in sciatica treatment is conducted, focusing on key comparisons such as microdiscectomy vs. conservative treatment and microdiscectomy vs. epidural steroid injections. Additionally, the potential of routinely collected data in clinical research is examined, highlighting its advantages and limitations. Finally, an outline of the overall structure and objectives of the thesis is presented.

1.2 Medical background – Sciatica and treatment options

The sciatic nerve is the longest and widest nerve in the human body and runs from the lower back down the back of each leg (Ropper and Zafonte, 2015). It controls several muscles in the legs and receives sensation signals from the skin. Symptoms of any irritation of this nerve include lower back and leg pain, as well as numbness. The term sciatica is often confused with general back pain but is a symptom, not a condition. The most common cause for sciatica is a herniated disc that is pressing on the nerve (in over 90% of the cases); often in the lumbar region of the spine (Koes et al., 2007). Sciatica affects over 3% of the UK population at any time (Wilby et al., 2021).

Duration and severity can vary and in 60-90% of patients, spontaneous regression occurs. Symptoms can then be treated with conservative methods such as physiotherapy and analgesics (Chen et al., 2018). Persisting pain (longer than 6 weeks) though, might require invasive methods. Most commonly performed is a surgical microdiscectomy. In this procedure the portion of the disc that is pressing on the nerve is removed. Although being an open surgery, a microdiscectomy can be done with relatively small incisions and minimal tissue damage. In most comparative clinical studies, the effectiveness of microdiscectomy and non-surgical treatment, such as physiotherapy or analgesics, has been investigated. However, results of those studies were inconclusive. Whereas some studies found significant superiority of surgery, others concluded that there is no difference in the long term (Atlas et al., 1996, Buttermann, 2004, Chen et al., 2018, Osterman et al., 2006, Weinstein et al., 2008). A meta-analysis has shown that the difference of treatment outcomes is not significant enough to

establish microdiscectomy as overall superior and therefore there exist no specific healthcare guidelines (Chen et al., 2018).

Epidural steroid injections (ESI) are a non-surgical treatment option for sciatica that may be used in combination with physiotherapy. During the procedure, a mixture of local anaesthetic and steroid medication is injected into the spine through one of three routes: caudal epidural, inter-laminar, or transforaminal epidural steroid injection (TFESI).

There are currently no specific recommendations for the treatment of sciatica in individual cases. Further research is needed to better understand the most effective treatment methods for optimizing healthcare in these cases. Before reviewing prior impactful trials in this medical field, it is necessary to examine and discuss commonly used methods for measuring pain in trials.

1.3 Outcome Measures for back and leg pain

1.3.1 Pain scores

Many clinical studies can rely on quite accurate instruments to obtain a certain measurement value e.g. blood pressure, body weight, blood sugar etc. There are no instruments though, to objectively measure pain, which has to be reported by the patients themselves. There are several techniques to assess a patient's pain and to capture any improvement of a certain treatment. Methods with which a patient indicates pain for example on a scale between "no pain" and "pain as bad as it could be" such as a Visual Analogue scale (VAS) (Price et al., 1983) are still commonly used.

1.3.2 Quality-of-life (QoL) questionnaires

For specific pathologies such as sciatica, there exist questionnaires that try to evaluate pain intensity and quality of life as informatively as possible, in order to quantify a treatment effect. It is common to assess a patient's quality of life by using a score that considers not only the intensity of pain, but also its impact on daily life. This approach aims to measure how the pain affects the patient's ability to perform daily activities and overall quality of life, rather than just the intensity of the pain itself.

The modified Roland Morris (MRM) questionnaire (Roland and Morris, 1983) is a tool used to assess the impact of low back pain on an individual's quality of life. It is a self-administered questionnaire that consists of 24 statements about daily activities that may be affected by low back pain. The individual is asked to indicate how often they experience difficulty with each activity due to their low back pain on a scale of 0 (never) to 5 (always). The total score is calculated by summing the responses to each statement, with a higher score indicating a greater impact of low back pain on quality of life. The MRMQ has been modified for use in assessing the impact of sciatica specifically (Kim et al., 2010).

Another method is the Oswestry Disability Index (ODI) (Fairbank and Pynsent, 2000). It consists of 10 items that cover several characteristics of low back and leg pain, e.g. intensity, standing, sleeping, sex life (if applicable) etc. Each item is scored from 0 to 5, with higher values representing greater disability. The overall score is then calculated as percentage of the scores of the applicable sections.

The Core Outcome Measure Index (COMI) is a questionnaire for assessing low-back and leg pain, patient satisfaction, and treatment complications (Mannion et al., 2016). It has been found to be highly correlated with other methods for measuring these outcomes. It is short but consistent, what makes it desirable for minimizing non-response and missing data.

Other commonly used measurements are the Sciatica Bothersome Index (SBI) and the Short Form 36 (SF-36) (Burholt and Nash, 2011, Grøvle et al., 2010). The SBI is a questionnaire that captures the impact of sciatica on quality of life. Items cover the frequency and intensity of sciatica-related symptoms, as well as the impact of these symptoms on daily activities. The SF-36 is a general health survey that assesses physical and mental health-related quality of life. It consists of 36 items that cover physical functioning, role physical, bodily pain, as well as mental health.

Chiarotto et al. aimed to generate a consensus of outcome measurements for low back pain and proposed ODI and Roland Morris for physical functioning and the numeric rating scale (NRS) for pain intensity (Chiarotto et al., 2018), but there exist no standards yet. One of the difficulties of comparing different studies that investigate the effect of different sciatica treatments is that outcome measures are only comparable up to a certain degree. Most of the measures correlate with each other quite strongly, but common measurement standards are needed. Whether or not the COMI can be used as core outcome measurement for sciatica patients has to be verified by further studies.

1.4 Clinical study designs

Before examining clinical trials for the treatment of sciatica, it is important to understand the different types of study designs that are commonly used in clinical research. Clinical studies can be classified as either observational or experimental, depending on whether the investigator intervenes in the study or simply observes and collects data. Experimental studies involve introducing an intervention and studying its effects, while observational studies do not involve any intervention. Clinical studies can also be classified based on the time frame in which data are collected, as either retrospective or prospective. Retrospective studies involve looking at data from the past, while prospective studies involve collecting data going forward from the start of the study. Observational studies can be either retrospective or prospective, while experimental studies are always prospective. When an

observational study covers a long period of time, it is called a cohort study (Ranganathan and Aggarwal, 2018).

When evaluating the effectiveness of a treatment in clinical studies, it's essential to make comparisons with other treatments or control groups, such as placebos. To obtain accurate results, one should maintain balance between the treatment groups, considering factors like age, sex, and severity of disease that might influence the measured outcome. The aim is to minimize any bias in the estimates that could lead to unreliable results. However, there are various ways that bias can be introduced, making it difficult to ensure completely bias-free results. Selection bias, for example, can occur when patients or clinicians select interventions based on personal preference. To prevent selection bias, some clinical studies therefore randomly assign recruited patients to a treatment (Randomised Controlled Trials) (Infante-Rivard and Cusson, 2018). However, when comparing interventions such as surgery and physiotherapy, it is not possible to ensure that neither the clinician nor the patient knows which treatment has been assigned (blinding).

1.4.1 Randomised controlled trials

Randomised controlled trials (RCTs) are considered the "gold standard" in clinical research because they aim to control as many confounding factors as possible. In RCTs, patients are randomly assigned to a treatment and can therefore only be conducted prospectively. However, RCTs can be expensive and time-consuming due to regulatory requirements. Due to strict protocols, it is also possible that they do not accurately reflect real-world practice. Observational and non-randomised studies can sometimes be a useful alternative (Spieth et al., 2016).

A potential issue of analysis of RCTs is cross-over. It occurs when patients are allowed to switch to another treatment arm after their assignment. One can analyse the data as treated (AT) or as intended to treat (ITT). ITT describes the analysis of patients using the treatment they were initially assigned to, instead of the treatment they received. The ITT method preserves the benefits of randomization, but the as-treated method is more intuitive. Patients may also drop out of a trial for various reasons, leading to missing data. High percentages of patients that changed treatment during the study or were lost to follow-up can impact statistical results. It is therefore crucial to consider which methods to use for analysis, to prevent incorrect conclusions (Tripepi et al., 2020).

RCTs are highly regulated, making them challenging to conduct. The approval and funding process also involve bureaucratic barriers, leading to high costs for conducting the trial (Hariton and Locascio, 2018). These factors may decrease the willingness of patients to participate. Additionally, strict criteria

for patient recruitment and the potential for randomisation cause a non-representative sample of the patient population. Blinding is used to address treatment preferences and protect against bias, but this is not always possible, e.g. in the case of treatments with distinctive procedures (e.g. TFESI compared to microdiscectomy).

1.4.2 Observational studies

Without the process of randomisation, the choice of treatment is very likely to be connected to the severity of the disease and presence of other conditions. Even if the statistical analysis methods account for potential confounding factors and differences between patients, these adjusted associations might still reflect residual confounding due to factors that were not assessed properly or due to unknown associated factors. Such potential biases might result in false conclusion, especially if the investigated treatment effect is rather moderate (Faraoni and Schaefer, 2016, Nørgaard et al., 2017).

Many reviews that compared treatment estimates from randomised trials and observational trials found those estimates to be significantly different from each other. For example, an observational study of the Danish Civil Registration System that tracks 98% of all incidents of cancer in Denmark reported that statin use in cancer patients is associated with reduced cancer-related mortality, even statistically adjusting for known potentially confounding factors (Nielsen et al., 2012). Other observational studies reported statin therapy being associated with reduced incidence of cancer. A later performed meta-analysis of RCTs including more than 10,000 patients reported no apparent effects of statins on incidence of cancer or death due to cancer (Collins et al., 2016).This shows that randomisation is in many cases a helpful tool to simplify the capture of unbiased cause-effect relationships.

Ultimately, RCTs are with good reason seen as ‘gold-standard’ for clinical research, especially if a study aims for the approval of a new intervention. In case of evaluating treatments that are routinely done though, the analysis of vast amount of collected data in registries can complement existing RCT results and add valuable insights. Concato et al. proposed that both types of studies are needed to assess the effect of treatments and that they should rather be seen as a compliments instead of one being superior over the other (Concato et al., 2010). Meta-analyses are seen as the highest level of quality in evidence-based research and can combine results from several former studies, understand underlying design issues and account for potential biases in the interpretation of the combined data (Colditz, 2010).

1.5 Literature review of impactful prior studies regarding sciatica treatment

In this section, some studies that have had a significant impact on the treatment of sciatica will be reviewed.

1.5.1 Microdiscectomy vs conservative treatment

In 1996, Atlas et al. conducted a study called "The Maine Lumbar Spine Study" that focussed on the outcomes of surgical and nonsurgical treatment for sciatica. It was a prospective cohort study that followed 507 patients recruited from medical practices in Maine (US) for one year. Instead of randomly assigning patients to treatment groups, the treatment was chosen by the patient and the physician (275 patients received surgery, while 232 received nonsurgical treatment). This study design may have led to a bias in the results, as it is likely that more severe cases of pain were treated with surgery. However, baseline characteristics such as age and sex were balanced between the two groups, although the nonsurgical group had a slightly higher percentage in patients that received of workers' compensation (40.1% in the nonsurgical group and 30.4% in the surgical group, p=0.02). Worker's compensation is a form of insurance that provides benefits to employees who suffer from work-related illness or injuries (U.S. Department of Labor). This could be a factor that affects the quality of treatment and possibly its outcomes. A total of 118 patients (23% of the total) dropped out of the study, but the characteristics of these "dropouts" were similar between the two groups. Therefore, only the remaining patients were included in the statistical analyses. When examining baseline clinical features such as pain severity and disability, there were many significant differences between the two groups. The Modified Roland Morris score and the number of disability days in the past month were significantly higher in the surgical group. In addition, the surgical group had a higher percentage of patients receiving narcotic treatment and a higher percentage of patients with worse leg pain compared to back pain. The significant differences in these clinical features suggest that the two treatment groups were not statistically comparable and the results of the study may be biased. The study found a significant superiority of surgery, with 71% of patients in the surgical group reporting definite improvement compared to 43% in the nonsurgical group. However, these results should be interpreted cautiously (Atlas et al., 1996).

A very impactful trial was the Spine Patient Outcomes Research Trial (SPORT) by Weinstein et al. It was a combination of both observational and randomised trial components, with 501 enrolled participants in the randomised trial and 743 participants in the observational cohort at 13 spine clinics in 11 US states.

The observational component involved enrolling participants into different treatment groups based on their preferences, clinical characteristics and other factors. These participants were followed over time, and their outcomes were assessed using standardized measures and patient-reported

questionnaires. It aimed to gather real-world data on the outcomes of different treatment approaches as they were chosen by the participants and their healthcare providers. In the randomised component of the study, participants were randomly assigned to receive either surgical intervention or non-surgical treatments. By combining these two components, SPORT aimed to provide a comprehensive assessment of the effectiveness of different treatment approaches for spinal disorders. This hybrid design allowed researchers to gather real-world evidence from observational data while also establishing causality and evaluating treatment effects through randomised controlled trials.

The study was conducted by the Department of Veterans Affairs and funded by the National Institutes of Health. The SPORT trial enrolled 2,437 patients with low back pain and sciatica at 13 clinical sites in the United States between 1999 and 2004. Participants that were part of the randomised component of the trial (501), were allocated to one of three treatment groups: surgical intervention (discectomy or laminectomy), non-surgical intervention (physical therapy, education, and medication), or "watchful waiting" (delayed treatment). The results of the SPORT trial were published in the New England Journal of Medicine in 2006. It also continued to collect data after the first publication, which was subsequently investigated in secondary analyses.

The study compared the effects of discectomy and non-operative care on outcome measures such as changes in the SF-36 Bodily Pain and Physical Function scales and the modified Oswestry Disability Questionnaire. These measures were assessed at 6 weeks, 3 and 6 months, and annually for four years. Due to a high rate of cross-over in the randomised trial, the statistical analysis was conducted using both the intention-to-treat method and the as-treated method. By the end of the four years, only 59% of the patients allocated to surgery had actually received it, while 45% of the patients allocated to non-operative care received surgery. Most patients who switched treatment arms did so within the first year (57% and 41% in the surgery and non-operative groups, respectively). The observational cohort had a significantly lower rate of cross-over (95% of patients who chose surgery received it, while 24% of patients who chose non-operative care received surgery). The two study groups had similar baseline measurements, although the observational cohort had slightly more symptoms and functional impairment. The results of both study groups showed significant benefits of surgery for all secondary measures except for work status, which showed a non-significant benefit. The primary outcomes of the observational study also showed a significant benefit for surgery. The as-treated analysis of the randomised trial produced results similar to those of the observational study, showing a significant benefit for surgery. However, the intention-to-treat analysis did not show a significant benefit (p-values of 0.15, 0.42, and 0.074 for bodily pain, physical function, and ODI, respectively). This highlights the challenges of statistical analysis in randomised trials with high rates of cross-over. An

intention-to-treat analysis may not detect benefits, while an as-treated analysis does not protect against confounding (Weinstein et al., 2008).

In 2018, Chen et al. published a systematic meta-analysis that included 19 randomised trials comparing surgery to non-operative treatment for lumbar disc herniation. Sample sizes of the included studies varied between 40 and 472 with a mean and standard deviation of 110.37 (s.d. 97.38). The total number of patients of all included trials was 2,272. The trials measured various outcomes, including self-reported pain (using visual and numeric rating scales, 11 studies), the ODI questionnaire (5 studies), and adverse events. These outcomes were divided into three categories: short-term (one to three months), mid-term (three to six months), and long-term (up to 12 months). Non-operative treatment included various approaches such as physiotherapy, home exercise instruction, medication, bed rest, and epidural steroid injections.

Compared with non-operative treatment, surgical treatment was more effective in lowering pain. For this, 12 trials were compared. Of these trials, 9 (961 patients) reported data at the short-term, three (293 patients) at the mid-term, and seven (725 patients) at the long-term follow-up periods. Two trials did not report the time points. A general concern about the quality of the data was raised. The sensitivity analysis indicated that the pooled result was unstable when the studies were removed one by one for the short-term follow-up period. Additionally, the funnel plot showed an asymmetrical distribution for self-related pain, suggesting the possibility of publication bias.

A total of 5 studies (842 patients) recorded the ODI questionnaire at the short-term, four (765) at the mid-term and four (762) at the long-term. Overall, surgical intervention was found to be more effective in improving disability than conservative treatment. The test for subgroup differences revealed that surgical treatment more effective than conservative treatment for mid-term and long-term periods. The significance of these findings did not change when the studies were removed one by one at the mid-term.

In terms of adverse events, no significant difference was observed between surgical and conservative treatments in 6 trials involving 1,060 patients. The sensitivity analysis indicated that the pooled result was not influenced by the individual trials.

Overall, the study concluded that it was not possible to make a firm recommendation based on the available data (Chen et al., 2018).

1.5.2 Microdiscectomy vs epidural steroid injection (ESI)

There are several studies that compare microdiscectomy to conservative treatments like physiotherapy, as well as studies that compare it specifically to epidural steroid injections. Epidural

steroid injections are a low-risk alternative to surgery, and their efficacy was investigated in a study by Buttermann et al. (Buttermann, 2004). This randomised prospective trial studied 100 patients who had not improved after at least six weeks of non-invasive treatment. The trial used interlaminar epidural steroid injections and assessed treatment improvement using the ODI questionnaire as a measure of quality of life. Both treatments resulted in a significant decrease in pain and disability ($p<0.0001$), but surgery was found to be superior to injections ($p=0.015$). The study concluded that epidural steroid injections can be a viable alternative, but are inferior to surgery. A retrospective case series by Manson et al. reported that surgery could be avoided in 56% of surgical candidates, while a retrospective study by Wang et al. reported that 77% of surgical candidates could avoid surgery (Wang et al., 2002, Manson et al., 2013).

It should be noted that the study by Buttermann et al. only considered the interlaminar approach for injections, and it is not known if other approaches would lead to different treatment results (Buttermann, 2004). A meta-analysis by Lee et al. reviewed seven randomised controlled trials and three prospective observational studies and compared the treatment effects of interlaminar and transforaminal steroid injections using pain intensity measurements such as the visual analogue scale (VAS), ODI, and numerical rating scales. The study found that transforaminal injections resulted in significantly better short-term outcomes in terms of pain control, and non-significantly more favourable long-term pain reduction (Lee et al., 2018).

In order to develop more evidence-based healthcare guidelines, there has been ongoing research comparing transforaminal epidural steroid injections (TFESI) to microdiscectomy. Wilby et al. conducted a prospective randomised controlled trial called Nerve root block versus surgery (NERVES), which directly compared these two treatments in 163 patients with sciatic pain that had not improved after at least six weeks of non-operative treatment. The primary outcome was the Oswestry Disability Index (ODI) questionnaire score at 18 weeks post-randomisation, and secondary outcomes included numerical pain ratings, the modified Roland-Morris score, and the COMI scale at 12-week intervals, as well as patient satisfaction at 54 weeks. During the study, 35% of the TFESI group underwent additional microdiscectomy. The study used the intention-to-treat method to control for confounding factors. It found no statistically significant difference between the two groups for the primary endpoint, with mean reductions in ODI scores of 26.74 in the surgical group and 24.52 in the TFESI group. However, the surgical group had four cases of serious adverse events, while there were none in the TFESI group. Additionally, the cost of the procedures was very different, with approximately £600 for TFESI and approximately £4,000 for microdiscectomy. This suggests that TFESI could be a suitable standard procedure for the treatment of sciatica, with surgery reserved for cases where TFESI is ineffective (Wilby et al., 2021).

1.6 Potential of the use of routinely collected data

Observational studies, can provide additional information to supplement results from randomised controlled trials (RCTs). However, it is important to note that observational studies may be subject to bias, especially when the treatment effects being studied are moderate (Faraoni and Schaefer, 2016). Routinely collected data from registries may differ in structure from data collected through RCTs, due to strict data regulation policies in RCTs. Registries are often designed for large amounts of data, potentially from multiple different sites. The sites' financial resources or treatment preferences can affect the reliability of the analysis.

However, the large amount of data collected by clinical registries can be used to identify rare events or to detect patterns of outcomes in certain subgroups of patients. In 2000, EUROSPINE, the Spine Society of Europe, and the Institute for Evaluative Research in Orthopaedic Surgery at the University of Bern, Switzerland developed the Spine Tango (ST) registry for outcomes regarding spinal surgeries. It was launched in 2002 and now includes over 700,000 collected forms (EUROSPINE, 2022b).

Several studies have found that observational studies, which involve collecting data on people who are already receiving a certain treatment, can provide valuable insights into real-world practice and outcomes. Clinical registries, which contain a large amount of data, can be used to describe patterns of care, understand variations in treatment and outcomes, and identify subgroups within a heterogeneous population of people with chronic low back pain (Hooff et al., 2015).

Most studies that aim for the approval of new drugs or techniques are randomised controlled trials (RCTs). However, observational studies can be very useful for the evaluation of the effectiveness of treatments in routine clinical practice outside of a research setting. These types of studies are relatively inexpensive and quick to conduct (if registries are already set up) and can provide valuable information on routine practice. Longitudinal data is particularly useful for market surveillance (Gilmartin-Thomas et al., 2018).

Clinical registries are being increasingly acknowledged as helpful data sources in clinical research and usually focus on a specific medical condition, population, or intervention. In the era of digitalization, the data available in registries offers valuable attributes for clinical research: cost-effectiveness (once established) and easy accessibility. Denmark, Sweden and the UK have some of the most complete national databases that are collecting data from patients in hospitals and health-care organizations, e.g. SWEDEHEART, SweSpine, and Spine Tango (James et al., 2015). The analysis of the data in such registries with a representative patient population helps assessing health care effectiveness and safety and evaluating prognostic factors (José and Edelman, 2017).

1.7 Thesis outline

The aim of this project was to examine how registry data is currently being used in the context of sciatica, and to consider how insights from routinely collected data can inform future clinical trials. To do this, data from the Spine Tango registry and the NERVES trial were available.

The first step was to conduct a literature review of both observational studies and randomised controlled trials (RCTs) in the sciatica patient population, comparing the characteristics of these studies in terms of missing data, collected outcomes, study length, number of patients, and use of registries. The results of this literature review provide an overview over how registry data is currently being used and its untapped potential. RCTs from other medical fields that successfully integrated a registry in the study design were analysed to identify ways in which it can be used to support future clinical studies in the sciatica-affected patient population.

Another aspect of the project was to compare the routinely collected data from the Spine Tango registry and the NERVES trial, looking at factors such as missing data and collected outcomes. The project also involved identifying correlations between patient characteristics and conducting a full descriptive analysis of both data sets.

Additionally, part of the project was to develop an appropriate method for imputing missing data in patient-reported outcome measures in routinely collected data. To do this, a simulation study using data from the Spine Tango registry was conducted. The study used patients with complete outcome and baseline questionnaires as a basis, and artificially introduced missing data at both the item and questionnaire levels. This simulation covered several parameters, such as the mechanism of missingness, the method of imputation (at the item and questionnaire score level), the percentage of missingness, and the cut-off points for calculating questionnaire scores.

Finally, the project involved the development of prognostic models to predict patient outcomes after surgery, using techniques such as regression and mixed-effect models. Outcomes that are considered included the COMI score and complications during surgery. The goal was to identify risk factors that could improve routine healthcare and decision-making, and to develop a prognostic model that can predict patient outcomes before an intervention.

The next chapter will offer an in-depth review of various research studies, including both observational studies and RCTs. This examination will focus on factors such as collected outcomes, study duration, sample size, missing data, and whether registries were utilized. If data from different sources show similarity, it would underscore the potential for integrating registry data into RCTs. Furthermore, the

chapter will explore how such integration can function and the limitations associated with this approach.

Chapter 2: Exploring the potential of the use of spinal registry data in clinical trials

2.1 Chapter Outline

The aim of this chapter is to conduct a literature review of both observational studies and RCTs in the sciatica patient population. The objective is to compare the characteristics of these studies, including aspects such as missing data, collected outcomes, study duration, patient numbers, and the potential utilization of registries. The investigation seeks to determine the level of comparability between these two types of publications.

The rationale behind comparing missing data is to assess data quality, with an initial expectation that observational studies may have a higher proportion of missing patient information. However, if this hypothesis is not true, the reliability of results could be equally strong.

Moreover, it is anticipated that observational studies, particularly those utilizing registries, would generally exhibit larger sample sizes. In cases where the collected outcomes, duration of follow-up, and dropout rates align with those of RCTs, the extensive pool of data collected can be readily compared with RCT data. Consequently, combining these sources may offer a comprehensive overview of the study population.

Finally, the results of this literature review provide an overview of how often registry data is currently being used in studies in this population and for which purposes. RCTs from other medical fields that successfully integrated a registry in the study design were analysed to identify ways in which it can be used to support future clinical studies in the sciatica-affected patient population.

2.2 Introduction

Clinical registries are databases that collect data from routine interventions in real world practice. They can be a valuable source of evidence for researchers, as well as practitioners, due to their capability to collect vast data over a long time. Depending on the scale of a registry, information about demographic sub-populations and rare events can be detected and risk can be assessed accurately.

One example of a large-scale clinical registry is the SWEDEHEART registry, which is a nationwide registry in Sweden that collects data on all patients with acute coronary syndrome and heart failure. It includes information on baseline patient characteristics, treatment details, as well as outcomes. It

has been used to study the effectiveness and safety of various interventions (Bäck et al., 2021, Figtree et al., 2022, Mars et al., 2021).

Another example is the Spine Tango registry, which is a global registry for spine surgery and is analysed in depth in Chapter 3 of this project. The registry collects data on patient characteristics, surgical procedures, and outcomes (EUROSPINE, 2022b).

Evidence from clinical registries can be used in a number of ways to inform clinical practice and research. Not only can registry data be used to identify patterns of treatment effectiveness, but also to identify potential adverse effects. With the large number of patients that are included in routinely collected data, rates of rare events can be assessed with greater accuracy than using data from RCTs, which often that do not include as many patients. The large scale of some registries can also facilitate the development of prediction models that can aid practitioners in their decision making.

In recent years, the use of registries has been increasingly investigated (Lauer and D'Agostino, 2013), but data quality in RCTs is unmatched, due to their randomisation and consistent follow-ups. The National Health Service (NHS) developed a payment model called the best practice tariff, which is a financial incentive system that rewards hospitals for delivering high-quality care according to specific guidelines for various medical conditions and procedures. The aim is to improve patient outcomes and reduce costs by encouraging healthcare providers to adhere to best practice guidelines. This also includes specific requirements for data collection and reporting to the British Spine Registry (BSR), which records patients and surgical data for all spinal procedures in the UK. A study by Habeebullah et al investigates the impact of this tariff by comparing patient data before and after its introduction. The authors found that the introduction significantly increased compliance with the BSR, with the percentage of cases entered into the registry improving from 70% to 97% (Habeebullah et al., 2021).

Registry-based randomised controlled trials (RCTs) are a type of RCT that incorporates a clinical registry (Li et al., 2016). A registry can be used to identify and enrol participants, and to collect data on patient characteristics, interventions, and outcomes. The implementation of an existing registry can reduce time and resources required to identify and enrol participants and therefore potentially lead to a more efficient trial in terms of outcome collection and costs (Dombkowski et al., 2014, Rao et al., 2014). Registry-based RCTs can also include a diverse group of patients who receive routine care in real-world settings. This can improve the generalisability of trial results to real-world settings. However, the registry data may not be as detailed as data collected just for the trial. Overall, registry-based RCTs are a promising approach to clinical trial design, and have the potential to provide valuable insight about interventions in real-world populations. However, it is important to carefully consider

the strengths and limitations of this approach, and to ensure that appropriate methods are used to minimize bias (Karanatsios et al., 2020).

A previous systematic review by van Hooff et al. (Hooff et al., 2015) focussed on spinal disorder registries and their impact on routine care. 25 spine registries were identified, representing 14 countries. However, it concluded that there was a lack of evidence that registries have significantly improved the quality of spine care. The purpose of this chapter is to review observational studies and randomized controlled trials (RCTs) regarding surgical interventions for disc herniations. The review will analyse various aspects of the studies, such as reported missing data, collected outcomes, study duration, sample size, and the use of registries. This analysis will offer an overview of how many studies, both observational and investigational, incorporate registries and the manner in which they are employed. By identifying the different ways in which registries are utilized, the chapter seeks to uncover any untapped potential. Moreover, successful RCTs that employed registries to enhance trial conduct will be examined to identify best practices for implementing registries in clinical research and to comprehend the potential advantages of this approach.

2.3 Review – Methods

.This chapter does not aim to provide the most comprehensive literature review possible, but instead aims to give an overview of the RCTs and observational studies conducted in this patient population, particularly focusing on:

a) missing data,
b) collected outcomes,
c) study length,
d) number of patients, and
e) the use of registries.

Therefore, the review was limited to the PubMed library Central® (PMC) of the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM) and focused on completed studies (excluding ongoing studies or trials). This literature review was conducted on the 22nd of July 2022. The following inclusion and exclusion criteria were applied.

2.3.1 Inclusion Criteria

- Patient Population: Patients diagnosed with sciatica resulting from lumbar disc herniation.
- Medical Procedure: Microdiscectomy for the treatment of lumbar disc herniation.

- Study Types: Two types of studies will be included:
    - Randomized Controlled Trials (RCTs)
    - Observational Studies (cohort studies, case-control studies, cross-sectional studies).

2.3.2 Exclusion Criteria

- Not Primarily Focused on Surgical Outcome: Publications that do not primarily investigate the surgical outcome, but rather focus on analgesics, anaesthetics, inflammation, diagnostic tools, pre-surgery, during-surgery, or post-surgery medication, etc.
- Not in Humans: Studies conducted on non-human subjects.
- Non-Surgical Interventions: Publications that investigate non-surgical interventions.
- Other surgical procedures: Studies that included patients with other procedures such as spinal fusions.
- Other medical indication: Studies that do not include patients with lumbar disc herniations.
- Other spinal regions: Studies that were focussed on other spinal regions (cervical, thoracic).
- Non-English Publications: Studies published in languages other than English.
- Wrong Publication Type: Literature reviews, meta-analyses, secondary analyses or protocols.
- Specific Subset of Patients: Studies specifically targeted to a subset of patients, such as those who underwent failed surgery, recurrent surgery, or amputees, will be excluded. The focus is on the general patient population.

In order to specify a search with multiple terms, they can be connected with logical “AND” and “OR” operators. The first step was to identify MesH and key terms that describe the patient population, which are listed in Table 2.1. A MeSH term (Medical Subject Heading) is a standardized and controlled vocabulary used by the National Library of Medicine (NLM) to categorize and index biomedical literature. It consists of specific terms or phrases that represent various medical concepts, conditions, treatments, and other relevant topics. MeSH terms are assigned to scientific articles and other resources to facilitate more efficient and accurate searching in databases like PubMed.

“disc herniation*”[tw]
“disk herniation*”[tw]
“herniated disc*”[tw]
“herniated disk*”[tw]
“Intervertebral Disc Displacement“[Mesh]
“slipped lumbar disc*”[tw]

“Sciatica”[Mesh]

Table 2.1: MesH and key terms describing the patient population.

By using the * symbol in the search term, the word stem can be looked for instead of listing all possible word endings. These were then connected with the OR operator and saved as search terms (1). To specify the type of medical intervention that this literature was focussed on, the MesH and key terms in Table 2.2 were used.

“Discectom*”[tw]
“Diskectom*”[tw]
“Diskectomy"[Mesh]
“Microdiscectom*”[tw]
“Microdiskectom*”[tw]
”Sciatica/surgery"[Mesh]
“Intervertebral Disc Displacement/surgery“[Mesh]
“spinal surger*”[tw]
“Spine/surgery“ [Mesh]
“Decompression, Surgical“[Mesh]

Table 2.2: MesH and key terms describing the medical intervention.

Again, these were then connected with the OR operator and saved as search terms (2).

To specify the research publication type for observational studies the MesH and key terms listed in Table 2.3 were used and connected with the OR operator. This category of search terms was not further expanded to terms such as “prospective”, since that would include publications that would not necessarily be observational studies.

“Observational Study“[Publication Type]
“observational”[tw]
“retrospective*”[tw]
“Retrospective Studies“[Mesh]

Table 2.3: MesH and key terms describing the publication type of retrospective observational.

These were then saved as search terms (3).

To specify the research publication type for RCTs the MesH and key terms listed in Table 2.4 were used and connected with the OR operator.

RCT[tw]

RCTs[tw]
“randomized controlled trial*”[tw]
"Randomized Controlled Trial"[Publication Type]
“Randomised Controlled Trial*”[tw]
“Clinical Trial*”[tw]

Table 2.4: MesH and key terms describing the publication type of RCTs.

These were then saved as search terms (4).

Afterwards, two searches were conducting by connecting (1), (2) and (3) for observational studies and (1), (2) and (4) for RCTs with “AND” operators.

After applying these search terms, the literature review was conducted in two rounds. The first round involved screening titles and abstracts, and the second round entailed a comprehensive reading of the remaining studies.

2.3.3 Summary techniques

In summarizing the amount of missing data across the studies (missing data defined as the number of patients that had to be excluded in the final analysis of the study), three techniques were employed. Firstly, the mean and standard deviation of the reported missing data were calculated for each study, without considering their sample sizes, ensuring equal weight for each mean value. Secondly, the weighted mean and standard deviation, accounting for the sample sizes, were determined to estimate the average missing data at the population level. Furthermore, missing data was categorised in the following intervals: 0-10%, 10-20%, 20-30%, 30-40% and 40-50%. When a study precisely reported 10% missing data, it was appropriately placed within the 0-10% category, and the same approach was taken for other specific percentages. Notably, no study included in the analysis had more than 50% missing data. Missingness was defined as the number of patients excluded from the study due to their lack of data in essential patient covariates or outcomes, such as instances of loss to follow-up. If indicated, missing data was measured for the primary outcome timepoint.

In order to summarise the sample sizes of the studies, they were grouped into distinct intervals: less than or equal to 50, 51 to 100, 101 to 200, 201 to 500, 501 to 1,000, 1,001 to 5,000, and more than 5,000 patients. An overall mean and standard deviation of sample size were also provided for the two included publication types.

To summarize the duration of the studies, they were categorized into specific intervals: less than or equal to six months, six months to one year, one to two years, two to five years, and more than five

years. Hereby, a study that reported a follow-up period of exactly one year, would be categorised into the “six months to one year” category. The same approach was taken for other specific follow-up periods. Additionally, an overall mean and standard deviation of study length were calculated for the two types of publications included.

In order to obtain a comprehensive overview of the outcomes gathered in the studies, we conducted a count of publications measuring each specific outcome. As studies commonly include multiple outcomes, it is important to note that this is not a simple one-to-one counting process. Instead, we accounted for the total number of publications for each outcome, and these results are visually presented through bar plots, providing a clear and informative representation of the overall findings.

2.4 Review – Results

The results of this review will be presented descriptively, providing a PRISMA-flow diagram of the two review rounds and categorised exclusion.

2.4.1 Observational studies

With the previously established combination of search terms (1), (2) and (3) 246 observational studies were identified regarding surgery due to disc herniations. After an initial screening of titles and abstracts 80 were excluded. At second screening at which papers were read in depth, a further 77 were excluded. Figure 2.1 shows a diagram of the exclusion of papers in the review rounds.

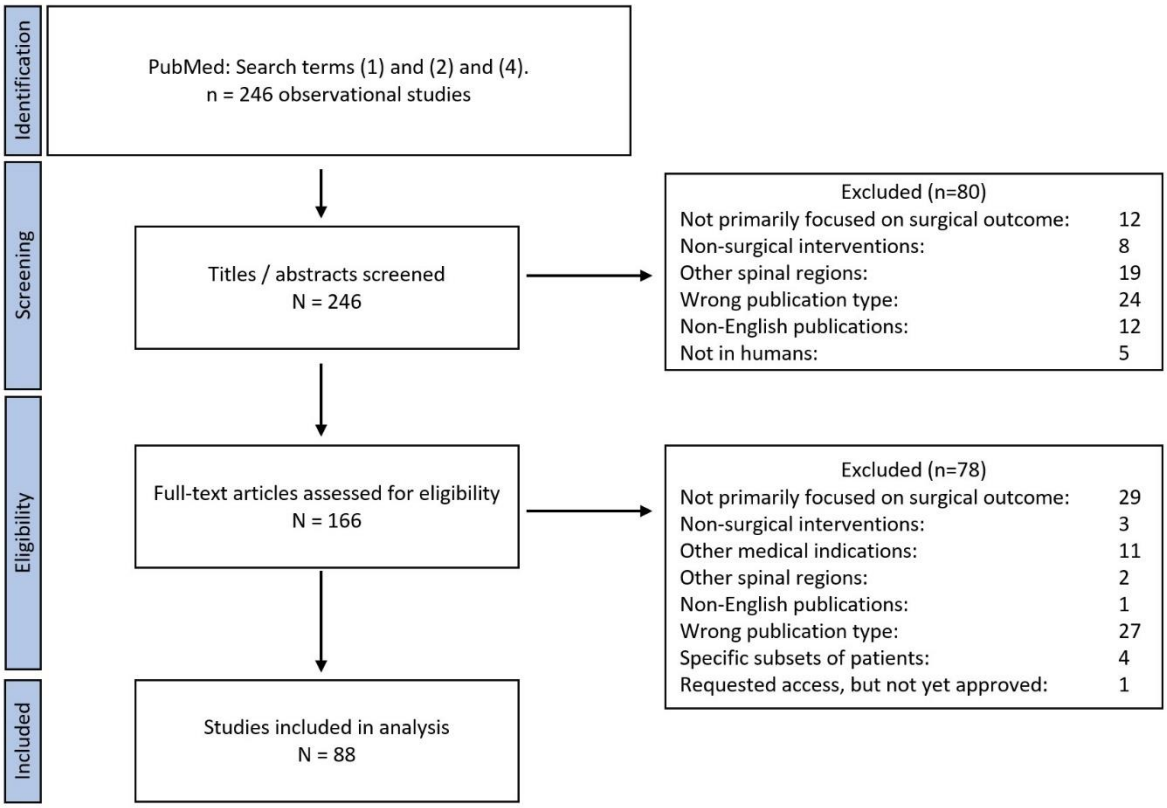


Figure 2.1: PRISMA flow diagram for exclusion of observational studies in review rounds.

The category of "Not primarily focussed on surgical outcome" included: drugs after surgery, anaesthetics, risk of site infections, depression/psychology, post-surgery treatment, extubating techniques during surgery, discharge time, impact of an HD videodisk program on patient satisfaction, treatment preferences, inflammation profile of herniated discs, degree of macrophage infiltration on disc material, costs due to fluid leaks and more. The category of "Specific subsets of patients" included publications that focussed on subsets of patients such as amputees, failed surgery and re-herniation.

The full text of 1 article could not be accessed. The first author of this paper was contacted to request access to the full text, but there was no response to date. This led to a total of 88 articles included in this review. A list of all included publications is listed in Appendix A.

2.4.2 RCTs

With the previously established combination of search terms (1), (2) and (3) 542 RCTs were identified regarding surgery due to disc herniations. After an initial screening of titles and abstracts 423 were excluded. At second screening at which papers were read in depth, a further 43 were excluded. Figure 2.2 shows a diagram of the exclusion of papers in the review rounds.

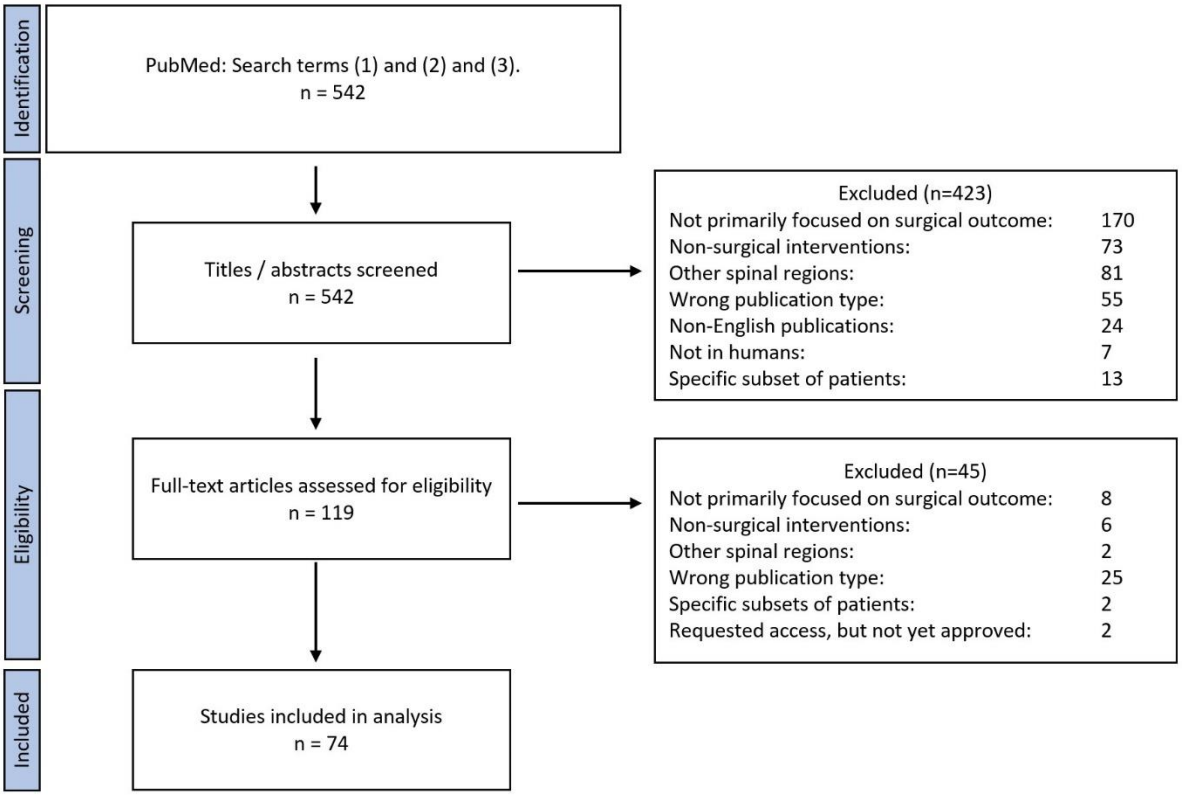


Figure 2.2: PRISMA flow diagram for exclusion of RCT publications in review rounds.

The category “Not about surgery effect” included publications that focussed on skin incisions, bladder needs, drain usage, scarring prevention, risk of site infections, psychology, MRI imaging, discharge time, pre-operative information-video effect on patient decision, bed rest decision or the effect of showing patients removed disc material. The full text of 2 of the publications could not be accessed. The authors of these papers were contact to request access to the full text, but there was no response to date. This led to a total of 74 articles in this review. A list of all included publications is listed in the Appendix B.

In the following, these RCTs are analysed regarding missing data, number of patients, length of study, outcomes collected and the use of any registry.

2.5 Review – Comparison of observational studies and RCTs

In this section, the included randomised controlled trials (RCTs) and observational studies will be compared with respect to the amount of missing data, the number of patients enrolled in the study, the length of follow-up, and the outcome measures that were collected. In **Error! Reference source not found.**, missing data percentages are compared.

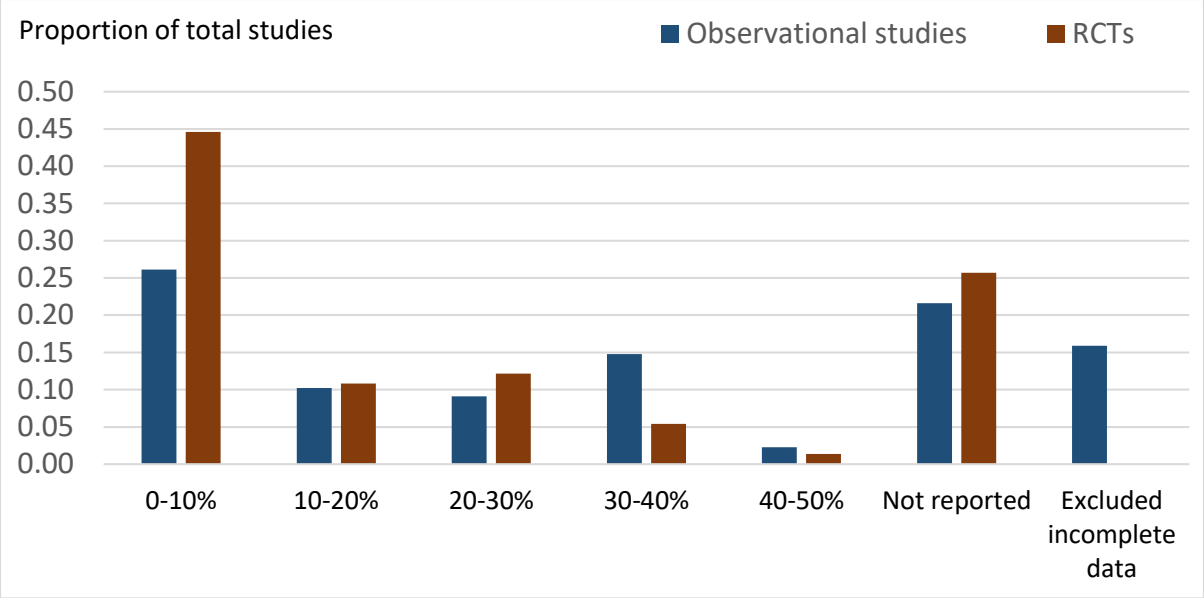


Figure 2.3: Bar-plot of proportion of studies / trials in each category of missing data. Both data series have been normalised by the number of total studies in the series.

As expected, a larger percentage of RCTs have low missing data than observational studies. In RCTs, missing data may occur due to participants dropping out of the study or not completing all study visits, while in observational trials, missing data may occur due to incomplete or missing data on participant characteristics or exposures. Clearly defined study protocols lead to more consistent data collection. However, there is a substantial number of publications in both RCTs and observational studies, that did not report missing data. Of all observational studies that reported missing data and did not specifically state that the analysis is by design on a complete data set, the mean and standard deviation of missing data was 12.95% (s.d. 15.82%). The weighted mean and standard deviation (using the "Mmisc" R-package), weighted with the number of patients normalised so that the sum of weights equals one, was 19.14% (s.d. 14.90%).

Figure 2.4 visualises percentages of studies (normalised to either total number of RCTs or observational studies) in each category of number of patients that were included.

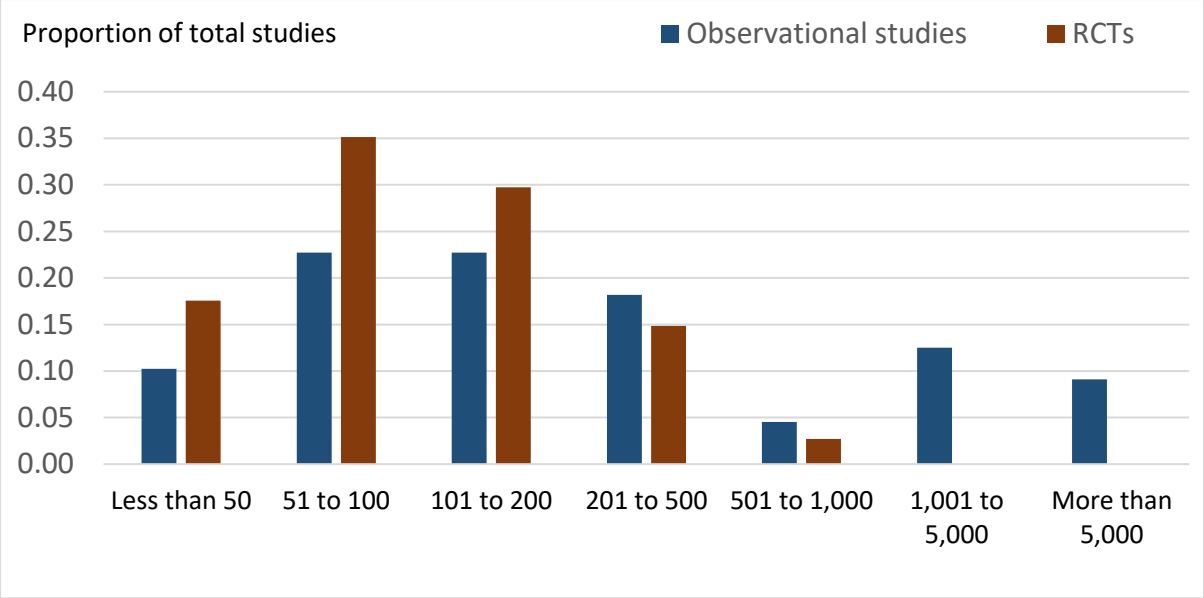


Figure 2.4: Bar-plot of proportion of studies / trials in each category of number of patients. Both data series have been normalised by the number of total studies in the series.

Only observational studies had more than 1,000 patients. It appears that sample size of patients in RCTs are smaller than in observational studies. A possible reason is that observational studies do not have the same strict eligibility criteria that RCTs have. In an RCT, participants must meet certain criteria to be eligible to participate, such as having a specific condition or being of a certain age. This is done to ensure that the results of the trial are as accurate as possible, by minimizing any potential confounding factors. On the other hand, observational studies often include a more diverse group of participants, and do not have the same strict eligibility criteria. This means that they may be able to include a larger number of participants, as it is easier to recruit people for the study. The mean and standard deviation of the sample size of observational studies and RCTs was 278.39 (s.d. 4,610.13) and 105.52 (s.d. 129.82) patients, respectively. The large sample size of some of the studies caused the large variation in sample sized in observational studies.

In Figure 2.5, percentages of studies (normalised to either total number of RCTs or observational studies) for each category of length of follow-up are visualised.

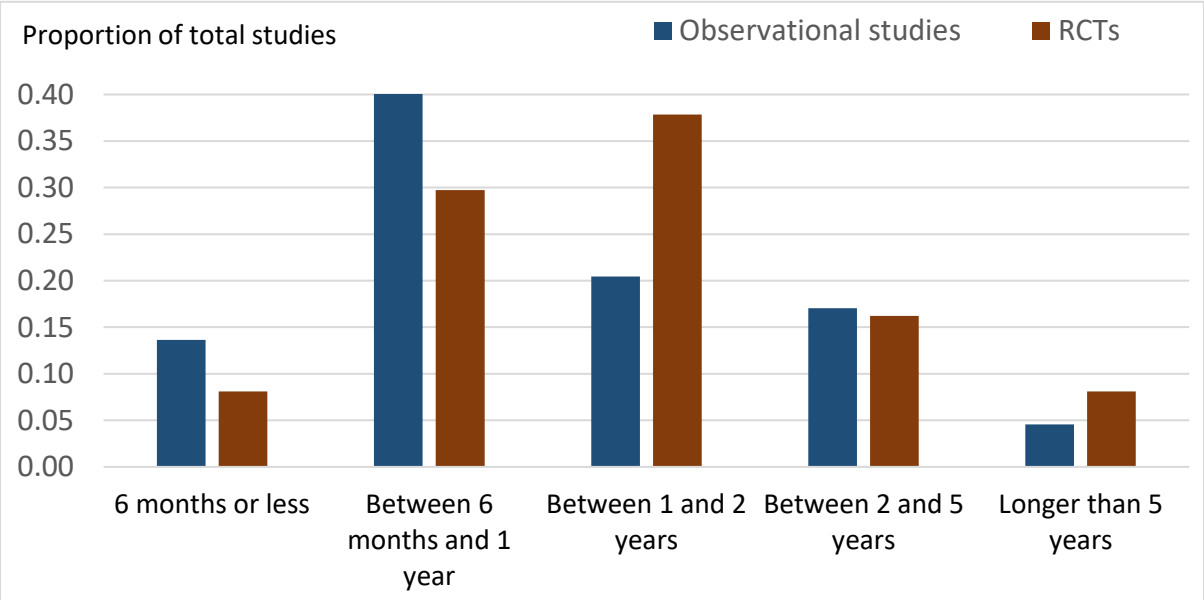


Figure 2.5: Bar-plot of proportion of studies / trials in each category of length of follow-up. Both data series have been normalised by the number of total studies in the series.

The two data series look similar. There appears to be more observational studies than RCTs with follow-up between 6 months and 1 year, and more RCTs than observational studies with follow-up between 1 and 2 years. However, there does not seem to be a systematic difference in length of follow-up over which either study type is conducted. The mean and standard deviation of the length of observational studies and RCTs was 1.25 (s.d. 2.40) and 1.70 (s.d. 3.41) years, respectively.

To get an overview of commonly used outcome measures and identify the most used ones, primary and secondary outcomes of each study were analysed. Since most studies collected more than one outcome, the sum of outcomes is higher than the sum of publications. In Figure 2.6, outcomes that were collected in both study types are visualised.

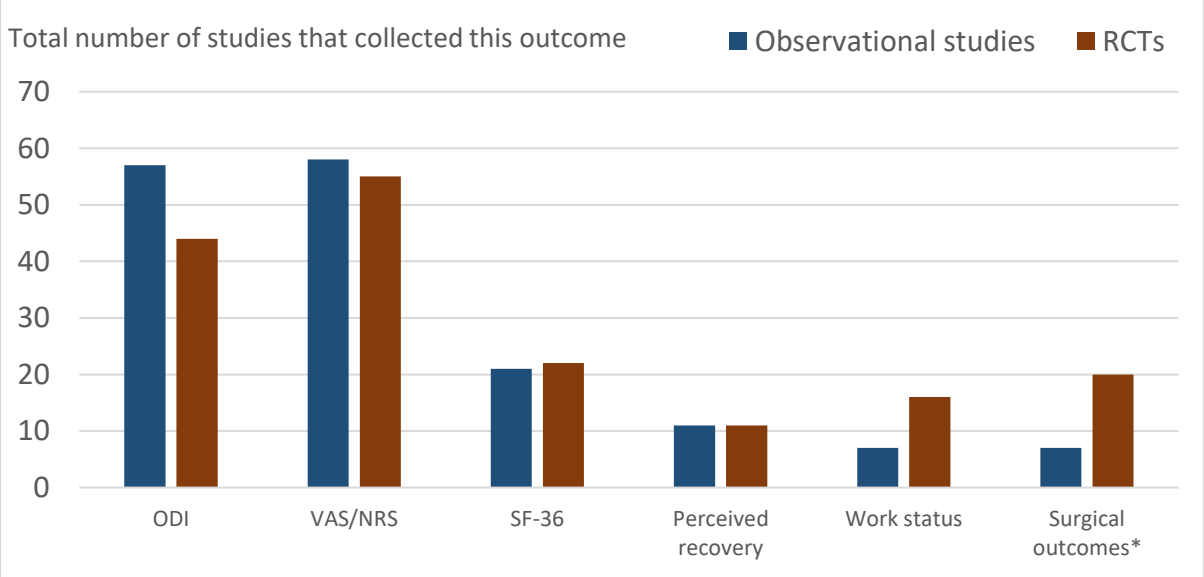


Figure 2.6: Bar-plot of total number of publications that collected each outcome. *Surgical outcomes include: neurological examinations, fusion rates, MRI, radiographic scans.

The main outcomes are similarly often collected by both RCTs and observational studies. However, there were a few outcomes that were only collected by one publication type. Some RCTs collected hospital stay, the Roland Morris Disability questionnaire or general examinations, whereas observational studies did not. Conversely, some observational studies collected the EQ-5D quality-of-life questionnaire, reoperation rates or the Macnab criteria. The Macnab criteria is a classification of global outcome of spinal interventions into the categories: excellent, good, fair, and poor (Ahn et al., 2018).

The main differences between RCTs and observational studies were the occurrence of missing data and number of included patients. Outcomes that are collected and follow-up times are mostly the same between the two publication types, which is why there is a lot of potential for observational studies to complement RCTs and deliver comparable data sets. These can be used to gather additional information of real-world practice and can potentially (due to larger sample size) lead to more accurate estimates, for

The most used outcomes for surgical interventions due to herniated lumbar discs in observational studies were ODI and visual analogue / numerical pain rating scales (VAS / NRS), followed by the SF-36 health related quality-of-life questionnaire.

Of the 88 included observational studies, 22 mentioned the use of a registry or a local database. The public registries used were: NORSpine (9 publications), SweSpine (5 publications), Spine Tango (2 publications), Danish spine register (1 publication), SWISSspine (1 publication) and others (NSN

database, including the CSORN registry, German Spine register, local clinical databases, Ohio Bureau of Workers Compensation database, BWC database, Quality and Outcomes Database QOD and local clinical databases. Registries were used for either data collection or patient monitoring.

The most used outcomes for surgical interventions due to herniated lumbar discs in RCTs were visual analogue and numerical pain scores and the ODI and SF-36 health related quality-of-life questionnaires.

Only 1 of the 74 included RCTs mentioned the use of a registry (Swespine). Three further trials used local databases for patient randomisation and data collection and storage. One trial used observational data from a not further specified database to complement their findings.

2.6 Lack of use of registries

Registries can potentially be used in both randomised controlled trials (RCTs) and observational studies. In RCTs, registries can be used to identify and enrol eligible participants, track their progress through the study, and collect data on outcomes. In observational studies, registries can be used to identify and follow a cohort of individuals over time and collect data on exposures, outcomes, and other variables of interest. Registries can be a valuable resource for researchers and healthcare professionals to gain insights into routine practice. For observational studies, there are multiple examples of where data collection of real-world evidence was obtained by clinical registries in similar fields such as spinal stenosis or recurrent surgeries (Fritzell et al., 2015, Möller et al., 2022, Sigmundsson et al., 2013). Such examples can motivate more such research to be done in the field of sciatica related microdiscectomies.

However, only one RCT used a spine registry (Swespine) for their study in the sciatica-affected patient population that had surgery due to a herniated disc. 22 observational studies mentioned the use of a registry. The most commonly used registries were NORSpine, SweSpine and Spine Tango. They were used for data collection and patient monitoring.

Registries can be expensive to set up and maintain. Therefore, studies do often not have the funding to set up a new registry for their specific purposes. However, the outcomes that were collected in many observational studies and RCTs are routinely collected on some of the existing registries. The infrastructure for data collection by using a registry exists. For studies that are focussed on outcomes that are routinely collected, it can be of financial benefit to conduct the study on a registry instead of developing a new data collection tool. In some cases, this method could even reduce the cost significantly and enable studies that were otherwise not possible. The potential of the use of registries can be seen in studies in other clinical areas. Payment schemes such as the spinal best practice tariff

on compliance with the British Spine Registry can be an effective way to improve not only the consistency of data collection via registries, but also the quality of the observational data (Habeebullah et al., 2021).

2.7 Registry-based RCTs – Potential

Registry-based randomised controlled trials (RCTs) are a modern trial design that combines the benefits of large-scale registries with treatment randomization. Registries can be used for various aspects of a trial, such as identifying patients, randomizing treatments, and collecting data. This approach allows for the enrolment of larger patient populations and long-term follow-up at relatively low cost. A registry can be used for various elements of a prospective trial, including identifying eligible patients, obtaining patient consent, randomizing treatments, collecting baseline data, and detecting and adjudicating clinical endpoints. By incorporating randomization into a clinical registry, some of the critical attributes of a prospective RCT can be combined with the practical features of a large-scale registry, including the benefits of consecutive enrolment and automated patient identification and follow-up. Registry-based RCTs are well-suited for open-label evaluations of commonly used therapeutic alternatives in settings with existing registries, but may not be appropriate for trials that require strict definitions of endpoints or comprehensive safety reporting. However, registry-based RCTs can still be useful for evaluating new indications for pharmaceutical agents and can offer benefits such as the ability to identify and enrol a larger proportion of patients and conduct long-term follow-up at low cost. An important advantage of registry-based RCTs is the ability to describe and follow up the entire reference population, including eligible non-randomised patients and non-eligible individuals. (James et al., 2015).

2.7.1 The TASTE-trial

The TASTE trial was a randomised controlled clinical trial that was conducted in Sweden (Lagerqvist et al., 2014). The trial compared two treatments for acute myocardial infarction, also known as a heart attack, and included a total of 7,244 patients. The SWEDEHEART (Swedish Web System for Enhancement and Development of Evidence-based Care in Heart Disease Evaluated According to Recommended Therapies) registry was used for the identification of patients, randomisation and collection of baseline and follow-up variables. The TASTE trial recruited a significant number of patients with ST-segment elevation myocardial infarction (STEMI) who were planning to undergo percutaneous coronary intervention and were able to provide oral informed consent. This means that the trial was representative of the overall population of STEMI patients in the region who undergo percutaneous coronary intervention. Hospitalizations for myocardial infarction are recorded with a

high level of accuracy. The use of personal identification numbers is required and helps to ensure that death registries in the Nordic countries are also complete, although it is not possible to distinguish between cardiac and non-cardiac causes of death. The use of automated personalized identification numbers in Sweden allowed the researchers to track all of the patients and ensure that none were lost during the study. The accuracy of the source data was checked against electronic health records and found to be in agreement 95% of the time (James et al., 2015). The cost of conducting a trial using the existing registry and willing investigators who provided their services for minimal pay was significantly lower than the cost of a traditional trial of the same size. Specifically, the cost of establishing and running the SCAAR/SWEDEHEART registries was approximately $400,000, while a traditional trial of the same size would have cost tens of millions of dollars. One potential limitation of the TASTE trial is that the outcomes were based on registry data rather than being systematically evaluated. This could potentially lead to less accurate results compared to a traditional randomised trial (James et al., 2015, Lagerqvist et al., 2014).

2.7.2 The DETO$_2$X-AMI trial

Another trial that integrated the SWEDEHEART registry into the study design was the DETO$_2$X-AMI trial. It tested the use of supplemental oxygen compared to normal air in patients with heart attack symptoms or a confirmed heart attack. The study included 6,600 patients who were randomly assigned to receive either supplemental oxygen or normal air. The study used the SWEDEHEART and other public registries for outcome collection. However, the study design had limitations, such as being an open-label design (no blinding), which can introduce bias. The results of the study found that using supplemental oxygen did not reduce the risk of death within one year (Hofmann et al., 2017, James et al., 2015).

2.7.3 The SORT OUT trials

In the SORT OUT trials, patients were randomly assigned to receive a stent using either a postal or interactive voice system. National registries were used to identify clinical events such as death, heart attacks, and revascularization. This trial design allowed for the systematic detection of clinical events in a real-world setting, without requiring additional patient visits (James et al., 2015, Thuesen et al., 2013).

2.7.4 The SAFE-PCI for Women trial in the USA

The SAFE-PCI for Women trial in the US used a registry-based trial methodology, which involved incorporating a randomised trial into the existing cardiovascular research infrastructure of the NIH National Cardiovascular Data Registry's CathPCI Registry. This trial was designed to compare radial and femoral artery access in women undergoing PCI, with the primary efficacy endpoint being a composite

of bleeding or vascular complications requiring intervention. The registry-based trial design had two main advantages: it enabled the identification of operators and sites that could include patients with a balanced risk of complications from both radial and femoral approaches, and it reduced the workload for site coordinators by about 65% per patient compared to traditional study forms. While the trial was successful, the cost savings were not as significant as in the TASTE trial due to the lack of full integration of the registry into clinical care (Hess et al., 2013, James et al., 2015, Moussa et al., 2013).

2.8 Summary

The comparison between RCTs and observational studies showed several key differences and similarities. RCTs tend to have lower missing data rates due to rigorous protocols, whereas missing data in observational studies is often linked to incomplete participant information. Study protocols also influence consistent data collection. Notably, both RCTs and observational studies sometimes omit reporting missing data. Observational studies tend to involve larger sample sizes, possibly due to their less strict eligibility criteria, compared to RCTs. The length of follow-up shows some variation, with more observational studies in the 6 months to 1-year range, and more RCTs in the 1 to 2-year range, but no systematic difference is evident.

Primary and secondary outcomes were analysed across studies, showing similarity in outcomes collected by both RCTs and observational studies. However, some outcomes were unique to either type. For instance, RCTs collected hospital stay and certain questionnaires, while observational studies gathered data on quality-of-life assessments and reoperation rates. The data series demonstrate similarity between the study types, indicating the potential for observational studies to supplement RCTs and provide valuable real-world insights.

Common outcome measures in both observational studies and RCTs are the ODI, numerical or visual pain rating scales, as well as the SF-36 questionnaire. Registries, however, see limited adoption in both RCTs (only 1 out of 74) and observational studies (22 out of 88). When employed, they were utilised for participant identification, data collection, and monitoring.

Registries, while beneficial, can be costly to establish and maintain. However, many outcomes collected in studies are already part of existing registries, suggesting financial benefits and increased feasibility for studies focused on routinely collected outcomes. Such use of registries has been successful in other clinical areas, such as the TASTE-trial, the DETO2X-AMI trial, the SORT OUT trials or

the SAFE-PCI for women trials. Additionally, Payment schemes tied to registry compliance can enhance data consistency and quality.

2.9 Discussion

The motivation behind this chapter was to analyse observational studies and randomized controlled trials (RCTs) within the sciatica-affected patient population that underwent microdiscectomy. Main aspects of the comparison of the two publication types were missing data, collected outcomes, study length, sample size and the use of a registry. The aim hereby was twofold: first, to assess the utilization of registries in these studies and, second, to examine whether there existed alignment in methods and collected outcomes between RCTs and observational studies. The identification of such alignment could potentially unlock the vast potential of routinely collected data. However, despite the advantages, there are several limitations and essential prerequisites for the effective integration of registries in RCTs.

To establish a registry for RCTs, it is crucial to incorporate standardized core measurement sets specific to the studied patient population, enabling uniform data collection across trials and registries. These registries should integrate with various data sources, including electronic health records and administrative databases, to enhance data accuracy and comprehensiveness. Comprehensive patient data collection is essential when cross-referencing is not feasible. Registries should possess the capability to track participants over time and collect follow-up data for evaluating long-term intervention outcomes. Collaboration with trial sponsors and adherence to study protocols, along with the potential for randomization techniques, enhance registry utility. Data comparability ensures merging of information from various sources, while robust data security measures are crucial due to the technical demands of clinical registries for RCTs.

The results of the literature review reveal alignment between the two publication types, particularly concerning the gathered outcomes. Despite the absence of a defined core outcome set for this patient population, key outcomes, including ODI, visual or analogue pain scales, and the SF-36, are collected in both publication types. Remarkably, registries used in certain observational studies routinely collect this data, supporting the assumption of under-utilisation of registries in RCTs. Noteworthy examples such as the TASTE-trial, the DETO2X-AMI trial, the SORT OUT trials, and the SAFE-PCI for women trials highlight how integrating routinely collected data infrastructure can significantly benefit RCTs.

In the subsequent chapter, a more detailed analysis will be conducted on data from a recent RCT and a registry (Spine Tango), including descriptive statistics of collected patient covariates, missing data,

and prognostic factors. Discovering further similarities between these two data sources would strengthen the recommendation for integrating registry data into RCTs whenever possible.

Chapter 3: Registry data vs RCTs: Insights from the Spine Tango registry and the NERVES trial

3.1 Chapter Outline

The purpose of this chapter is to compare the routinely collected data from the Spine Tango registry and the NERVES trial, looking at factors such as missing data and collected outcomes. This chapter also involves identifying correlations between patient characteristics and conducting a full descriptive analysis of both data sets. The aim was to investigate if the data sources are comparable in terms of baseline patient characteristics and collected outcomes. Discrepancies between the two sets would suggest that registry data potentially collects outcomes not measured in RCTs. These differences might then suggest that insights might either be more biased since the collected data is not randomized or that there are differences in the patient population. If the two data sources are very similar, this would support the use of registry data in the use of clinical trials.

3.2 Introduction

RCTs are considered the gold standard for evaluating the effectiveness of a medical treatment. In an RCT, participants are randomly assigned to a treatment arm. This helps to control for potential confounding factors and ensures that bias is minimised. Routinely collected registry data is data that is collected as part of routine clinical care, rather than for the purpose of a specific research study. One key difference between RCTs and registry data is the level of control over the data collection process. In an RCT, the data collection process is carefully controlled by research protocols. In contrast, the data collected through registries is observational and could therefore be affected by specific treatment preferences of patients and clinicians. This can lead to differences in the quality of the data, with RCTs generally considered to have higher-quality data (Collins et al., 2020).

Another difference between RCTs and registry data is the type of information that is collected. RCTs are typically designed to answer specific research questions. As a result, they focus on the variables that are directly relevant to the research question. In contrast, registry data tends to include a broader range of information about the patient, including both clinical and demographic data (Collins et al., 2016, Grootendorst et al., 2010)

Although some registries have built-in randomization tools and can be used to collect and analyse RCT data such as the SWEDEHEART registry (Jernberg et al., 2010), data from registries is typically observational and therefore of structural difference compared to data from RCTs. Entries are often recorded by several research sites, that can differ in various characteristics such as financial budget, therapy preferences etc., which is why registry data is considered to be more biased (Concato et al., 2010).

Nevertheless, the collection of data in registries can lead to additional and crucial insights for healthcare, since the vast amounts of data in clinical registries can detect rare adverse events or benefits and identify patterns of outcomes for subsets of patients (Benson and Hartz, 2000). Most RCTs are conducted over a specific timeframe and with regard to their individual research question. Strict protocols and surveillance over long follow-up periods require numerous health professionals to be involved, which can result in high costs to conduct such a trial. Additionally, strict patient recruitment regarding inclusion and exclusion criteria and the reluctance regarding randomization might lead to a non-representative sample of the patient population and therefore selection bias of the resulting data.

Ultimately, RCTs are with good reason seen as 'gold-standard' for clinical research, especially if a study aims for the approval of a new intervention. In case of evaluating treatments that are routinely done however, the analysis of vast amount of collected data in registries can complement existing RCT results and add valuable insights.

In terms of comparing the descriptive statistics of RCT data and registry data, it is important to consider the specific measures that are being used. For example, if both sets of data are reporting on the same outcome, such as the rate of hospitalizations, then the descriptive statistics for the two sets of data can be directly compared. However, if the two sets of data are reporting on different outcomes, then it may be more difficult to compare their descriptive statistics.

In this chapter data from the recently conducted nerve root block versus surgery (NERVES) trial and data from the international registry Spine Tango (Wilby et al., 2021) will be summarised. This will be done in the framework of sciatica patients as study population who underwent a microdiscectomy. Access to the Spine Tango data was obtained through the submission of a study protocol to the Spine Tango committee. A data sharing agreement was established between Spine Tango and the University of Liverpool, with the data being provided in a fully anonymized format. As a result, obtaining approval from the ethics committee was not necessary.

One of the main outcomes in both data sources and focus in this chapter is the Core Outcome Measures Index (COMI), which was recorded pre-surgery and at follow-up visits. It is a quality-of-life (QoL) questionnaire based on the items proposed by an expert group for the use in clinical routine, quality management and research (Deyo et al., 1998). It covers not only pain intensity and its effect on quality of life, but also allows patients to report complications, overall satisfaction, and further surgeries.

3.3 The Core Outcome Measures Index (COMI)

There are different versions of the COMI questionnaire and the version that was used in the NERVES trial was different from the one that is routinely collected in Spine Tango. Changes in the design of questionnaires could potentially impact the comparability of summarized scores. However, for simplicity, it is assumed that scores are the same, if the same patient filled out both. This assumption however, has not been examined.

The questionnaire used in the Spine Tango registry can be found in Appendix A: COMI version in the Spine Tango registry. The answers of these questions will be summarised in a score, which is calculated with the following method. From question 2a and 2b, the higher of the values is selected, which indicates the more intense pain. For questions 3-7, the 5 ordered sections will assigned values 0-4, depending on the severity and then transformed to a 0-10 scale, by multiplying with 2.5. To obtain the overall score, the mean over all questions 2-7 will be computed, were the value of question 2 is the higher of the two values of 2a and 2b (EUROSPINE, 2022a).

Mannion et al. showed that COMI showed similar external responsiveness to the common questionnaire SRS-22. “It is well able to detect important change. Coupled with its brevity, which minimizes patient burden, these favourable psychometric properties suggest the COMI-back is a suitable instrument for use in registries and can serve as a valid instrument in clinical studies emerging from such data pools.” (Mannion et al., 2016)

The questionnaire used in the NERVES trial can be found in Appendix B: COMI version in NERVES trial. The first 3 questions were answered on a 5-point scale, whereas the last two questions were answered in total days (0-28) and afterwards categorized into groups (0 days = 1 point, 1-7 days = 2 points, 8-14 days = 3 points, 15-21 days = 4 points, more than 21 days = 5 points). In the trial the score was computed as average over all points of questions 1-5 (scale 1-5), however, this score was re-scaled to a score from 0 to 10, by allocating a point system in the following way: 0 days = 0 points, 1-7 days = 2.5 points, 8-14 days = 5 points, 15-21 days = 7.5 points, more than 22 days = 10 points). The same scaling (0, 2.5, 5, 7.5, 10) is applied to category items with 5 possible answers. Afterwards, a score is

defined as mean over these questions 1-5. For questionnaires that were completed the mean will be over questions 1-7. The statistical analysis plan (SAP) recommended, that a score is only calculated when all items are present.

3.4 Nerve root block versus surgery (NERVES) trial

The NERVES trial was a phase 3, multicentre, open-label, randomised controlled trial that compared surgical microdiscectomy to transforaminal epidural steroid injection in patients with sciatica secondary to herniated lumbar disc. There are various approaches for the treatment of this condition e.g. microdiscectomy, conservative non-surgical treatment or epidural injections, but controversy over the optimal treatment remains. In most comparative clinical studies, the effectiveness of microdiscectomy and non-invasive treatment in form of physiotherapy or analgesics etc. has been investigated. Results of those studies were inconclusive (Atlas et al., 1996, Buttermann, 2004, Osterman et al., 2006, Weinstein et al., 2008). A meta-analysis has shown that the difference of treatment outcomes is not significant enough to establish microdiscectomy as overall superior and therefore there exist no specific healthcare guidelines (Chen et al., 2018). Few former studies directly compared microdiscectomy to epidural steroid injections (ESI) via interlaminar approach and concluded that microdiscectomy has a better effect on pain reduction, but that ESI can often prevent the need for surgery (Buttermann, 2004, Wang et al., 2002).

The NERVES trial focused on the direct comparison of the transforaminal (TFESI) approach versus surgery regarding clinical and cost-effectiveness of these options for management of radicular pain due to herniated lumbar disc (Wilby et al., 2021). The trial was conducted at 11 spinal units across the UK, where eligible patients were aged 16-65 years, had MRI-confirmed non-emergency sciatica with symptom duration between 6 weeks and 12 months, and had leg pain that was not responsive to non-invasive treatment. Patients with prior spinal surgeries at same disc level, serious neurological deficit, known to be pregnant or patients who did not attempt any form of conservative treatment or have contraindication for surgery and/or injection were not included. Patients were randomly allocated to a treatment by an online randomization system that was stratified by centre with random permuted blocks. Primary outcome was the Oswestry Disability Index (ODI) 18 weeks after randomization and all patients who completed a valid questionnaire at baseline were included. Secondary outcome measures included ODI at 30, 42 and 54 weeks, the Core Outcome Measurement Index (COMI), numerical rating scores, and the Modified Roland-Morris questionnaire (MRM). Baseline characteristics included gender, age, weight, height, BMI, number of weeks of symptoms, employment status, inability to work due to sciatica, estimated volume of canal occupied by disc prolapse and level of disc prolapse.

A total of 163 patients enrolled with a total of 80 (49%) assigned to the TFESI group and 83 (51%) to the surgery group. Analysis has been made according to intent-to-treat concept. ODI mean improvement in the TFESI group was 24.52 points (scale 0-100) and 26.74 points for the surgery group (less than 10 points are not considered clinically significant improvement). The pain reduction of the two treatment approaches was therefore similar, but there were four serious adverse events in four participants associated with surgery and none with TFESI. Although ODI was the primary outcome in this study the focus will be on COMI scores, because the number of patients who completed COMI in the Spine Tango registry (11,093) was significantly higher than the patients who completed ODI (3,019).

3.4.1 Descriptive statistics of baseline patient/surgery characteristics

The patient characteristics that were collected in the trial included: sex, age, estimated volume of canal occupied by disc prolapse, treatment, BMI, duration of symptoms in weeks and level of spine. A summary of these variables is shown in Table 3.1.

Characteristic		N=163
Sex	Female	86 (52.76%)
	Male	77 (47.24%)
	Missing	0 (0%)
Age		Mean 42.83 (s.d. 9.28)
	Missing	0 (0%)
Volume of Canal	Less than 25%	87 (53.37%)
	Between 25% and 50%	70 (42.94%)
	Greater than 50%	6 (3.68%)
	Missing	0 (0%)
Allocated Treatment	Surgical microdiscectomy	83 (50.92%)
	TFESI	80 (49.08%)
	Missing	0 (0%)
BMI		27.72 (s.d. 5.86)
	Missing	21 (12.88%)
Weeks of Symptoms		Mean 41.83 (s.d. 10.91)
	Missing	0 (0%)
Level of Spine	L5 / S1	92 (56.44%)
	L4 / L5	52 (31.90%)
	L3 / L4	3 (1.84%)

Other	1 (0.61%)
Missing	15 (9.20%)

Table 3.1: Descriptive statistics of patient characteristics that were collected in the NERVES trial. Data is complete unless otherwise indicated.

For further details, visualisations of the distribution of the continuous variables were generated as depicted in Figure 3.1**Error! Reference source not found.** – 3.3**Error! Reference source not found.**. This will allow for later comparison with the Spine Tango patient population to determine if the two data sources contain comparable patient population data.

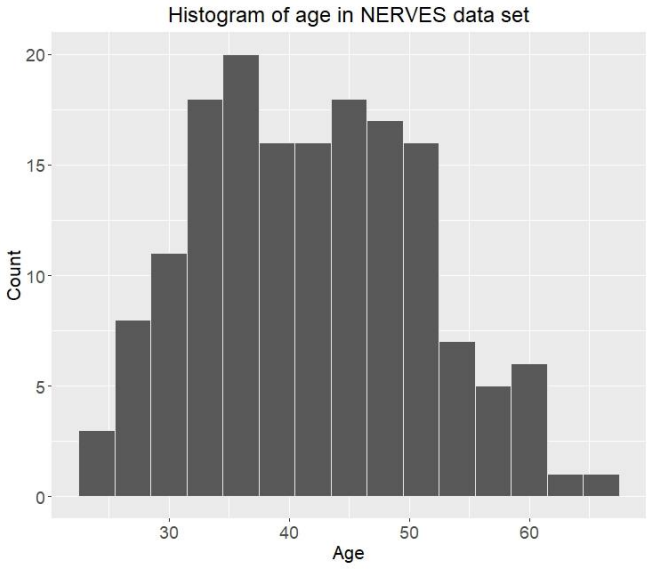


Figure 3.1: Histogram of age in NERVES patient population

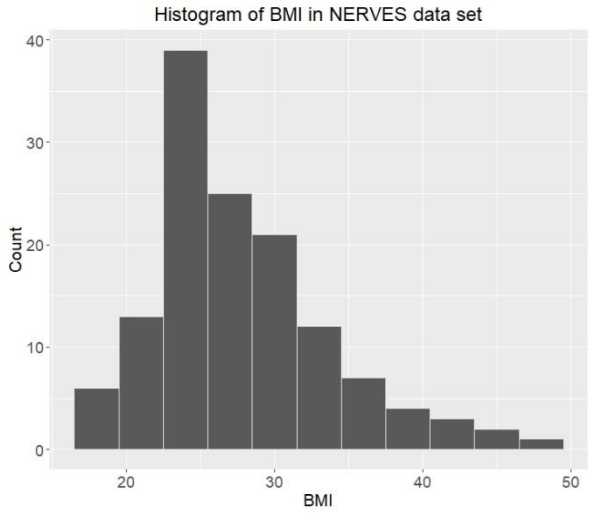


Figure 3.2: Histogram of BMI in NERVES patient population

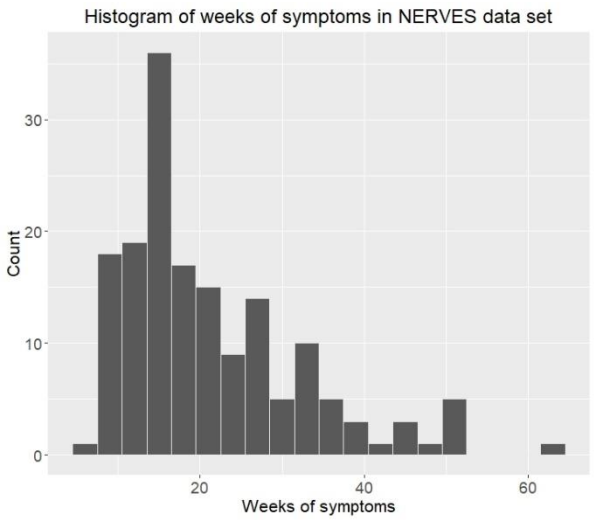


Figure 3.3: Histogram of weeks of symptoms in NERVES patient population

3.4.2 Dependencies between baseline variables

In analysing pairs of continuous variables, the Pearson-correlations were computed using the `cor()` –function in R, along with a scatter plot. For a combination of a continuous and categorical variable, medians and ranges were analysed using boxplots. Additionally, two-sample Kolmogorov-Smirnov tests were used to compare the underlying distribution of the two samples, with a p-value below 0.05 indicating that they are likely from different distributions. For analysing relationships between two categorical variables, grouped bar plots were used to display the frequencies of the variables, and chi-square tests were used to quantify dependency.

The issue of multiplicity was acknowledged and considered. Multiplicity refers to the potential increase in Type 1 error rate when multiple statistical tests are conducted on a single dataset, leading to a higher likelihood of false positive findings. While recognizing the importance of adjusting for multiplicity to mitigate this risk, the decision was made not to make explicit adjustments in this study for the following reasons. This is due to the exploratory nature of this analysis, where the goal was to identify potential relationships or dependencies between variables that might warrant further investigation. This exploratory analysis was rather meant to be hypothesis-generating than hypothesis-testing.

3.4.2.1 Pairs of continuous patient covariates

In this section, the relationship between continuous variables will be analysed using scatter plots and the `cor()` –function in R. The continuous variables in this data set are age, BMI and weeks of symptoms, which results in 3 distinct pairs. The aim of analysing the correlations between these variables is to gain a better understanding of the relationships between them and identify any patterns

or trends that may exist. Scatter plots of the three combinations of age, BMI and weeks of symptoms are shown in Figure 3.4.

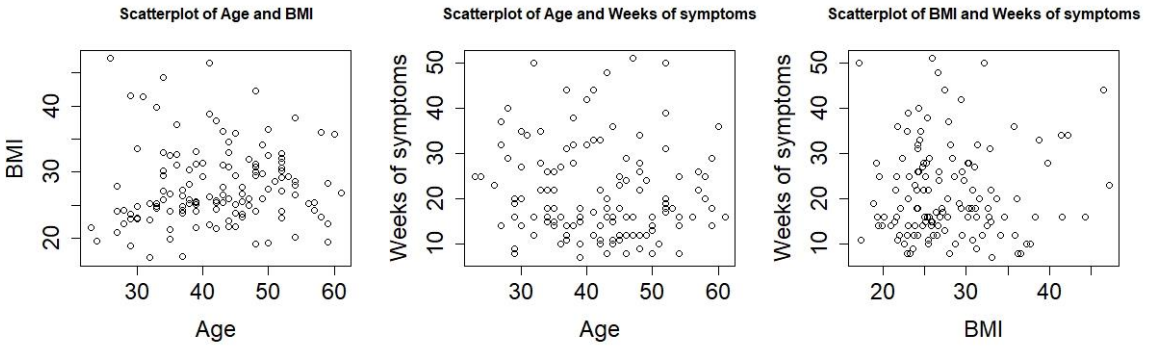


Figure 3.4: Scatter plots of combinations of BMI, age and weeks of symptoms.

The correlation between age vs BMI, age vs weeks of symptoms and BMI vs weeks of symptoms were 0.068, -0.11 and 0.04 respectively, which indicates that correlation is very low in each combination.

3.4.2.2 Pairs of continuous and categorical patient covariates

In the following section, the relationship between continuous and categorical variables will be examined using box plots and the Kolmogorov-Smirnov (KS) test. Box plots will be used to visualize the distribution of the continuous variables for different categories of the categorical variables, and the KS-test to determine whether there is a statistically significant difference between the distributions of the continuous variables for different categories of the categorical variables. The aim is to gain a better understanding of the relationships between these variables by analysing the dependencies between them. The result of KS-tests depends on the sample size. Significant differences that are detected can therefore be minor from a clinical point of view and should be interpreted with caution.

Figure 3.5 displays group the median, range, as well as lower and upper quartiles of age, grouped by sex.

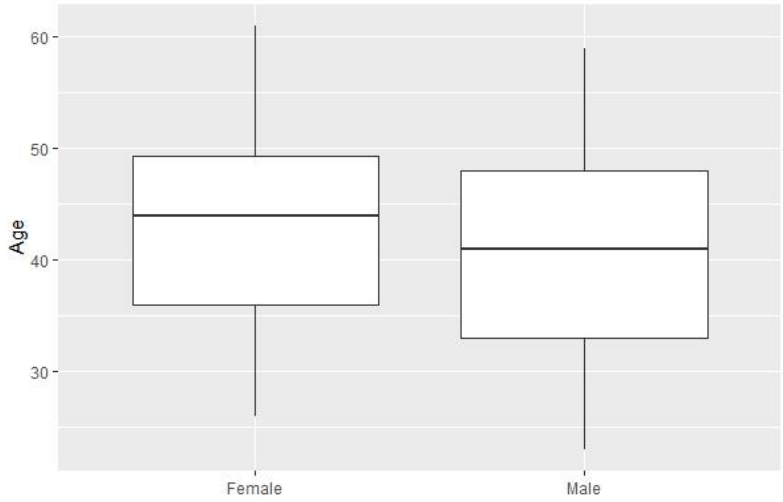


Figure 3.5: Boxplots of age grouped by sex.

It appears that men were slightly younger in this patient sample. Group means of men and women were 40.97 (s.d. 9.32) and 43.00 (s.d. 8.67) respectively. However, the two-sample KS-test resulted in a p-value of 0.44, which indicates that the null-hypothesis that they are from the same distribution cannot be discarded. Figure 3.6 shows boxplots for BMI for both men and women.

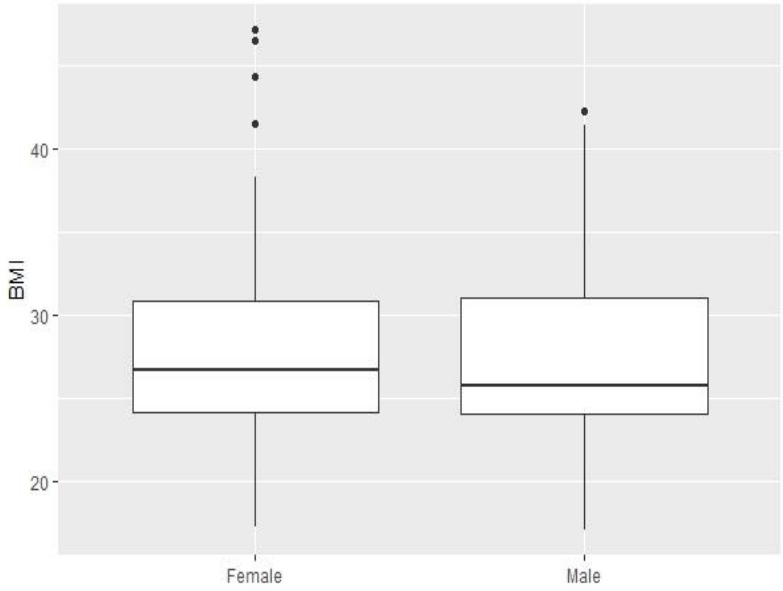


Figure 3.6: Boxplots of BMI, grouped by sex.

The KS test resulted in a p-value of 0.63, which indicates that the null-hypothesis that they are from the same distribution cannot be discarded. Figure 3.7shows boxplots for weeks of symptoms for both men and women.

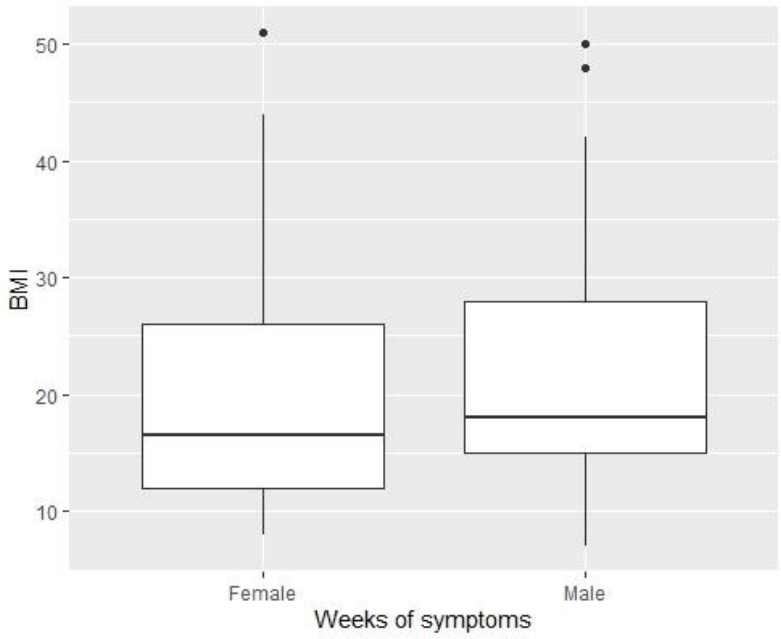


Figure 3.7: Boxplots of weeks of symptoms, grouped by sex.

The KS test resulted in a p-value of 0.34, which indicates that the null-hypothesis that they are from the same distribution cannot be discarded. Figure 3.8 shows boxplots for age, grouped by estimated volume of canal occupied by disc prolapse.

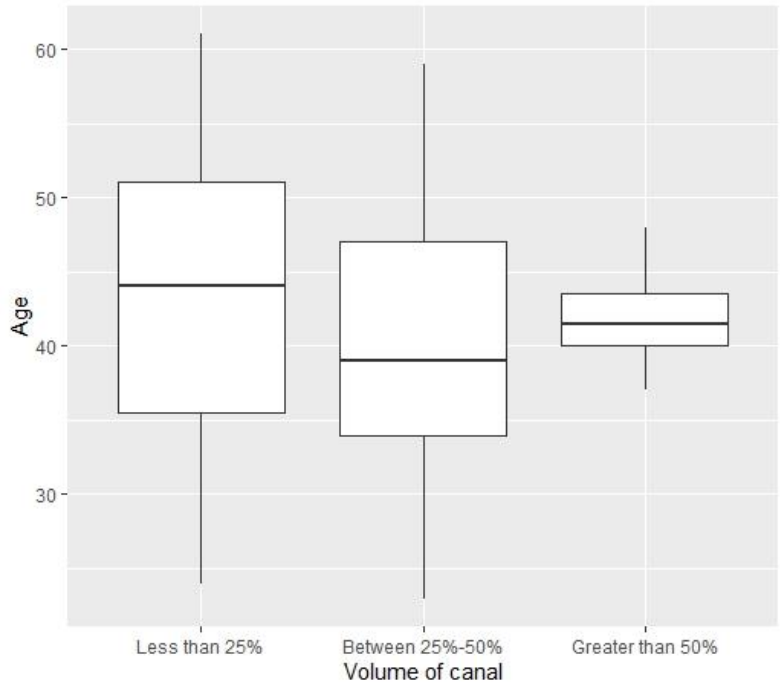


Figure 3.8: Boxplots of age, grouped by estimated volume of canal occupied by prolapsed disc.

It seems that patients with 25%-50% volume of canal were younger than patients with less than 25%.

Group means of “less than 25%” and “between 25% and 50%” were 43.01 (s.d. 9.57) and 40.37 (s.d. 8.41) respectively. However, the two-sample Kolmogorov-Smirnov (KS) test resulted in a p-value of 0.12, which indicates that the null-hypothesis that they are from the same distribution cannot be discarded. Tests between the group “greater than 50%” and others was not performed, since this group had a very low number of patients. Figure 3.9shows boxplots for BMI, grouped by estimated volume of canal occupied by disc prolapse.

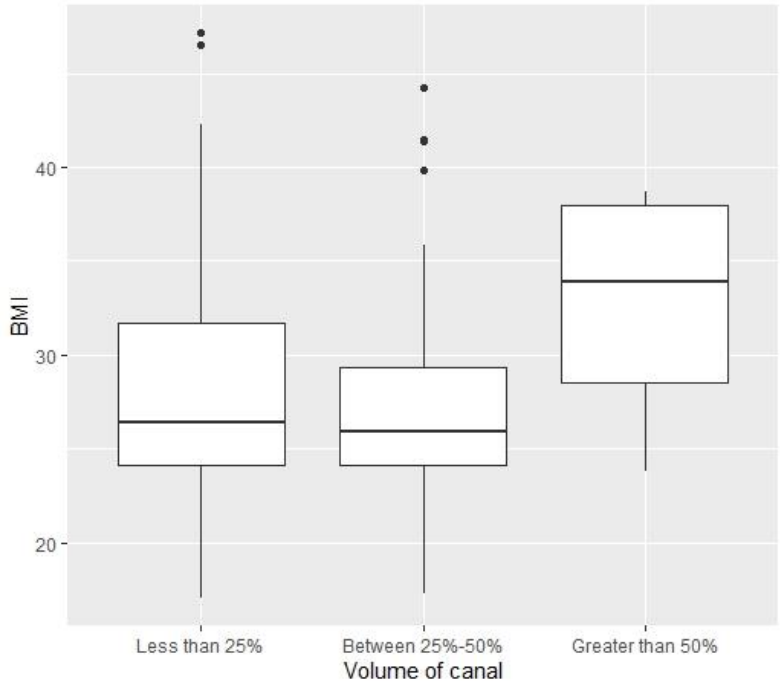


Figure 3.9: Boxplots of BMI, grouped by estimated volume of canal occupied by prolapsed disc.

Group means of “less than 25%”, “between 25% and 50%” and “greater than 50%” were 28.03 (s.d. 6.01), 27.15 (s.d. 5.79) and 32.59 (s.d. 7.02) respectively. It seems that patients that had more than 50% of volume of canal occupied by prolapsed disc had higher BMI. However, there were only 6 patients in this group (no missingness in this variable). KS tests between the group “greater than 50%” and other groups did not have significant p-values and the null-hypothesis could not be discarded. It would be interesting to see if there actually is a significant difference if there are the same variables available from data sets with larger sample size. Figure 3.10 shows boxplots for weeks of symptoms, grouped by estimated volume of canal occupied by disc prolapse.

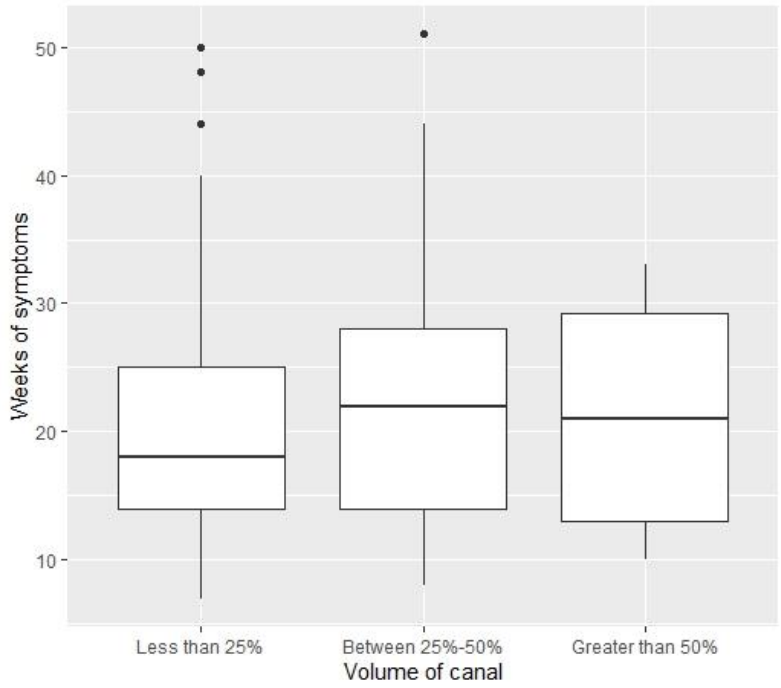


Figure 3.10: Boxplots of BMI, grouped by estimated volume of canal occupied by prolapsed disc.

KS tests between these groups did not have significant p-values and the null-hypothesis could not be discarded. There seems to be no correlation between these two patient covariates.

In the following, this procedure will be applied to the combinations level of spine and age, BMI and weeks of symptoms.

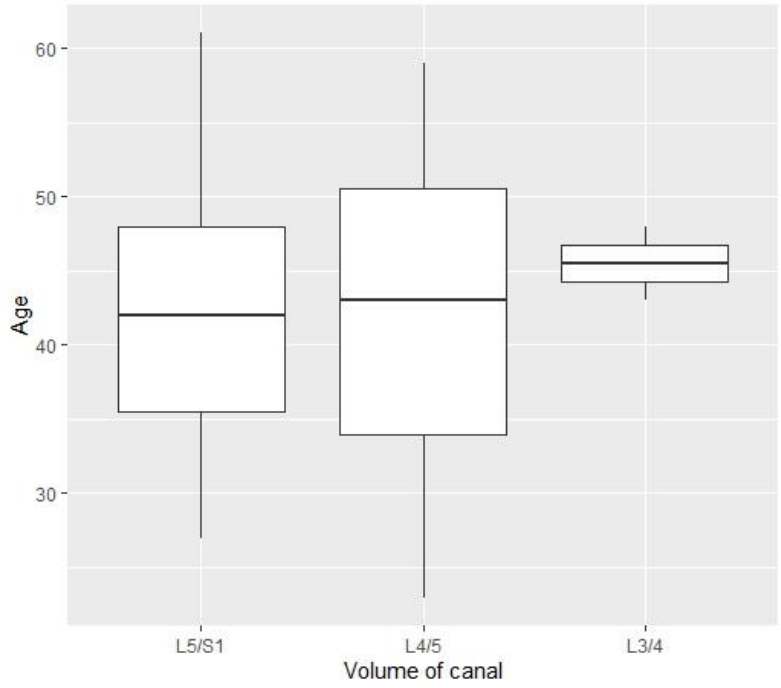


Figure 3.11: Boxplots of age, grouped by level of spine.

Group means of age of patients that had surgery at level “L5/S1” and “L4/L5” were 41.83 (s.d. 8.71) and 41.96 (s.d. 9.80) respectively. A KS test could not detect any significance. Tests including “L3/L4” were discarded, since there were only 2 patients in this sub-group. Analysis of level of spine in combination with both BMI and weeks of symptoms resulted in the same conclusion. Group means were close to each other and KS-tests could detect no significant dependence.

3.4.2.3 Pairs of categorical patient covariates

The categorical patient covariates were sex, level of spine and estimated volume of canal occupied by prolapsed disc. In order to analyse a possible correlation between sex and estimated volume of canal occupied by disc prolapse the frequencies of the categories of volume of canal for both men and women will be visualised.

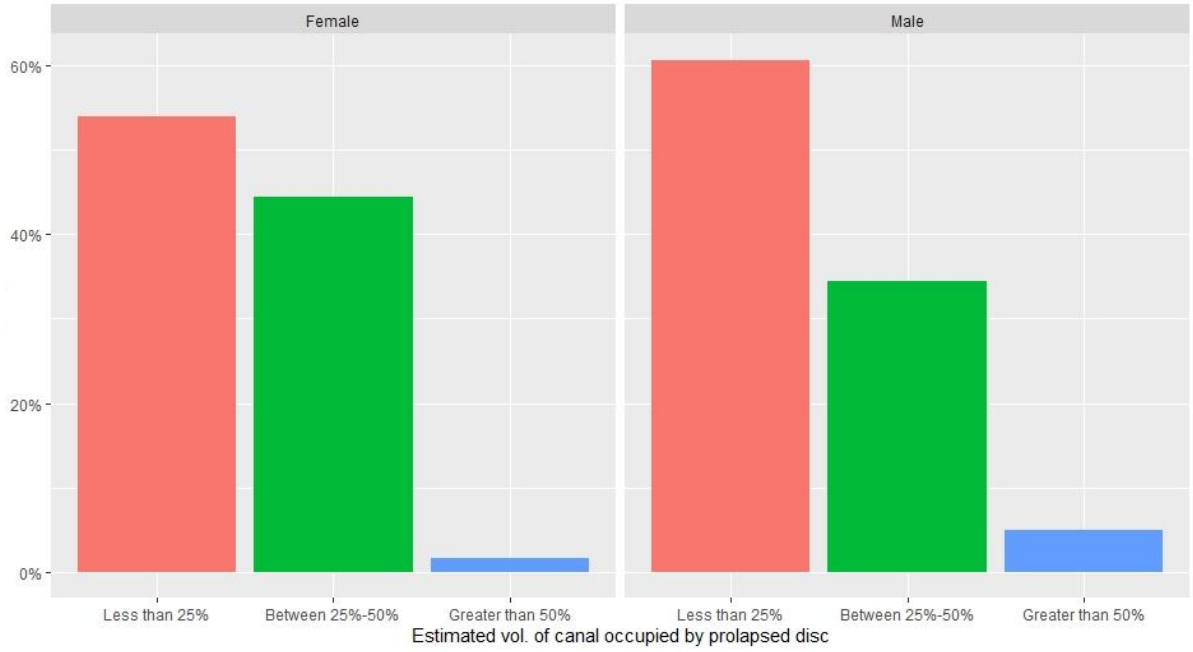


Figure 3.12: Percentages of volume of canal occupied by prolapsed disc grouped by sex.

Figure 3.12 shows that in males, the percentage of a volume of 25%-50% is less frequent than in females, whereas the frequency of less than 25% is higher. These two patient covariates could be correlated. A Chi-square test resulted in a test statistic (X-squared) of 2.10 (degrees of freedom = 2) and a p-value of 0.35, which indicates that there is insufficient evidence to reject the null hypothesis of independence. The same visualisation for sex and level of spine is shown in Figure 3.13.

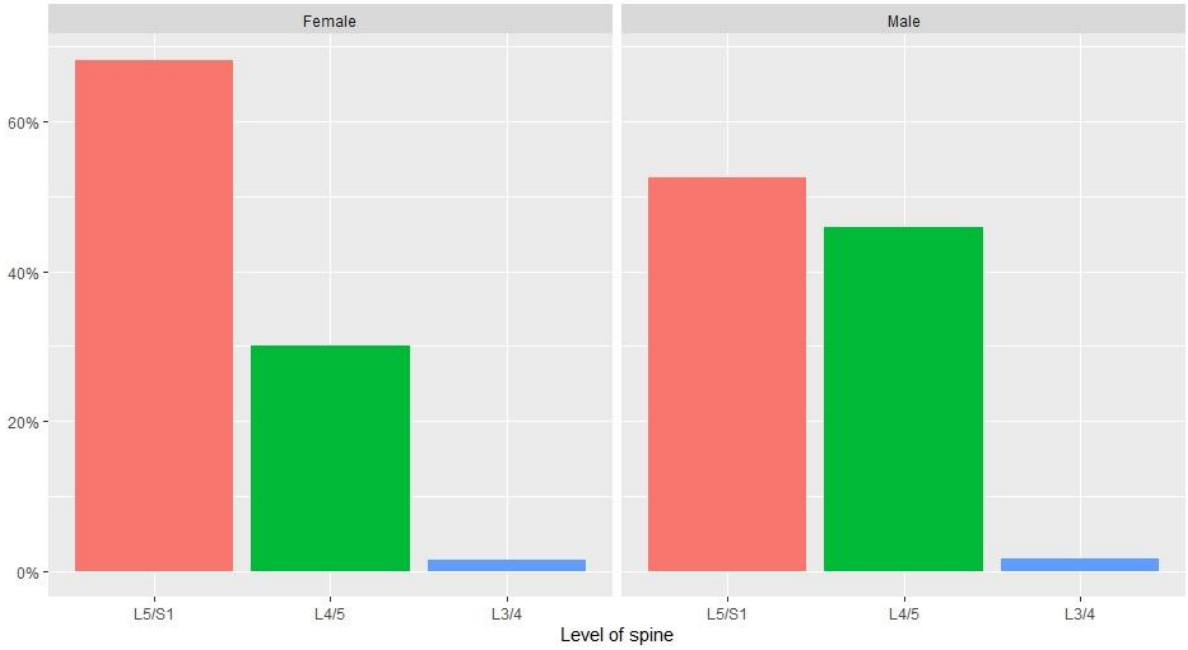


Figure 3.13: Percentage of level of spine, grouped by sex.

It seems that more women had surgery at the L5 / S1 level and more men had surgery at L4 / L5 level. Level of disc and sex seems to be slightly correlated. The Chi-square test resulted in a test statistic (X-squared) of 3.31 (degrees of freedom = 2) and a p-value of 0.19, which indicates that there is insufficient evidence to reject the null hypothesis of independence. The same visualisation for level of spine and volume of canal is shown in Figure 3.14.

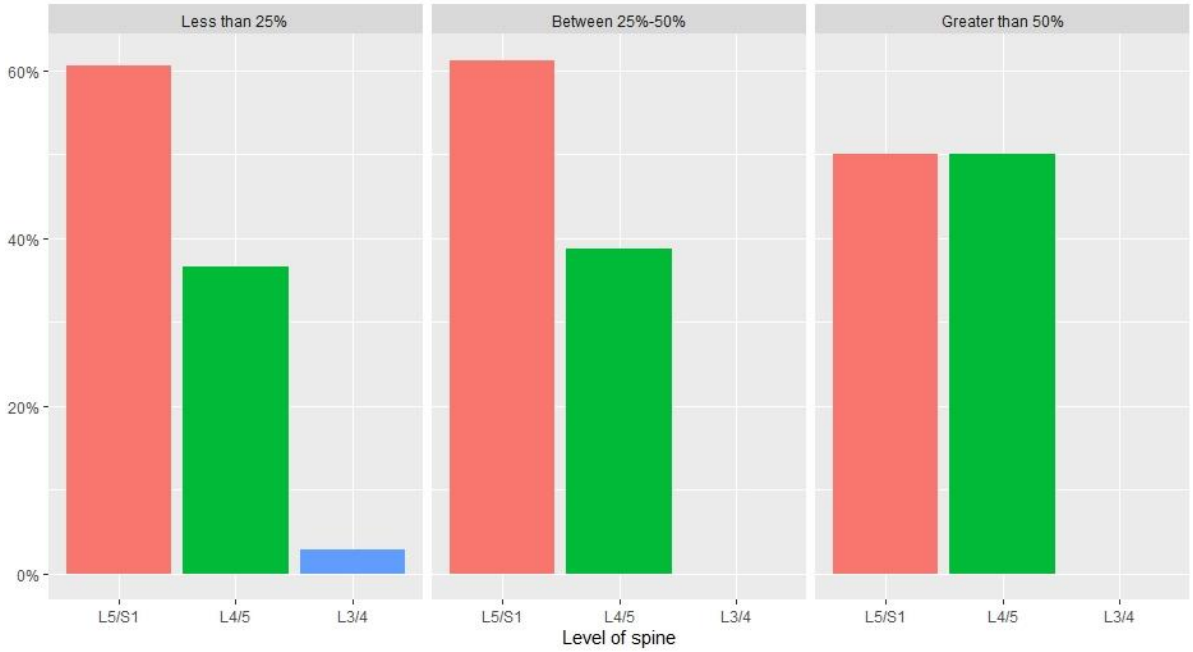


Figure 3.14: Percentage of level of spine, grouped by estimated volume of canal occupied by prolapsed disc.

The distribution of the two groups “less than 25%” and “between 25% and 50%” show a very similar distribution, which indicates that the variables estimated volume of canal and level of spine are not correlated. The Chi-square test resulted in an X-squared of 1.77 (degrees of freedom = 4) and a p-value of 0.78, which means that there is no evidence that these two variables are dependent.

3.4.3 Descriptive statistics of COMI questionnaires

Missingness of items or entire questionnaires at baseline was low overall, with 13 of 163 scores missing. Details about item missingness is displayed in Table 3.2.

Item	Q1 a)	Q1 b)	Q2	Q3	Q4	Q5	Q6*	Q7*	score
Missingness	3 (1.84%)	1 (0.61%)	3 (1.84%)	1 (0.61%)	8 (4.91%)	8 (4.91%)	163 (100%)	163 (100%)	13(7.98%)

Table 3.2: Missingness of items in baseline COMI questionnaires in the NERVES trial. * Items Q6 and Q7 were only applicable for follow-up questionnaire and are therefore completely missing at baseline.

3.4.3.1 Baseline scores

In the following it will be investigated if baseline scores have any associations with other baseline patient covariates. It could be for example, that the baseline COMI scores are higher, and the quality of life therefore lower, for patients with a long duration of symptoms or older age. Identifying such associations helps understanding the indication and the overall patient population. The baseline score distribution is displayed in Figure 3.15.

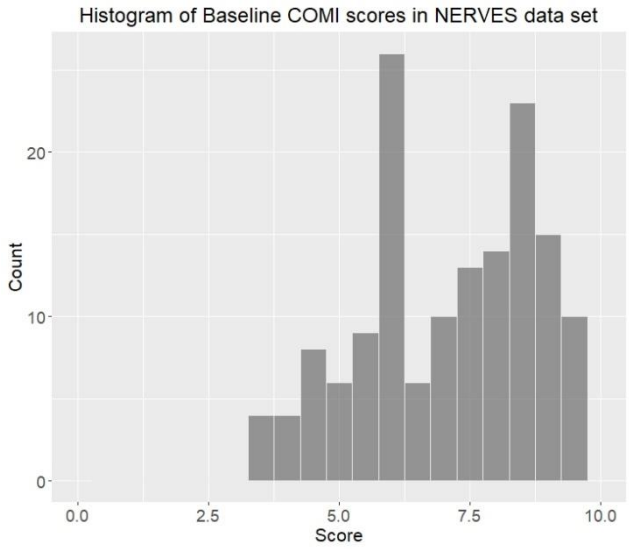


Figure 3.15: Histogram of baseline COMI scores for all patients for whom it was available.

For continuous patient covariates the Pearson-correlation was computed using the `cor()` –function in R. For age, BMI and weeks of symptoms, those correlation values were 0.084, -0.026 and -0.015 respectively. Scatter plots of these three tests are displayed in Figure 3.16.

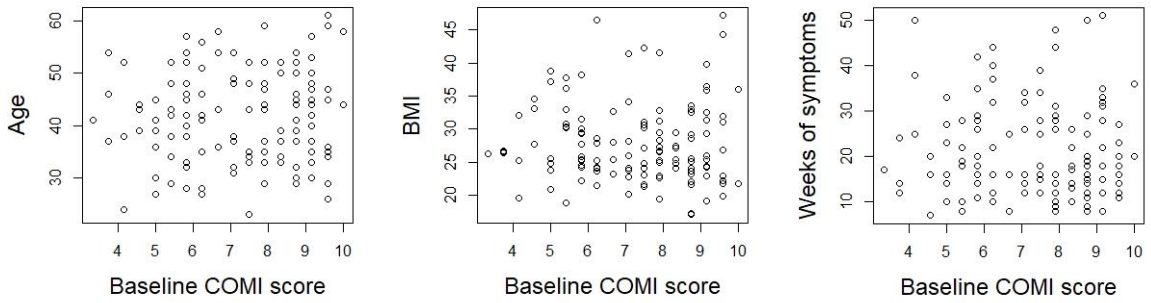


Figure 3.16 a), b) and c): Scatter plots of baseline COMI scores vs a) Age (left), b) BMI (middle) and c) weeks of symptoms (right).

Considering the low Pearson-correlation values and that there are no visible connections between baseline scores and the other three variables, it can be assumed that each are not dependent on each other.

For categorical patient covariates sex, level of spine and estimated volume of canal occupied by prolapsed disc, box plots were used to visualise the means and standard deviations in each subcategory, which are shown in Figure 3.17.

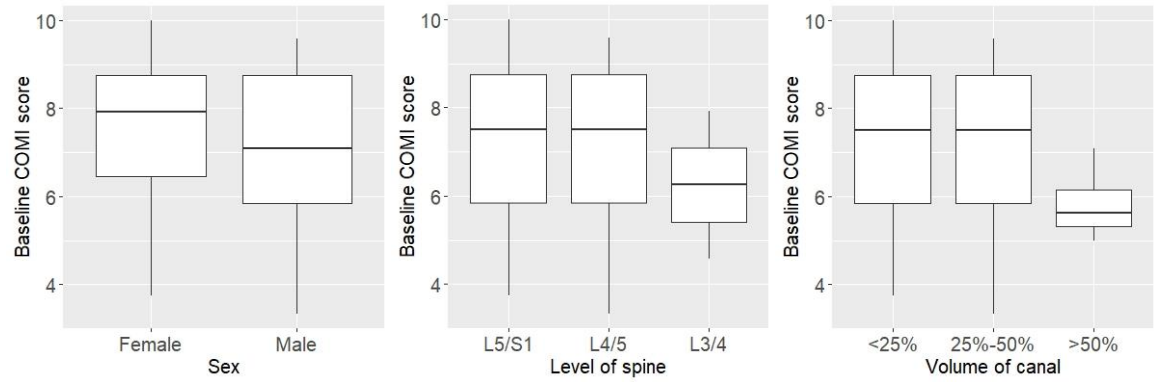


Figure 3.17 a), b) and c): Box plots of baseline COMI scores vs a) sex (left), b) level of spine (middle) and c) volume of canal (right).

For each of the combinations, KS-tests were performed, the p-values of which are summarised in Table 3.3.

Subset of patients for which the COMI scores were tested with the KS-test	p-value

Female and Male patients	0.212
Volume of canal less than 25% and 25% to 50%	0.883
Volume of canal less than 25% and greater than 50%	0.163
Volume of canal less than 25% to 50% and greater than 50%	0.217
Level of Spine S1/L5 and L4/L5	0.911
Level of Spine S1/L5 and L3/L4	0.832
Level of Spine L4/L5 and L3/L4	0.928

Table 3.3: Results of KS-test for sub-categories of baseline COMI scores and categorical patient covariates in the NERVES trial data set.

None of the KS-test showed significant p-values, which indicates that no dependencies between baseline COMI scores and either of these categorical variables could be detected.

Patients were split into the two treatment groups TFESI and microdiscectomy. Since they were randomised, it is expected that the distribution of COMI baseline scores is similar between the groups.

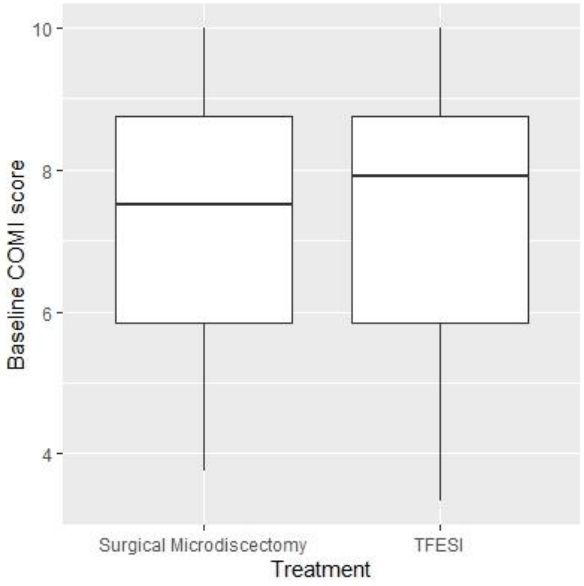


Figure 3.18: Boxplots for baseline COMI scores, grouped by allocated treatment.

The KS-test results indicate that there is no significant dependency (p-value = 0.833) between the type of treatment and the baseline COMI scores. This suggests that there is relatively little difference in the distribution of COMI scores between the treatment groups. This outcome aligns with expectations, as the treatment allocation was randomized, meaning that patients were assigned to their respective treatment groups without bias.

3.4.3.2 Scores past surgery

In Table 3.4 the missingness of the follow-up COMI questionnaires are summarised.

Item	Q1 a)	Q1 b)	Q2	Q3	Q4	Q5	Q6	Q7	Score
Week 18	30 (18.40%)	29 (17.78%)	29 (17.78%)	28 (17.18%)	30 (18.40%)	31 (19.02%)	33 (20.25%)	32 (19.63%)	38 (23.31%)
Week 30	68 (41.72%)	71 (43.56%)	67 (41.10%)	67 (41.10%)	73 (44.79%)	78 (47.85%)	72 (44.17%)	72 (44.17%)	87 (53.37%)
Week 42	68 (41.72%)	69 (42.33%)	68 (41.72%)	68 (41.72%)	73 (44.79%)	71 (43.56%)	68 (41.72%)	68 (41.72%)	86 (41.72%)
Week 54	41 (25.15)	40 (24.54%)	40 (24.54%)	40 (24.54%)	44 (26.99%)	47 (28.83%)	44 (26.99%)	44 (26.99%)	60 (36.81%)

Table 3.4: Missingness of items in follow-up COMI questionnaires in the NERVES trial in total numbers and percent.

Worth mentioning is that intervals of 18- and 54-weeks past randomisation had much lower missingness overall than intervals of 30- and 42-weeks past randomisation. This is because 18- and 54-week measurements were collected during hospital visits, whereas 30- and 42-week measurements were postal. Scores of each interval are summarised as box-plot in Figure 3.19.

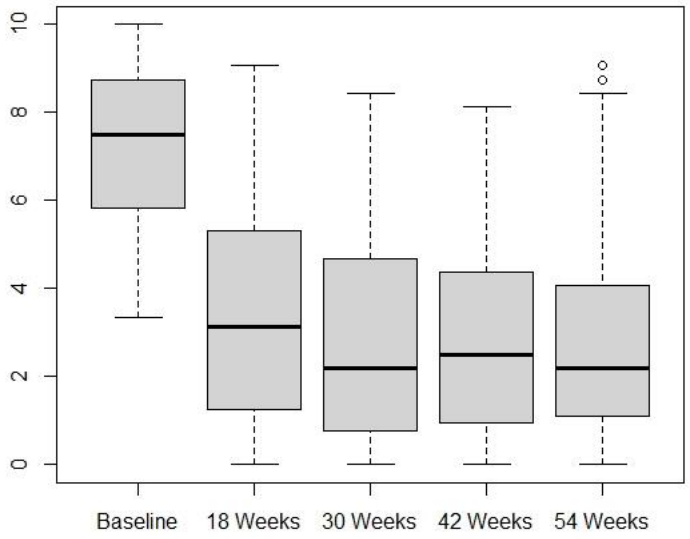


Figure 3.19: Box-plots of COMI scores at each interval in the NERVES trial data.

For each of the combinations, KS-tests were performed, the p-values of which are summarised in Table 3.5.

KS-test	p-value
COMI at Baseline vs COMI at 18 weeks	<0.001

COMI at Baseline vs COMI at 30 weeks	<0.001
COMI at Baseline vs COMI at 42 weeks	<0.001
COMI at Baseline vs COMI at 54 weeks	<0.001
COMI at 18 weeks vs COMI at 30 weeks	0.756
COMI at 18 weeks vs COMI at 42 weeks	0.455
COMI at 18 weeks vs COMI at 54 weeks	0.082
COMI at 30 weeks vs COMI at 42 weeks	0.961
COMI at 30 weeks vs COMI at 54 weeks	0.982
COMI at 42 weeks vs COMI at 54 weeks	0.817

Table 3.5: Results of KS-test for pairs of time-points of COMI scores in the NERVES trial data set.

KS-tests showed that baseline scores come from a different underlying distribution than any of the other outcome scores. P-values of combinations between outcome score intervals after surgery were not significant. This suggests that there is little progression over time between the first and the last time point of follow-up and that improvement takes place between surgery and the first follow-up.

Since in this trial, patients were allocated randomly to two different treatments, microdiscectomy and TFESI, baseline scores and scores at 18 weeks past randomisation (primary outcome time point) are analysed regarding their treatment group. The treatment difference between 18 weeks past surgery and baseline is analysed regarding each of the following variables: treatment type, sex, volume of canal occupied by prolapsed disc, level of spinal disc, age, BMI and weeks of symptoms. Figure 3.20 shows box plots of the outcome difference and categorical baseline variables.

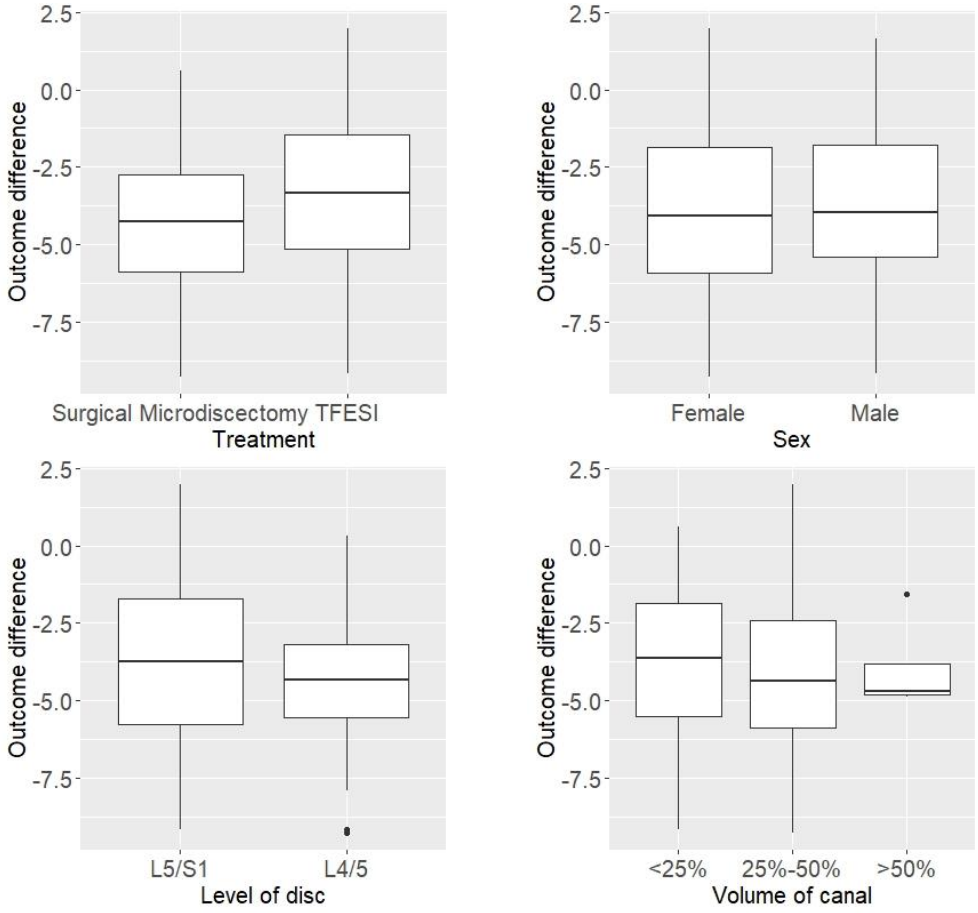


Figure 3.20 a), b), c) and d): Boxplots for COMI differences; subtraction of baseline scores from 18-weeks past surgery scores, grouped by a) allocated treatment (top left), b) sex (top right), c) level of disc (bottom left) and d) estimated volume of canal occupied by prolapsed disc (bottom right).

For each of the combinations, KS-tests were performed, the p-values of which are summarised in Table 3.6.

KS-test	p-value
Microdiscectomy vs TFESI	0.165
Female vs Male	0.476
Level of Spine L5/S1 vs L4/L5	0.057
Volume of Canal <25% vs 25-50%	0.325
Volume of Canal <25% vs >50%	0.595
Volume of Canal 25-50% vs >50%	0.742

Table 3.6: Results of KS-test for pairs of baseline sub-groups and their difference in COMI scores in the NERVES trial data set.

None of the KS-tests yielded statistically significant differences between the pairs of variables. Notably, among the patients who underwent surgery at level L5/S1, there was a relatively small

reduction in COMI scores compared to those who had surgery at level L4/5, with a mean difference of -3.439 (standard deviation: 2.795) for the former and -4.614 (standard deviation: 2.242) for the latter.

It's important to mention that tests involving subgroups of patients who had surgery at different levels were excluded from the analysis due to the limited number of patients in these specific groups.

Figure 3.21 Shows scatter plots of the outcome difference and continuous baseline variables.

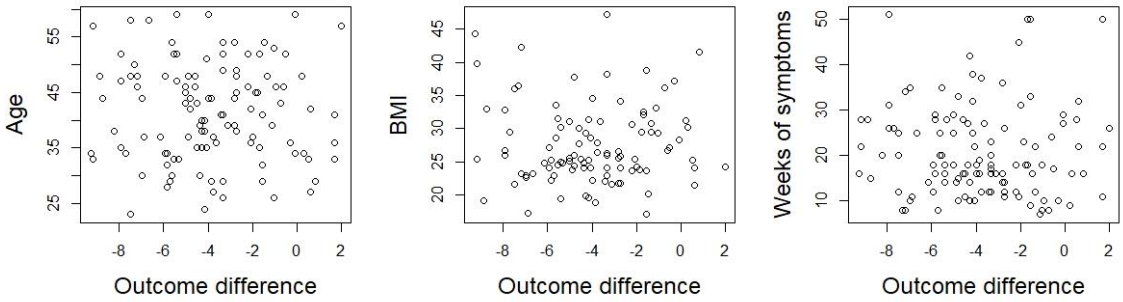


Figure 3.21 a), b) and c): Scatter plots for COMI differences; subtraction of baseline scores from 18-weeks past surgery scores and a) age (left), b) BMI (middle), and c) weeks of symptoms (right).

The Pearson-correlation coefficients, computed with the R-function `cor()`, were -0.04, -0.01 and -0.01 for age, BMI and weeks of symptoms respectively, which all indicate that there is very low correlation between the pairs of variables.

Another way to analyse if any of the patient characteristics were associated with treatment outcome, is to fit a multivariable linear regression model with 18 weeks past randomisation COMI scores as outcome measure. The coefficients of each variable and their 95% confidence interval of this linear regression model are listed in Table 3.7.

Characteristic		Coefficient estimate	95% CI	p-Value
Sex	Female	Reference		
	Male	0.044	[-0.968, 1.057]	0.931
Age		0.004	[-0.051, 0.059]	0.884
BMI		-0.005	[-0.089, 0.080]	0.913
Weeks of Symptoms		-0.019	[-0.071, 0.033]	0.462
Vol. Canal:	< 25%	Reference		
	25% - 50%	-0.273	[-1.310, 0.765]	0.603
	> 50%	-0.720	[-3.302, 1.861]	0.581

Spine Level:	L5 / S1	Reference		
	L4 / L5	-0.876	[-1.882, 0.130]	0.087
	L3 / L4	1.910	[-3.090, 6.910]	0.449
Treatment:	Microdiscectomy	Reference		
	TFESI	0.833	[-0.159, 1.825]	0.099
Baseline COMI score		**0.489**	**[0.259, 0.767]**	**0.001***

Table 3.7: Estimates of coefficients, 95%-confidence intervals and p-values for covariates in multivariable linear regression model with COMI scores at 18 weeks past randomisation as outcome.

The only patient characteristics to be connected to treatment outcome is baseline COMI scores. The R^2 was picked to measure the explained proportion of variance in the outcome that is explained by the covariates. The R^2 of 0.151 shows that only little outcome variation could be explained by the included input variables. This means that there might be unmeasured variables that could possibly explain outcome variation. As seen, the treatment method did not have a significant p-value, and there is therefore a lack of evidence for a treatment difference. However, there were four serious adverse events in four participants associated with surgery, and none with TFESI. Moreover, TFESI showed better cost-effectiveness, which is why Wilby et al. recommend TFESI as first invasive treatment option (Wilby et al., 2021).

In summary, the analysis of the RCT data set showed relatively low levels of missingness in patient covariates, with only a small fraction of the data missing for most variables. Outcomes, on the other hand, exhibited a higher proportion of missing values compared to the covariates. These findings are consistent with the publication by Wilby et al., which included a descriptive analysis of patient covariates and COMI scores at baseline, as well as over time. In that publication, a longitudinal mixed model was used to assess the treatment effect on COMI scores, yielding an estimate of -0.77 (-1.58, 0.03). This aligns with the results of the multivariable linear regression model in this chapter, indicating that the treatment was not statistically significant (although it approached significance with a p-value close to the threshold).

This section added a comprehensive exploration of dependencies between baseline variables, accompanied by visualizations that should be considered during analyses. Furthermore, it establishes the groundwork for a comparison of these dependencies with the Spine Tango dataset, where patient data was routinely collected.

3.5 The Spine Tango registry (EUROSPINE)

The vast amount of data in clinical registries can be used to detect rare adverse events or benefits and identify patterns of outcomes for subsets of patients. The growing demand for outcome measurement and quality assurance in the field of low-back and leg pain led to the development of such a registry in 2000. The so-called Spine Tango (ST) registry was built by EUROSPINE, the Spine Society of Europe in collaboration with the Institute for Evaluative Research in Orthopaedic Surgery at the University of Bern, Switzerland. It was launched in 2002 and to date, over 750,000 forms (134,458 surgery forms) from five continents have been collected (EUROSPINE, 2022b). Several studies concluded that in a heterogeneous group of conditions, observational studies obtain valid insights into real-world practice and outcome research, as well as help optimizing health service and quality assurance. Results from well-designed observational studies can be similarly trustworthy as results from RCTs (Benson and Hartz, 2000, Colditz, 2010, Concato et al., 2010, Concato et al., 2000).Containing such a vast amount of entries, "registries can describe care patterns, appropriateness of care, understand variations in treatment and outcomes, identify and select subgroups in the heterogeneous chronic low back pain population with a probability of poor or successful outcome" (Hooff et al., 2015).

When it comes to the evaluation of therapeutic effectiveness of treatments in routine clinical practice in non-research settings, well-performed observational studies can be of crucial importance and are relatively inexpensive and fast to conduct. Sweden for example has some of the most complete national databases that are collecting data from patients in hospitals and health-care organizations, e.g. SWEDEHEART and SweSpine (James et al., 2015). The analysis of the data in such registries with a representative patient population helps assessing health care effectiveness and safety and evaluating prognostic factors (José and Edelman, 2017). Studies that are based on registry data must make sure to include an appropriate and representative population of patients, regarding their condition and characteristics, so that the inferences drawn from the statistical analysis are valid.

Data access was granted for the Spine Tango registry regarding the patient population affected by sciatica secondary to prolapsed disc. The registry mostly collected data of surgical procedures, a further investigation about the difference of treatment effects between microdiscectomy and TFESI is therefore difficult.

Data have been collected with various forms, such as outcome questionnaires at several time points, surgery and follow-up forms. Each patient has a unique anonymized patient ID by which all available information for each patient can be matched. The surgery forms have been updated over the years and collected slightly different information, therefore there are some characteristics that have only been measured over specific years. There are four different surgery forms, indicated by the year from which they were updated (2005, 2006, 2011 and 2017). Table 3.8 indicates the availability of the

characteristics which were considered as potentially connected to treatment outcome, as well as information about adverse events and in which version of the surgery forms they were collected.

Variable	Surgery sheet 2005	Surgery sheet 2006	Surgery sheet 2011	Surgery sheet 2017
Age	x	x	x	x
Gender	x	x	x	x
Surgeon credentials	x	x	x	x
Level of disc	x	x	x	x
Country ID	x	x	x	x
ASA Morbidity status	x	x	x	x
Smoker			x	x
BMI			x	x
Previous treatment	x	x	x	
Duration of symptoms				x
No. of previous surgeries	x	x	x	
Adverse events	x	x	x	x
Blood loss	x	x	x	x

Table 3.8: Availability of variables in each version of the surgery form

The total number of patients in the surgery sheets are 292 (2005 form), 4,615 (2006 form), 10,936 (2011 form) and 2,340 (2017 form). Some patients had multiple surgery entries in the registry. Later analyses require an assumption of independence of individual entries. Therefore, only the earliest of the procedures for each patient was considered, in case there were multiple. This, together with deleting duplicate rows, led to a total of 17,252 patients. After reducing the data to the first surgery per patient (if multiple) and deleting duplicates, independence between data rows was assumed, however not further investigated. ASA Morbidity denotes the ASA (American Society of Anesthesiologists) physical status classification system for assessing the fitness of patients before surgery (Anesthesiologists, 2020). In further notation the categories will be abbreviated with ASA 1-4 (there are 6 categories, but no patient in the data was classified in category 5 or 6).

The most commonly patient reported outcome measures for quality of life and lower back and leg pain were the Oswestry Disability Index (ODI) and the Core Outcome Measures Index (COMI), which were recorded pre-surgery and at follow-up visits. The focus in this project is on the COMI questionnaire.

3.5.1 Descriptive statistics on patient/surgery characteristics

Table 3.9 shows descriptive statistics each variable. Since instead of BMI the 2017 forms had height and weight, BMI values were calculated according to the formula BMI = kg/m^2 and categorized into “<20”, “20-25”, “26-30”, “31-35” and “>35”, like in the 2011 form. Countries were anonymised and

can therefore not be identified further. For simplicity ID were renamed in descending order by number of patients using alphabetic lettering. Countries with less than 100 entries were grouped into “other” for simplicity.

Variable		N = 17,252 patients
Sex	Female	8,033 (46.56%)
	Male	9,219 (54.44%)
	Missing	0 (0%)
Age		Mean 47.38 (s.d. 14.12)
	Missing	18 (0.10%)
Surgeon credentials	Board certified neurosurgeon	6,950 (40.29%)
	Specialized spine surgeon	6,902 (40.01%)
	Neurosurgeon in training	1,814 (10.51%)
	Board certified orthopedic surgeon	815 (4.72%)
	Orthopedic surgeon in training	260 (1.51%)
	Other	191 (1.11%)
	Missing	320 (1.85%)
Country ID	A	8,062 (46.73%)
	B	3,779 (21.90%)
	C	1,545 (8.96%)
	D	1,010 (5.85%)
	E	627 (3.63%)
	F	325 (1.88%)
	G	170 (0.96%)
	H	125 (0.72%)
	Other	384 (2.23%)
	Missing	1,068 (6.19%)
Level of spine	L5/S1	7,510 (43.53%)
	L4/L5	7,571 (43.88%)
	L3/L4	1,370 (7.91%)
	L2/L3	442 (2.56%)
	L1/L2	30 (0.17%)
	Other	329 (1.91%)
	Missing	0 (0%)
ASA Morbidity	1	7,010 (40.63%)

2	6,114 (35.44%)
3	989 (5.73%)
4	26 (0.15%)
Missing	3,113 (18.04%)

Table 3.9: Descriptive statistics of all variables, the collection of which was consistent over the changes in surgery forms.

Figure 3.22 shows a histogram of the age in this patient population.

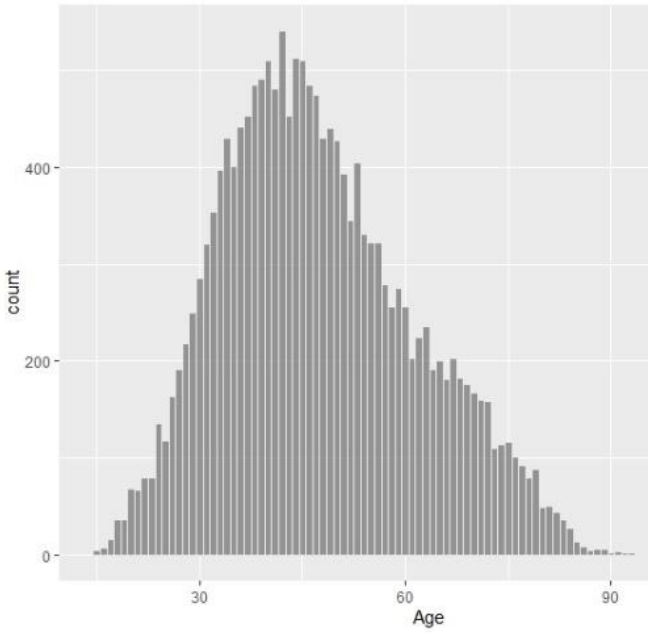


Figure 3.22: Histogram of age in this patient population in the Spine Tango registry (of a total of 17,252 data points out of which 18 were missing)

With a mean of 47.38 (s.d. 14.12) this population is on average older than participants of the NERVES trial, which had a mean of 42.83 (s.d. 9.28).

As seen in Table 3.8, some patient characteristics were not consistently part of surgery forms. For example, previous treatment was stopped in 2017 and instead replaced by duration of symptoms. Smoking status and BMI have been introduced in 2011. This change results in large percentages of missing data and makes more advanced modelling approaches such as prognostic modelling challenging. In the following, descriptive statistics of these variables are presented regarding the subset of surgery forms that collected them.

3.5.1.1 Smoking status and BMI

In this subset there were 12,485 patients. Variable	N = 12,485 patients

Smoking status	Smoker	1,703 (13.64%)
	Non-smoker	7,432 (59.52%)
	Missing	3,350 (26.83%)
BMI	≤ 20	452 (3.62%)
	> 20 to ≤25	2,864 (22.94%)
	>25to ≤30	3,349 (26.82%)
	>30 to ≤35	1,427 (11.43%)
	> 35	626 (5.01%)
	Missing	3,767 (30.17%)

Table 3.10: Descriptive statistics of BMI and smoking status.

3.5.1.2 Previous treatment

Previous treatment was collected in surgery forms from 2005, 2006 and 2011, but were discontinued. Therefore, there are only 15,094 total patients for which this item was available.

None	< 3 months conservative	3 – 6 months conservative	6 – 12 months conservative	> 12 months conservative	Surgical treatment	Missing
3,865, (25.61%)	3,082 (20.42%)	3,145 (2.84%)	2,280 (15.11%)	1,643 (10.89%)	604 (4.00%)	475 (3.15%)

Table 3.11: Descriptive statistics of previous treatment (N=15,094).

3.5.1.3 Duration of symptoms

After previous treatment was discontinued to be collected, it was replaced by duration of symptoms, included in surgery forms since 2017. Therefore, there are only 2,158 patients for which this item was available.

< 3 months	3-12 months	> 12 months	Missing
656 (30.40%)	1,110 (51.44%)	392 (18.16%)	0 (0%)

Table 3.12: Descriptive statistics of duration of symptoms (2,158).

3.5.1.4 Complications

Complications were also consistently entered in surgery forms, with an occurrence rate of 4.33% (747 complications). The type of complication and their occurrences are summarised inTable 3.13.

Complication type	Occurrence
Dural lesion	650 (87.01%)
Nerve root damage	38 (5.09%)
Bleeding in spinal canal	11 (1.47%)

Bleeding outside spinal canal	12 (1.61%)
Cauda equina damage	2 (0.27%)
Wound infection	8 (1.07%)
Vascular injury	4 (0.54%)
Other	20 (2.68%)
Wrong level	2 (0.27%)
Missing	236 (1.37% of all 17,252 patients)

Table 3.13: Descriptive statistics of complications during surgery (N = 747)

In the following it will be analysed if there are any dependencies between patient covariates. By analysing dependencies between patient covariates, one can identify potential confounding factors that may affect the outcomes of interest. In both registry data and RCTs, there might be hidden variables that influence the outcome, leading to biased conclusions. Recognizing these confounding factors allows to adjust for them appropriately during analysis, improving the accuracy of the results.

Understanding how patient covariates relate to treatment outcomes can help in patient stratification. Identifying subgroups of patients who respond better to the intervention can lead to personalized or tailored treatment plans, optimizing patient care and potentially improving overall outcomes.

For both registry data sets and RCTs, understanding the relationships between covariates can help assess the generalizability of the findings. If certain patient characteristics consistently influence outcomes across different datasets or study designs, it adds to the robustness of the conclusions and enhances the external validity of results.

Insights gained from the analysis of dependencies can be valuable in designing future studies. Researchers can use this information to adjust sample size, stratify randomization, or consider other factors that might impact the outcomes of interest, ultimately leading to more efficient and effective studies.

The main rationale behind this analysis is to determine if there are any dependencies between patient covariates that need to be considered for later analyses. If any dependencies are found, the goal is to ascertain whether they align with the data set from the NERVES data set.

3.5.2 Dependencies

In this study, all possible combinations of two patient covariates are tested in order to identify any potential dependency. For combinations of a continuous and a categorical variable, group means and standard deviation were further analysed by using boxplots. Additionally, two-sample KS-tests are used to investigate if two samples are from the same underlying distribution. If the test results in a p-

value smaller than 0.05, it can be assumed that they are not from the same distribution. For two categorical variables grouped bar plots, split by the categorical variables and displaying the frequencies of the other variable, were used. In order to quantify association, Chi-square test were used. However, due to very large sample sizes, even the smallest differences in group means are detected by either test. Results of such tests are therefore interpreted with caution.

Given the vast number of combinations, only those exhibiting detectable associations are presented, while those without detected associations are summarized in Appendix E. The primary objective is to comprehensively examine the impact of these factors on patient outcomes by analysing the interrelationships between various patient characteristics. Similar to the analyses conducted with the dataset of the NERVES trial, the issue of multiplicity was acknowledged, however, the decision was made not to make explicit adjustments in this study. Again, this is due to the exploratory nature of this analysis, where the goal was to identify potential relationships or dependencies between variables that might warrant further investigation. This exploratory analysis was rather meant to be hypothesis-generating than hypothesis-testing.

The age distribution for each category of level of spine is shown in Figure 3.23.

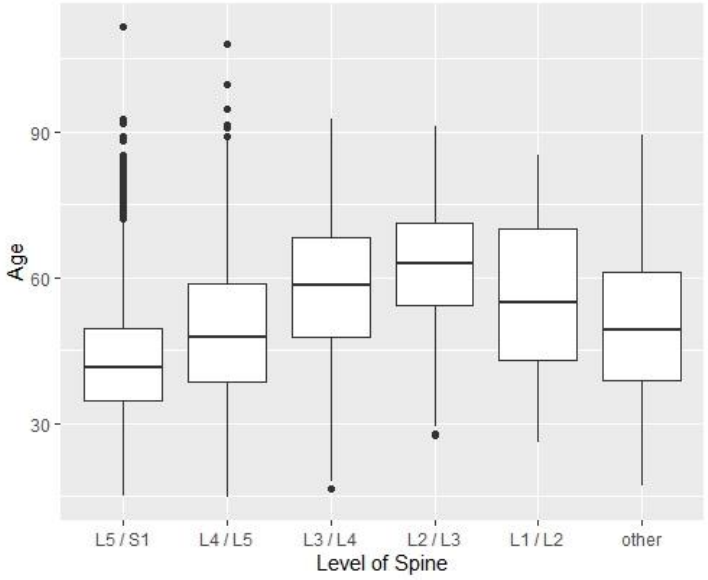


Figure 3.23: Box plots for of age in each sub-category of level of spine in the Spine Tango registry.

Surgeries at levels “L4 / L5”, “L3 / L4” and “L2 / L3” apparently were in an older sub-population of patients than at level “L5 / S1”. Group means were 42.96 (s.d. 12.04), 48.87 (s.d. 14.22), 57.46 (s.d. 13.97) and 62.47 (s.d. 12.15) for “L5 / S1”, “L4 / L5”, “L3 / L4” and “L2 / L3” respectively.

The subgroup “L5/S1” was significantly younger than any other sub-group (KS-test p-values <0.05). The only pairs of subgroups that did not result in significant p-values were “L1/L2” and “other”, “L3/L4” and “L1/L2”, as well as “L4/L5” and “other”.

Additionally, there seems to be an association between age and ASA morbidity status. This is visualised in Figure 3.24.

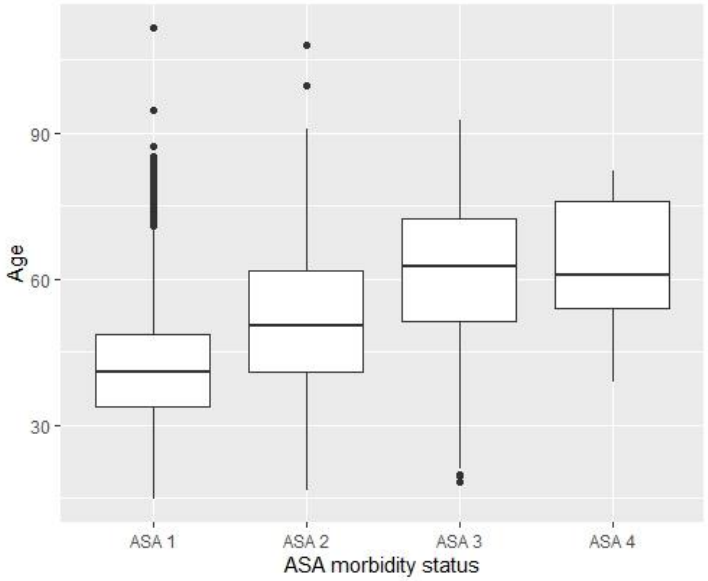


Figure 3.24: Box plots for age in each sub-category of ASA morbidity status.

Patients that were classified as ASA morbidity status 2, 3 and 4 apparently were in an older sub-population of patients than with status 1. Group means were 41.96 (s.d. 11.56), 51.32 (s.d. 14.13), 61.14 (s.d. 14.56) and 63.68 (s.d. 13.49) for status groups 1,2,3 and 4 respectively. The difference in group means showed significance (KS-test p-value smaller than 0.05) for each of the pairs, but “ASA 3” and “ASA 4”.

There are numerous countries contributing to Spine Tango, but country A and B account for roughly 68% of the data set in this patient population. Routine practice might be different in each country, which especially reflects in annotation of surgeon credentials. A summary is displayed in Figure 3.25.

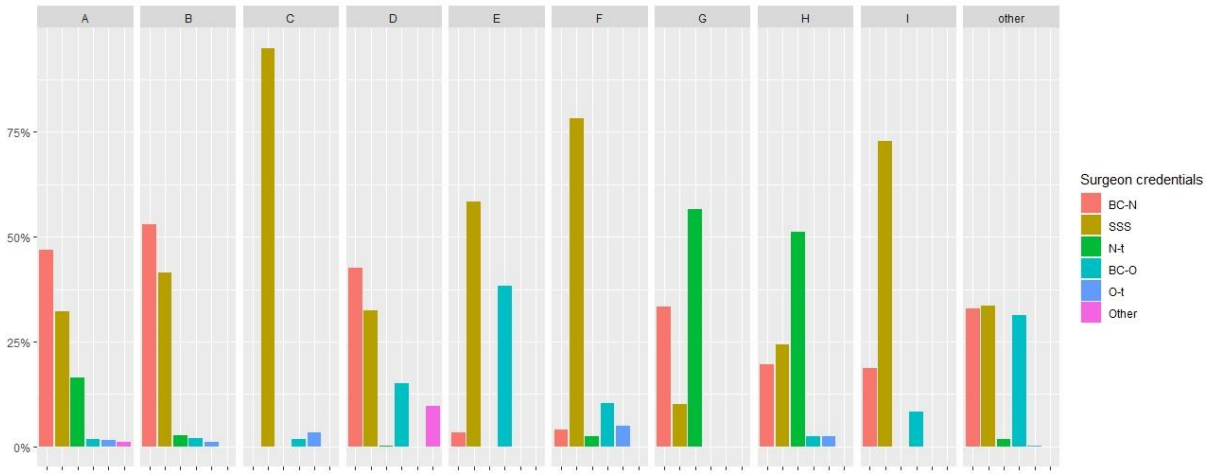


Figure 3.25: Bar plot of percentage of surgeon types, split by country ID. Abbreviations in legend: BC-N = Board-certified neurosurgeon, SSS = specialized spinal surgeon, N-t = Neurosurgeon in training, BC-O = Board-certified orthopedic surgeon, O-t = Orthopedic surgeon in training.

One example is country C, which does not report any board-certified neurosurgeons or neurosurgeons in training (abbreviated by BC-N and N-t respectively). Similar differences can be seen for other countries, which leads to an association between country and surgeon credentials, just due to the practice of reporting. This needs to be kept in mind in later prediction modelling approaches. The Chi-square test resulted in a test statistic (X-squared) of 7,244.7 (degrees of freedom = 45) and a p-value < 0.001, which demonstrates that there is an association between these two variables.

Differences between countries were also apparent for different types of previous treatment, see Figure 3.26.

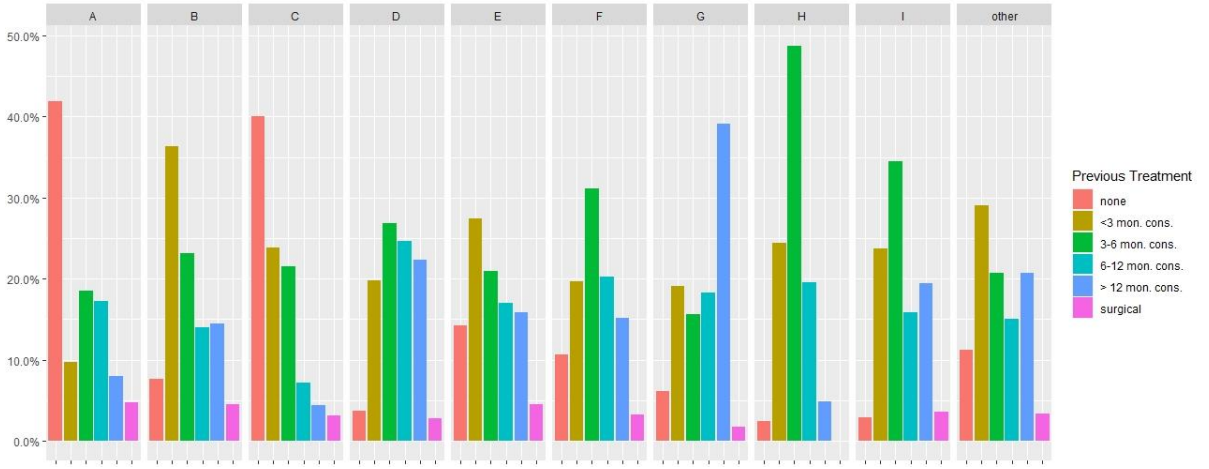


Figure 3.26: Bar plot of percentages of different types of previous treatment, split by country ID.

In country A and C, a larger percentage of patients had no received previous treatment than in other countries. Country H had a higher percentage of patients that had between 3 and 6 months prior conservative treatment than other countries.

These discrepancies may be due to variations in treatment practices between the different countries. The Chi-square test resulted in a test statistic (X-squared) of 2,852.7 (degrees of freedom = 45) and a p-value < 0.001, which shows that there is an association between these two variables.

Interestingly, some countries had a different distribution of patients regarding their ASA morbidity classification, which is displayed in Figure 3.27.

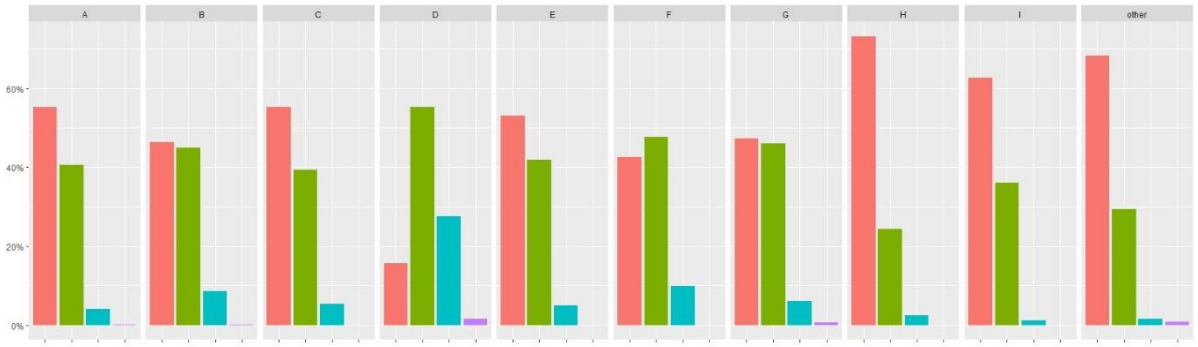


Figure 3.27: Bar plot of percentages of different ASA morbidity statuses, split by country ID. In each country sub-plot ASA statuses are numbered 1-4 from left to right (red, green, blue, purple).

Country D had a higher percentage of patients with ASA status 2 and 3, compared to countries A, B or C. Countries with ID H, I and the small countries that were grouped into “other”, had higher percentages of patients with ASA status 1. The Chi-square test resulted in a test statistic (X-squared) of 1,070.8 (degrees of freedom = 27) and a p-value < 0.001, which shows that there is a dependency between these two variables.

Additionally, there were slight differences of frequencies of ASA Morbidity state in different groups of previous treatment, as illustrated in Figure 3.28.

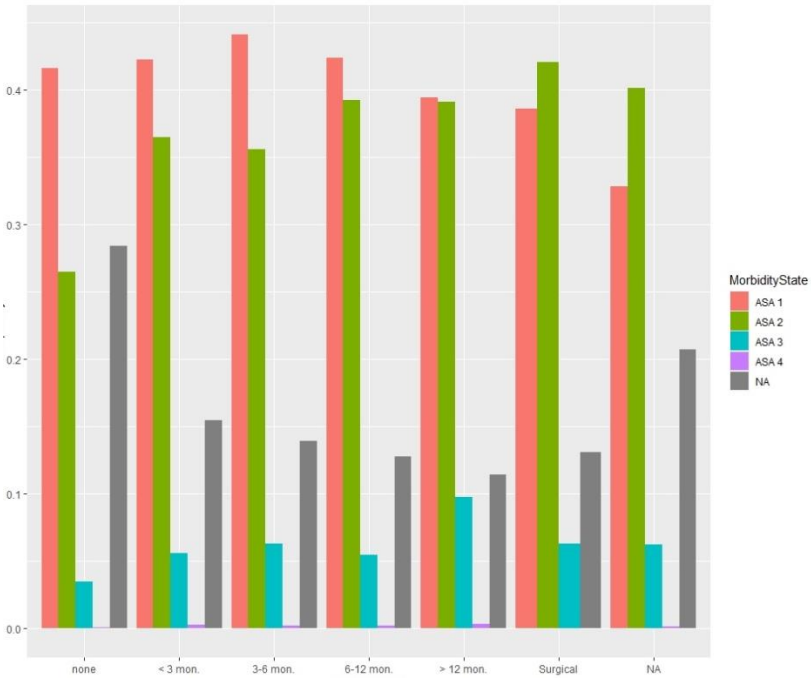


Figure 3.28: Bar-plot of percentages (y-axis) of ASA Morbidity state categories split by the sub-categories of previous treatment (x-axis).

It appears that ASA Morbidity status of 1 was more frequent in patients that had no previous treatment or previous treatment of less than 3 or 3 to 6 months, compared to patients that had prior surgery. Additionally, patients with no previous treatment more often had no ASA morbidity status available.

Even if differences are slight, Figure 3.29 shows that more patients that had prior surgery had complications during surgery, compared to conservative or no prior treatment.

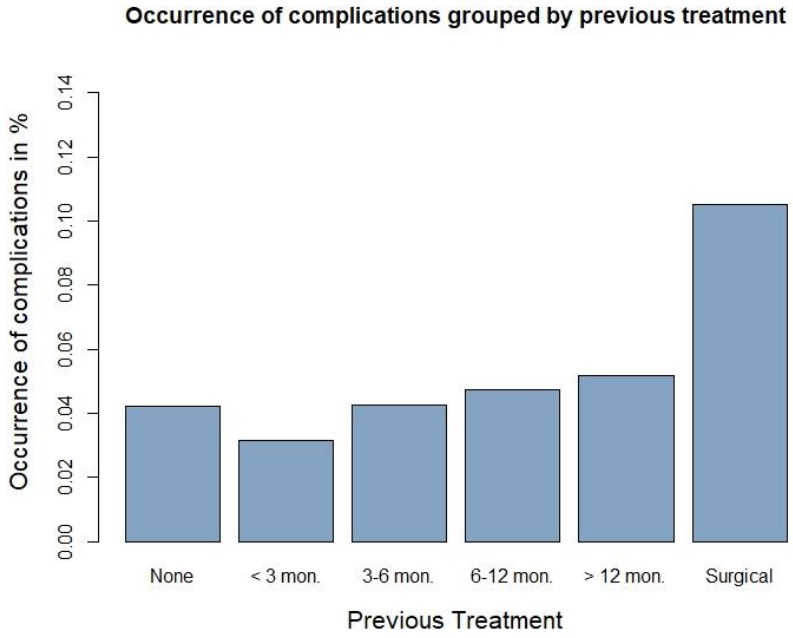


Figure 3.29: Bar plot of percentage of complications, grouped by previous treatment. Categories indicated in length of months (mon.) stand for conservative treatment of the indicated length.

It seems that prior surgery is a risk factor, which will be analysed later in Chapter 5 in further depth. The Chi-square test resulted in a test statistic (X-squared) of 66.85 (degrees of freedom = 5) and a p-value < 0.001, which shows that there is an association between these two variables.

ASA morbidity status appears to be correlated to BMI as seen in Figure 3.30.

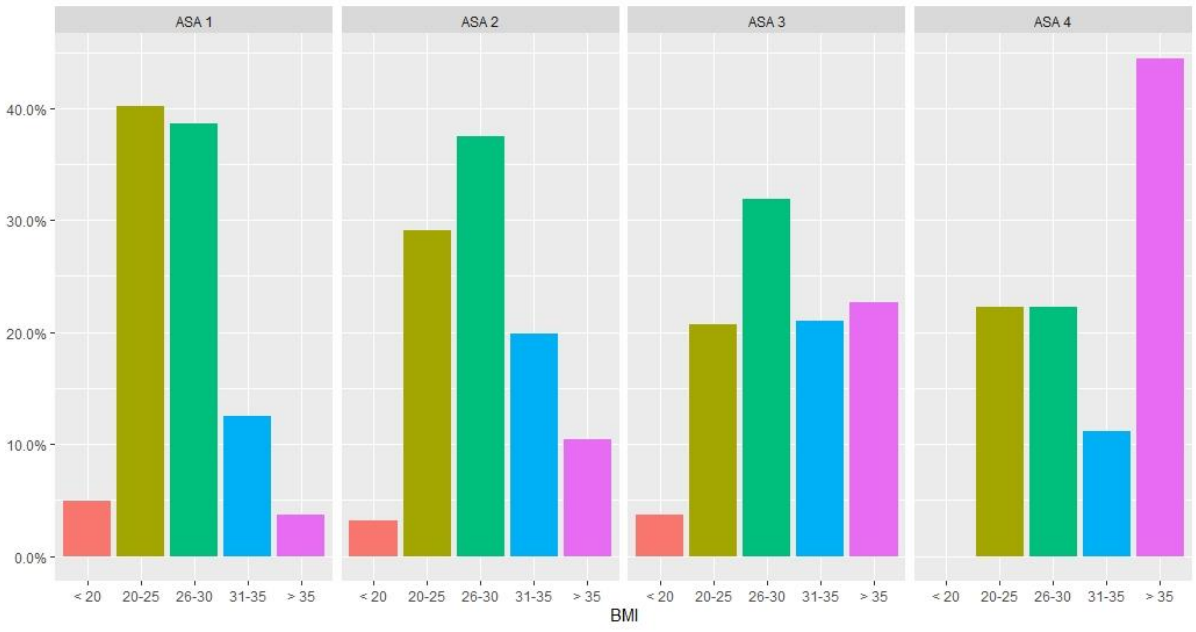


Figure 3.30: Bar plot of percentage of BMI categories, split by ASA morbidity statuses.

Patients with ASA morbidity status 3 or 4 more often had BMIs of 31-35 and >35. The Chi-square test resulted in a test statistic (X-squared) of 315.5 (degrees of freedom = 12) and a p-value < 0.001, which shows that there is an association between these two variables.

In this population, there were some associations between patient covariates. In particular, differences in routine practice between countries regarding surgeon reporting and previous treatment were correlated. This suggests that variations in these factors may impact patient outcomes and should be considered in future modelling approaches. Also apparent were associations between age and ASA morbidity as well as age and level of spine at which the surgery took place. Prior surgery could be a risk factor that clinicians should be aware of, which will further be investigated in prediction modelling approaches. It should be remembered, that the data set was very large, and even the slightest differences can be significantly detected by both the Chi-square and the KS-test. These p-values therefore need to be interpreted with caution.

3.5.3 Outcomes in Spine Tango

The Spine Tango registry collects multiple patient-reported outcome measures. The Oswestry Disability Index and COMI questionnaires are the main quality-of-life measures that are routinely collected. In the following, the amount of missing data in the COMI questionnaire and possible dependencies of its reporting with other variables are investigated.

The number of patients that had at least 1 answered COMI entry is 11,093 (64%), but questionnaires are categorized in many different time intervals and the amount of missingness is much higher than in the NERVES trial. Of all patients that have at least one answered COMI questionnaire, only 54% (6,008) have a baseline score and of those who have, only a further 59% (3,530) have at least one follow-up score. Table 3.14 shows the percentages of entries in the available time spans after surgery of all patients with at least one COMI entry.

Before surgery	4 weeks	6 weeks	2 months	3 months	6 months	9 months	1 year	2 years	3 years
54% (6,008)	1% (144)	4% (441)	0.4% (44)	49% (5,395)	10% (1,157)	6% (697)	46% (5,094)	38% (4,232)	0.5% (51)

Table 3.14: Percentages of entries in the available time spans after surgery of all patients with at least one COMI entry.

Before surgery, 3 months, 1 year and 2 years after surgery seem to be standard check-in follow-up times, whereas the other times are much less frequent. In the forms there are also the exact days of completion, which makes it possible to use all available patients with a baseline entry and any follow-

up entry for a longitudinal model. Due to the observational nature of the data, the classification of these intervals is not necessarily the same for each country or clinic. It is expected that there are time windows of a specific number of days for each of the intervals, however further details were not available.

In order to check that the questionnaires were allocated correctly to these time spans, the days after surgery were plotted and grouped according to their category of time span, the result of which is shown in Figure 3.31. Although there is some variance of the days after surgery at which forms were submitted, the groups are clearly separated and therefore categorized correctly.

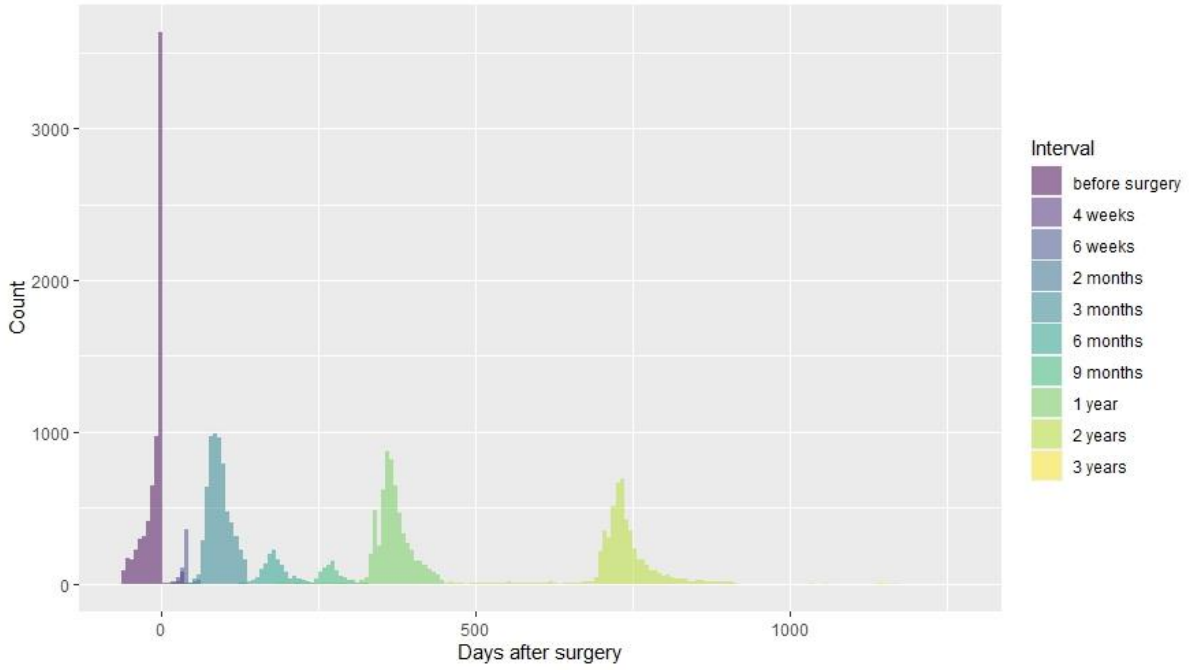


Figure 3.31: Frequencies of days after surgery regarding the groups into which they have been categorized in the registry excerpts. Overlap did not appear often, but is visible due to a degree of opacity of the color-scheme.

Figure 3.32 shows from which surgery form and therefore also from which time span in years these collected forms are.

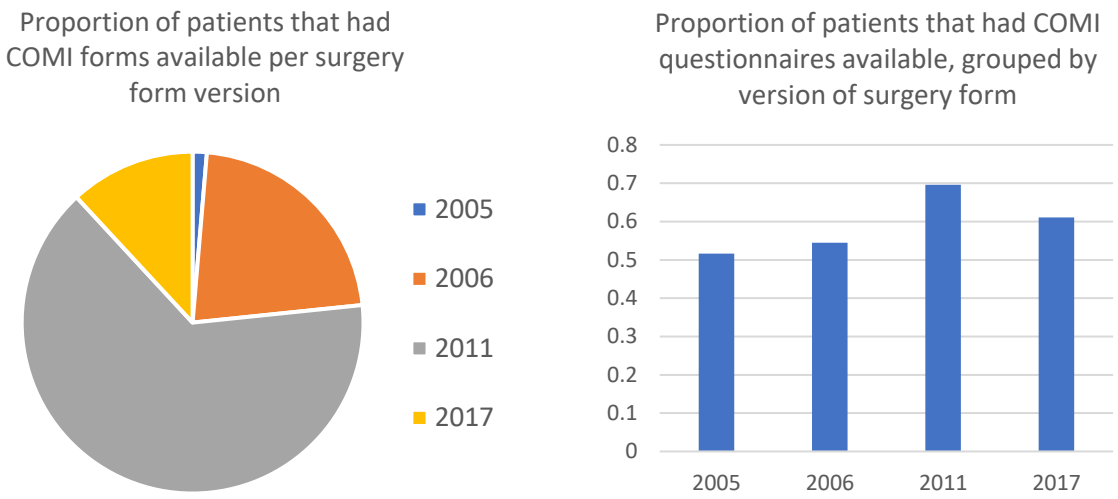


Figure 3.32 a) and b): a) Visualisation of how many patients had COMI questionnaires available in each surgery form (left) b) fractions of patients with available COMI for each surgery form (right).

Most of the forms were collected from patients from the 2011 form, which is mainly due to the size of the data set. However, it also seems that a higher percentage of patient entries of the 2011 form had COMI available (69.5%), compared to other forms (51.6%, 54.5% and 61.1% for surgery forms 2005, 2006 and 2017 respectively).

Routine practice can differ especially regarding reporting. This was already seen in differences in surgeon credentials, but there are also differences in the reporting of COMI questionnaires between countries.

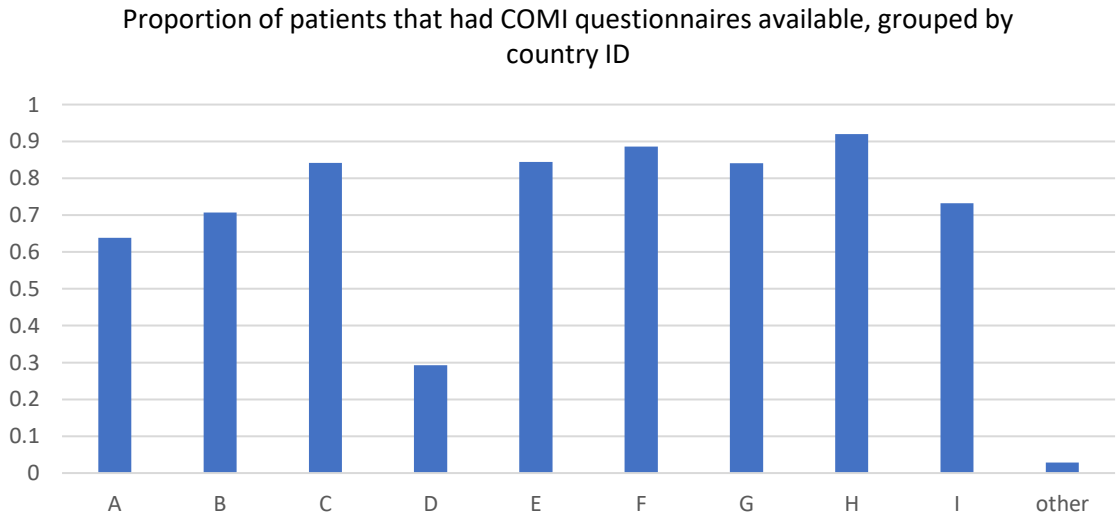


Figure 3.33: Bar plot of COMI availability, grouped by country ID.

Country D and countries that were in the group “other” had much lower percentages of COMI availability.

In the following, similar to the procedure of the NERVES trial data, it will be analysed if COMI baseline scores had any associations to other baseline patient characteristics. For continuous patient covariates Pearson-correlation was computed using the `cor()` −function in R and scatter plots.

For categorical patient covariates, box-plots for each sub-category of each characteristic were considered. Due to large sample size however, KS tests could detect significant p-values and lead to the assumption that two sample sizes are not from the same underlying distribution. Therefore, means and standard deviations of the sub-groups were also taken into consideration.

Figure 3.34 shows a scatter plot between age and baseline COMI scores.

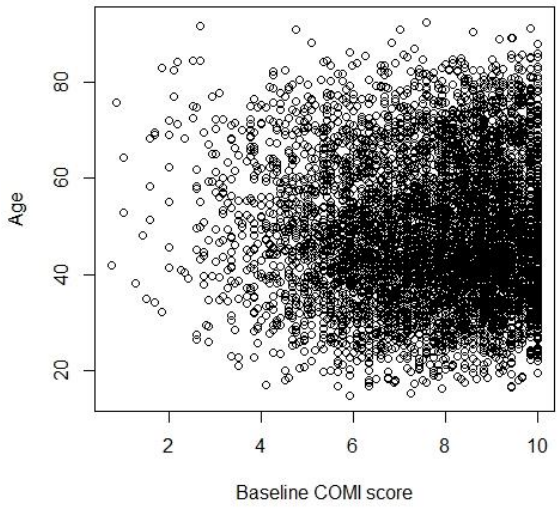


Figure 3.34: Scatter plot of baseline COMI scores on x-axis vs age on y-axis.

More baseline COMI scores are on the upper end of the scale (between than 5 and 10) than on the lower half (between 0 and 5). However, there is no visible pattern in this plot that indicates a relationship between COMI baseline scores and age. This is supported by a Pearson-correlation measure of -0.006. It is evident that these two patient covariates are not dependent.

All other patient covariates were categorical and will be analysed using box plots and KS tests. However, it needs to be remembered that even small differences can be detected with significant p-values due to large sample size. Each pair of variables is analysed on a complete data set regarding the two columns.

In Figure 3.35, the median, range, as well as lower and upper quartiles of baseline COMI scores are visualised, grouped by sex.

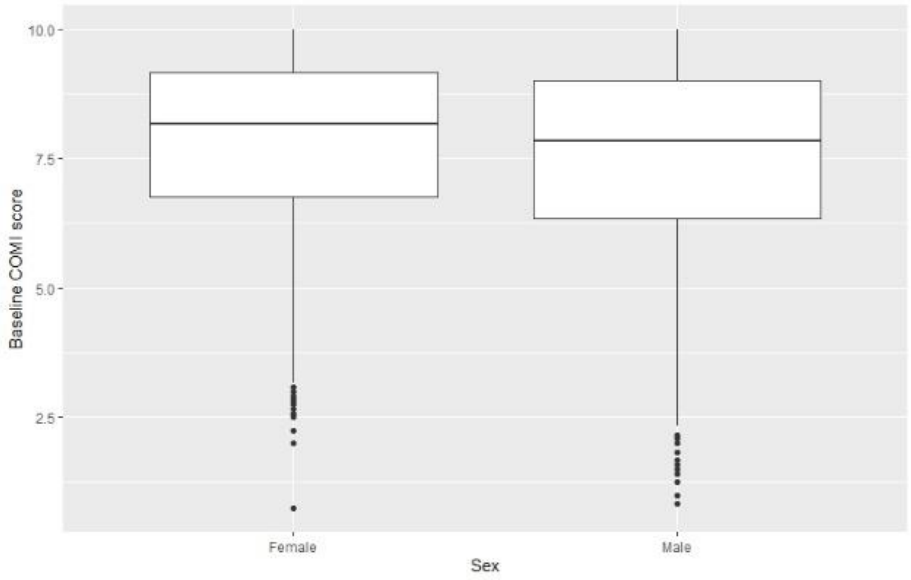


Figure 3.35: Box plot of baseline COMI scores, grouped by sex.

The KS test rejected the null-hypothesis with a p-value smaller than 0.001. However, group means of women and men were 7.88 (s.d. 1.64) and 7.53 (s.d. 1.78) respectively. The clinically important difference of COMI scores for improvement is 2.2 (between 2.0 and 2.5). In a clinical point of view, these two patient samples, with a difference of 0.35 are therefore not considered significantly different. In Figure 3.36, the median, range, as well as lower and upper quartiles of baseline COMI scores are visualised, grouped by surgeon credentials.

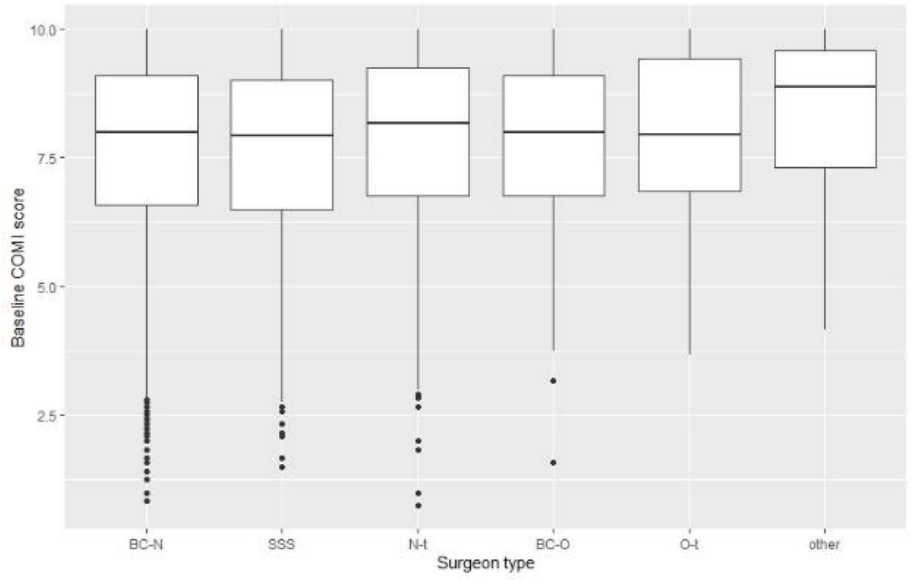


Figure 3.36: Box plot of baseline COMI scores, grouped by surgeon credentials.

Baseline scores between most surgeon credentials are similar, with the exception of “other”. Group means are summarised in Table 3.15.

Surgeon type	BC-N	SSS	N-t	BC-O	O-t	Other
Mean (s.d.)	7.69 (1.74)	7.64 (1.70)	7.79 (1.78)	7.80 (1.65)	7.86 (1.70)	8.24 (1.78)

Table 3.15: Baseline COMI score mean and standard deviations of each subgroup of patients, regarding surgeon credentials.

Even though baseline scores seem to be higher for surgeons in group “Other”, it is not assumed to be dependent from a clinical point of view. In Figure 3.37, the median, range, as well as lower and upper quartiles of baseline COMI scores are visualised, grouped by country ID.

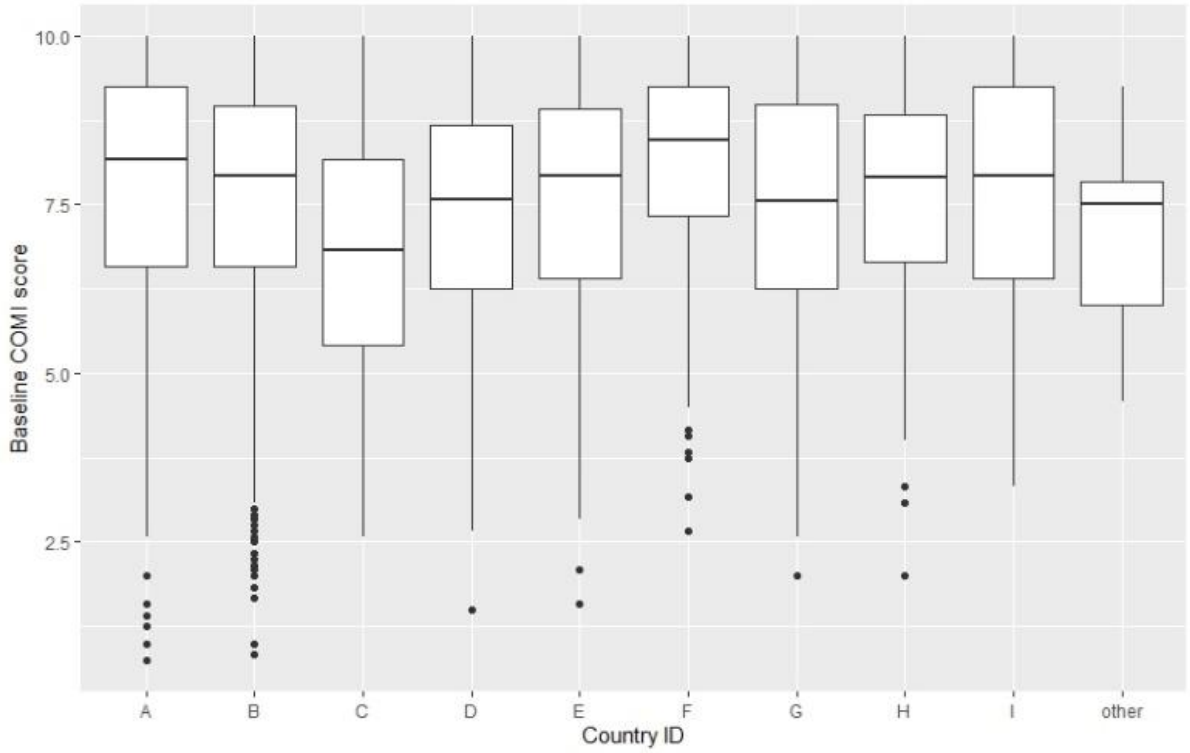


Figure 3.37: Box plot of baseline COMI scores, grouped by country ID.

In this case, there are some pairs of countries, between which there seems to be a significant difference in baseline COMI scores (country C and F). Means and standard deviations are summarised inTable 3.16.

Country ID	A	B	C	D	E	F	G	H	I	Other
Mean (s.d.)	7.81 (1.70)	7.60 (1.72)	6.68 (2.06)	7.34 (1.63)	7.57 (1.72)	8.10 (1.51)	7.36 (1.91)	7.55 (1.67)	7.75 (1.77)	7.03 (1.79)

Table 3.16: Baseline COMI score means and standard deviations of each subgroup of patients, regarding country IDs.

There seem to be some variations in the baseline COMI scores between different countries, the reason of which needs to be investigated further. In Figure 3.38, the median, range, as well as lower and upper quartiles of baseline COMI scores are visualised, grouped by level of spine.

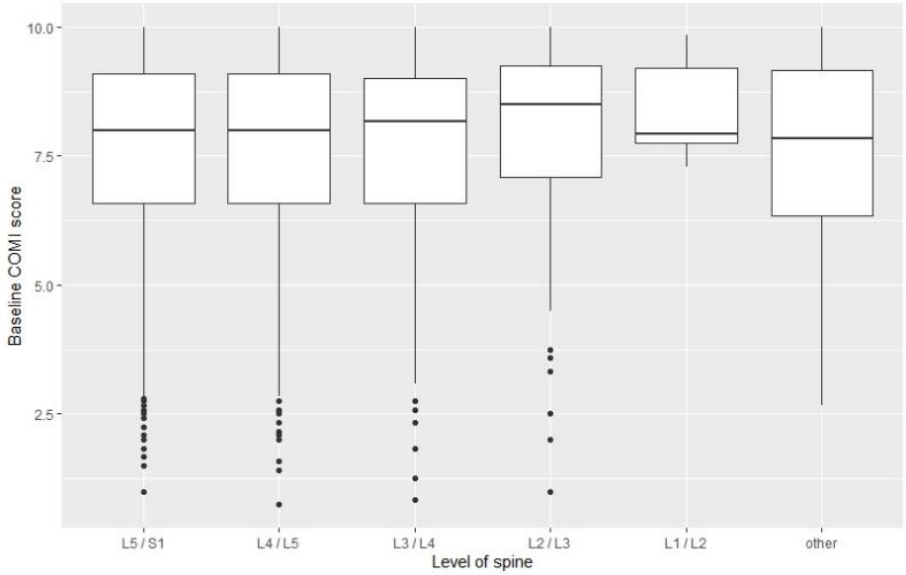


Figure 3.38: Box plot of baseline COMI scores, grouped by level of spine.

Patients that had surgery at the L2/L3 level had slightly higher baseline COMI scores than patients that had surgery at any other level. The difference however is not large enough to consider COMI scores and level of spine to be correlated. In Figure 3.39, the median, range, as well as lower and upper quartiles of baseline COMI scores are visualised, grouped by ASA morbidity status.

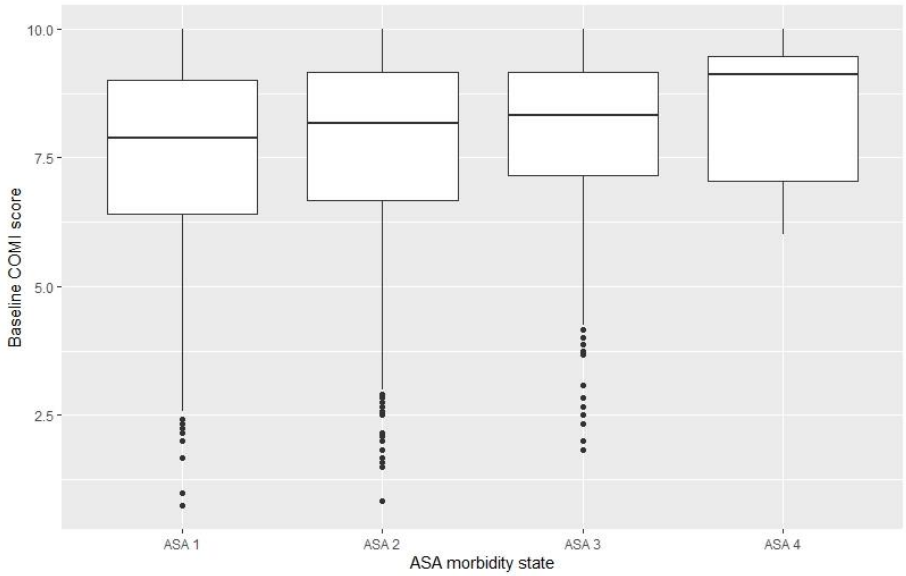


Figure 3.39: Box plot of baseline COMI scores, grouped by ASA morbidity status.

Patients with ASA morbidity 4 had higher baseline COMI scores than patients of the other categories. However, standard deviation in this group was also very large, which is why there is no reason to assume that COMI scores and ASA morbidity are connected. In Figure 3.40, the median, range, as well as lower and upper quartiles of baseline COMI scores are visualised, grouped by BMI.

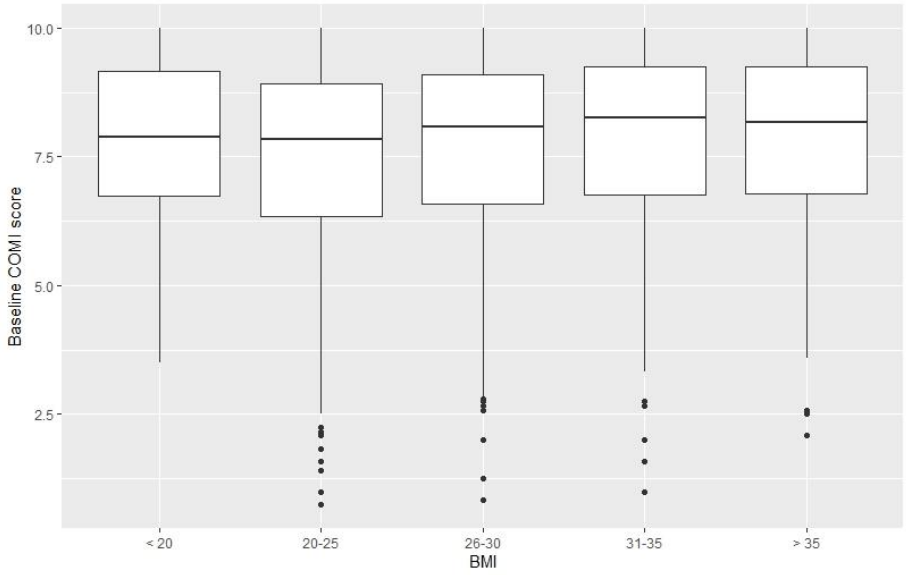


Figure 3.40: Box plot of baseline COMI scores, grouped by BMI.

Differences in mean were not large enough to assume that COMI scores and BMI are correlated. InFigure 3.41, medians, ranges as well as lower and upper quartiles of baseline COMI scores are visualised, grouped by previous treatment.

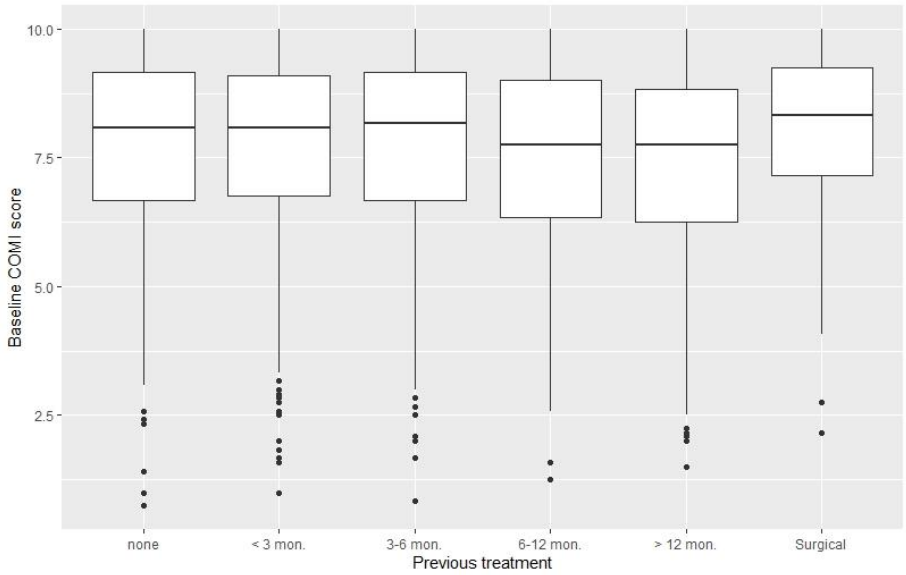


Figure 3.41: Box plot of baseline COMI scores, grouped by previous treatment.

There are slight differences between mean and standard variation of COMI scores between categories of previous treatment. Patients that had prior surgery have larger baseline COMI scores than patients with no prior surgery. Considering standard deviation however, differences were not large enough to assume that COMI scores and prior treatment were correlated. In Figure 3.42, medians, ranges, as well as lower and upper quartiles of of baseline COMI scores are visualised, grouped by smoking status.

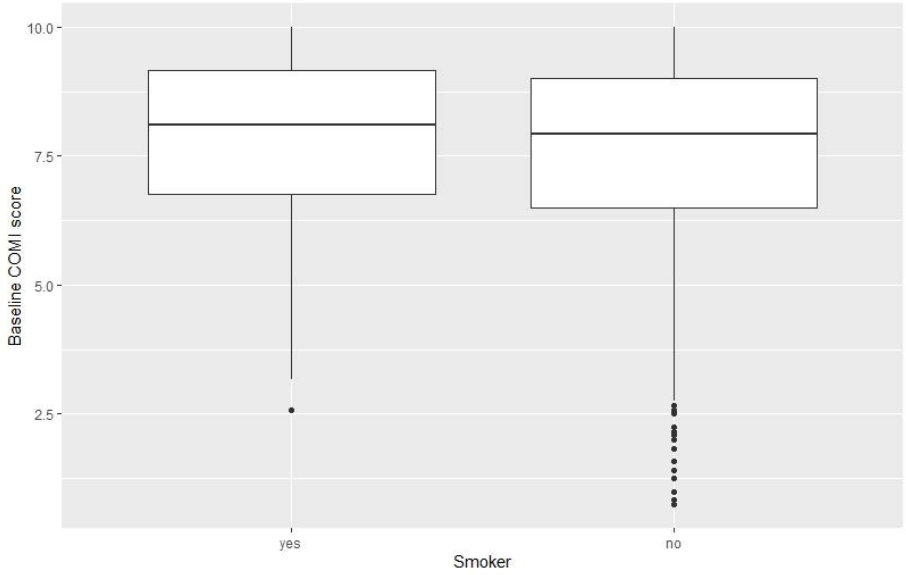


Figure 3.42: Box plot of baseline COMI scores, grouped by smoking status.

It seems that non-smokers have slightly lower baseline COMI scores, however the difference in means is very low (0.21 points on COMI score scale). It cannot be assumed that there is a dependency

between baseline COMI scores and smoking status. In Figure 3.43, median, range, as well as lower and upper quartiles of baseline COMI scores are visualised, grouped by the occurrence of complications during surgery.

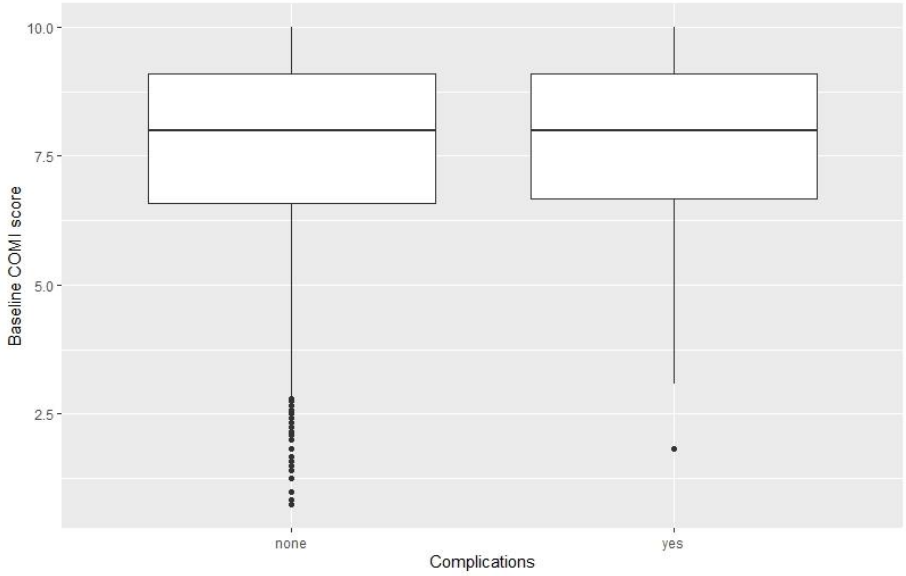


Figure 3.43: Box plot of baseline COMI scores, grouped by complications.

Similar to previous cases, it cannot be assumed that there is a dependency between baseline COMI scores and the occurrence of complications during surgery.

In summary, the only considerable possibilities of associations between baseline COMI scores and other variables were for country IDs, ASA morbidity and surgeon credentials. This needs to be kept in mind in later analyses.

Analyses if any of the patient characteristics were associated with treatment outcome, is subject of Chapter 5.

3.6 Comparison of patient characteristics and COMI outcomes between the NERVES trial and the Spine Tango registry

The following variables are directly comparable between the two data sources: age, sex, weeks of symptoms, level of disc, BMI, and COMI scores. These are compared in Table 3.17.

Variable		NERVES trial	Spine Tango registry
Age	mean (s.d.)	41.94 (9.046)	47.38 (14.122)
Sex	female	63 (50.81%)	8,033 (46.56%)
	Male	61 (49.19%)	9,219 (53.44%)
Weeks of symptoms	mean (s.d.)	21.38 (10.164)	NA
	<3 months	27 (21.77%)	727 (31.07%)
	3-12 months	97 (78.23%)	415 (17.74%)
	>12 months		1,198 (51.20%)
Level of Spine	L5 / S1	75 (60.48%)	7,510 (43.53%)
	L4 / L5	47 (37.90%)	7,571 (43.88%)
	L3 / L4	2 (1.61%)	1,370 (7.94%)
	L2 / L3		442 (2.56%)
	L1 / L2		30 (0.17%)
	Other		329 (1.91%)
BMI		27.83 (5.982)	28.394 (8.590)
COMI score	Baseline	7.158 (s.d. 1.686)	7.688 (s.d. 1.744)
		18 weeks: 3.333 (s.d. 2.581)	3 mon.: 4.642 (s.d. 2.937)
		30 weeks: 2.878 (s.d. 2.423)	6 mon.: 4.518 (s.d. 2.814)
		42 weeks: 2.796 (s.d. 2.144)	9 mon.: 4.459 (s.d. 2.858)
		54 weeks: 2.779 (s.d. 2.434)	1 year: 3.935 (s.d. 3.049)

Table 3.17: Comparable statistics from the NERVES trial and the Spine Tango registries. Each statistic is based on a complete case analysis.

Patients in the Spine Tango registry were slightly older than in the NERVES trial. In the registry data set, there were more men than women, whereas in the NERVES trial there were more women than men. However, both sources were close to a 1-1 distribution of sex. The Spine Tango registry categorized the duration of symptoms, whereas the NERVES data set did not. As a result, the mean and standard deviation could not be computed for the registry data. However, the NERVES data was categorized to enable a comparison of the distribution. In the NERVES trial, patients were selected based on the duration of their symptoms, and those with symptoms lasting more than 12 months were excluded. Consequently, the distribution of categories differs significantly between the two data sources. In both data sources, S1/L5 and L5/L4 were the most frequent locations.

In the NERVES trial, COMI outcomes were collected in specific time points (baseline, 18 weeks, 30 weeks past randomisation etc), whereas in the registry data set, they were categorised in weeks or years past surgery. However, the following pairs of intervals were regarded as similar and comparable:

18 weeks and 3 months, 30 weeks and 6 months, 42 weeks and 9 months, and 54 weeks and 1 year. The means and 95% confidence intervals for these intervals are compared in Figure 3.44.

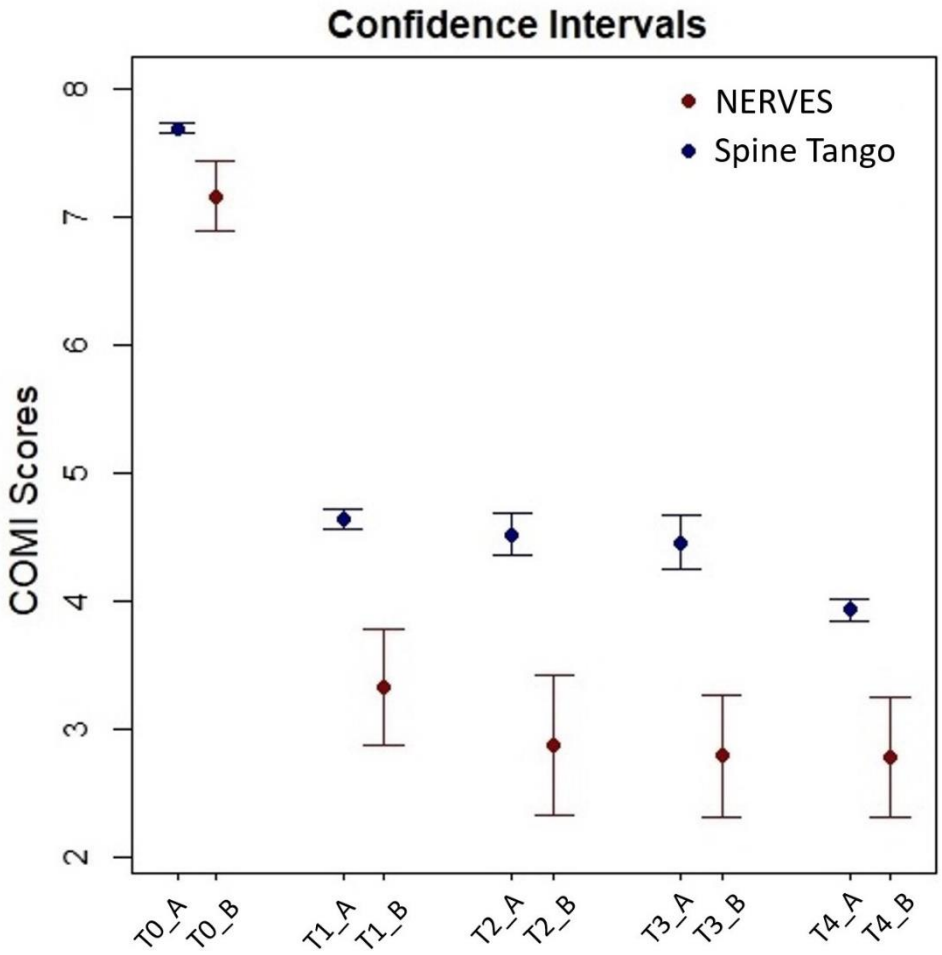


Figure 3.44: Means and 95% confidence intervals of COMI scores at each collected timepoint (blue = Spine Tango, red = NERVES. Time-point coding: T0_A = Baseline of Spine Tango data set, T0_B = Baseline of NERVES dataset, T1_A = 3 months in ST, T1_B = 18 weeks in NERVES, T2_A = 6 months in ST, T2_B = 30 weeks in NERVES, T3_A = 9 months in ST, T3_B = 42 weeks in NERVES, T4_A = 1 year in ST, T4_B = 54 weeks in NERVES.

Because of the larger sample size in Spine Tango (even for the less frequently collected time points) the confidence intervals are smaller than for the NERVES data set. However, in Table 3.17 shows that variation in the Spine Tango data set was higher. Both baseline and follow-up means are higher in the Spine Tango data set, compared to the RCT environment of the NERVES trial. The reasons for this should be addressed in future research. There are a number of possible reasons for this, including difference in patient population (international real-world practice vs RCT with inclusion criteria), which can be seen in the difference in the duration of symptoms. Additionally, the patient population in the NERVES trial was UK only, whereas Spine Tango is an international registry.

3.7 Summary

The purpose of this chapter was to compare the routinely collected data from the Spine Tango registry and the NERVES trial. The focus was on a thorough descriptive analysis, as well as an analysis of dependencies between collected patient covariates in each of the sources. The data from the Spine Tango registry and the NERVES trial differed in several ways. Already in their set-up there is one obvious difference, which is that the NERVES trial compared outcomes between microdiscectomies and transforaminal epidural steroid injections (TFESI). Hence, the study population was randomised and split into two treatment arms. The Spine Tango registry is a database on which observational data from surgeries are routinely collected.

Regarding the collection of variables, the NERVES trial collected data on sex, age, prolapsed disc volume, BMI, symptom duration, and spinal level. The Spine Tango registry encompassed a broader spectrum of variables, including sex, age, surgeon credentials, country ID, spinal level, ASA morbidity status, smoking status, and BMI.

Upon analysing these datasets, notable observations emerged. Both datasets featured an equal distribution of male and female patients, along with similar mean BMI values across the entire population. However, there were distinct variations in the distribution of spinal surgery levels between the NERVES trial and the Spine Tango registry. Moreover, patients within the Spine Tango registry exhibited higher average ages and COMI scores. Both datasets captured key outcomes like COMI, ODI, and complications, aligning with the findings in Chapter 2. Interestingly, in the Spine Tango dataset, variations in collected outcomes were observed between countries, possibly linked to reporting guidelines.

Regarding data completeness in outcomes, the NERVES trial displayed a low rate of missing data, with only 17.18% of COMI questionnaires missing for the 18 weeks after randomization. In contrast, the registry data had 46% missingness for baseline questionnaires, and 41% of those with available baseline data had no follow-up data. Questionnaires were collected in more time intervals in the registry dataset, although there seem to be main check-in times such as 3 months, 1 year and 2 years after surgery. Within the Spine Tango dataset, a distinct pattern of missingness in patient covariates (BMI, smoking status, duration of symptoms, previous treatment) was observed. These gaps in the covariate data were attributed to changes in the data collection forms over time, leading to inconsistent reporting.

While complications were rare in the NERVES trial, the extensive data in the registry allowed for a more in-depth exploration of complication occurrence, potentially associated with prior surgeries.

This chapter's findings emphasize the comparability not only in terms of patient population but also the collected measurements and their corresponding descriptive statistics. However, the differences in collected covariates between the two sources highlight the importance of establishing a core set of covariates to ensure comparability across data types within this patient population.

Importantly, this chapter lays the groundwork for subsequent analyses, highlighting the registry's potential in more complex models to predict patient outcomes. Up to this point, associations between patient covariates and COMI scores at baseline were investigated. Dependencies between variables such as country ID and previous treatment or surgeon credentials (potentially due to differences in guidelines and reporting standards between countries) as well as age and level of spine or ASA morbidity, need to be considered in further analyses. The large amount of data in the registry allows for the use of more complex models that can consider a wider range of variables and provide more accurate prognostic factor analysis. In Chapter 5, further analysis will be conducted about COMI scores and their dependency on baseline variables using prognostic factor analysis.

As seen in this chapter, outcomes in the Spine Tango registry were not collected consistently, since patients are not followed up as rigorously as in a RCT. It is therefore of crucial importance to explore the performance of imputation methods that could deal with this missing data. The following chapter will conduct a simulation study, using real-world data from the Spine Tango registry. It aims to determine if data should be imputed at the item-level of the COMI questionnaires or at the score-level. Multiple missingness scenarios will be considered, along with varying cut-off points to identify when a questionnaire should be considered missing. By determining the optimal approach for handling missing data, valuable recommendations can be derived for sciatica-affected patients that consider undergoing microdiscectomy.

Chapter 4: How to handle missingness of values in data from registries regarding Patient-Reported Outcome Measures (PROMs)

4.1 Chapter Outline

The aim of this chapter is to identify an appropriate method for imputing missing data in patient-reported outcome measures in routinely collected data. It is expected that data collection could be affected by substantial missing data in both outcomes and baseline characteristics. These gaps would reduce a complete case analysis (CCA) sample size significantly. Especially missingness of items or complete questionnaires of the assessed quality of life is investigated. To do this, a simulation study using data from the Spine Tango registry was conducted. The study used patients with complete outcome and baseline questionnaires as a basis, and artificially introduced missing data at both the item and questionnaire levels. This simulation covered several parameters, such as the mechanism of missingness, the method of imputation (at the item and questionnaire score level). Investigating the precision of imputation methods of questionnaire data provides insights of the data collection in this registry and can support a recommended cut-off point of questionnaire items. By identifying a recommended method for imputing questionnaire data, subsequent analysis can be done on larger sample sizes.

4.2 Introduction

Data collected from registries may have a higher rate of missing data in comparison to data gathered from randomized controlled trials (RCTs). This is possibly due to the collection of data over extended time periods. Any changes in collected patient characteristics, protocols, or the addition of new countries or sites participating in the registry can result in inconsistent variables within the dataset. Especially regarding outcomes this could potentially reduce the available sample size. In RCTs, outcomes are collected at previously defined time points, whereas in registries, timepoints may be dependent on preferences of patients and clinical sites.

When only small percentages of values are missing, analysing the subset of complete observations (Complete Case Analysis) can be valid. However, it assumes the missingness of data points to be missing completely at random (MCAR), which can be ruled out in most clinical trials (Pedersen et al., 2017). Imputation is a statistical method that is used to estimate missing data. It involves using available data to make reasonable assumptions about the missing data, and then using these assumptions to fill in the gaps in the data. There are several different methods of missing data imputation, each of which has its own strengths and limitations. Some common methods include: mean imputation, single imputation, or multiple imputation (Austin et al., 2021, Mehrotra et al., 2017, Zhang, 2016). In general, missing data imputation can help to reduce the impact of missing data on the accuracy and reliability of PROMs, and can provide a more complete and more accurate picture of the patient's condition and treatment response. If the number of complete cases is less than 90%, imputation of missing data leads to more accurate statistical results (Eekhout et al., 2014).

Throughout this work, data from the spinal registry Spine Tango is used not only for analysis and the development of prognostic models, but also to study methods of imputation in the setting of routinely collected data. To date, over 750,000 forms (134,458 surgery forms) from five continents have been collected (EUROSPINE, 2022b). The most commonly patient reported outcome measure for quality of life and lower back and leg pain was the Core Outcome Measures Index (COMI), which was recorded pre-surgery and at follow-up visits. It is a quality-of-life (QoL) questionnaire based on the items proposed by an expert group for the use in clinical routine, quality management and research (Deyo et al., 1998). It covers not only pain intensity and its effect on quality of life, but also allows patients to report complications, overall satisfaction and further surgeries.

The population in focus are patients that underwent microdiscectomy due to disc herniations. Although being one of the most measured outcomes, COMI scores had high percentages of missingness, in baseline, but also past surgery.

The focus of this chapter is, whether one should impute single missing items in questionnaires and then compute the total score of the patients (even if all items are unanswered), or if one should impute the total score directly. Commonly, the evaluation of questionnaires allows for a few items to be missing and still calculate a percentage score of answered items, but it will consider the complete outcome as missing if item-missingness is too large. Several simulations will be performed to investigate how to choose such cut-off points appropriately, with a variety of missing data percentages. There are three types of mechanisms of missing data, missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), which will be explained in detail. The simulation procedure will be performed for each of these mechanisms to find out if there

is an overall superior imputation method. The focus of this chapter is to investigate imputation of baseline questionnaires. The general goal was to explore to what extent this could result in accurate imputations and to investigate if the imputation of single items could reduce the overall missing data of questionnaire scores, especially in cases where the cut-off threshold is sensitive. Outcome questionnaire imputation was additionally explored, although this can be considered as ethically controversial and will probably not find application in real-world scenarios.

4.3 Mechanisms of missingness

Missing data can have a significant impact on the accuracy and reliability of patient-reported outcome measures (PROMs), as it can lead to bias and distorted results. This can be particularly problematic in the context of low-back pain, where the condition can be complex and difficult to assess accurately.

There are several mechanisms that can cause missing data in patient-reported outcome measures (PROMs) and understanding these mechanisms can help to identify the appropriate methods for handling missing data. In general, three types of mechanisms are defined by the following classifications.

4.3.1 Missing completely at random (MCAR)

The mechanism causing missing data does not depend on either observed or unobserved data. In this case, a subset of the data would be representative for the entire dataset. Analysis results would have larger standard deviations, but there would not be any systematic error. This mechanism is very rare though, especially in clinical research (Arnab, 2017).

4.3.2 Missing at random (MAR)

In this mechanism, missing values are not directly related to the variable being measured, but rather are influenced by other observed variables. For example, if data is collected on patients' ages and genders during a trial and some patients have missing age data, this mechanism suggests that the cause of the missing values may not be related to the age variable itself, but rather may be related to the patient's gender. If this type of missing data is not properly accounted for in subsequent analyses, it could lead to biased results. Techniques such as multiple imputation can help to reduce this bias, but it is not always possible to confirm that the data meet the criteria for this mechanism (MAR). As a result, sensitivity analyses may be needed to evaluate the potential impact of missing not at random (MNAR) data on the estimated results” (Jakobsen et al., 2017).

4.3.3 Missing not at random (MNAR)

When missing values depend not only on other observed variables, but also on the variable itself that has missing values, the data are classified as missing not at random (MNAR). Using the previous example, this would mean that a missing value for the age variable may be influenced by both the patient's gender and their age. It is not possible to definitively determine, based on the observed data alone, whether the data meet the criteria for MNAR or MAR. No statistical method can completely account for the potential bias that may be introduced by MNAR data, as these methods rely on assumptions that cannot be tested using only the observed data (Jakobsen et al., 2017).

MAR and MNAR are common mechanisms of missing data in clinical studies, and it is important to properly address these issues to avoid bias in the analysis results. If these mechanisms are not handled appropriately, the estimated benefits and harms of a treatment may be inaccurate. While advanced imputation methods can help to mitigate the impact of missing data, a high percentage of missing data can still reduce the representativeness of the sample and bias any estimates produced through analysis.

The appropriate method for handling missing data will depend on the mechanism of missingness. For example, if the missing data is missing completely at random (MCAR), then simple methods such as mean or median imputation may be sufficient. However, if the missing data is missing at random (MAR) or missing not at random (MNAR), then more complex methods such as multiple imputation may be needed to accurately estimate the missing data. It is important to carefully consider the mechanism of missingness when developing a plan for handling missing data in PROMs.

Even with strict trial conduct and protocols, as it is the case in prospective randomised clinical trials (RCT), there usually occurs some missing data which can have several reasons. For example, unreadable answers that could not be transcribed, patients not answering questions due to individual reasons or that patients could not be contacted anymore. Especially difficult to handle are missing data that are missing due to an underlying cause which is directly connected to the trial purpose. As example, consider a survey about depression. Regarding the patients who did not answer some of the survey questions or did not respond at all, it is likely that their level of depression is associated with their response. Missing data can therefore seriously compromise analysis results if not handled appropriately. The higher the percentage of missing values, the more biased a complete case analysis would be, and in the above example the effect of depression would be underestimated, because rather severe cases were left out (Jakobsen et al., 2017). When dealing with routinely collected data, percentages of missingness could be even higher. It is therefore important to use appropriate techniques to impute missing data. Methods of imputation

One generally distinguishes between “single imputation” and “multiple imputation”. Single imputation is a method used to handle missing data by replacing the missing values with single estimated values. One commonly used approach in single imputation is to impute the missing value with the mean, median, or mode of the non-missing values for that variable. However, this can result in an underestimation of variance of the variable, which is why this method should be used with caution (Pedersen et al., 2017). The most common approaches of single imputation will be reviewed in the following.

4.3.4 Single imputation

4.3.4.1 Last observation carried forward

This method imputes a missing value with the last observation of the individual and is therefore specific for longitudinal data. It assumes though, that the observation has not changed since the last measured observation, which is most often unrealistic, and leads to an underestimation of treatment effects.

4.3.4.2 Single imputation: Regression imputation

In this method, the imputed value is predicted from a regression equation that is obtained by all complete observations. This implies though, that imputed values fall directly on a regression line with non-zero slope and therefore a correlation of 1 between predictors and outcome. This can be adjusted by adding a small error term to the equation which preserves variability and estimated parameters are less biased.

4.3.4.3 Single imputation: Mean imputation

Mean imputation replaces a missing value by the mean of all other available cases. This is a straightforward and easy method but leads to underestimated variances of the observations, as well as covariances and correlations. Therefore, this method often causes biased estimates, irrespective of the underlying missing data mechanism (Eekhout et al., 2014).

4.3.5 Multiple Imputation

Multiple imputation is another statistical method used to handle missing data by generating multiple plausible values for the missing observations. It involves generating multiple datasets, each with differing imputed values. Each dataset containing imputed values is analysed separately and the results are then combined using Rubin’s Rules. One popular method within MI is the Chained Equations approach. This approach iteratively imputes each missing variable conditional on the

observed values of the other variables, allowing for complex relationships between variables to be captured during the imputation process.

In the following, a brief overview of the technique using chained equations (as used in the R-package `'mice'`) is provided, as detailed by Azur et al. (Azur et al., 2011).

- Step 1: Initially, a simple imputation technique, such as imputing the mean or median, is applied to all missing values in the dataset, creating 'place holder' imputations.
- Step 2: Subsequently, the 'place holder' imputations for one variable ('var') are reverted to missing.
- Step 3: The values of the variable 'var,' as of Step 2, undergo a regression analysis against the remaining variables in the dataset.
- Step 4: Predictions (imputations) for the missing values of 'var' are derived from the regression model. These imputed values are used when 'var' is employed as an independent variable in subsequent regression models.
- Step 5: Steps 2–4 are repeated for each variable with missing data, constituting one iteration or 'cycle.' After one cycle, all missing values are replaced with imputed values generated from regressions reflecting the observed data relationships.
- Step 6: Steps 2–4 are reiterated for several cycles, with imputations being updated at each cycle.

It's important to note that while Multiple Imputation can be a powerful method for handling missing data, it doesn't guarantee perfect imputations. Careful consideration of the underlying data and the assumptions of the imputation models is essential for accurate and valid results.

In many cases the multiple imputation technique is considered to be more accurate and reliable than other methods of missing data imputation, because it considers the uncertainty associated with the missing data (Jakobsen et al., 2017). It is particularly useful in situations where the missing data is not missing completely at random, as it can provide a more accurate estimate of the missing data by considering the relationships between the observed and missing data. There exist many different types of multiple imputation methods and covering all of them would exceed the scope of this project. Most of the methods differ in their choice of “plausible values” regarding the type of variable (continuous, binary, categorical etc).

The multiple imputation package `'mice'` in R has standard techniques for numeric data, factor data with 2 levels, factor data with 2 or more unordered levels and factor data with 2 or more ordered levels (van Buuren and Groothuis-Oudshoorn, 2011). It detects the type of variable, and uses an

appropriate method, but this can be specified and individualized. The number of plausible values and iterations can also be adjusted. There are several methods that can be selected such as classification and regression trees, random forest imputation or predictive mean matching (PMM).

Predictive mean matching is a technique for filling in missing values in a dataset by using values from other observations that are similar (Morris et al., 2014). It is a specific method for Steps 3 and 4 in the previous overview. The process involves calculating the mean of each variable for each individual in the dataset and using these means to predict the missing values for each individual. To fill in a missing value, the method searches for other individuals in the dataset with similar mean values for the variables that are not missing, and then takes a random sample of these similar individuals. The observed values for the missing variable from this sample are used to fill in the missing value. This method preserves any skewedness of the imputed variable, boundaries of its values and detects if the variable is discrete or continuous. It includes the use of the linear regression model of the form

$$Y = X\beta + \varepsilon.$$

The variables of this formula are explained as follows:

- Y is a vector of outcomes of length N (N is the number of total patients).
- X is a p-dimensional vector of patient covariates where p is the number of included patient baseline covariates (including an intercept).
- β is p-dimensional vector of coefficients.
- ε is a N-dimensional vector of error terms, each of which are each assumed to be normally distributed (all ε_i are assumed from the same distribution).

The following description outlines the essential steps of this process, which have been adapted from the following sources: (Morris et al., 2014, StatisticsGlobe, 2022, Vink et al., 2014).

Step 1: Begin by estimating a linear regression model as follows:

- Utilize the variable to be imputed, denoted as Y, along with a well-selected set of predictors, denoted as X.
- Limit the analysis to complete cases, employing X and Y to estimate the model and derive the coefficients represented by b.

Step 2: Proceed by drawing random samples from the posterior predictive distribution of b, yielding a new set of coefficients denoted as b^*. For a comprehensive explanation of this Bayesian step, please refer to (Yuan, 2005).

Step 3: Calculate predicted values for both observed and missing values of Y:

- Use the b coefficients to compute predicted values for the observed Y.
- Employ the b* coefficients to compute predicted values for missing Y.

Step 4: For each case where Y is missing, identify the closest predicted values (typically three) among cases where Y is observed. To illustrate, consider the following example.

- Y_i is missing, with its predicted value calculated as 10 based on b*.
- The dataset includes five observed cases of Y with values 6, 3, 22, 7 and 12.
- Predicted values based on b for these five observed cases are 7, 2, 20, 9 and 13.
- The algorithm selects the three closest values to the missing Y_i, which are 7, 9 and 13.

Step 5: Randomly select one of these three close cases and impute the missing value Y_i with the observed value of this close case. For example:

- The algorithm randomly draws from 6, 7, and 12 (the observed values corresponding to the predicted values 7, 9, and 13).
- The algorithm selects 12 and substitutes this value for Y_i.

Marshall et al. (Marshall et al., 2010) compared a variety of imputation methods. They concluded, that multiple imputation using PMM was the preferred choice, at least for missingness up to 50%. This method will be used in the following simulation study in which imputation of items in questionnaires and imputing scores are compared in different scenarios of mechanism of missingness, probability of missingness and cut-off points of questionnaire score calculation.

4.4 Literature review

Eekhout et al. conducted a literature review regarding the reporting of missing data in questionnaires and the methods used for analysis. It was found that 78% lacked clear information about the measurement instruments, and although advanced techniques such as multiple imputation are available, CCA was the most frequently reported method (81%). It is a viable method if the missing data percentages are low, but for higher percentages, it reduces the power of analysis since the sample size is reduced. CCA also assumes that missing data occurs completely at random (MCAR), which can often be ruled out depending on the clinical background. The comparison between missing data methods for item-level and total score-level missingness in questionnaire data is seldom made in a single study (Eekhout et al., 2012).

In 2010, Marshall et al., conducted a simulation study regarding missing covariate data techniques in the development of prognostic models. For multiple scenarios of missingness percentages, data were generated and then imputed using single imputation, multiple imputation or CCA. A Cox proportional

hazard model fit was used to compare the performance of the imputation techniques. CCA showed unbiased regression estimates, but inflated standard errors, which affected the significance of the covariates in the model. It was shown that single imputation techniques underestimate variability. Multiple imputation using the predictive mean matching (PMM) method produced the least biased estimates. However, when 50% or more cases had missing data, underlying the missing completely at random (MCAR) or missing at random (MAR) mechanism, regression coefficients were still biased. As for missing not at random (MNAR) this bias occurred where 10% or more cases were incomplete (Marshall et al., 2010). Several studies focused only on imputation of items but did not compare them to methods of imputation of complete scores (Burns et al., 2011, Buuren, 2010, Hawthorne and Elliott, 2005, Roth et al., 1999). For example, Burns et al. investigated item imputation of a questionnaire for cognitive status in dementia, where missing item-level data are frequently reported. A simulation was conducted, which found multiple imputation (MI) to be the superior method to estimate missing items, although "serious decrements in estimation occurred when 50% or more of item-level data were missing" (Burns et al., 2011).

Eekhout et al. conducted a simulation study to compare item-level versus score level imputation (Eekhout et al., 2014). Data from an RCT regarding low-back pain, specifically the questionnaire Pain Coping Inventory (PCI), was used to create data sets for the simulation. It was found that MI methods at item-level outperformed models applied to total scores.

This study focusses on data from an international registry, therefore observational and routinely collected data and specifically the low-back/leg pain questionnaire COMI for patients with sciatica.

4.5 Design of simulations

Simulation studies are used in order to investigate the behaviour of statistical methods. They use generated data sets and allow to quantify bias and resilience of the used methods in different scenarios (varying sample size or other parameters). Although the design, analysis, presentation and reporting of such simulation studies should be done rigorously in medical data science, many pointed out that researchers still fail to do (Burton et al., 2006, Hauck and Anderson, 1984, Hoaglin and Andrews, 1975).

Morris et al. recommend to systematically approach a simulation study by using the 'ADEMP' structure (**A**ims, **D**ata-generating mechanism, **E**stimands, **M**ethods, **P**erformance measures) in order to cover all important aspects (Morris et al., 2019). The aims are typically about estimating the performance for different sample sizes, variance estimations, robustness or misspecification of different methods, but could also be proof-of-concept or to investigate extreme cases in which methods fail. Data-

generating mechanisms should be as close to the real-life data for which the methods will be used later and can either be an appropriate underlying probability distribution from which new data will be sampled, or drawing with replacement and therefore bootstrapping actual real-life data to the desired sample size. The choice of mechanism depends on the projects aims and the availability of a parametric model. Estimands are quantities that are used to compare the performance of different methods. These quantities may be model parameters, outcome measures, hypothesis power, prognostic ability, or some other metric depending on the goals of the project. 'Methods' is a rather generic term that can refer to a model for analysis or some procedure such as a decision rule or as in our case, the method of imputation of missing data. Performance measures are numerical quantities used to assess the performance of a method. These measures may include the variance of estimates, bias, coverage, degrees of freedom, and others. One important performance measure is the Monte Carlo error, which is defined as the standard deviation of the simulated estimates divided by the square root of the number of simulations. This measure should be low (e.g., less than 0.05) in order to ensure that the number of simulations was sufficient.

The aim of this simulation study is to identify an appropriate method for imputing missing data in patient-reported outcome measures in routinely collected data. Especially missingness of items or complete questionnaires of the assessed quality of life is investigated. It will be investigated if COMI questionnaires should be imputed (if at all) on an item- or score-level. This imputation is considered for baseline questionnaires. Additionally, even the performance for the imputation of 3-month outcome questionnaires is considered. However, this was solely out of curiosity and would raise ethical concerns if applied in real-life.

Instead of creating a simulation dataset from scratch, the dataset of patients from the Spine Tango registry that had complete data regarding COMI questionnaires at baseline (6,008 patients) was used. This assures that the investigated methods are applicable to a real-world scenario. Missing data is then artificially introduced on both the item and questionnaire level. This missingness can be created in several ways. The parameters that were considered were the probability of missingness in an individual and the mechanism of missingness (MCAR, MAR, MNAR). The computation of COMI scores often allows for single items to be missing, so that a COMI score would still be calculated if only e.g. 7 items were answered (by averaging over the number of answered items). Some studies define a cut-off point of possible missing items, the excess of which would result in the entire questionnaire to be considered missing. If sensitive, this cut-off point can reduce the sample size of available scores and therefore the power of analyses performed on the data set. However, score calculation on a small subset of items could lead to biased scores. This choice of cut-off point is also subject of investigation in this study, to identify a recommended cut-off point for the COMI questionnaire.

The main estimand is the accuracy of the methods, namely item-wise and score-wise imputation, in terms of their ability to recover missing items and maintain the overall population statistics, such as the mean and standard deviation of the baseline COMI (Core Outcome Measure Index) scores. Specifically, this accuracy is assessed by the root-mean-square error (RMSE). This measure is defined as the square root of the mean of the squared difference between the estimated and true COMI scores. The RMSE is computed for each simulation iteration, and subsequently, the mean RMSE across all iterations is determined for each combination of scenarios (of probability of missingness, mechanism of missingness and cut-off point). Furthermore, we compare the mean and standard errors of population COMI scores, averaged over all simulation iterations within a given scenario, to the true mean and standard error values obtained before introducing artificial missingness.

As performance measure the Monte Carlo error was considered, to make sure that the number of simulations was sufficient for reliable results. Additionally, the feasibility of the methods was considered in terms of time of computation, which can be an important factor in large data sets.

For each mechanism of missingness N simulations for each combination of probability of missingness and choice of cut-off point will be run. For each of these iterations, missing data will be introduced using the `ampute()` function of the `'mice'` R-package and then imputed by both item-wise and score-wise imputation of COMI. Imputation will be performed using the `mice()` function of the `'mice'` R-package, in which the number of computed datasets with imputed data can be specified by the parameter `m`. This parameter will be set to the number N of simulation iterations.

When imputing item-wise, all items of the questionnaire, independent of the number of missing items is imputed and the score afterwards calculated. In this method, the cut-off point is not relevant. For the score-wise imputation, the scores will first be calculated regarding the cut-off point of the iteration. For those individuals who had too much missingness of a questionnaire, the score will be considered missing and imputed directly.

An overview of the scenarios of this missingness is provided in Figure 4.1.

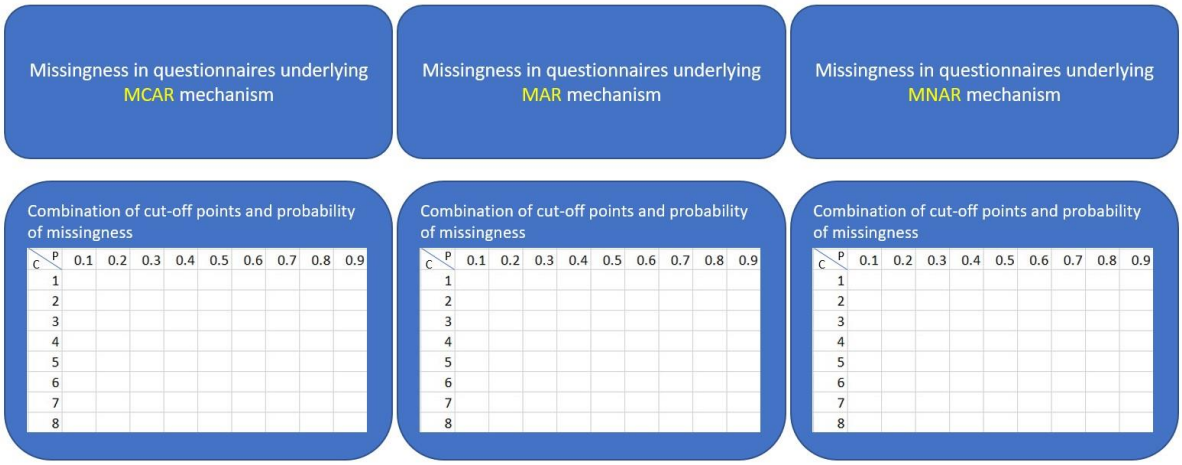


Figure 4.1: Overview of simulation scenarios. C = Cut-off point for the calculation of COMI scores, P = probability of missingness in an individual.

4.5.1 Patient population

Not all patients had COMI questionnaires available. For details of the total patient population that was identified according to the clinical definition of sciatica, see Chapter 3. The subset of patients that had complete baseline COMI questionnaires available was 6,008 patients. However, missing values of patient covariates were present. Only considering patients with complete data, would reduce the sample size further to 4,312 patients. To maintain the larger dataset for this simulation study, the missing values of the patient characteristics were imputed using predictive mean matching. In this case, MI was used to create one single dataset without missing data, by only iterating the "`mice()`" function once (`m=1`). It cannot be assumed that the results of any subsequent analyses are not affected by using this technique to create this initial dataset. Nevertheless, this was used to maintain the sample size. The same simulations were performed on the dataset of patients that had all variable complete (4,312) as a sensitivity analysis to support the robustness of the results. The result of the same simulations on the complete case analysis set are included in Appendix F.

An overview in form of descriptive statistics of the dataset before and after imputing missing patient covariates is shown in Table 4.1.

Variable		Before imputation	After imputation
Sex	Female	2,822 (46.97%)	2,822 (46.97%)
	Male	3,186 (53.03%)	3,186 (53.03%)
Age		Mean 49.02 (s.d. 14.57)	Mean 49.02 (sd 14.57)

Surgeon cred.	Board-certified neurosurgeon	2,785 (46.35%)	2,860 (47.60%)
	Specialized spine surgeon	2,139 (35.60%)	2,217 (36.90%)
	Neurosurgeon in training	589 (9.80%)	603 (10.04%)
	Board-certified orthopedic surgeon	212 (3.53%)	217 (3.61%)
	Orthopedic surgeon in training	80 (1.33%)	84 (1.40%)
	Other	26 (0.43%)	27 (0.45%)
	Missing	177 (2.95%)	
Country ID	A	1,945 (32.37%)	2,210 (36.78%)
	B	2,381 (39.63%)	2,520 (41.94%)
	C	23 (0.38%)	27 (0.45%)
	D	426 (7.09%)	431 (7.17%)
	E	248 (4.13%)	254 (4.23%)
	F	194 (3.23%)	196 (3.26%)
	G	138 (2.30%)	144 (2.40%)
	H	108 (1.80%)	112 (1.86%)
	Other	110 (1.83%)	114 (1.90%)
	Missing	177 (2.95%)	
Level of Spine	L5/S1	2,637 (43.89%)	2,637 (43.89%)
	L4/L5	2,260 (37.62%)	2,260 (37.62%)
	L3/L4	480 (7.99%)	480 (7.99%)
	L2/L3	162 (2.70%)	162 (2.70%)
	L1/L2	7 (0.12%)	7 (0.12%)
	Other	102 (1.70%)	102 (1.70%)
Prev. Treat.	None	835 (13.90%)	993 (16.53%)
	<3 mon. conservative	1,181 (19.66%)	1,347 (22.42%)
	3-6 mon. conservative	1,335 (22.22%)	1,518 (25.27%)
	6-12 mon. conservative	854 (14.21%)	996 (16.58%)
	>12 mon. conservative	767 (12.77%)	884 (14.71%)
	Surgical	227 (3.78%)	270 (4.49%)
	Missing	809 (13.47%)	
ASA Morbidity	1	2,564 (42.68%)	2,932 (48.80%)
	2	2,325 (38.70%)	2,661 (44.29%)
	3	361 (6.01%)	407 (6.77%)
	4	7 (0.12%)	8 (0.13%)

Missing	751 (12.50%)	

Table 4.1: Descriptive statistics of baseline patient characteristics before and after imputation (N=6,008).

4.5.2 Introducing missing data

In each simulation iteration, missing data will be generated in items from COMI questionnaires using different probabilities and mechanisms of missingness, which can be specified using the `ampute()`-function in the `mice`-package in R. Amputing data is a term used in this package that generates missing data in a complete dataset.

This multivariate amputation procedure was programmed for general use by Schouten et al. (Schouten et al., 2018). Figure 4.2 shows an overview of the process with which missing values are generated.

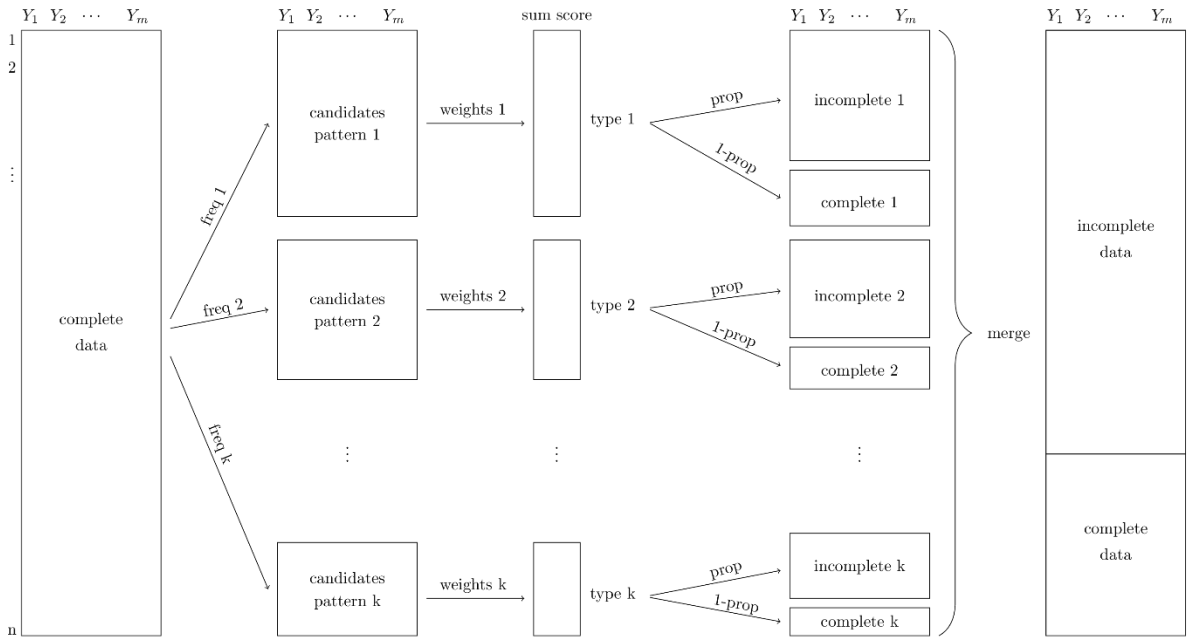


Figure 4.2: Schematic overview of `ampute`-function in the R-package mice. Figure taken from (Schouten et al., 2018) in the section "Multivariate amputation".

The dataset will be divided into k subsets, where the value of k is determined by the number of different patterns of missing data. This value can be customized to meet specific needs. In this case, a pattern for every possible way that an item could be missing in the dataset was created. For example, if there are 8 items in a questionnaire, there are 8 different ways that a single item could be missing.

If two items are missing, there are 28 different patterns of missingness. The formula for calculating the number of patterns is:

$$\frac{n!}{(n-k)!\,k!}$$

where n is the total number of items and k is the number of missing items.

In this simulation design, there were 255 patterns of missingness. It was not assumed that any particular pattern was more likely to occur than any other, so each pattern had a frequency of 1/255. However, the possibility that the entire questionnaire could be missing was also considered, which was assumed to be more likely to occur than any specific pattern of missingness. It was assumed that the total frequency of missingness for the entire questionnaire was approximately 10%. As a result, the frequency of each pattern was specified as 1/280 and the frequency of complete missingness as 25/280 (rounded for simplicity when coding frequencies).

Specifying weights is only interesting when dealing with the MAR and MNAR mechanism and will be explained in depth later. Each subset will then create missingness with the probability p regarding the pattern that specifies the subset. The other rows will be left complete. Afterwards, all subsets are merged again. This will produce a data set that has an expected number of rows (patients) of (p x number of patients) to have missingness in any way. In order to cover a wide range of scenarios, this value will be ranging between 0.1 and 0.9 with increments of 0.1 (9 scenarios). Scores will then be computed with respect to the currently defined cut-off point of when a questionnaire is considered missing. This cut-off point will have a range of 1 to 8 (8 scenarios). A cut-off point of 1 means that a questionnaire is considered missing if one or more items are missing, whereas a cut-off point of 8 means that a questionnaire is only then considered missing if all 8 items are missing. This results in a total of 72 scenarios.

4.5.3 Simulation set-up

For each scenario with different cut-off points, probabilities of missingness, and mechanisms of missingness, a total of N simulations will be conducted. The mean and a 95% confidence interval of the RMSE will be calculated for both score-based and item-based imputation. In the imputation process, the number of iterations in the "`mice()`"-function can be set to m=N in order to obtain N different datasets, each using PMM. Initial simulations showed that N=50 simulations were sufficient to achieve a Monte Carlo error smaller than 0.05.

In short, the process of each simulation is summarised in the following steps in pseudo-code:

```
# loop through each cut-off point
```

```
# loop through each probability of missingness
            # loop through each simulation
                  # generate missing data
                  Generate_missingdata(mechanism, probability)

                  # impute data set with both methods
                  impute_data_method1()
                  impute_data_method2()

                  # calculate RMSE for both methods
                  rmse1 = calc_rmse(method1)
                  rmse2 = calc_rmse(method2)

            # calculate mean and standard deviation of RMSE of
            # N simulation iterations
            mean_rmse1 = calc_mean(rmse1)
            stddev_rmse1 = calc_stddev(rmse1)
            mean_rmse2 = calc_mean(rmse2)
            stddev_rmse2 = calc_stddev(rmse2)
```

When presenting tables of RMSEs or population means, conditional formatting is used to colour high errors (large RMSEs or large error from true population mean) appropriately. Colouring boundaries are chosen to be the same for each simulation set-up, in order to provide visual aid for comparing methods. Hereby, darker colours stand for more inaccurate values. To clearly distinguish between tables displaying RMSE values and those displaying population mean estimates, the colour scheme has been differentiated, with the former represented in red and the latter in yellow. For RMSE tables, the darkest colour possible is reached at the highest error of all simulation results, whereas a theoretical RMSE value of 0 would not be coloured. For population means, the largest error of all simulations was a difference of 1.59 to the true population mean (maximum value of both baseline

and 3-month outcome simulations). The colour-code is designed so that a correct estimation would not be coloured and the higher the error from the true value, the darker the colour. Hereby the colour-code is symmetric (same increase in colouring for both under- and overestimation). Figure 4.3 shows the colour coding for both RMSE and population means.

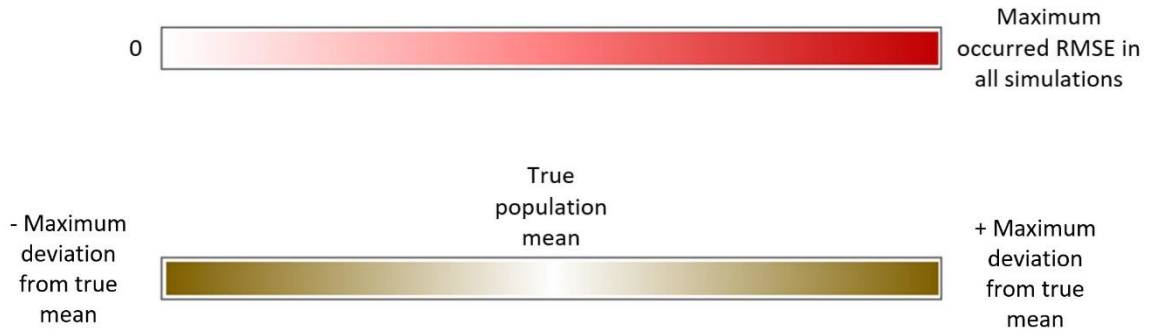


Figure 4.3: Colour coding scales for RMSE (red) and population mean (yellow).

4.6 Questionnaires at baseline for missing data underlying MCAR mechanism

In order to generate missing items in a given data set, the function `ampute()` is used. In this function the mechanism of missingness can be specified, which in this simulation will be missing completely at random (MCAR). The procedure of the simulation is designed according to the script above, with a number of simulations of N=50 (as previously mentioned this was sufficient to achieve a Monte Carlo error smaller than 0.05) for each scenario combination (72) of cut-off points and probability of missingness.

The results regarding RMSE are shown in Table 4-2 and 4-3. Each table shows each combination of scenarios for the RMSE using item-based imputation and score-based imputation.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	0.42	0.59	0.71	0.83	0.92	1.01	1.09	1.17	1.25
2	0.41	0.58	0.71	0.83	0.92	1.01	1.09	1.17	1.24
3	0.41	0.58	0.71	0.83	0.93	1.01	1.10	1.17	1.24
4	0.42	0.58	0.72	0.82	0.92	1.01	1.09	1.17	1.24
5	0.41	0.58	0.71	0.82	0.92	1.01	1.09	1.17	1.24
6	0.42	0.58	0.72	0.82	0.92	1.02	1.09	1.18	1.24
7	0.41	0.58	0.71	0.83	0.92	1.01	1.09	1.17	1.24
8	0.41	0.59	0.71	0.82	0.93	1.01	1.09	1.16	1.24

Table 4.2: Mean of RMSE of each combination of probability of amputed missingness and cut-off point, using item-based imputation method. RMSE was averaged over number of simulations (N=50).

In Table 4.2 the means of the RMSE are shown for the item-based imputation method. It becomes clear that the cut-off point of questionnaires does not matter in this method, since every item is imputed and scores are calculated afterwards. Standard deviations of the RMSE (simulation estimand) were rather low, with a range of 0.014 and 0.024. The goal was to achieve a Monte Carlo error (standard deviation of estimand divided by number of simulations) of less than 0.05. Having 50 simulation iteration makes sure that this is achieved, even if standard deviations become much larger.

In Table 4.3, the RMSE mean of each combination is shown for the score-based imputation.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	0.75	1.08	1.31	1.52	1.71	1.86	2.02	2.16	2.30
2	0.71	1.02	1.25	1.45	1.62	1.79	1.92	2.07	2.21
3	0.64	0.92	1.12	1.31	1.47	1.62	1.75	1.89	2.02
4	0.55	0.78	0.96	1.11	1.26	1.39	1.52	1.63	1.75
5	0.48	0.68	0.84	0.98	1.10	1.21	1.33	1.42	1.53
6	0.46	0.65	0.81	0.94	1.05	1.16	1.25	1.36	1.44
7	0.46	0.65	0.79	0.93	1.04	1.15	1.25	1.34	1.43
8	0.45	0.66	0.80	0.92	1.05	1.15	1.25	1.34	1.43

Table 4.3: Mean of RMSE of each combination of probability of amputed missingness and cut-off point, using score-based imputation method. RMSE was averaged over number of simulations (N=50).

There are two types of pattern visible for the score-based imputation, when analysing the table of mean RMSEs. Errors become larger, the higher probability of missingness (similar to item-based imputation), but additionally, the errors also become larger, the smaller the cut-off point of questionnaires. This is due to the lost information that is introduced by disregarding available items of uncompleted questionnaires. Score imputation does not accurately impute these questionnaires and it would be better to take scores based on the items that are available. Standard deviations of the RMSE (simulation estimand) were again low with a range of 0.018 to 0.046.

How these two methods performed for the calculation of the population mean and standard deviation of baseline scores will now be analysed. For each combination the population mean was calculated and presented in Table 4.4 and 4.5. It needs to be reminded, that the true population baseline COMI score mean was 7.69 (s.d. 1.74).

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	7.69	7.69	7.69	7.69	7.69	7.69	7.70	7.70	7.71
2	7.69	7.69	7.69	7.69	7.69	7.69	7.69	7.69	7.70
3	7.69	7.69	7.69	7.69	7.70	7.69	7.69	7.69	7.70
4	7.69	7.69	7.69	7.69	7.70	7.70	7.70	7.70	7.70
5	7.69	7.69	7.69	7.69	7.70	7.69	7.70	7.70	7.70
6	7.69	7.69	7.69	7.70	7.69	7.69	7.70	7.69	7.70
7	7.69	7.69	7.69	7.69	7.69	7.70	7.70	7.70	7.70
8	7.69	7.69	7.69	7.69	7.69	7.69	7.70	7.70	7.70

Table 4.4: Estimated population mean of COMI baseline scores for each scenario, using item-based imputation.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	7.69	7.69	7.68	7.68	7.68	7.68	7.69	7.67	7.68
2	7.69	7.68	7.68	7.68	7.68	7.68	7.68	7.68	7.67
3	7.69	7.68	7.69	7.68	7.69	7.68	7.68	7.69	7.70
4	7.69	7.68	7.68	7.68	7.69	7.70	7.68	7.68	7.69
5	7.68	7.69	7.68	7.68	7.69	7.69	7.68	7.68	7.69
6	7.69	7.68	7.68	7.68	7.68	7.69	7.68	7.68	7.69
7	7.68	7.69	7.68	7.69	7.68	7.69	7.68	7.68	7.69
8	7.69	7.68	7.68	7.69	7.68	7.69	7.68	7.69	7.68

Table 4.5: Estimated population mean of COMI baseline scores for each scenario, using score-based imputation.

The range of the population mean was accurately restored by both methods, with a range of [7.69 – 7.71] for the item-wise imputation and a range of [7.67 – 7.70] for the score-wise imputation. The ranges of estimated standard deviation of population mean baseline COMI scores were [1.74 – 1.75] and [1.75 – 2.15] for item- and score-based imputation respectively. Score-based imputation tended more to slightly overestimating the true standard deviation of 1.74.

4.7 Questionnaires at baseline for missing data underlying MAR mechanism

Data sets were again generated using the `ampute()` function, but the mechanism of missingness was now indicated as MAR. This means that the missingness can depend on other measured variables. One can specify weights, so that some measured variables have a higher effect on the missingness than others. The computation of the probability of a missing value is then a linear combination of each weight and variable.

For simplicity, the vector of weights for each pattern was equally distributed over the non-missing variables. For future simulation studies, this can be further explored by adapting reasonable assumptions about this distribution in a given patient population. For further information see (Schouten et al., 2018). It is expected that both methods to have higher RMSEs than for missingness

under the MCAR mechanism, especially for higher percentages of total missingness. Otherwise, the scenarios are the same as in the previous simulation study, with each having N=50 simulation iterations. Specifically, taking into account the possible patterns of missingness for a given cut-off point calculation and probability of missingness, amputation of missing data in each simulation iteration is created by specifying the mechanism to “MAR” in the ampute() function. The imputation however, is the same over all scenarios.

The results regarding RMSE are shown in Table 4-6 and 4-7. Each table shows each combination of scenarios for the RMSE using item-based imputation and score-based imputation.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	0.41	0.58	0.71	0.82	0.92	1.01	1.09	1.18	1.24
2	0.40	0.58	0.71	0.82	0.92	1.01	1.09	1.17	1.24
3	0.42	0.57	0.71	0.82	0.92	1.01	1.09	1.17	1.25
4	0.41	0.58	0.71	0.82	0.92	1.01	1.10	1.17	1.25
5	0.41	0.58	0.71	0.82	0.92	1.01	1.09	1.17	1.25
6	0.41	0.58	0.71	0.82	0.92	1.01	1.09	1.17	1.25
7	0.41	0.58	0.70	0.82	0.92	1.01	1.10	1.17	1.25
8	0.42	0.58	0.71	0.82	0.92	1.01	1.09	1.17	1.25

Table 4.6: Mean of RMSE of each combination of probability of amputed missingness and cut-off point, using item-based imputation method. RMSE was averaged over number of simulations (N=50).

In Table 4.6 the mean of the RMSE is shown for the item-based imputation method. It becomes clear that the cut-off point of questionnaires does not matter in this method, since every item is imputed and scores are calculated afterwards. Standard deviations of the RMSE (simulation estimand) were low with a range of 0.015 to 0.026.

In Table 4.7, the means and standard deviations of each combinations are shown for the score-based imputation.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	0.75	1.07	1.32	1.54	1.73	1.92	2.09	2.25	2.41
2	0.72	1.02	1.25	1.45	1.64	1.81	1.96	2.11	2.27
3	0.64	0.90	1.12	1.29	1.46	1.60	1.75	1.88	2.02
4	0.53	0.77	0.95	1.10	1.23	1.37	1.50	1.62	1.74
5	0.46	0.66	0.82	0.95	1.07	1.20	1.30	1.41	1.52
6	0.43	0.62	0.77	0.90	1.01	1.13	1.23	1.33	1.43
7	0.43	0.62	0.76	0.89	1.02	1.12	1.23	1.32	1.42
8	0.44	0.62	0.77	0.90	1.01	1.12	1.22	1.32	1.42

Table 4.7: Mean of RMSE of each combination of probability of amputed missingness and cut-off point, using score-based imputation method. RMSE was averaged over number of simulations (N=50).

The same pattern as in the MCAR mechanism could be observed that score-based imputation was worse than item-based imputation for scenarios with high missingness and small cut-off points. Standard deviations of the RMSE (simulation estimand) were low with a range of 0.017 to 0.051.

How these two methods performed for the calculation of the population mean and standard deviation of baseline scores will now be analysed. For each combination the population mean (and standard deviation) was calculated and is presented in Tables 4.8 and 4.9. It needs to be reminded, that the true population baseline COMI score mean was 7.69 (s.d. 1.74).

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	7.69	7.69	7.69	7.69	7.69	7.69	7.69	7.69	7.70
2	7.69	7.69	7.69	7.69	7.69	7.69	7.69	7.69	7.70
3	7.69	7.69	7.69	7.69	7.69	7.69	7.69	7.69	7.70
4	7.69	7.69	7.69	7.69	7.70	7.69	7.69	7.70	7.70
5	7.69	7.69	7.69	7.69	7.69	7.70	7.70	7.70	7.70
6	7.69	7.69	7.69	7.69	7.69	7.70	7.70	7.70	7.70
7	7.68	7.69	7.69	7.69	7.69	7.69	7.70	7.69	7.70
8	7.69	7.69	7.69	7.69	7.69	7.69	7.69	7.69	7.69

Table 4.8: Estimated population mean of COMI baseline scores for each scenario, using item-based imputation.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	7.64	7.59	7.53	7.48	7.43	7.37	7.31	7.28	7.26
2	7.65	7.61	7.57	7.53	7.49	7.47	7.44	7.45	7.47
3	7.66	7.64	7.62	7.60	7.59	7.59	7.59	7.59	7.62
4	7.68	7.67	7.67	7.67	7.67	7.66	7.66	7.67	7.67
5	7.69	7.69	7.70	7.70	7.70	7.70	7.70	7.69	7.69
6	7.69	7.70	7.71	7.72	7.72	7.71	7.71	7.71	7.70
7	7.70	7.70	7.71	7.71	7.71	7.72	7.71	7.71	7.70
8	7.70	7.70	7.71	7.71	7.72	7.72	7.71	7.70	7.70

Table 4.9: Estimated population mean of COMI baseline scores for each scenario, using score-based imputation.

Both methods were able to accurately restore the true population mean of baseline COMI scores. However, item-based imputation was more accurate, with a range of [7.68 – 7.70], compared to the score-based imputation, which had a range of [7.26 – 7.72]. The ranges of estimated standard deviation of population mean baseline COMI scores were [1.74 – 1.75] and [1.75 – 2.15] for item- and score-based imputation respectively. Score-based imputation tended more to slightly overestimating the true standard deviation of 1.74.

4.8 Questionnaires at baseline for missing data underlying MNAR mechanism

In this scenario it is assumed that missingness can be dependent not only on other variables, but on the missing value itself. In the extreme case for which the missingness of a variable is only dependent on itself, this means that the weight vector of a specific pattern is the inverted vector of the pattern, with ones for the variables that are amputed as missing and zeros for all other variables that are not amputed. However, mixed patterns are also possible. In this case, to test the robustness of the imputation methods, the extreme case is assumed. This should make the imputation less precise with both methods, which is why it is expected that imputation methods fail for lower percentages of missingness, have higher RMSEs and imprecise estimates of baseline score mean. The results regarding RMSE are shown in Table 4.10 and 4.11.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	0.38	0.55	0.68	0.79	0.89	0.98	1.06	1.14	1.23
2	0.39	0.55	0.69	0.79	0.90	0.98	1.06	1.14	1.22
3	0.39	0.55	0.69	0.78	0.89	0.97	1.07	1.15	1.23
4	0.38	0.55	0.68	0.79	0.89	0.97	1.07	1.15	1.22
5	0.38	0.55	0.68	0.79	0.89	0.98	1.06	1.14	1.22
6	0.38	0.55	0.68	0.79	0.88	0.98	1.06	1.14	1.22
7	0.38	0.55	0.69	0.79	0.89	0.98	1.07	1.14	1.23
8	0.38	0.55	0.68	0.79	0.89	0.98	1.06	1.14	1.22

Table 4.10: Mean of RMSE of each combination of probability of amputed missingness and cut-off point, using item-based imputation method. RMSE was averaged over number of simulations (N=50).

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	0.72	1.05	1.32	1.56	1.78	2.00	2.20	2.41	2.63
2	0.70	1.01	1.25	1.47	1.68	1.87	2.04	2.21	2.36
3	0.61	0.89	1.12	1.30	1.47	1.63	1.79	1.93	2.05
4	0.51	0.74	0.93	1.08	1.22	1.36	1.49	1.62	1.73
5	0.43	0.63	0.78	0.92	1.05	1.17	1.27	1.38	1.49
6	0.40	0.59	0.73	0.86	0.98	1.09	1.20	1.29	1.41
7	0.40	0.58	0.74	0.86	0.98	1.08	1.19	1.30	1.40
8	0.39	0.58	0.73	0.85	0.98	1.08	1.18	1.29	1.39

Table 4.11: Mean of RMSE of each combination of probability of amputed missingness and cut-off point, using item-based imputation method. RMSE was averaged over number of simulations (N=50).

Standard deviations of the RMSE estimates for item-based imputation ranged from 0.013 to 0.028 and from 0.021 to 0.069. Even though score-based imputation had larger standard deviation, the main difference lies between their mean RMSEs. Even missing data underlying the MNAR mechanism could be handled by the item-based imputation, whereas score-based imputation performed much more inaccurate, especially in scenarios with high missingness and small cut-off points.

Again, the mean and standard deviation of baseline scores in each scenario for item-based and score-based imputation were investigated and compared to the true baseline score mean and standard deviation, which were 7.686 and 1.744 respectively.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	7.65	7.62	7.59	7.57	7.56	7.55	7.56	7.59	7.63
2	7.65	7.62	7.59	7.57	7.55	7.55	7.56	7.58	7.62
3	7.65	7.62	7.59	7.57	7.55	7.55	7.56	7.58	7.62
4	7.65	7.62	7.59	7.57	7.55	7.55	7.56	7.58	7.63
5	7.65	7.62	7.59	7.57	7.56	7.55	7.56	7.58	7.62
6	7.65	7.62	7.59	7.57	7.56	7.55	7.56	7.58	7.62
7	7.65	7.62	7.58	7.57	7.55	7.55	7.56	7.59	7.63
8	7.65	7.62	7.59	7.57	7.56	7.55	7.56	7.58	7.63

Table 4.12: Estimated population mean of COMI baseline scores for each scenario, using item-based imputation.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	7.59	7.50	7.40	7.28	7.16	7.04	6.91	6.79	6.71
2	7.60	7.51	7.42	7.34	7.23	7.16	7.09	7.08	7.15
3	7.62	7.55	7.48	7.42	7.37	7.34	7.32	7.33	7.44
4	7.64	7.59	7.55	7.51	7.49	7.48	7.47	7.50	7.57
5	7.65	7.62	7.60	7.57	7.56	7.56	7.56	7.59	7.62
6	7.66	7.64	7.62	7.60	7.59	7.59	7.60	7.61	7.63
7	7.66	7.64	7.62	7.61	7.60	7.60	7.61	7.61	7.64
8	7.66	7.64	7.62	7.61	7.60	7.59	7.60	7.61	7.65

Table 4.13: Estimated population mean of COMI baseline scores for each scenario, using score-based imputation.

Overall, the performance of both methods when amputing missing data using the MAR mechanism was very similar to the performance when missing data was underlying the MCAR mechanism. The `mice()` function using predictive mean matching can recover data reliably, but item-based imputation could restore population means more accurately and had lower RMSEs than score-based imputation. It was surprising that imputing data that is missing under the MNAR mechanism did not perform much worse than for other mechanisms. In fact, in some scenarios it even had (slightly) lower mean RMSEs. Method performance regarding both RMSEs and population means of baseline COMI scores showed, that item-based imputation was superior over score-based imputation. This can be explained by the loss of information that is introduced by disregarding the answered items of questionnaires that had high item-missingness. Item-based imputation was better suited to recover the missing items and produced lower RMSEs and more precise estimates of population COMI baseline score means.

To further test the limits of the methods, the total missingness will be increased. The missingness probabilities will remain the same (0.1 to 0.9), but the pattern of missingness will be changed so that, in case of missingness, all items are missing. This also means that cut-off points will not be relevant, as there will only be either complete or completely missing questionnaires. The missingness will be applied under MCAR, MAR, and MNAR as before, and the score-based imputation will be compared to the item-based imputation again.

4.9 MCAR, MAR and MNAR in scenarios of high questionnaire-missingness

To further test the limits of the methods, the total missingness will be increased. The missingness probabilities will remain the same (0.1 to 0.9), but the pattern of missingness will be changed so that, in case of missingness, all items are missing. This also means that cut-off points will not be relevant, as there will only be either complete or completely missing questionnaires. The missingness will be applied under MCAR, MAR, and MNAR as before, and the score-based imputation will be compared to the item-based imputation again. This will be done by comparing RMSEs of estimates and the accuracy to restore the true population mean of baseline scores. In Table 4.14 the RMSEs are summarised for each mechanism of missingness and method of imputation.

Imputation type \ Probability		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
MCAR	item-based	0.77	1.09	1.34	1.55	1.73	1.89	2.05	2.19	2.30
	score-based	0.77	1.09	1.34	1.54	1.72	1.89	2.04	2.18	2.32
MAR	item-based	0.79	1.11	1.36	1.56	1.74	1.90	2.06	2.19	2.34
	score-based	0.78	1.11	1.34	1.54	1.73	1.89	2.05	2.18	2.33
MNAR	item-based	0.73	1.08	1.36	1.62	1.85	2.09	2.33	2.59	2.90
	score-based	0.74	1.08	1.36	1.62	1.86	2.10	2.33	2.60	2.94

Table 4.14: RMSEs of both imputation methods for all mechanisms of missingness. Columns are ordered regarding the probability of missingness (complete questionnaire missingness).

Standard deviations of these estimates did not exceed 0.09 and were similar for both imputation techniques for each probability of missingness. Both methods performed very similar for each scenario and produced larger RMSEs with larger amounts of missingness. RMSE values reached up to 2.94, so over the minimal clinically important difference of the COMI score. When missing data is too substantial, neither method could restore the data accurately. When questionnaires are either complete or completely missing, it seems that the methods do not differ much in performance.

In Table 4.15 the estimated population means are summarised for each mechanism of missingness and method of imputation.

Imputation type \ Probability		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
MCAR	Item-based	7.69	7.69	7.70	7.69	7.69	7.69	7.70	7.68	7.69
	Score-based	7.69	7.68	7.68	7.69	7.68	7.68	7.68	7.69	7.68
MAR	Item-based	7.69	7.69	7.68	7.68	7.68	7.68	7.67	7.66	7.55
	Score-based	7.69	7.68	7.68	7.68	7.68	7.67	7.67	7.67	7.66
MNAR	Item-based	7.58	7.46	7.33	7.19	7.03	6.86	6.66	6.43	6.13
	Score-based	7.58	7.46	7.32	7.18	7.03	6.85	6.66	6.42	6.10

Table 4.15: Estimated population means of baseline COMI scores for both imputation methods and all mechanisms of missingness. Columns are ordered regarding the probability of missingness (complete questionnaire missingness).

When the data was missing due to MCAR or MAR mechanisms, the population mean of the baseline COMI scores could be restored, even when a large amount of data was missing. This might be due to the large sample size, so that even for 90% of questionnaire missingness, there were still enough patients to get a realistic estimate. However, when the data was missing due to an MNAR mechanism, the population mean was underestimated. Standard deviations of these estimates were similar for both imputation methods and ranged between 1.75 and 2.02 for MNAR (higher standard deviations were observed for higher probabilities of missingness). Standard deviations therefore were systematically overestimated in both methods, whereas population means were underestimated.

4.10 Missing data in outcomes at 3 months past surgery

This next simulation study will focus on identifying the most effective imputation technique for handling missing data in outcome questionnaires among patients who had baseline questionnaires available. This experimental design, if applied in real-world, raises ethical concerns, but nevertheless was explored, to investigate how the imputation methods would perform for outcome questionnaires. Hereby, the items of the baseline questionnaires were available and, in case of MAR and MNAR, are also considered as possibly connected to missingness in outcome items for the creation of missing data.

This simulation was based on the subset of the previous set of patients, who had not only had baseline but also 3-month outcome COMI questionnaires available. This led to an inclusion of 2,128 patients. Simulations are set-up as previously, for MCAR, MAR and MNAR for several scenarios of probability of missingness and cut-off points. The main question is, if outcome questionnaires could also reliably be restored by either imputation method, and for which scenarios one has to expect bias to be introduced. Baseline items of the COMI score were included in both the creation of missingness, as well as part of the imputation.

Previously, it was found that item-based imputation was superior to score-based imputation for handling missing data in baseline questionnaires. Only when questionnaires were either complete or completely missing, the performance of both methods was equal. Both methods struggled to get precise estimates per patient when missing data was high (large RMSEs). However, when data is missing underlying the MCAR or MAR mechanism, population means of baseline scores could be restored. When data is missing underlying the MNAR mechanism, bias is introduced and estimates inaccurate.

The following six tables (4.16 – 4.21) show the mean RMSEs of both imputation methods over 50 simulations for each scenario for MCAR, MAR and MNAR.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	0.52	0.73	0.89	1.03	1.14	1.29	1.37	1.46	1.54
2	0.50	0.74	0.88	1.05	1.15	1.27	1.37	1.46	1.55
3	0.51	0.73	0.89	1.04	1.15	1.26	1.36	1.47	1.56
4	0.52	0.73	0.89	1.01	1.15	1.28	1.36	1.47	1.54
5	0.52	0.72	0.88	1.03	1.17	1.28	1.35	1.46	1.55
6	0.51	0.72	0.89	1.02	1.15	1.24	1.36	1.47	1.57
7	0.52	0.72	0.88	1.03	1.15	1.25	1.35	1.46	1.55
8	0.51	0.73	0.88	1.03	1.16	1.26	1.37	1.46	1.55

Table 4.16: Mean of RMSE of each combination of probability of amputed missingness and cut-off point, using item-based imputation method and MCAR missingness. RMSE was averaged over number of simulations (N=50).

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	1.21	1.72	2.11	2.41	2.70	2.97	3.20	3.47	3.85
2	1.13	1.62	1.98	2.29	2.55	2.80	3.05	3.26	3.47
3	1.01	1.46	1.76	2.01	2.27	2.48	2.70	2.91	3.07
4	0.81	1.15	1.41	1.63	1.81	2.02	2.20	2.36	2.52
5	0.64	0.89	1.10	1.27	1.45	1.62	1.74	1.84	1.98
6	0.55	0.76	0.95	1.09	1.23	1.35	1.46	1.59	1.69
7	0.52	0.74	0.89	1.03	1.19	1.30	1.41	1.53	1.60
8	0.53	0.76	0.91	1.05	1.20	1.31	1.41	1.52	1.63

Table 4.17: Mean of RMSE of each combination of probability of amputed missingness and cut-off point, using score-based imputation method and MCAR missingness. RMSE was averaged over number of simulations (N=50).

Probability / Cut-off point	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	0.53	0.75	0.92	1.06	1.18	1.28	1.40	1.48	1.58
2	0.55	0.75	0.92	1.07	1.19	1.28	1.39	1.48	1.60
3	0.53	0.73	0.91	1.06	1.19	1.30	1.39	1.48	1.56
4	0.53	0.76	0.92	1.08	1.19	1.30	1.39	1.49	1.58
5	0.53	0.76	0.91	1.06	1.19	1.30	1.40	1.49	1.56
6	0.52	0.75	0.92	1.07	1.18	1.30	1.40	1.48	1.57
7	0.54	0.75	0.92	1.08	1.19	1.29	1.40	1.50	1.57
8	0.51	0.75	0.92	1.07	1.20	1.29	1.41	1.48	1.57

Table 4.18: Mean of RMSE of each combination of probability of amputed missingness and cut-off point, using item-based imputation method and MAR missingness. RMSE was averaged over number of simulations (N=50).

Probability / Cut-off point	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	1.22	1.73	2.13	2.44	2.73	2.98	3.20	3.43	3.66
2	1.14	1.66	2.03	2.31	2.59	2.84	3.05	3.27	3.49
3	1.01	1.45	1.77	2.06	2.30	2.53	2.72	2.91	3.10
4	0.81	1.20	1.45	1.68	1.90	2.07	2.23	2.36	2.53
5	0.65	0.93	1.13	1.32	1.48	1.62	1.74	1.88	1.95
6	0.55	0.79	0.97	1.13	1.25	1.37	1.48	1.57	1.69
7	0.55	0.76	0.94	1.10	1.19	1.30	1.44	1.52	1.62
8	0.53	0.75	0.94	1.09	1.23	1.30	1.42	1.53	1.61

Table 4.19: Mean of RMSE of each combination of probability of amputed missingness and cut-off point, using score-based imputation method and MAR missingness. RMSE was averaged over number of simulations (N=50).

Probability / Cut-off point	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	0.59	0.84	0.96	1.09	1.22	1.29	1.40	1.48	1.57
2	0.58	0.79	0.98	1.10	1.22	1.33	1.40	1.49	1.58
3	0.59	0.81	0.97	1.10	1.21	1.32	1.41	1.51	1.58
4	0.57	0.79	0.95	1.10	1.22	1.30	1.40	1.49	1.59
5	0.58	0.81	0.98	1.09	1.22	1.31	1.42	1.52	1.55
6	0.58	0.80	0.98	1.12	1.23	1.30	1.42	1.49	1.58
7	0.59	0.80	0.96	1.10	1.22	1.31	1.39	1.50	1.57
8	0.61	0.80	0.95	1.11	1.22	1.31	1.39	1.47	1.57

Table 4.20: Mean of RMSE of each combination of probability of amputed missingness and cut-off point, using item-based imputation method and MNAR missingness. RMSE was averaged over number of simulations (N=50).

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	1.31	1.86	2.27	2.57	2.87	3.12	3.34	3.55	3.74
2	1.23	1.77	2.13	2.45	2.71	2.97	3.15	3.35	3.51
3	1.11	1.55	1.85	2.13	2.40	2.58	2.75	2.95	3.12
4	0.88	1.24	1.52	1.72	1.91	2.10	2.26	2.38	2.54
5	0.70	0.96	1.19	1.36	1.51	1.62	1.74	1.87	1.96
6	0.60	0.82	1.01	1.16	1.29	1.38	1.51	1.58	1.71
7	0.56	0.80	0.97	1.11	1.23	1.32	1.43	1.54	1.63
8	0.58	0.80	0.96	1.10	1.23	1.32	1.44	1.52	1.62

Table 4.21: Mean of RMSE of each combination of probability of amputed missingness and cut-off point, using score-based imputation method and MNAR missingness. RMSE was averaged over number of simulations (N=50).

Interestingly, the methods performed similarly for each mechanism of missingness. For both imputation methods the mean RMSE values increase with increasing probability of missingness. Item-based imputation again showed the same pattern as previously, when baseline COMI questionnaire items were imputed. Cut-off points do not play a big role, since items are imputed anyway. For score-based imputation, not only high probability of missingness, but also smaller cut-off points introduce errors. It is therefore concluded, that item-based imputation is superior over score-based imputation, regarding mean RMSEs of outcome questionnaires. Overall, RMSE values were larger than for baseline COMI imputation, regardless of the method.

The following six tables (4.22 – 4.27) show the population mean of COMI outcome scores at 3 months past surgery over 50 simulations for each scenario. It should be reminded, that the population mean of COMI scores at 3 months past surgery is 4.11 (s.d. 2.91).

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	4.11	4.10	4.11	4.11	4.11	4.11	4.12	4.11	4.13
2	4.11	4.11	4.11	4.10	4.11	4.11	4.10	4.13	4.13
3	4.11	4.10	4.11	4.11	4.11	4.11	4.12	4.11	4.12
4	4.11	4.11	4.11	4.10	4.12	4.11	4.12	4.14	4.11
5	4.11	4.10	4.11	4.10	4.12	4.10	4.11	4.12	4.12
6	4.11	4.11	4.11	4.10	4.11	4.10	4.11	4.11	4.12
7	4.11	4.10	4.11	4.10	4.11	4.12	4.11	4.11	4.12
8	4.11	4.11	4.11	4.11	4.12	4.10	4.11	4.12	4.14

Table 4.22: Estimated population mean of COMI baseline scores for each scenario, using item-based imputation and MCAR missingness.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	4.11	4.10	4.13	4.11	4.12	4.08	4.06	4.09	4.45
2	4.11	4.12	4.11	4.12	4.13	4.13	4.14	4.12	4.12
3	4.11	4.11	4.10	4.11	4.09	4.09	4.12	4.09	4.10
4	4.11	4.11	4.11	4.12	4.12	4.11	4.13	4.14	4.12
5	4.11	4.11	4.11	4.11	4.13	4.12	4.12	4.13	4.14
6	4.11	4.11	4.12	4.11	4.11	4.12	4.12	4.12	4.12
7	4.11	4.11	4.12	4.11	4.12	4.12	4.12	4.13	4.13
8	4.11	4.12	4.12	4.12	4.12	4.12	4.13	4.13	4.14

Table 4.23: Estimated population mean of COMI baseline scores for each scenario, using score-based imputation and MCAR missingness.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	4.10	4.10	4.09	4.09	4.09	4.09	4.09	4.10	4.09
2	4.10	4.09	4.09	4.09	4.09	4.09	4.10	4.10	4.10
3	4.10	4.10	4.09	4.09	4.09	4.09	4.10	4.10	4.11
4	4.10	4.09	4.09	4.08	4.08	4.10	4.09	4.10	4.11
5	4.10	4.09	4.09	4.09	4.08	4.09	4.08	4.09	4.11
6	4.10	4.09	4.10	4.09	4.09	4.09	4.10	4.09	4.11
7	4.10	4.10	4.09	4.09	4.08	4.07	4.08	4.10	4.10
8	4.10	4.10	4.09	4.09	4.08	4.09	4.09	4.09	4.11

Table 4.24: Estimated population mean of COMI baseline scores for each scenario, using item-based imputation and MAR missingness.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	4.09	4.08	4.06	4.05	4.01	4.01	3.95	4.02	4.10
2	4.09	4.09	4.09	4.04	4.04	4.03	4.04	4.03	4.08
3	4.10	4.09	4.09	4.07	4.08	4.07	4.09	4.09	4.12
4	4.11	4.10	4.09	4.09	4.09	4.09	4.10	4.11	4.13
5	4.11	4.11	4.11	4.12	4.11	4.10	4.11	4.13	4.13
6	4.11	4.10	4.11	4.12	4.11	4.11	4.12	4.12	4.12
7	4.11	4.12	4.11	4.12	4.11	4.10	4.11	4.12	4.11
8	4.11	4.11	4.11	4.12	4.12	4.12	4.11	4.13	4.13

Table 4.25: Estimated population mean of COMI baseline scores for each scenario, using score-based imputation and MAR missingness.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	4.06	4.02	3.99	3.98	3.98	3.98	4.00	4.04	4.06
2	4.06	4.03	4.00	3.98	3.97	3.99	4.00	4.03	4.07
3	4.06	4.02	3.99	3.99	3.98	3.99	3.98	4.02	4.06
4	4.06	4.02	4.00	3.98	3.98	3.99	3.99	4.02	4.06
5	4.06	4.02	4.00	3.98	3.98	3.98	4.00	4.02	4.07
6	4.06	4.02	4.00	3.98	3.98	3.99	4.00	4.03	4.06
7	4.06	4.02	4.00	3.99	3.97	3.98	3.99	4.03	4.07
8	4.06	4.02	4.00	3.98	3.98	3.98	4.00	4.03	4.06

Table 4.26: Estimated population mean of COMI baseline scores for each scenario, using item-based imputation and MNAR missingness.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	3.92	3.74	3.56	3.40	3.26	3.14	3.02	2.92	2.99
2	3.95	3.78	3.62	3.49	3.39	3.30	3.30	3.32	3.52
3	3.98	3.86	3.75	3.69	3.61	3.64	3.64	3.73	3.92
4	4.03	3.95	3.89	3.86	3.85	3.84	3.88	3.94	4.01
5	4.06	4.02	3.99	3.96	3.97	3.98	4.00	4.01	4.09
6	4.07	4.05	4.03	4.01	4.02	4.03	4.05	4.06	4.10
7	4.08	4.05	4.04	4.03	4.03	4.04	4.04	4.07	4.11
8	4.08	4.05	4.04	4.03	4.03	4.05	4.05	4.07	4.10

Table 4.27: Estimated population mean of COMI baseline scores for each scenario, using score-based imputation and MNAR missingness.

In most cases, the population mean could be restored accurately, since even in scenarios of high percentages of missingness, due to the large sample size, there are still enough values left to obtain a realistic estimate. When data is missing not at random (MNAR) errors are introduced the higher the percentage of missingness. However, by choosing a higher cut-off point, this can be prevented. Interestingly, population means rather tended to be underestimated for both baseline and outcome questionnaires.

4.11 Discussion

In general, imputing missing data in questionnaires is more effective when using item-based imputation rather than score-based imputation. This finding holds value beyond the scope of this study and could potentially apply to similar datasets in various clinical areas. The robustness of item-based imputation in handling missing data across specific items within the questionnaire suggests a broader applicability, especially when maintaining the accuracy of imputed values is paramount.

It's important to note that if missing data is only present in the overall questionnaire, but not in specific items within the questionnaire, then either method will perform equally well. In such scenarios, selecting between the two approaches could consider computational efficiency, making score-based imputation an attractive option due to its lower computational cost.

However, in cases where there is missing data in specific items within the questionnaire, our results emphasize the superiority of item-based imputation. This method consistently demonstrated lower root mean squared error (RMSE) values and effectively restored population means (COMI scores), enhancing the accuracy of imputation outcomes. It is worth highlighting that when employing score-based imputation, the choice of high cut-off points is pivotal to avoid information loss and minimize the introduction of errors.

Importantly, the conclusions drawn from this study extend to both baseline and outcome questionnaires. Although the imputation of outcomes was found to be comparatively less accurate than that of baseline questionnaires, the overarching trend remains consistent. Researchers and practitioners engaging with similar datasets should consider these findings as a valuable reference for guiding their imputation strategies.

Considering these findings, it is prudent to acknowledge that while the current study offers insights specific to our dataset, the principles underlying imputation efficacy could extend to other comparable datasets and clinical contexts. Future simulation studies could provide additional validation and insight into selecting appropriate imputation methods, which should be taken into consideration prior to implementing imputation techniques. By drawing attention to these factors, this study contributes to a broader understanding of imputation strategies and their implications, not only within our specific domain but potentially across related domains as well.

While the concept of conducting a simulation study to evaluate imputation models on a given dataset is undeniably valuable, it's important to acknowledge the practical considerations that can influence its implementation. A simulation study demands a substantial investment of time and effort, often surpassing the scope of a specific project. Furthermore, while the insights acquired from such studies might have the potential for broader relevance, their direct applicability to distinct questionnaires or datasets could be limited.

The importance of comprehending the underlying mechanisms of missingness should not be underestimated for any study. By discerning whether data is missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR), researchers gain a deeper understanding of their dataset. The identification of missingness patterns not only aids in understanding the patient population but can also illuminate potential biases or the reliability of results, if imputation methods are employed.

In the earlier Chapter 3, the analysis focussed on examining associations among patient covariates within the Spine Tango registry. In the upcoming chapter, this dataset will be utilised to employ more sophisticated techniques in order to uncover risk factors linked to unfavourable outcomes related to COMI scores or surgical complications. While a method to impute missing data in questionnaires was explored, the substantial patient sample size within the registry, even when accounting for those with available questionnaires, led to the choice of conducting a complete case analysis. In essence, the identification of risk factors constitutes a significant focus within this study.

Chapter 5: Predictive Modelling with Spine Tango data

5.1 Chapter Outline

The aim of this chapter is to perform a prognostic factor analysis of the Spine Tango data, using techniques such as regression and mixed-effect models. Outcomes that are considered included the COMI score and complications during surgery. The objective was to identify risk factors that can help optimise individual treatments, thereby enhancing routine healthcare and decision-making.

5.2 Introduction

Large datasets can give further insights into real-world practice and allow for a thorough prognostic factor research. Observational registries allow for continuous data collection over indefinite time for numerous patients. One can therefore gain additional insight in subgroup demographics of the patient population and rare events. This chapter will focus on the sciatica affected patient population of the Spine Tango registry and explore several model approaches in order to analyse prognostic factors associated with quality of life improvement after surgery. As primary outcome the focus will be on the Core Outcome Measures Index (COMI). For each of the commonly measured timepoints post-surgery a linear regression model will be fitted in order to identify predictive factors. Sensitivity analyses will be carried out to assess the stability of the model fit.

Another model approach that was explored is logistic regression. According to the minimal clinically important difference (MCID) of COMI scores, outcomes will be dichotomized into “significant improvement” or “no significant improvement”. Similar to the procedure of model exploration, this will be done for 3-month, 1-year and 2-year outcome timepoints.

Since patient-reported outcome measures (PROMs) in the Spine Tango registry had high amounts of missingness and inconsistency in their follow-ups, a longitudinal model was considered to integrate follow-up data from every patient that had a baseline measurement available.

Complications are rare, but the amount of data still allows for a logistic regression models to be fitted on the data with complication occurrence as outcome. For this, complications were grouped into several categories according to a clinician’s expertise. Afterwards, logistic regression models were applied to each of these categories to identify risk factors pre-surgery. The identification of such risk

factors can inform individual treatment decisions, potentially supporting or discouraging surgery or alternative treatments. Additionally, this knowledge contributes to improved decision-making in routine healthcare.

5.3 Literature Review

Several studies have shown that in a heterogeneous group of conditions, observational studies can provide useful insights into the outcomes of interventions that have been implemented into every day clinical practice. Results from well-designed observational studies can be similarly trustworthy as results from randomised controlled trials (RCTs) (Benson and Hartz, 2000, Colditz, 2010). In 2000, EUROSPINE developed a registry for the collection of spinal surgery data in collaboration with the University of Bern. To date, over 750,000 forms (134,458 surgery forms) from five continents have been collected (EUROSPINE, 2022b). The most commonly patient reported outcome measure for quality of life and lower back and leg pain was the COMI, which was recorded pre-surgery and at follow-up visits.

Sobottke et al. identified ASA morbidity status, age and blood loss as risk factors for adverse events in patients with spinal stenosis within the Spine Tango registry (Sobottke et al., 2012). Zehnder et al. performed multiple logistic regression models on general and surgical complications and identified ASA and prior surgery at the same level as predictive factors for patients with lumbar degenerative diseases that underwent surgery within the Spine Tango registry (Zehnder et al., 2021). Sunderland et al. analysed the success of lumbar decompression surgery by using COMI score improvement, but did and identified ASA morbidity status 3, age, lateral stenosis (pathological factor), revision surgery, and surgeon in training as prognostic factors (Sunderland et al., 2021). Sobottke et al. used COMI scores as outcomes in order to identify predictors for the improvement of QoL for patients with lumbar spinal canal stenosis that underwent open decompression surgery (Sobottke et al., 2017). The main predictor was baseline COMI scores, but the number of prior surgeries, lower patient comorbidity and rigid or dynamic stabilization also had partially prognostic influence. The preoperative status of each outcome was a prognostic factor for its own postoperative outcome. Fewer previous surgeries, rigid or dynamic stabilization, and lower patient comorbidity also had a partially prognostic influence for one or the other outcome. Staub, L.P., et al. developed predictive models for 1-year clinical outcome after decompression surgery using data from the Spine Tango registry, using linear regression and LASSO. Although model accuracy was good overall, considerable uncertainty on individual level was pointed out (Staub et al., 2020). Aghayev, E., et al. determined risk factors for negative global treatment outcomes as self-assessed by patients undergoing surgical treatment for lumbar spinal stenosis and could identify high baseline, department-level and potentially country-level factors as associated with

treatment outcome (Aghayev et al., 2020). No literature was found that utilizes the Spine Tango registry within the subset of the sciatica-affected patient population that underwent microdiscectomy.

5.4 Methods

Throughout this chapter, data from the Spine Tango registry was used. In Chapter 3, 3,530 patients were identified that fit the sciatica population and had both baseline and follow-up baseline COMI scores available. Of all the measured patient characteristics, the following were included in the model approaches after consultation with the multi-disciplinary supervisor team (neurologist, neuro surgeon, biostatistician): sex, age, surgeon credentials, country ID, level of spine, ASA morbidity status, BMI, smoking status, baseline scores and previous treatment. Most of them showed very low amount of missingness (less than 3% of patients). There was 12% of missing data for previous treatment and 17% for ASA morbidity. To not further reduce the sample size by analysing on a complete case dataset, each of these characteristics were imputed using the multiple imputation toolbox 'mice' in R, but with m=1 iteration since only one complete dataset is required. Specifically, predictive mean matching), was chosen for filling these, gaps, which is applicable to different data types (continuous and categorical), as well as suitable for large data sets. It is a flexible approach to be considered when missing data is possible not missing completely at random (MCAR) and preserves original data structure and variability (Bailey et al., 2020, StatisticsGlobe, 2022). BMI and smoking status had high missingness, largely due to change in collection forms throughout the years, and were not imputed, but analysed in their respective subset (48% patients had missing smoking status and 33% missing BMI in this patient set).

An exact sample size calculation was not performed. However, the general rule of thumb, which recommends a minimum of 10 events per predictor parameter (EPP), was followed. Although this approach is sometimes criticised, the study's robustness is supported by a substantial sample size of 3,530 patients, ensuring sufficient statistical power for the conducted analyses (Riley et al., 2019).

As seen in Chapter 3, there are a few covariates that have categories with very low numbers of patients, such as spinal disc level "L1 / L2" or ASA morbidity classification of 4. These will be considered with caution and checked if they were the only reason why a parameter was considered significant by the model fit. In some cases, when the number of patients in sub-categories was small, these patients were excluded. It needs to be reminded, that outcome collection was dependent on the country. Therefore, there are no measurements for 2-year outcomes, e.g. for country G. To simplify the categorisation, patients who received treatment for less than three months or between three to six

months were combined and referred to as "less than 6 months." Similarly, those who received treatment between six to twelve months and over twelve months were grouped together as "6 to twelve months" and "more than twelve months," respectively. A sensitivity analysis of this recategorization has not been done in the scope of this project. However, Figure 3.41 shows that the treatment outcome was very similar in the grouped categories.

Country C did not have any measurements in this subset of patients (the 3,530 patients that have both baseline and follow-up COMI scores available) and was therefore excluded. There was correlation between categories of surgeon credentials and country ID. Countries H, D and E did not report surgeon in training and the only country that listed “other” as surgeon credential was country A. This will be examined further, in case surgeon credentials shows significance regarding treatment outcome. Model approaches included in this chapter are:

- linear regression with COMI scores at 3 months, 1 year and 2 years (additionally done on the subset of patients that had BMI and smoking status available),
- Logistic regression with dichotomised outcome (significant improvement of COMI scores and no significant improvement of COMI scores),
- Longitudinal mixed model approach (additionally done cut-off data at 3 months, 1 year and 2 years for comparison with previous models),
- Joint model approach, and
- Logistic regression with complication as outcome.

For all modelling approaches that used COMI scores on its continuous scale, the root mean squared error (RMSE) is computed (root of mean square difference between true scores and estimated score by the model) in order to compare models. Additionally, all estimated effects, 95% confidence intervals and p-values are presented.

For linear regression, the R^2, which is a measure for the outcome variation that could be explained with the model, was used to assess the model fit. For logistic regression approaches the area under the receiver operating characteristic (ROC) curve was used to assess the model fit. It shows the true positive rate against the true negative rate for various thresholds. The area under this curve (AUC) is used to measure the model's ability to predict outcomes. For mixed-model approaches, there are several formulas that are somewhat an equivalent to the R^2 of linear regression models. One of the most common ones was developed by D. Zhang and extends the proportion of explained variance (Zhang, 2020). Moreover, it defines this proportion into variation explained by the whole model, fixed effects only, and random effects only. This measure is implemented in the “rsq”-package in R. To compare the joint model approach with the mixed-model, the longitudinal sub-model was compared

regarding the estimated effects, 95% confidence intervals and p-values. For the logistic regression with complication as outcome, considering the small incidence of complications, COMI baseline scores were not included as predictive factor.

5.5 Linear regression approaches

When intending to develop a prognostic model, the first approach often is a multivariate linear regression model of the form

$$Y_i = X_i\beta + \varepsilon_i \qquad (1)$$

The variables of this formula are explained as follows:

- Y_i is the outcome for the i-th patient, where i ranges from 1 to N (N is the number of total patients).
- X_i is a (p+1)-dimensional vector of patient covariates (+1 extra value for model intercept).
- $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ is a (p+1)-dimensional vector of coefficients, where p is the number of included patient baseline covariates. β_0 is the value of the intercept of the linear model.
- ε_i is an error term, which are each assumed to be normally distributed (all ε_i are assumed from the same distribution).

Ordered categorical characteristics such as level of spine, previous treatment and ASA morbidity status were ordered accordingly where previous treatment “none”, level of spine “L5 / S1” and ASA morbidity status “ASA 1” were set as reference category. For the other characteristics, the most common category was set as reference. There were three main outcome time points available, as seen in Figure 5.1.

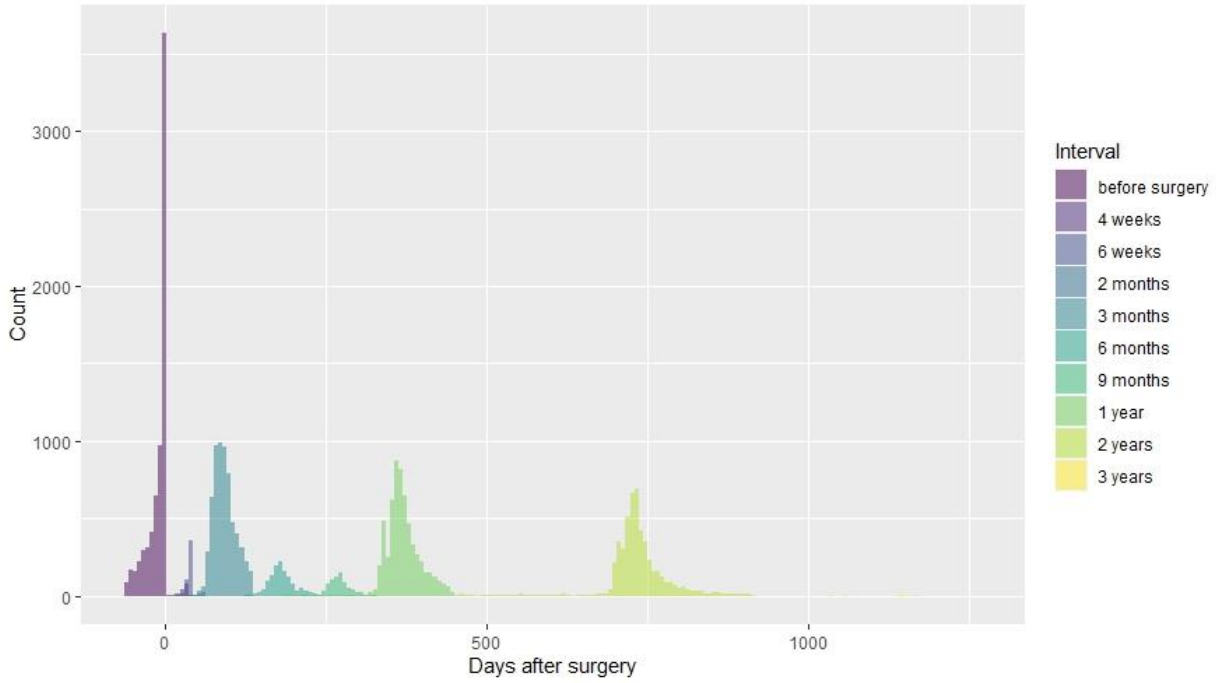


Figure 5.1: Number of COMI questionnaires collected on each day before/after surgery.

The number of patients with 3-month, 1-year and 2-year outcome available were 2,127, 2,190 and 1,670 respectively. For patients that had more than one measurement in the same interval (for example if a patient had answered the COMI questionnaire multiple times at the 3-month time interval), the measurements were averaged. To preserve the assumption of independence between data rows (1 per patient) it was required that one patient could not have more than one measurement in each interval. Considering their close temporal proximity, it is presumed that the scores are similar. Averaging was employed to avoid any random selection bias. For each of these outcome time points a model of the form (1) was fitted to identify prognostic factors and quantify their association with the outcome.

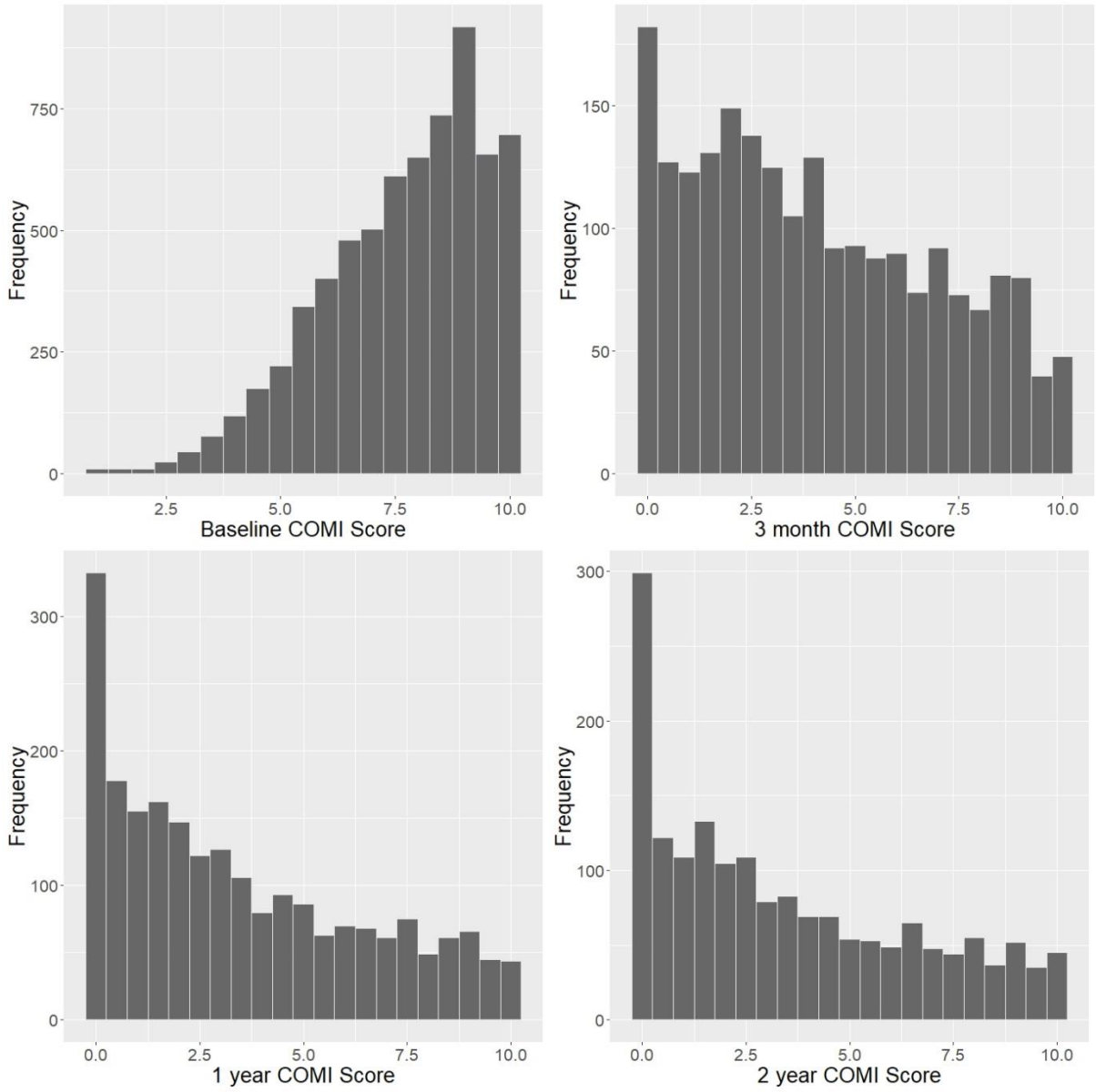


Figure 5.2: Histograms of COMI scores at baseline (top left), 3 months (top right), 1 year (bottom left) and 2 years (bottom right) after surgery.

The histograms in Figure 5.2 show that the COMI scores are skewed to the maximum score (10) for baseline and towards the minimum score (0) for scores after surgery.

5.5.1 **Linear regression – COMI Scores at three months past surgery**

The first model approach considers three-month COMI scores as primary outcome and includes 2,127 patients in total. The model of the form (1) was fitted using the `lm()` function in R. The table of coefficients, their 95% confidence interval and p-values are displayed in Table 5.1.

Variable	Estimate	95% Confidence Interval	p-Value
Intercept	**1.496**	**[0.697, 2.295]**	**<0.001**

Sex	Female	Reference		
	Male	-0.059385	[-0.301, 0.182]	0.505
Age		-0.008098	[-0.018, 0.002]	0.203
Surgeon Credentials	BC-N	Reference		
	SSS	-0.149	[-0.434, 0.136]	0.413
	N-t	-0.065	[-0.461, 0.331]	0.770
	BC-O	0.198	[-0.923, 1.412]	0.714
	O-t	-0.210	[-1.122, 0.702]	0.970
	Other	0.244	[-0.923, 1.412]	0.814
Country ID	A	Reference		
	B	**-0.570**	**[-0.841 -0.300]**	**<0.001**
	D	1.301	[-0.130, 2.732]	0.069
	E	-0.320	[-0.9778, 0.336]	0.325
	F	**-0.914**	**[-1.763, -0.065]**	**0.034**
	G	**-2.456**	**[-4.124, -0.788]**	**0.004**
	H	-1.248	[-2.778, 0.280]	0.116
	Other	-0.421	[-1.305, 0.440]	0.507
Previous Treatment	None	Reference		
	<6 mon. cons.	-0.192	[-0.529, 0.144]	0.323
	>6 mon. cons.	0.299	[-0.050, 0.647]	0.093
	Surgical	**0.625**	**[0.042, 1.208]**	**0.035**
Level of Spine	L5/S1	Reference		
	L4/L5	-0.031	[-0.291, 0.228]	0.927
	L3/L4	-0.284	[-0.775, 0.205]	0.325
	L2/L3	0.208	[-0.576, 0.993]	0.567
	L1/L2	4.095	[-1.394, 9.584]	0.155
	Other	-0.087	[-1.008, 0.833]	0.803
ASA Morbidity	1	Reference		
	2	**0.518**	**[0.251, 0.786]**	**<0.001**
	3	**0.754**	**[0.204, 1.303]**	**0.007**
	4	**4.539**	**[0.669, 8.410]**	**0.022**
Baseline COMI score		**0.400**	**[0.331, 0.470]**	**<0.001**

Table 5.1: Coefficient estimates, 95% confidence intervals and p-values of variables in the linear regression model with 3-month COMI scores.

The country in which the intervention took place, previous treatment (especially prior surgery), baseline COMI scores and ASA morbidity status seem to be associated with treatment outcome for COMI scores at 3-months past surgery. Specifically, COMI outcome scores from country B, F and G were lower than from the reference country A. Prior surgery and ASA morbidity scores of 2, 3 or 4 were associated with higher COMI outcome scores.

The R^2 of a given model is a goodness-of-fit measure for linear models. It identifies the variance in the outcome that is explained by the included input parameters, in this case patient characteristics. It is computed by dividing the residual mean square error by the total mean square error. The result is subtracted from 1. Values close to 1 represent a good fit and show that the model explains large proportions of outcome variation, whereas values close to zero show the opposite. The linear regression model with 3-month COMI scores as outcome had an R^2 of 0.098 and could therefore only explain a small fraction of the outcome variability.

Prediction accuracy was assessed using the root mean squared error (RMSE), which can be described as the average distance between real values and the corresponding predicted value on the regression line. The model after AIC model selection has a RMSE of 2.618, which is even higher than the minimal clinically important difference (MCID) of the COMI questionnaire score, which is 2.2. It can therefore be said that prediction accuracy is low.

Only a small percentage of outcome variation could be explained by the model which demonstrates a poor accuracy when used for individual predictions. However, significance in some of the patient covariates (previous treatment, ASA morbidity status, baseline COMI scores and country) shows that prognostic factors could be identified.

In order to investigate if these variables were consistently predictive for COMI outcomes, the same method of model fitting were applied to other outcome time points.

5.5.2 Linear regression – COMI Scores at one-year past surgery

This approach will be the same as for 3-month outcomes, the only difference being that 1-year outcomes of COMI scores, and therefore a set of 2,190 patients will be used. A total of 1,389 of these patients were also part of the 3-month outcome set. At this time point there were no patients that had surgery at the level "L1 / L2". Again `lm()` was used to fit this model and included the same patient characteristics. A table of estimates, 95%-confidence intervals and p-values is displayed in Table 5.2.

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		**1.129**	**[0.345, 1.914]**	**0.005**
Sex	Female	Reference		

	Male	-0.089	[-0.330, 0.152]	0.470
Age		-0.008	[-0.017, 0.002]	0.135
Surgeon credentials	BC-N	Reference		
	SSS	-0.036	[-0.330, 0.257]	0.809
	N-t	0.496	[-0.026, 0.916]	0.081
	BC-O	-0.170	[-0.867, 0.526]	0.631
	O-t	0.022	[-1.042, 1.087]	0.822
	Other	-0.091	[-1.282, 1.099]	0.750
Country ID	A	Reference		
	B	**-0.793**	[-1.073, -0.512]	**<0.001**
	D	1.756	[-0.365, 3.878]	0.154
	E	-0.229	[-0.727, 0.270]	0.419
	F	-0.670	[-1.427, 0.086]	0.089
	G	-0.518	[-1.679, 0.643]	0.443
	H	-0.425	[-2.293, 1.442]	0.655
	Other	-0.271	[-1.673, 1.131]	0.773
Previous Treatment	None	Reference		
	<6 mon. cons.	-0.069	[-0.414, 0.275]	0.692
	>6 mon. cons.	0.226	[-0.128, 0.581]	0.211
	Surgical	**1.533**	[0.913, 2.153]	**<0.001**
Level of spine	L5/S1	Reference		
	L4/L5	-0.221	[-0.482, 0.040]	0.107
	L3/L4	**-0.606**	**[-1.095, -0.117]**	**0.020**
	L2/L3	0.199	[-0.570, 0.968]	0.666
	Other	0.274	[-0.729, 1.277]	0.558
ASA Morbidity	1	Reference		
	2	**0.408**	[0.130, 0.669]	**0.003**
	3	**0.863**	[0.234, 1.316]	**0.002**
	4	1.654	[-2.777, 5.173]	0.317
Baseline COMI score		**0.379**	**[0.309, 0.445]**	**<0.001**

Table 5.2: Coefficient estimates, 95% confidence intervals and p-values of variables in linear regression with 1-year COMI scores.

There are some differences between the model fit of 1-year and 3-month outcomes. Level of spine of "L3 / L4" had a significant p-value and had lower COMI outcome scores, which was not the case in the 3-month model.

The reverse can be observed for countries F and G, for which now there are no significant p-values. ASA morbidity of stage 4 did not show a significant p-value either. However, it has to be considered that there were very few observations in this subgroup (3). The covariates with considerable significance in terms of p-value and confidence intervals were the same among both outcome times, namely country B, prior surgery, baseline COMI scores and ASA morbidity (ASA 1 was reference category).

Prediction accuracy was again assessed using RMSE. For the model after AIC model selection, this error is 2.712 and therefore similarly large as the error in the 3-month outcome model. The R^2 of this model was 0.095. The prediction accuracy and goodness of fit therefore remain poor. Finally, the same will be done for 2-year outcomes.

5.5.3 Linear regression – COMI Scores at two years past surgery

Again, the approach will be the same, now including 1,679 patients. A total of 1,138 and 1,352 of these patients were also part of the 3-month and 1-year outcome set, respectively. Again `lm()` was used to fit this model and included the same patient characteristics and applied the AIC model selection algorithm. A table of estimates, 95%-confidence intervals and p-values is displayed in Table 5.3.

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		**1.321**	**[0.386, 2.256]**	**0.005**
Sex	Female	Reference		
	Male	**-0.323**	**[-0.603, -0.042]**	**0.024**
Age		-0.007	[-0.018, 0.003]	0.223
Surgeon credentials	BC-N	Reference		
	SSS	-0.332	[-0.667, 0.003]	0.055
	N-t	0.250	[-0.180, 0.679]	0.254
	BC-O	-0.427	[-1.300, 0.445]	0.333
	O-t	0.165	[-0.964, 1.294]	0.774
	Other	-0.080	[-1.386, 1.225]	0.904
Country ID	A	Reference		
	B	**-0.860**	**[-1.181, -0.540]**	**<0.001**
	D	-2.500	[-6.140, 1.140]	0.385

	E	-0.206	[-0.822, 0.410]	0.511
	F	-0.209	[-1.220, 0.802]	0.686
	H	-2.341	[-5.621, 0.939]	0.161
Previous Treatment	None	Reference		
	<6 mon. cons.	-0.185	[-0.571, 0.201]	0.349
	>6 mon. cons.	0.190	[-0.211, 0.591]	0.351
	Surgical	**0.753**	**[0.036, 1.471]**	**0.040**
Level of spine	L5/S1	Reference		
	L4/L5	0.128	[-0.169, 0.424]	0.405
	L3/L4	-0.224	[-0.798, 0.349]	0.444
	L2/L3	0.693	[-0.751, 2.138]	0.149
	Other	0.628	[-0.601, 1.857]	0.315
ASA Morbidity	1	Reference		
	2	**0.554**	**[0.243, 0.865]**	**0.001**
	3	**1.481**	**[0.842, 2.121]**	**<0.001**
Baseline COMI score		0.346	**[0.265, 0.428]**	**<0.001**

Table 5.3: Coefficient estimates, 95% confidence intervals and p-values for linear regression with 2-years COMI scores.

Counties in category G and "other" did not have 2-year outcomes available. Additionally, no patients that were classified as ASA category 4 had 2-year outcomes available. At this timepoint, male patients appear to have performed significantly better than female patients at 2 years after surgery. Similar to the 3-month outcomes, country B was associated with lower COMI outcome scores than in country A. Again, surgery at level L3/L4 was associated with lower COMI outcome scores than at level L5/S1.

To summarize, the following variables were consistently detected as correlated to outcome (although significances of categories of these variables were not always consistent): country ID, baseline COMI scores, ASA morbidity and prior surgery. In the 1- and 2-year outcomes, surgery at L3/L4 was detected to perform better than the reference category L5/S1.

The RMSE and R^2 of the model are 2.734 and 0.089 respectively. The model approach therefore shows a similarly poor prediction accuracy, similar to the 3-month and 1-year outcomes. A summary of the model fit statistics at all time points is given in Table 5.4.

Time interval / Model statistics	3 months	1 year	2 years
RMSE	2.618	2.712	2.734
Significant variables	Country, Previous treatment, ASA morbidity, Baseline COMI scores	Country, Previous treatment, ASA morbidity, Baseline COMI scores, Level of spine	Country, Previous treatment, ASA morbidity, Baseline COMI scores, Sex
R^2	0.098	0.095	0.089
Number of patients	2,127	2,190	1,679

Table 5.4: Summary of model fit statistics of linear regression approaches. Column headers are the time point of the primary outcome in the regression model.

None of the models had an R^2 of more than 0.1 and can be regarded as poor fit in terms of their ability to explain outcome variability and predict individual treatment outcome.

5.5.4 Linear regression – Subsets of BMI and Smoking status

BMI and smoking status were available only to a subset of patients. Initial missing data percentage was large, which is why these two variables were not imputed. In this section the subsets of patients that had these variables available are analysed to investigate if these two variables could help improve goodness of fit. It must be reminded throughout the following analyses, that there were no measurements at 1 and 2 years available for country D and countries in category "other". For country G there was no 2-year outcome available and only 6 and 9 measurements for the BMI and smoking status subset respectively. Estimates for these cases should therefore be interpreted with caution.

The details about estimates, 95%-confidence intervals, and p-values for the model with COMI scores as outcome are displayed in Tables 5.5-5.7.

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		**1.637**	**[0.468, 2.806]**	**0.006**
Sex	Female	Reference		
	Male	0.137	[-0.157, 0.431]	0.357
Age		-0.009	[-0.021, 0.003]	0.163
Surgeon credentials	BC-N	Reference		
	SSS	-0.201	[-0.550, 0.148]	0.259
	N-t	-0.051	[-0.540, 0.439]	0.837
	BC-O	-0.256	[-1.484, 0.972]	0.680
	O-t	0.055	[-1.174, 1.283]	0.930

	Other	0.044	[-0.760, 0.848]	0.942
Country ID	A	Reference		
	B	**-0.398**	**[-0.760, -0.036]**	**0.031**
	E	-0.265	[-1.053, 0.524]	0.509
	F	-0.943	[-2.117, 0.231]	0.111
	G	-2.677	[-4.567, -0.787]	0.005
	H	-1.125	[-2.676, 0.426]	0.153
	Other	-0.444	[-3.651, 2.761]	0.786
Previous Treatment	None	Reference		
	<6 mon. cons.	-0.217	[-0.605, 0.170]	0.273
	>6 mon. cons.	0.328	[-0.076, 0.732]	0.112
	Surgical	0.640	[-0.069, 1.350]	0.074
Level of spine	L5/S1	Reference		
	L4/L5	-0.069	[-0.383, 0.245]	0.665
	L3/L4	-0.217	[-0.816, 0.382]	0.476
	L2/L3	-0.042	[-0.966, 0.882]	0.928
	Other	0.033	[-0.987, 1.052]	0.950
ASA Morbidity	1	Reference		
	2	**0.457**	**[0.130, 0.785]**	**0.006**
	3	**0.706**	**[0.001, 1.412]**	**0.050**
BMI	<20	Reference		
	20 - 25	-0.509	[-1.246, 0.228]	0.174
	25 - 30	-0.385	[-1.130, 0.359]	0.304
	30 – 35	-0.003	[-0.797, 0.791]	0.944
	>35	-0.487	[-1.414, 0.439]	0.301
Baseline COMI score		**0.425**	**[0.341, 0.510]**	**<0.001**

Table 5.5: Coefficient estimates, 95% confidence intervals and p-values for linear regression with 3-month COMI scores on a patient subset that included BMI. Cases for which BMI was exactly 25 were included in "20 – 25". The same method applies to other categories.

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		**1.343**	**[0.191, 2.495]**	**0.022**
Sex	Female	Reference		
	Male	0.100	[-0.201, 0.400]	0.516

Age		**-0.014**	**[-0.027, -0.002]**	**0.030**
Surgeon credentials	BC-N	Reference		
	SSS	-0.175	[-0.535, 0.184]	0.259
	N-t	0.259	[-0.246, 0.764]	0.313
	BC-O	-0.585	[-1.439, 0.269]	0.178
	O-t	-0.359	[-1.953, 1.236]	0.656
	Other	-0.190	[-1.417, 1.036]	0.761
Country ID	A	Reference		
	B	**-0.907**	**[-1.287, -0.528]**	**<0.001**
	E	0.057	[-0.538, 0.653]	0.852
	F	0.643	[-1.105, 2.391]	0.469
	G	-0.258	[-1.579, 1.062]	0.701
	H	-0.290	[-2.199, 1.619]	0.765
Previous Treatment	None	Reference		
	<6 mon. cons.	-0.077	[-0.482, 0.329]	0.710
	>6 mon. cons.	0.140	[-0.277, 0.557]	0.509
	Surgical	**1.393**	**[0.647, 2.139]**	**<0.001**
Level of spine	L5/S1	Reference		
	L4/L5	-0.301	[-0.383, 0.245]	0.665
	L3/L4	-0.447	[-0.816, 0.382]	0.476
	L2/L3	0.233	[-0.966, 0.882]	0.928
	Other	0.118	[-0.987, 1.052]	0.950
ASA Morbidity	1	Reference		
	2	**0.431**	**[0.097, 0.766]**	**0.012**
	3	**0.923**	**[0.212, 1.634]**	**0.011**
	4	-3.297	[-8.099, 1.505]	0.262
BMI	<20	Reference		
	20 - 25	-0.141	[-0.875, 0.592]	0.707
	25 - 30	-0.104	[-0.843, 0.634]	0.781
	30 – 35	0.31	[-0.482, 1.102]	0.441
	>35	0.034	[-0.920, 0.989]	0.944
Baseline COMI score		**0.388**	**[0.302, 0.475]**	**<0.001**

Table 5.6: Coefficient estimates, 95% confidence intervals and p-values for linear regression with 1-year COMI scores on a patient subset that included BMI. Cases for which BMI was exactly 25 were included in "20 – 25". The same method applies to other categories.

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		0.749	[-0.176, 1.674]	0.291
Sex	Female	Reference		
	Male	-0.198	[-0.563, 0.168]	0.288
Age		**-0.016**	**[-0.030, -0.001]**	**0.044**
Surgeon credentials	BC-N	Reference		
	SSS	-0.632	[-1.707, 0.443]	0.247
	N-t	-0.405	[-0.845, 0.035]	0.073
	BC-O	0.256	[-0.305, 0.817]	0.370
	O-t	-0.500	[-2.703, 1.704]	0.657
	Other	-0.002	[-1.358, 1.355]	0.998
Country ID	A	Reference		
	B	-0.21	[-0.758, 0.339]	0.454
	E	0.168	[-0.603, 0.938]	0.667
	F	0.633	[-1.216, 2.482]	0.502
	H	-2.144	[-5.519, 1.231]	0.211
Previous Treatment	None	Reference		
	<6 mon. cons.	-0.198	[-0.663, 0.267]	0.404
	>6 mon. cons.	-0.025	[-0.509, 0.460]	0.921
	Surgical	0.496	[-0.460, 1.452]	0.305
Level of spine	L5/S1	Reference		
	L4/L5	0.233	[-0.157, 0.623]	0.239
	L3/L4	0.033	[-0.768, 0.835]	0.935
	L2/L3	0.365	[-0.896, 1.626]	0.566
	Other	0.835	[-0.500, 2.171]	0.218
ASA Morbidity	1	Reference		
	2	**0.615**	**[-0.009, 1.238]**	**0.003**
	3	**1.671**	**[0.822, 2.520]**	**<0.001**

BMI	<20	Reference		
	20 - 25	0.092	[-0.806, 0.989]	0.840
	25 - 30	0.429	[-0.471, 1.329]	0.344
	30 – 35	0.576	[-0.376, 1.528]	0.233
	>35	0.444	[-0.699, 1.587]	0.442
Baseline COMI score		**0.405**	**[0.300, 0.509]**	**<0.001**

Table 5.7: Coefficient estimates, 95% confidence intervals and p-values for linear regression with 2-year COMI scores on a patient subset that included BMI. Cases for which BMI was exactly 25 were included in "20 – 25". The same method applies to other categories.

Overall, variables and categories that had significant p-values were similar to the analysis approaches without the inclusion of BMI, the only difference being that age was associated with lower COMI outcomes at 1- and 2-years after surgery. None of the models indicated that BMI was associated with treatment outcome.

The details about estimates, 95%-confidence intervals and p-values for the model on the subset of patients that had smoking status available are displayed in Tables 5.8-5.10.

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		**1.541**	**[0.426, 2.656]**	**0.006**
Sex	Female	Reference		
	Male	0.151	[-0.176, 0.479]	0.372
Age		-0.011	[-0.024, 0.001]	0.122
Surgeon credentials	BC-N	Reference		
	SSS	0.18	[-1.411, 1.772]	0.824
	N-t	-0.209	[-0.614, 0.195]	0.308
	BC-O	-0.1	[-0.680, 0.481]	0.735
	O-t	-0.03	[-1.513, 1.454]	0.969
	Other	2.225	[-0.043, 4.494]	0.054
Country ID	A	Reference		
	B	**-0.59**	**[-1.021, -0.159]**	**0.003**
	E	-0.44	[-1.279, 0.398]	0.301
	F	-1.315	[-2.715, 0.085]	0.066
	G	**-3.098**	**[-5.330, -0.867]**	**0.007**
	H	-1.001	[-2.593, 0.590]	0.219
	Other	-0.646	[-3.828, 2.536]	0.691

Previous Treatment	None	Reference		
	<6 mon. cons.	-0.239	[-0.695, 0.217]	0.305
	>6 mon. cons.	0.374	[-0.101, 0.849]	0.121
	Surgical	0.797	[-0.019, 1.614]	0.056
Level of spine	L5/S1	Reference		
	L4/L5	-0.077	[-0.436, 0.283]	0.673
	L3/L4	-0.246	[-0.909, 0.415]	0.467
	L2/L3	0.256	[-0.757, 1.270]	0.619
	Other	-0.284	[-1.386, 0.818]	0.614
ASA Morbidity	1	Reference		
	2	**0.453**	**[0.084, 0.822]**	**0.016**
	3	0.651	[-0.146, 1.448]	0.108
Smoking status	Non-smoker	Reference		
	Smoker	**0.747**	**[0.338, 1.156]**	**<0.001**
Baseline COMI score		**0.395**	**[0.299, 0.491]**	**<0.001**

Table 5.8: Coefficient estimates, 95% confidence intervals and p-values for linear regression with 3-month COMI scores on a patient subset that included smoking status.

There was only one measurement for country D, which was therefore disregarded. In the following subset of 1-year outcomes there were no measurements for country group D or "other".

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		**1.5**	**[0.386, 2.614]**	**0.007**
Sex	Female	Reference		
	Male	-0.086	[-0.431, 0.258]	0.625
Age		-0.014	[-0.029, 0.001]	0.052
Surgeon credentials	BC-N	Reference		
	SSS	-0.556	[-1.511, 0.398]	0.254
	N-t	-0.238	[-0.667, 0.190]	0.276
	BC-O	0.405	[-0.200, 1.009]	0.190
	O-t	0.034	[-2.268, 2.336]	0.977
	Other	0.869	[-1.728, 3.466]	0.505
Country ID	A	Reference		
	B	**-1.038**	**[-1.455, -0.622]**	**<0.001**
	E	-0.042	[-0.674, 0.590]	0.895

	F	0.944	[-1.098, 2.987]	0.363
	G	-0.661	[-2.146, 0.824]	0.382
	H	-0.356	[-2.271, 1.560]	0.715
Previous Treatment	None	Reference		
	<6 mon. cons.	-0.129	[-0.595, 0.337]	0.590
	>6 mon. cons.	0.254	[-0.230, 0.737]	0.300
	Surgical	**1.905**	**[0.997, 2.813]**	**<0.001**
Level of spine	L5/S1	Reference		
	L4/L5	-0.352	[-0.723, 0.018]	0.062
	L3/L4	-0.261	[-0.987, 0.465]	0.479
	L2/L3	0.359	[-0.656, 1.375]	0.484
	Other	0.013	[-1.183, 1.209]	0.983
ASA Morbidity	1	Reference		
	2	**0.578**	**[0.196, 0.961]**	**0.003**
	3	**0.935**	**[0.144, 1.726]**	**0.021**
Smoking status	Non-smoker	Reference		
	Smoker	**0.751**	**[0.324, 1.179]**	**0.001**
Baseline COMI score		**0.357**	**[0.259, 0.456]**	**<0.001**

Table 5.9: Coefficient estimates, 95% confidence intervals and p-values for linear regression with 1-year COMI scores on a patient subset that included smoking status.

In the following subset of 2-year outcomes there were no measurements for country group D, G or "Other".

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		**1.543**	**[0.193, 2.894]**	**0.025**
Sex	Female	Reference		
	Male	-0.31	[-0.727, 0.106]	0.147
Age		**-0.021**	**[-0.038, -0.003]**	**0.019**
Surgeon credentials	BC-N	Reference		
	SSS	-0.282	[-1.439, 0.876]	0.629
	N-t	-0.326	[-0.836, 0.185]	0.211
	BC-O	0.219	[-0.446, 0.884]	0.519
	O-t	-1.37	[-4.036, 1.296]	0.308
	Other	0.187	[-2.479, 2.853]	0.890

Country ID	A	Reference		
	B	-0.497	[-1.074, 0.080]	0.093
	E	-0.269	[-1.066, 0.527]	0.504
	F	0.921	[-1.272, 3.115]	0.407
	H	-2.199	[-5.050, 0.652]	0.203
Previous Treatment	None	Reference		
	<6 mon. cons.	-0.437	[-0.980, 0.106]	0.113
	>6 mon. cons.	-0.082	[-0.635, 0.471]	0.771
	Surgical	0.749	[-0.339, 1.837]	0.176
Level of spine	L5/S1	Reference		
	L4/L5	0.146	[-0.300, 0.592]	0.520
	L3/L4	0.107	[-0.831, 1.045]	0.822
	L2/L3	0.518	[-0.894, 1.929]	0.472
	Other	0.027	[-1.428, 1.482]	0.971
ASA Morbidity	**1**	**Reference**		
	2	**0.83**	**[0.362, 1.298]**	**0.001**
	3	**1.793**	**[0.834, 2.751]**	**<0.001**
Smoking status	**Non-smoker**	**Reference**		
	Smoker	**0.635**	**[0.098, 1.173]**	**0.020**
Baseline COMI score		**0.395**	**[0.277, 0.513]**	**<0.001**

Table 5.10: Coefficient estimates, 95% confidence intervals and p-values for linear regression with 2-year COMI scores on a patient subset that included smoking status.

Results are very similar to prior approaches that did not include smoking status. However, smoking status was significant in each of the outcome time points for COMI scores. More precisely, smokers were associated with higher COMI outcome scores than non-smokers.

R^2 values of the models including BMI were 0.108, 0.109 and 0.106 for the 3-month, 1-year and 2-year model respectively. R^2 values of the models including smoking status were 0.120, 0.129 and 0.127 for the 3-month, 1-year and 2-year model respectively.

RMSEs of the models including BMI were 2.878, 2.791 and 2.811 respectively and for the models including smoking status RMSEs were 2.561, 2.617 and 2.697 for the 3-month, 1-year and 2-year model respectively.

Each of the models had R^2 values and root mean squared errors similar to the linear model approaches with the full set of available patients for each time point and can therefore be described having a poor

fit. It should be pointed out though, that smoking status is consistently correlated to treatment outcome and should be considered in modelling approaches if available.

Overall, even though patient characteristics could be identified as associated to treatment outcome, model fits in terms of R^2 and RMSE is poor. There remains large outcome variation that cannot be explained by this model approach, which is why other approaches were explored. As illustrated in Figure 5.2, COMI scores deviate from normality and are constrained within the range of 0 and 10. Linear regression assumes the outcome to be unbounded, which is a limitation of the application to this outcome and opens the possibility for future research to incorporate advanced analytical techniques, such as tobit regression, to accommodate this distribution. It's important to note that while this analysis holds promise, given the current scope of this project, it was not pursued. Instead, the exploration of logistic regression was pursued. The reasoning was to simplify the outcome, before exploring more complex models such as mixed models. By transforming COMI to a binary outcome and using this approach, it was hoped to identify factors that are associated with successful treatment outcome (significant improvement), as opposed to unsuccessful (no significant improvement).

5.6 Logistic regression approaches

Another common approach that uses one specific outcome measurement is a logistic regression model. For this, the outcome needs to be binary. Therefore, the treatment outcome was dichotomized into “significant improvement” or “no significant improvement”. The threshold for the treatment outcome to be successful was decided based on the minimal clinically important difference of the COMI score, which was 2.2 points (Mannion et al., 2016). The model approach is based on formula (2).

$$\mathrm{Logit}(\mathrm{p_i}) = \ln\left(\frac{\mathrm{p_i}}{1-\mathrm{p_i}}\right) = X_i\beta + \varepsilon_i \qquad (2)$$

Where $\mathrm{p_i}$ is the probability of the outcome of patient i being successful. The rest of the annotation is defined in the same way as in formula (1). For all further model the `glm()` function in R with the specification family = “binomial” was used.

5.6.1 Logistic regression – Treatment success using COMI scores at three months, 1 year and 2 years after surgery

This model approach classified successful treatment based on a point decrease of greater or equal 2.2 on the COMI score scale (0-10). For each of the time points a logistic model of the form (2) was fitted

and a stepwise AIC model selection algorithm applied in order to eliminate non-significant variables. The individuals who fell under categories ASA 4 and country IDs D and H were excluded from the respective analyses of 1-year and 2-year outcomes due to their low representation in the datasets. Additionally, there were no individuals from country G and "other" countries in the 2-year dataset.

A table of Odds ratios, 95%-confidence intervals and p-values for 3-month outcomes is displayed in Table 5.11.

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		**-1.12**	**[-1.723, -0.510]**	**<0.001**
Gender	Female	Reference		
	Male	-0.004	[-0.191, 0.183]	0.969
Age		<0.001	[-0.008, 0.008]	0.984
Surgeon credentials	BC-N	Reference		
	SSS	0.129	[-0.091, 0.351]	0.251
	N-t	0.082	[-0.222, 0.391]	0.598
	BC-O	0.283	[-0.447, 1.072]	0.461
	O-t	0.190	[-0.515, 0.959]	0.609
	Other	0.140	[-0.728, 1.055]	0.756
Level of spine	L5/S1	Reference		
	L4/L5	0.014	[-0.187, 0.215]	0.894
	L3/L4	0.119	[-0.260, 0.507]	0.543
	L2/L3	0.052	[-0.646, 0.568]	0.887
	Other	0.005	[-0.686, 0.730]	0.989
Country ID	A	Reference		
	B	**0.263**	**[0.054, 0.473]**	**0.014**
	D	-0.756	[-1.845, 0.313]	0.162
	E	0.101	[-0.404, 0.626]	0.700
	F	0.480	[-0.194, 1.216]	0.179
	G	**2.462**	**[0.763, 5.388]**	**0.021**
	H	0.742	[-0.479, 2.261]	0.272
	Other	0.293	[-0.371, 1.007]	0.400
Previous Treatment	None	Reference		
	<6 mon. cons.	0.067	[-0.197, 0.328]	0.618
	>6 mon. cons.	**-0.273**	**[-0.541, -0.006]**	**0.045**

	Surgical	**-0.536**	**[-0.973, -0.095]**	**0.017**
ASA Morbidity	1	Reference		
	2	**-0.236**	**[-0.443, -0.029]**	**0.025**
	3	-0.257	[-0.676, 0.173]	0.256
Baseline COMI score		**0.227**	**[0.174, 0.281]**	**<0.001**

Table 5.11: Odds ratios, 95% confidence intervals and p-values of variables of logistic regression using 3-month successful treatment, based on clinically significant COMI score changes.

A table of Odds ratios, 95%-confidence intervals and p-values for 1-year outcomes is displayed in Table 5.12.

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		**-1.056**	**[-1.674, -0.438]**	**0.001**
Gender	Female	Reference		
	Male	0.035	[-0.160, 0.230]	0.725
Age		0.004	[-0.004, 0.011]	0.333
Surgeon credentials	BC-N	Reference		
	SSS	0.07	[-0.491, 0.631]	0.807
	N-t	**-0.356**	**[-0.662, -0.051]**	**0.022**
	BC-O	0.018	[-0.220, 0.257]	0.883
	O-t	-0.017	[-0.919, 0.885]	0.971
	Other	0.633	[-0.419, 1.685]	0.233
Level of spine	L5/S1	Reference		
	L4/L5	0.159	[-0.052, 0.369]	0.139
	L3/L4	0.251	[-0.162, 0.665]	0.230
	L2/L3	-0.292	[-0.898, 0.315]	0.345
	Other	-0.136	[-0.899, 0.628]	0.725
Country ID	A	Reference		
	B	**0.592**	**[0.361, 0.823]**	**<0.001**
	D	-0.367	[-1.910, 1.176]	0.640
	E	0.147	[-0.248, 0.543]	0.465
	F	**0.667**	**[0.003, 1.331]**	**0.049**
	G	0.272	[-0.619, 1.164]	0.547
	H	0.732	[-0.896, 2.360]	0.373
	Other	0.138	[-0.968, 1.244]	0.804

Previous Treatment	None	Reference		
	<6 mon. cons.	0.137	[-0.138, 0.412]	0.330
	>6 mon. cons.	-0.118	[-0.395, 0.160]	0.405
	Surgical	**-0.809**	**[-1.266, -0.352]**	**0.001**
ASA Morbidity	1	Reference		
	2	-0.212	[-0.429, 0.003]	0.054
	3	**-0.567**	**[-0.982, -0.152]**	**0.009**
Baseline COMI score		**0.215**	**[0.161, 0.269]**	**<0.001**

Table 5.12: Odds ratios, 95% confidence intervals and p-values of variables of logistic regression using 2-year successful treatment, based on clinically significant COMI score changes.

A table of Odds ratios, 95%-confidence intervals and p-values for 2-year outcomes is displayed in Table 5.13.

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		**-0.923**	**[-1.643, -0.205]**	**<0.012**
Gender	Female	Reference		
	Male	**0.215**	**[-0.008, 0.440]**	**0.059**
Age		0.003	[-0.006, 0.012]	0.549
Surgeon credentials	BC-N	Reference		
	SSS	0.185	[-0.087, 0.461]	0.183
	N-t	**-0.346**	**[-0.669, -0.021]**	**0.036**
	BC-O	0.165	[-0.547, 0.915]	0.655
	O-t	-0.188	[-1.055, 0.786]	0.683
	Other	0.331	[-0.667, 1.491]	0.538
Level of spine	L5/S1	Reference		
	L4/L5	0.019	[-0.219, 0.259]	0.872
	L3/L4	0.165	[-0.299, 0.651]	0.495
	L2/L3	-0.570	[-1.265, 0.160]	0.114
	Other	-0.863	[-1.763, 0.037]	0.057
Country ID	A	Reference		
	B	**0.437**	**[0.177, 0.701]**	**0.001**
	E	0.461	[-0.053, 1.003]	0.086
	F	-0.089	[-0.848, 0.724]	0.822
Previous Treatment	None	Reference		

	<6 mon. cons.	0.109	[-0.203, 0.418]	0.489
	>6 mon. cons.	-0.131	[-0.445, 0.177]	0.405
	Surgical	-0.258	[-0.807, 0.311]	0.363
ASA Morbidity	1	Reference		
	2	**-0.282**	**[-0.531, -0.035]**	**0.025**
	3	**-0.729**	**[-1.215, -0.235]**	**0.003**
Baseline COMI score		**0.214**	**[0.152, 0.277]**	**<0.001**

Table 5.13: Odds ratios, 95% confidence intervals and p-values of variables of logistic regression using 2-year successful treatment, based on clinically significant COMI score changes.

The models used to predict outcomes at three different time points all included country ID, ASA morbidity, and baseline COMI scores as important factors. Again, country B was associated with better treatment outcomes. However, in the logistic regression the estimates are Odds ratios, with a positive value indicating a higher chance of having significant improvement, in comparison to the reference category. Similarly, patients with ASA morbidity status of 2 or 3 had a lower chance for significant improvement, compared to patients with ASA morbidity of 1. A one-unit increase in COMI baseline score is associated with an increase of the chance of not having a significant improvement.

The 2-year outcome model was the only one in which previous treatment was not considered during model selection, which raises questions for further discussion. The linear regression model used at 1 year after surgery included the level of the spine as a factor, while the logistic regression model did not. Overall, the main predictive factors were similar across both linear regression and logistic regression approaches.

There are measures that attempt to quantify something similar to the R^2 in linear models like the Cox-Snell R^2 or the McFadden R^2, but the most common model performance measure for a logistic regression is the receiver operating characteristic (ROC) curve, which shows the true positive rate against the true negative rate for various thresholds. The area under this curve (AUC) is used to measure the model's ability to predict outcomes. The closer the AUC value is to one, the better the fit of the model and the worst fit would be similar to a coin toss at 0.5.

The ROC-curves for the above models are displayed in Figure 5.3.

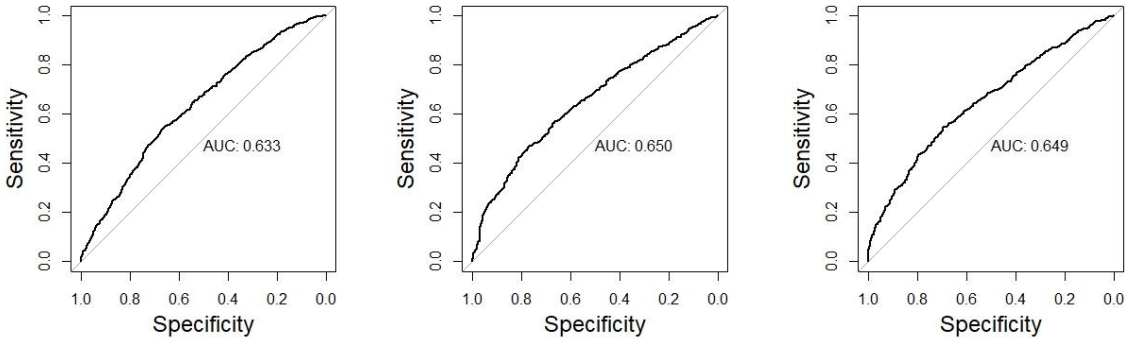


Figure 5.3: Areas under ROC-curves of logistic regression models for each time point past surgery.

95%-confidence intervals of the areas under ROC-curves were 0.633 [0.608, 0.657], 0.650 [0.626, 0.675] and 0.649 [0.621, 0.677] for dichotomised 3-month, 1-year and 2-year outcomes respectively. Neither of these ROC-curves represent a good predictive ability of the model regarding the precision of individual predictions or a good model fit. For the purpose of identifying prognostic factors however, these were mostly consistent with the results of the linear regression analyses that were performed previously. Due to the similarities of the model fit and the factors that were identified to be associated with treatment outcome, a subset of analysis of patients that had BMI or smoking status available was not performed.

5.6.2 Summary

Overall, linear and logistic regression with one outcome time point had poor fit and therefore also poor ability to predict outcomes for patient when only baseline data is available. There seems to be a lot of unexplained outcome variation left, which is why other modelling approaches were explored. The main take away is that a few consistent patient characteristics were identified that were correlated to treatment outcome, even if the fit was poor. ASA morbidity status, prior treatment, where prior surgery had the lowest p-values, baseline COMI scores and country were considered to be correlated in most of the models and smoking status was correlated consistently in the models that used the subset of patients that answered it. The identification of a few country IDs for which outcomes were different from the reference country 6 is also helpful for the Spine Tango registry and clinicians in these countries. It was found out that BMI was not significant regarding outcomes, which is why it was not considered in future modelling approaches. While smoking status was found to be significant, the subset of patients with available data on smoking status was notably smaller. As such, there is a need for more consistent measurement of smoking status and its inclusion in the development of a core outcome set for this patient population and treatment. However, for future model approaches, the focus will be on all available patients, without incorporating BMI or smoking status.

Every model approach so far only focussed on one specific outcome timepoint and therefore did not use the rest of the available data for the model fit. The next step is to explore if models that incorporate all available data points past surgery will be have a better fit.

5.7 Longitudinal mixed-effects model – COMI scores

In prior model approaches, outcome measures were defined at one chosen time point. Therefore, all other available data points were disregarded and information is lost. A longitudinal model that includes all available data points for each patient, could potentially achieve better model fit and be able to explain treatment outcome and changes over time. The approach works similar to a linear multivariate model, with the difference being that the outcome variable is time dependent. Instead of categorising COMI scores in “3 months”, “1 year” and “2 years”, exact values of weeks past surgery were used and treated as continuous variable. This led to a dataset containing 6,704 measurements of 3,530 patients. Again, factor levels were ordered either by size (country ID, surgeon credentials) so that the largest category is reference, or clinically (ASA morbidity, level of spine, previous treatment). Similar to previous model approaches, previous treatment was regrouped into “none”, “less than 6 months conservative treatment”, “more than 6 months conservative treatment” and “surgical”, where patients that had combinations such as “conservative and surgical treatment” were regarded into the “surgical” category. No previous treatment was chosen as reference category. Level of spine was ranked as following: “L5/S1” (reference category), “L4/L5”, “L3/L4” or “L2/L3” and “other”, where “other” also included “L1/L2” due to low count. ASA morbidity categories were ranked ascending with “ASA 1” being reference category. All patient baseline covariates were considered as constant over time. Although some of them might change over time, there are no repeat measurements available and therefore not considered time-dependent in the model approaches. The only time-dependent variables were the outcome scores and the time variable (in weeks) itself as input variable. Figure 5.4 shows a random sample of patients and their progression in COMI scores over time, to give a first impression of treatment improvement.

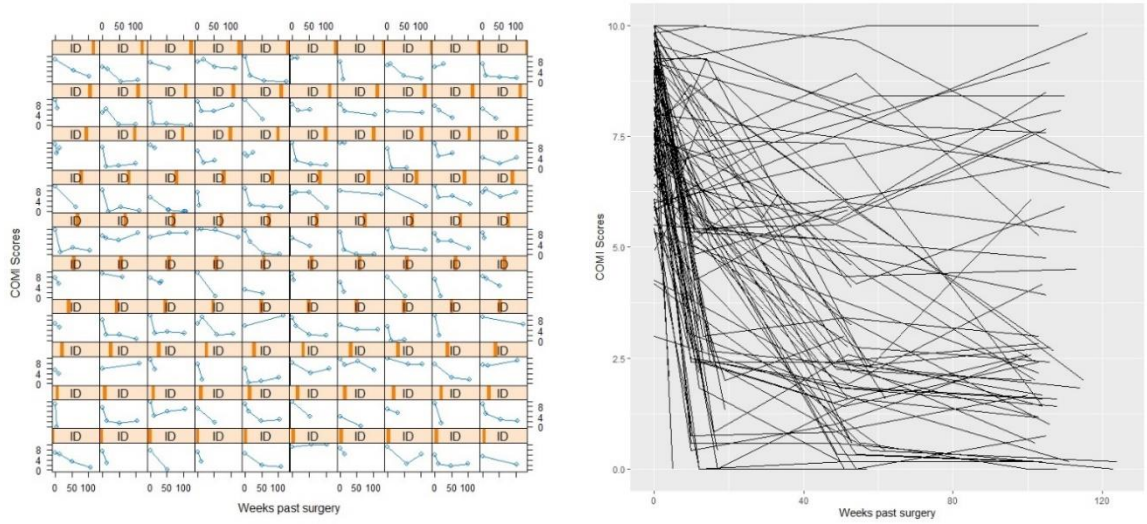


Figure 5.4 a) and b): a) COMI score progression over time for 100 random patients of data set individually (left) and b) overlapping (right).

Most patients show an initial improvement shortly after surgery. COMI scores then often stay constant for the following time intervals. However, there are some patients for who it continues to improve and also patients for who COMI scores become worse again after the initial improvement. Most patients have baseline scores of more than 5, but it seems that there is a wide spread of initial values between 5 and 10 on the score range before surgery.

Mixed effects models contain both fixed and random effects and are useful for datasets with repeated measurements on the same statistical units (in this case patients). Fixed effects are constant across all individuals, whereas random effects vary for each individual. This allows for a large number of potential models to be explored, but the focus is on the most straight-forward approaches, namely including random intercepts, slopes and both. The model is based on formula (3).

$$Y_i(t) = X_i(t)\beta + Z_i(t)u_i + \varepsilon_i \qquad (3)$$

In this formula β and ε_i are defined in the same way as in the linear regression model (1). $Y_i(t)$ and $X_i(t)$ are similarly defined as in the linear regression model, with the only difference that they are time-dependent, with t denoting the time. $Z_i(t)$ is a time-dependent $(1 \times q)$-dimensional design vector of the q random effects that are considered in the model that can be seen as equivalent to $X_i(t)$. u_i is a (q x 1)-dimensional vector of random effect coefficients, which can be seen as equivalent to β. However, this vector is individually defined for each patient. It is assumed to all u_i are independent

and identically distributed of the normal distribution $\sim N(0, D)$. For further details on the distribution of random effects, see (Daniels and Zhao, 2003).

When adding random effects, one has to include the desired variable as both fixed and random effect in the model specification. The coefficients will then automatically be fitted and if an effect is purely random, the fixed coefficient will be zero and if it is purely a fixed effect then the random coefficient will be zero.

Comparing this model approach to others is complex because it incorporates more time points. As a result, traditional measures such as loglikelihood may not be comparable. The model also allows for the inclusion of more patients, which is expected to lead to a better model fit. To evaluate the effectiveness of this approach, it will be compared to linear regression using the same subsets of patients. The prediction accuracy will be determined by comparing the root mean square errors of both methods. Additionally, the root mean square errors will be calculated for all available patients to evaluate the improvement from including more patients in the model. For example, there are patients in the mixed model at 1 year, that did not have a 1-year outcome and are therefore not included in the 1-year linear regression model. However, they had outcomes at different intervals (between surgery and 1 year) that can be used to fit the mixed model. Therefore, the mixed model can potentially include more patients at each time-point.

Additionally, there are several formulas that are somewhat an equivalent to the R^2 of linear regression models. One of the most common ones was developed by D. Zhang and extends the proportion of explained variance (Zhang, 2020). Moreover, it defines this proportion into variation explained by the whole model, fixed effects only, and random effects only. This measure is implemented in the "`rsq`"-package in R and will be used for all subsequent mixed-effects models. It will from now on be notated as R_z^2. All mixed-model analyses were done using the `lme()` function of the "`nlme`"-package in R (Pinheiro et al., 2013).

5.7.1 Mixed model in comparison to linear regression model at 3 months past surgery

The first model approach includes a random intercept. This allows each patient to have an individual intercept, which makes a lot of sense due to the variance of baseline values. In order to obtain a fair comparison with prior linear regression models, this model considers the same set of patients that were used in the linear regression model at 3 months past surgery and disregards later measurements. However, there were measurements for these patients between surgery and the 3-month outcome. The data set included 2,188 measurements from 2,127 patients. Only a few patients have more than one measurement in this dataset, since the 3-month interval is one of the first and later outcomes should not be included in the model fit when prediction accuracy is examined. Therefore, this data set

is nearly identical to the data set of the linear regression with 3-month outcomes. Model results regarding fixed effects are displayed in Table 5.14.

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		**2.441**	**[1.468, 3.414]**	**<0.001**
Weeks past surgery		**-0.067**	**[-0.108, -0.026]**	**0.002**
Sex	Female	Reference		
	Male	-0.068	[-0.308, 0.171]	0.580
Age		-0.009	[-0.018, 0.000]	0.073
Surgeon Credentials	BC-N	Reference		
	SSS	-0.197	[-0.482, 0.087]	0.178
	N-t	-0.155	[-0.552, 0.242]	0.445
	BC-O	-0.03	[-0.936, 0.876]	0.949
	O-t	0.008	[-0.898, 0.914]	0.986
	Other	0.048	[-0.741, 0.837]	0.936
Country ID	A	Reference		
	B	**-0.596**	**[-0.866, -0.327]**	**<0.001**
	D	**1.636**	**[0.187, 3.084]**	**0.027**
	E	-0.116	[-0.775, 0.543]	0.730
	F	-0.622	[-1.461, 0.217]	0.144
	G	**-2.224**	**[-3.822, -0.625]**	**0.006**
	H	-0.768	[-2.266, 0.729]	0.306
	Other	-0.137	[-1.007, 0.733]	0.756
Previous Treatment	None	Reference		
	<6 mon. cons.	-0.19	[-0.531, 0.150]	0.273
	>6 mon. cons.	0.319	[-0.033, 0.670]	0.081
	Surgical	**0.927**	**[0.338, 1.516]**	**0.002**
Level of Spine	L5/S1	Reference		
	L4/L5	-0.012	[-0.271, 0.247]	0.930
	L3/L4	-0.226	[-0.712, 0.260]	0.361
	L2/L3	0.207	[-0.576, 0.990]	0.603
	Other	0.096	[-0.815, 1.007]	0.835

ASA Morbidity	1	Reference		
	2	**0.519**	**[0.255, 0.784]**	**<0.001**
	3	**0.888**	**[0.343, 1.433]**	**0.002**
	4	**4.762**	**[0.879, 8.646]**	**0.016**
Baseline COMI score		**0.399**	**[0.329, 0.468]**	**<0.001**

Table 5.14: Coefficient estimates of fixed effects, 95% confidence intervals and p-values for mixed-effects model including random intercepts. Dataset of the same patients as in linear regression up to 3-month data.

The variables with significant p-values are the same as in previous modelling approaches, the only difference being that weeks past surgery additionally has a significant p-value. A one-week increase is associated with a lower COMI score (by 0.067 points).

For this current model fit the R_z^2 of the total model was 0.638, consisting of the fixed component of 0.103 and the random component of 0.535. The fixed component is very close to the R^2 that was found in linear regression models, which underlines, that linear regression models struggle to explain outcome variation. The addition of random intercepts led to an increase of proportion of explained outcome variation by 0.540 (0.638-0.098) and could therefore improve the model fit. However, this random component can only aid in individual predictions, if the coefficients of the random effects have been estimated. This estimation is possible only when post-surgery measurements are available for a given patient. Therefore, while the model fits more precisely by adjusting for individual progression, it does not necessarily imply that pre-surgery predictions can be made more accurately.

Including only random intercepts assumes that there is a linear decrease in COMI scores after surgery that is the same for the entire patient population. Figure 5.4 however shows that progression can be very different for patients, which is why the model was extended by adding random slopes. This allows each patient to have an individual intercept and slope instead of having the same estimates for the full population. Details of the fixed effects of the resulting model can be found in Table 5.15.

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		**2.444**	**[1.466, 3.422]**	**<0.001**
Weeks past surgery		**-0.067**	**[-0.109, -0.026]**	**0.002**
Sex	Female	Reference		
	Male	-0.065	[-0.305, 0.176]	0.596
Age		-0.009	[-0.018, 0.000]	0.070
Surgeon Credentials	BC-N	Reference		

	SSS	-0.196	[-0.479, 0.086]	0.181
	N-t	-0.156	[-0.550, 0.238]	0.441
	BC-O	-0.032	[-0.936, 0.872]	0.945
	O-t	0.001	[-0.907, 0.909]	0.999
	Other	0.051	[-0.736, 0.838]	0.932
Country ID	A	Reference		
	B	**-0.596**	**[-0.865, -0.327]**	**<0.001**
	D	**1.626**	**[0.185, 3.067]**	**0.027**
	E	-0.117	[-0.776, 0.541]	0.728
	F	-0.62	[-1.461, 0.221]	0.146
	G	**-2.226**	**[-3.828, -0.625]**	**0.006**
	H	-0.766	[-2.266, 0.733]	0.307
	Other	-0.136	[-1.005, 0.732]	0.758
Previous Treatment	None	Reference		
	<6 mon. cons.	-0.189	[-0.530, 0.151]	0.276
	>6 mon. cons.	0.32	[-0.029, 0.669]	0.080
	Surgical	**0.926**	**[0.339, 1.513]**	**0.003**
Level of Spine	L5/S1	Reference		
	L4/L5	-0.012	[-0.270, 0.247]	0.930
	L3/L4	-0.225	[-0.709, 0.258]	0.365
	L2/L3	0.205	[-0.574, 0.984]	0.607
	Other	0.111	[-0.799, 1.022]	0.811
ASA Morbidity	1	Reference		
	2	**0.52**	**[0.254, 0.786]**	**<0.001**
	3	**0.889**	**[0.343, 1.434]**	**0.002**
	4	**4.765**	**[0.821, 8.709]**	**0.016**
Baseline COMI score		**0.399**	**[0.330, 0.468]**	**<0.001**

Table 5.15: Coefficient estimates of fixed effects, 95% confidence intervals and p-values for mixed-effects model including random intercepts and slopes. Dataset of the same patients as in linear regression up to 3-month data.

The patient characteristics with significant p-values are the same as in the random intercept model, which shows consistency of these variables to be connected to treatment outcome, even after including additional random effects. The total model R_z^2 of 0.644 is split in the fixed effects component of 0.103 and the random effects component of 0.541. In order to compare this model to the prior less

complex model that included only random intercepts, an ANOVA was performed. Using ANOVA to compare models tests the hypothesis if the residual sums of squares are different between the models and quantifies the p-value for the decision. Comparison could have also been made using the AIC values of each model; however, ANOVA additionally implies a hypothesis of superiority of one model and calculates the p-value of that hypothesis. A p-value of 0.818 showed, that the model fit was not significantly improved by including random slopes.

Looking closer at the progression of COMI scores, one finds that many patients have an L-shaped curve consisting of an initial improvement and constant progression after. Linear graphs to fit the data is therefore not capturing this trajectory. Additionally, there is doubt that some of the linear graphs are accurate, since they resulted from patients that only had 2 measurements available. There is reason to assume, that more measurements would also reveal a non-linear progression. It is also crucial to assume, that COMI scores cannot be negative. Linear models with negative slopes could potentially lead to negative estimates, especially when long-term outcomes are computed. It is therefore reasonable to assume that COMI-score progression is non-linear, especially for long-term outcome modelling. To approach this non-linearity polynomial functions of the patient covariates can be considered; however, combinations are infinite and therefore non-linear time terms that fit the shape of the progression curve were explored.

In the following model approaches a random intercept and a random slope were included. Furthermore, several non-linear time terms were explored as both random effects and fixed effects. The most stable approach was including the term $f(t) = {}^{1}/_{t}$, where t denotes time in weeks past surgery. Several other terms, such as $f_1(t) = {}^{1}/_{t^2}$ or $f_2(t) = \exp(-2t)$ were considered, but none improved the model fit more than $f(t) = {}^{1}/_{t}$, which is why $f(t)$ was picked for further investigations. Details of the fixed effects of the resulting model can be found in Table 5.16.

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		0.499	[-1.367, 2.365]	0.596
Weeks past surgery		0.018	[-0.062, 0.098]	0.667
$\boldsymbol{f(t)}$		**9.877**	**[1.531, 18.223]**	**0.019**
Sex	Female	Reference		
	Male	-0.065	[-0.305, 0.176]	0.599
Age		-0.009	[-0.018, 0.000]	0.091
Surgeon Credentials	BC-N	Reference		
	SSS	-0.198	[-0.482, 0.085]	0.176
	N-t	-0.144	[-0.545, 0.257]	0.478

	BC-O	-0.045	[-0.954, 0.863]	0.922
	O-t	-0.041	[-0.900, 0.818]	0.930
	Other	0.101	[-0.686, 0.888]	0.866
Country ID	A	Reference		
	B	**-0.602**	**[-0.872, -0.332]**	**<0.001**
	D	**1.632**	**[0.183, 3.080]**	**0.026**
	E	-0.133	[-0.790, 0.525]	0.693
	F	-0.647	[-1.489, 0.195]	0.129
	G	**-2.457**	**[-4.259, -0.655]**	**0.003**
	H	-0.76	[-2.254, 0.734]	0.311
	Other	-0.18	[-1.047, 0.687]	0.682
Previous Treatment	None	Reference		
	<6 mon. cons.	-0.168	[-0.507, 0.171]	0.332
	>6 mon. cons.	0.335	[-0.018, 0.688]	0.067
	Surgical	**0.95**	**[0.370, 1.530]**	**0.002**
Level of Spine	L5/S1	Reference		
	L4/L5	-0.017	[-0.277, 0.242]	0.896
	L3/L4	-0.237	[-0.723, 0.249]	0.339
	L2/L3	0.218	[-0.566, 1.003]	0.582
	Other	0.096	[-0.817, 1.008]	0.835
ASA Morbidity	1	Reference		
	2	**0.516**	**[0.251, 0.781]**	**<0.001**
	3	**0.864**	**[0.319, 1.409]**	**0.003**
	4	**4.692**	**[0.785, 8.599]**	**0.018**
Baseline COMI score		**0.400**	**[0.331, 0.469]**	**<0.001**

Table 5.16: Coefficient estimates of fixed effects, 95% confidence intervals and p-values for mixed-effects model including random intercepts, slopes and non-linear time terms. Dataset of the same patients as in linear regression up to 3-month data.

An increase of the time term t was associated with a decrease in COMI scores (considering that the term $f(t)$ is decreases with increasing t) . The model including $f(t)$ had a R_z^2 of 0.651, consisting of a fixed effects component of 0.107 and a random effects component of 0.544. Weeks past surgery and therefore the linear slope did not have a significant p-value after including the non-linear time term, which supports our assumption, that COMI score progression is not linear. However, the R_z^2 could not

be significantly increased and an ANOVA also revealed (p = 0.487) that this model did not fit the data better than the model that only considered a random intercept.

Since the R_z^2 can only be interpreted as roughly similar to the R^2 of the previous linear regression models, RMSE values were also computed. Table 5.17 shows that RMSE values were smaller when considering random effects.

Model statistics \ Time intervals			3 months
Linear regression		RMSE	2.75
		Significant variables	Country, Previous treatment, ASA morbidity, Baseline COMI scores
		R^2	0.098
Mixed-model	Random intercept	RMSE	1.117
		Significant variables	Country, Previous treatment, ASA morbidity, Baseline COMI scores, Weeks past surgery
		R_z^2	0.638
	Random intercept and slope	RMSE	1.301
		Significant variables	Country, Previous treatment, ASA morbidity, Baseline COMI scores, Weeks past surgery
		R_z^2	0.644
	Random intercept, slope and non-linear time term	RMSE	1.387
		Significant variables	Country, Previous treatment, ASA morbidity, Baseline COMI scores, Non-linear time term
		R_z^2	0.651
Number of included patients			2,127

Table 5.17: Model fit statistics of linear regression and mixed modelling approach for COMI scores at three months past surgery.

One major limitation of these model approaches is, that they barely consider more data points than the linear regression model, since the 3-month outcome is one of the earliest in the data set. This means that most patient progressions consist of two data points and therefore linear. However, this linearity is possibly misleading. Insights regarding the effect of random slopes or non-linear time terms could differ, when analysing datasets with longer time intervals.

5.7.2 Mixed model in comparison to linear regression model at one year past surgery

Again, the first model approach includes a random intercept. In order to obtain a fair comparison with prior linear regression models, this model considers the same set of patients that were used in the linear regression model at 1-year past surgery and disregards later measurements. The data set included 3,129 measurements from 2,190 patients. Model details regarding fixed effects are displayed in Table 5.18.

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		**1.899**	**[1.179, 2.620]**	**<0.001**
Weeks past surgery		**-0.013**	**[-0.016, -0.010]**	**<0.001**
Sex	Female	Reference		
	Male	-0.063	[-0.282, 0.157]	0.579
Age		**-0.011**	**[-0.021, -0.002]**	**0.013**
Surgeon Credentials	BC-N	Reference		
	SSS	-0.119	[-0.393, 0.155]	0.376
	N-t	0.279	[-0.079, 0.637]	0.127
	BC-O	-0.067	[-0.706, 0.573]	0.838
	O-t	0.202	[-0.751, 1.155]	0.677
	Other	-0.100	[-1.170, 0.969]	0.854
Country ID	A	Reference		
	B	**-0.79**	**[-1.042, -0.537]**	**<0.001**
	D	1.474	[-0.462, 3.410]	0.133
	E	-0.287	[-0.756, 0.181]	0.229
	F	-0.48	[-1.187, 0.227]	0.183
	G	-0.238	[-1.380, 0.904]	0.682
	H	-0.769	[-2.530, 0.993]	0.393
	Other	0.343	[-0.862, 1.548]	0.573
Previous Treatment	None	Reference		
	<6 mon. cons.	-0.063	[-0.359, 0.234]	0.679
	>6 mon. cons.	**0.348**	**[0.042, 0.654]**	**0.026**
	Surgical	**1.188**	**[0.639, 1.737]**	**<0.001**
Level of Spine	L5/S1	Reference		
	L4/L5	-0.233	[-0.471, 0.004]	0.055
	L3/L4	-0.416	[-0.864, 0.033]	0.068

	L2/L3	0.096	[-0.611, 0.804]	0.787
	Other	0.279	[-0.638, 1.195]	0.551
ASA Morbidity	1	Reference		
	2	**0.554**	**[0.317, 0.790]**	**<0.001**
	3	**1.013**	**[0.540, 1.487]**	**<0.001**
	4	1.635	[-1.111, 4.382]	0.240
Baseline COMI score		**0.378**	**[0.316, 0.439]**	**<0.001**

Table 5.18: Coefficient estimates of fixed effects, 95% confidence intervals and p-values for mixed-effects model including random intercepts. Dataset of the same patients as in linear regression up to 1-year data.

The variables with significant p-values are the same as in in the 3-month dataset. For this current model fit the R_z^2 of the total model was 0.651, consisting of the fixed component of 0.100 and the random component of 0.551, which is also very close to the 3-month approach. In addition to the 3-month model, higher age was considered with smaller COMI outcome scores.

A summary of the fixed effects of the random intercept and random slope model is summarised in Table 5.19.

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		**1.853**	**[0.576, 3.130]**	**0.004**
Weeks past surgery		**-0.013**	**[-0.018, -0.009]**	**<0.001**
Sex	Female	Reference		
	Male	-0.247	[-0.655, 0.161]	0.235
Age		-0.01	[-0.026, 0.006]	0.253
Surgeon Credentials	BC-N	Reference		
	SSS	-0.035	[-0.519, 0.448]	0.887
	N-t	0.339	[-0.322, 1.000]	0.319
	BC-O	0.436	[-0.747, 1.620]	0.468
	O-t	0.616	[-0.682, 1.915]	0.521
	Other	0.179	[-1.778, 2.136]	0.876
Country ID	A	Reference		
	B	-0.397	[-0.863, 0.069]	0.096
	D	**3.833**	**[0.667, 6.999]**	**0.018**
	E	-0.391	[-1.303, 0.520]	0.399
	F	0.014	[-1.360, 1.388]	0.984

	G	2.354	[-1.329, 6.037]	0.085
	H	1.332	[-0.389, 3.053]	0.332
	Other	0.912	[-1.302, 3.126]	0.418
Previous Treatment	None	Reference		
	<6 mon. cons.	-0.143	[-0.683, 0.396]	0.601
	>6 mon. cons.	-0.191	[-0.757, 0.374]	0.506
	Surgical	**1.038**	**[0.051, 2.025]**	**0.038**
Level of Spine	L5/S1	Reference		
	L4/L5	-0.425	[-0.874, 0.023]	0.062
	L3/L4	-0.624	[-1.438, 0.190]	0.130
	L2/L3	-0.066	[-1.265, 1.132]	0.914
	Other	-0.346	[-2.200, 1.507]	0.699
ASA Morbidity	1	Reference		
	2	0.283	[-0.155, 0.722]	0.204
	3	0.223	[-0.626, 1.073]	0.606
	4	-2.907	[-6.478, 0.664]	0.108
Baseline COMI score		**0.420**	**[0.308, 0.533]**	**<0.001**

Table 5.19: Coefficient estimates of fixed effects, 95% confidence intervals and p-values for mixed-effects model including random intercepts and slopes. Dataset of the same patients as in linear regression up to 1-year data.

The patient characteristics with significant p-values are the same as in the random intercept model, with the only difference being that age was not considered significantly associated with outcome scores. The total model R_z^2 of 0.698 is split in the fixed effects component of 0.100 and the random effects component of 0.598. This value increased after adding random slopes to patients and resulted in a better model fit. Again, ANOVA was used to decide if this more complex model is significantly better. A p-value of 0.031 indicates, that the model fit was significantly improved by including random slopes.

Considering the L-shaped progression of COMI scores of many patients, a non-linear time term of the form $f(t) = {}^{1}/_{t}$ was included. The resulting fixed effects are summarised in Table 5.20.

Variable	Estimate	95% Confidence Interval	p-Value
Intercept	0.658	[-0.203, 1.518]	0.133
Weeks past surgery	0.007	[-0.001, 0.015]	0.083
$\boldsymbol{f(t)}$	**12.079**	**[7.476, 16.683]**	**<0.001**

Sex	Female	Reference		
	Male	-0.062	[-0.281, 0.156]	0.580
Age		**-0.012**	**[-0.021, -0.002]**	**0.012**
Surgeon Credentials	BC-N	Reference		
	SSS	-0.112	[-0.374, 0.150]	0.404
	N-t	0.259	[-0.099, 0.617]	0.156
	BC-O	-0.124	[-0.769, 0.520]	0.706
	O-t	0.211	[-0.740, 1.163]	0.662
	Other	-0.115	[-1.176, 0.945]	0.832
Country ID	A	Reference		
	B	**-0.794**	**[-1.046, -0.542]**	**<0.001**
	D	1.267	[-0.662, 3.197]	0.197
	E	-0.293	[-0.764, 0.178]	0.222
	F	-0.659	[-1.371, 0.054]	0.069
	G	-0.423	[-1.583, 0.737]	0.475
	H	-0.805	[-2.588, 0.978]	0.373
	Other	0.085	[-1.113, 1.282]	0.890
Previous Treatment	None	Reference		
	<6 mon. cons.	-0.071	[-0.367, 0.224]	0.636
	>6 mon. cons.	**0.341**	**[0.035, 0.648]**	**0.029**
	Surgical	**1.181**	**[0.633, 1.729]**	**<0.001**
Level of Spine	L5/S1	Reference		
	L4/L5	-0.238	[-0.477, 0.000]	0.051
	L3/L4	-0.383	[-0.829, 0.063]	0.093
	L2/L3	0.099	[-0.599, 0.797]	0.780
	Other	0.278	[-0.640, 1.197]	0.552
ASA Morbidity	1	Reference		
	2	**0.559**	**[0.324, 0.794]**	**<0.001**
	3	**1.015**	**[0.544, 1.486]**	**<0.001**
	4	2.247	[-1.322, 5.816]	0.099
Baseline COMI score		**0.377**	**[0.315, 0.440]**	**<0.001**

Table 5.20: Coefficient estimates of fixed effects, 95% confidence intervals and p-values for variables of the mixed-effects model including random intercepts, slopes and non-linear time terms. Dataset of the same patients as in linear regression up to 1-year data.

In this model, higher age was considered to be associated with lower outcome scores again, similar to the approach with random intercepts. The model including $f(t)$ had a R_z^2 of 0.740, consisting of a fixed effects component of 0.101 and a random effects component of 0.639. Weeks past surgery and therefore the linear slope did not have a significant p-value after including the non-linear time term, which supports our assumption, that COMI score progression is not linear. However, the R_z^2 was only increased by 0.006. Again, ANOVA was used to decide if this more complex model is significantly better. A p-value of <0.042 indicates, that the model fit was significantly improved by including the non-linear time term.

A summary of all descriptive statistics, including the RMSEs are shown in Table 5.21.

Model statistics \ Time intervals			3 months	1 year
Linear regression		RMSE	2.75	2.812
		Significant variables	Country, Previous treatment, ASA morbidity, Baseline COMI scores	Country, Previous treatment, ASA morbidity, Baseline COMI scores
		R^2	0.098	0.095
Mixed-model	Random intercept	RMSE	1.117	1.301
		Significant variables	Country, Previous treatment, ASA morbidity, Baseline COMI scores, Weeks past surgery	Country, Previous treatment, ASA morbidity, Baseline COMI scores, Weeks past surgery, Age
		R_z^2	0.638	0.651
	Random intercept and slope	RMSE	1.301	1.09
		Significant variables	Country, Previous treatment, ASA morbidity, Baseline COMI scores, Weeks past surgery	Country, Previous treatment, Baseline COMI scores, Weeks past surgery,
		R_z^2	0.644	0.698
	Random intercept, slope and non-linear time term	RMSE	1.387	0.951
		Significant variables	Country, Previous treatment, ASA morbidity, Baseline COMI scores, Non-linear time term	Country, Previous treatment, ASA morbidity, Baseline COMI scores, Age, Non-linear time term
		R_z^2	0.651	0.704
Number of included patients			2,127	2,190

Table 5.21: Model fit statistics of linear regression and mixed modelling approach for COMI scores at three months and 1-year past surgery.

It becomes clear that the inclusion of random effects, compared to linear regression, produces a much better model fit regarding RMSEs. Even though the R_z^2 is not directly comparable to the R^2 in linear

regression, it seems that the proportion of explained outcome variation was much higher in mixed-models than in linear regression. At the 1-year interval, the non-linear time term as random effect fit the best, suggesting that the progression of COMI scores is not linear. This was not the case in the 3-month model. However, this was likely due to the fact that most patients had only 1 measurement after surgery and progression therefore appeared linear.

5.7.3 Mixed model in comparison to linear regression model at 2 years past surgery

Again, the first model approach includes a random intercept. In order to obtain a fair comparison with prior linear regression models, this model considers the same set of patients that were used in the linear regression model at 2 years past surgery and disregards later measurements. The data set included 4,268 measurements from 1,678 patients. Similar to the analyses of prior datasets, model specifications with random intercepts, random intercepts and slopes, as well as the inclusion of a non-linear time term of the form $f(t)$ was performed. The following tables (Table 5.22 – 5.24) show details of the fixed effects of these models.

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		**1.682**	**[0.882, 2.482]**	**<0.001**
Weeks past surgery		**-0.006**	**[-0.007, -0.006]**	**<0.001**
Sex	Female	Reference		
	Male	-0.178	[-0.423, 0.068]	0.153
Age		-0.009	[-0.019, 0.001]	0.082
Surgeon Credentials	BC-N	Reference		
	SSS	-0.244	[-0.536, 0.049]	0.102
	N-t	0.169	[-0.204, 0.543]	0.373
	BC-O	-0.396	[-1.165, 0.373]	0.310
	O-t	0.164	[-0.826, 1.153]	0.744
	Other	-0.009	[-1.126, 1.108]	0.987
Country ID	A	Reference		
	B	**-0.786**	**[-1.062, -0.509]**	**<0.001**
	D	-2.627	[-5.655, 0.402]	0.303
	E	-0.317	[-0.865, 0.231]	0.256
	F	-0.502	[-1.443, 0.439]	0.295
	H	-2.133	[-4.186, 0.921]	0.170
Previous Treatment	None	Reference		
	<6 mon. cons.	-0.184	[-0.503, 0.136]	0.257

	>6 mon. cons.	0.279	[-0.051, 0.610]	0.096
	Surgical	**0.933**	**[0.345, 1.522]**	**0.002**
Level of Spine	L5/S1	Reference		
	L4/L5	0.005	[-0.258, 0.269]	0.971
	L3/L4	-0.366	[-0.866, 0.135]	0.151
	L2/L3	0.425	[-0.391, 1.241]	0.304
	Other	0.389	[-0.689, 1.467]	0.478
ASA Morbidity	1	Reference		
	2	**0.445**	**[0.188, 0.701]**	**<0.001**
	3	**1.116**	**[0.600, 1.632]**	**<0.001**
Baseline COMI score		**0.387**	**[0.317, 0.456]**	**<0.001**

Table 5.22: Coefficient estimates of fixed effects, 95% confidence intervals and p-values of variables of the mixed-effects model including random intercepts. Dataset of the same patients as in linear regression up to 2-year data.

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		**1.607**	**[0.811, 2.404]**	**<0.001**
Weeks past surgery		**-0.006**	**[-0.007, -0.006]**	**<0.001**
Sex	Female	Reference		
	Male	-0.158	[-0.400, 0.084]	0.202
Age		-0.009	[-0.019, 0.001]	0.073
Surgeon Credentials	BC-N	Reference		
	SSS	-0.236	[-0.528, 0.055]	0.113
	N-t	0.155	[-0.215, 0.525]	0.410
	BC-O	-0.399	[-1.170, 0.372]	0.307
	O-t	0.134	[-0.859, 1.127]	0.790
	Other	0.008	[-1.089, 1.105]	0.989
Country ID	A	Reference		
	B	**-0.778**	**[-1.052, -0.505]**	**<0.001**
	D	-2.665	[-5.604, 0.273]	0.287
	E	-0.341	[-0.895, 0.213]	0.225
	F	-0.523	[-1.474, 0.429]	0.280
	H	-2.134	[-4.293, 0.024]	0.180
Previous Treatment	None	Reference		

	<6 mon. cons.	-0.175	[-0.492, 0.142]	0.279
	>6 mon. cons.	0.313	[-0.013, 0.640]	0.060
	Surgical	**0.955**	**[0.370, 1.540]**	**0.001**
Level of Spine	L5/S1	Reference		
	L4/L5	-0.006	[-0.269, 0.257]	0.962
	L3/L4	-0.364	[-0.861, 0.134]	0.150
	L2/L3	0.374	[-0.436, 1.184]	0.363
	Other	0.306	[-0.774, 1.385]	0.576
ASA Morbidity	**1**	**Reference**		
	2	**0.438**	**[0.181, 0.695]**	**<0.001**
	3	**1.081**	**[0.568, 1.594]**	**<0.001**
Baseline COMI score		**0.393**	**[0.324, 0.463]**	**<0.001**

Table 5.23: Coefficient estimates of fixed effects, 95% confidence intervals and p-values of variables of mixed-effects model including random intercepts and slopes. Dataset of the same patients as in linear regression up to 2-year data.

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		**1.009**	**[0.182, 1.836]**	**0.017**
Weeks past surgery		0	[-0.002, 0.001]	0.889
$\boldsymbol{f(t)}$		**8.318**	**[5.224, 11.412]**	**<0.001**
Sex	Female	Reference		
	Male	-0.157	[-0.396, 0.081]	0.204
Age		-0.009	[-0.019, 0.000]	0.074
Surgeon Credentials	BC-N	Reference		
	SSS	-0.247	[-0.538, 0.045]	0.096
	N-t	0.145	[-0.226, 0.517]	0.442
	BC-O	-0.399	[-1.173, 0.375]	0.309
	O-t	0.12	[-0.868, 1.108]	0.811
	Other	-0.014	[-1.104, 1.075]	0.981
Country ID	A	Reference		
	B	**-0.774**	**[-1.048, -0.500]**	**<0.001**
	D	-3.086	[-7.836, 1.663]	0.195
	E	-0.294	[-0.846, 0.257]	0.300
	F	-0.587	[-1.527, 0.353]	0.222

	H	-2.238	[-4.349, -0.127]	0.155
Previous Treatment	None	Reference		
	<6 mon. cons.	-0.186	[-0.500, 0.127]	0.247
	>6 mon. cons.	0.298	[-0.028, 0.625]	0.073
	Surgical	**0.888**	**[0.308, 1.467]**	**0.003**
Level of Spine	L5/S1	Reference		
	L4/L5	0.009	[-0.253, 0.272]	0.949
	L3/L4	-0.318	[-0.814, 0.177]	0.209
	L2/L3	0.362	[-0.446, 1.170]	0.378
	Other	0.318	[-0.756, 1.391]	0.560
ASA Morbidity	1	Reference		
	2	**0.439**	**[0.183, 0.696]**	**<0.001**
	3	**1.104**	**[0.593, 1.614]**	**<0.001**
Baseline COMI score		**0.389**	**[0.320, 0.459]**	**<0.001**

Table 5.24: Coefficient estimates of fixed effects, 95% confidence intervals and p-values of variables of mixed-effects model including random intercepts, slopes and non-linear time terms. Dataset of the same patients as in linear regression up to 2-year data.

Similarly to the models with 1-year outcomes, the inclusion of non-linear time term random effects perfomed best regarding R_z^2 values. A comparison table of all the model statistics is shown in Table 5.25.

Model statistics \ Time intervals			3 months	1 year	2 years
Linear regression		RMSE	2.75	2.812	2.85
		Significant variables	Country, Previous treatment, ASA morbidity, Baseline COMI scores	Country, Previous treatment, ASA morbidity, Baseline COMI scores, Level of spine	Country, Previous treatment, ASA morbidity, Baseline COMI scores, Sex
		R^2	0.098	0.095	0.087
Mixed-model	Random intercept	RMSE	1.117	1.301	1.387
		Significant variables	Country, Previous treatment, ASA morbidity, Baseline COMI scores, Weeks past surgery	Country, Previous treatment, ASA morbidity, Baseline COMI scores, Weeks past surgery, Age	Country, Previous treatment, ASA morbidity, Baseline COMI scores, Weeks past surgery,
		R_z^2	0.638	0.651	0.670
	Random intercept and slope	RMSE	1.301	1.09	0.972
		Significant variables	Country, Previous treatment, ASA morbidity, Baseline COMI scores, Weeks past surgery	Country, Previous treatment, ASA morbidity, Baseline COMI scores, Weeks past surgery	Country, Previous treatment, ASA morbidity, Baseline COMI scores, Weeks past surgery
		R_z^2	0.644	0.698	0.735
	Random intercept, slope and non-linear time term	RMSE	1.387	0.951	0.967
		Significant variables	Country, Previous treatment, ASA morbidity, Baseline COMI scores, Non-linear time term	Country, Previous treatment, ASA morbidity, Baseline COMI scores, Age, Non-linear time term	Country, Previous treatment, ASA morbidity, Baseline COMI scores, Non-linear time term
		R_z^2	0.651	0.704	0.764
Number of included patients			2,127	2,190	1,678

Table 5.25: Model fit statistics of all linear regression and mixed model approaches.

For this direct comparison with linear regressions, the same patients were considered for each time point.

Overall, mixed models had lower RMSE values, meaning better prediction accuracy. This is due to random effects that allow each patient to have an individual intercept slope or non-linear time term that can describe the individual outcome progression more accurately.

However, this approach only considered the same patient data set that was used in the linear regression model, in order to get a fair comparison. There are more patients available, that do not

have 2-year outcomes, but measurements between surgery and two years. These could be included in order to improve model fit and incorporate all available data.

5.7.4 Mixed model at 2 years past surgery including all available patients

This approach used data from patients who had undergone surgery and had information available for up to 2 years. Data beyond 3 years was not included due to a lack of observations. This resulted in a dataset of 6,681 measurements from 3,520 patients, which is larger than the previous 2-year model that had 4,268 measurements from 1,678 patients. However, the Root Mean Squared Error (RMSE) could only be calculated using the same patients as before, as the true values of 2-year outcomes were needed. shows the summary of the fixed effects coefficients of the model including random intercepts, slopes and non-linear time term.

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		**1.401**	**[0.808, 1.994]**	**<0.001**
Weeks past surgery		-0.001	[-0.003, 0.001]	0.588
$\boldsymbol{f(t)}$		**8.283**	**[6.644, 9.923]**	**<0.001**
Sex	Female	Reference		
	Male	-0.14	[-0.313, 0.033]	0.116
Age		**-0.012**	**[-0.018, -0.005]**	**0.001**
Surgeon Credentials	BC-N	Reference		
	SSS	-0.217	[-0.426, 0.008]	0.103
	N-t	0.056	[-0.246, 0.359]	0.716
	BC-O	-0.206	[-0.758, 0.346]	0.432
	O-t	-0.064	[-0.779, 0.650]	0.860
	Other	0.225	[-0.739, 1.188]	0.647
Country ID	A	Reference		
	B	**-0.584**	**[-0.790, -0.377]**	**<0.001**
	D	-0.092	[-0.713, 0.528]	0.771
	E	-0.243	[-0.626, 0.140]	0.214
	F	**-0.808**	**[-1.276, -0.340]**	**0.001**
	G	-0.723	[-1.596, 0.149]	0.105
	H	**-1.149**	**[-2.143, -0.155]**	**0.024**
	Other	-0.464	[-1.106, 0.178]	0.160
Previous Treatment	None	Reference		
	<6 mon. cons.	-0.114	[-0.350, 0.123]	0.345

	>6 mon. cons.	0.089	[-0.157, 0.336]	0.477
	Surgical	0.864	[0.450, 1.279]	**<0.001**
Level of Spine	L5/S1	Reference		
	L4/L5	-0.068	[-0.255, 0.119]	0.476
	L3/L4	-0.273	[-0.632, 0.087]	0.137
	L2/L3	0.361	[-0.410, 1.132]	0.217
	L1/L2	3.26	[-0.524, 7.044]	0.090
	Other	0.19	[-0.533, 0.913]	0.607
ASA Morbidity	1	Reference		
	2	**0.445**	**[0.260, 0.629]**	**<0.001**
	3	**0.65**	**[0.282, 1.018]**	**0.001**
	4	0.371	[-1.814, 2.556]	0.740
Baseline COMI score		**0.382**	**[0.332, 0.431]**	**<0.001**

Table 5.26: Coefficient estimates of fixed effects, 95% confidence intervals and p-values of variables of the mixed-effects model including random intercepts, slopes and non-linear time terms. This dataset includes all available patient data up to 2 years of follow-up.

Significant variables and their estimates are similar to the model with the smaller sample size of patients. However, higher age was again significantly associated with smaller COMI outcome scores.

The patient covariates that were included in the model were the same as in the prior 2-year model. The R_z^2 of the model was 0.740, with a fixed component of 0.107 and a random component of 0.633. The model fit did not improve, with a slightly higher RMSE (0.990) than the previous 2-year model (0.967).

5.7.5 Inclusion of smoking status and BMI

Smoking status was only available for 1,841 and BMI only for 2,358 from a previous total of 3,520 patients. The most recent model specification was fitted on the data of patients up to two years of follow-up, for each of the subsets of patients that had smoking status/BMI available. Details of the fixed effects are shown in Table 5.27 and Table 5.28.

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		**1.634**	**[0.769, 2.499]**	**<0.001**
Weeks past surgery		-0.001	[-0.004, 0.001]	0.631
$\boldsymbol{f(t)}$		**8.705**	**[6.631, 10.779]**	**<0.001**
Sex	Female	Reference		

	Male	0.039	[-0.176, 0.254]	0.723
Age		**-0.016**	**[-0.024, -0.007]**	**<0.001**
Surgeon Credentials	BC-N	Reference		
	SSS	-0.229	[-0.498, 0.039]	0.096
	N-t	0.094	[-0.304, 0.492]	0.641
	BC-O	-0.443	[-1.067, 0.181]	0.164
	O-t	-0.04	[-0.991, 0.911]	0.933
	Other	0.093	[-0.919, 1.105]	0.856
Country ID	A	Reference		
	B	**-0.452**	**[-0.721, -0.183]**	**0.001**
	D	-0.369	[-1.476, 0.738]	0.512
	E	-0.103	[-0.561, 0.355]	0.660
	F	-0.726	[-1.459, 0.006]	0.052
	G	-0.58	[-1.560, 0.399]	0.242
	H	-0.995	[-2.083, 0.093]	0.069
	Other	-0.517	[-2.845, 1.811]	0.662
Previous Treatment	None	Reference		
	<6 mon. cons.	-0.25	[-0.525, 0.025]	0.076
	>6 mon. cons.	-0.037	[-0.327, 0.253]	0.803
	Surgical	**0.667**	**[0.168, 1.165]**	**0.009**
Level of Spine	L5/S1	Reference		
	L4/L5	-0.157	[-0.390, 0.077]	0.183
	L3/L4	-0.265	[-0.720, 0.190]	0.250
	L2/L3	0.101	[-0.380, 0.583]	0.772
	Other	0.163	[-0.624, 0.949]	0.689
ASA Morbidity	1	Reference		
	2	**0.562**	**[0.330, 0.794]**	**<0.001**
	3	**1.103**	**[0.616, 1.591]**	**<0.001**
BMI	<20	Reference		
	20 - 25	-0.387	[-0.925, 0.151]	0.158
	25 - 30	-0.324	[-0.860, 0.212]	0.237
	30 – 35	-0.022	[-0.600, 0.556]	0.940
	>35	-0.367	[-1.041, 0.306]	0.288
Baseline COMI score		**0.400**	**[0.339, 0.460]**	**<0.001**

Table 5.27: Coefficient estimates, 95% confidence intervals and p-values of mixed-effect model including random intercepts, slopes and non-linear time terms. Data is from a subset of patients that had BMI available. Cases for which BMI was exactly 25 were included in "20 – 25". The same method applies to other categories.

Variable		Estimate	95% Confidence Interval	p-Value
Intercept		**1.482**	**[0.667, 2.297]**	**<0.001**
Weeks past surgery		0	[-0.002, 0.002]	0.959
$\boldsymbol{f(t)}$		**8.723**	**[6.457, 10.989]**	**<0.001**
Sex	Female	Reference		
	Male	-0.013	[-0.313, 0.033]	0.914
Age		**-0.015**	**[-0.024, -0.006]**	**0.003**
Surgeon Credentials	BC-N	Reference		
	SSS	-0.313	[-1.004, 0.377]	0.375
	N-t	0.064	[-0.414, 0.542]	0.793
	BC-O	0.286	[-0.892, 1.464]	0.631
	O-t	1.723	[-0.301, 3.746]	0.093
	Other	-0.245	[-0.559, 0.069]	0.128
Country ID	A	Reference		
	B	**-0.536**	**[-0.835, -0.236]**	**<0.001**
	D	-0.327	[-1.396, 0.741]	0.549
	E	-0.178	[-0.663, 0.307]	0.474
	F	-0.337	[-1.223, 0.549]	0.451
	G	**-1.174**	**[-2.293, -0.055]**	**0.039**
	H	-0.967	[-2.057, 0.124]	0.080
	Other	0.098	[-2.393, 2.589]	0.938
Previous Treatment	None	Reference		
	<6 mon. cons.	**-0.352**	**[-0.682, -0.022]**	**0.036**
	>6 mon. cons.	-0.121	[-0.459, 0.217]	0.483
	Surgical	**0.831**	**[0.243, 1.419]**	**0.006**
Level of Spine	L5/S1	Reference		
	L4/L5	-0.229	[-0.488, 0.029]	0.085
	L3/L4	-0.345	[-0.845, 0.155]	0.178
		0.159	[-0.573, 0.890]	0.672

	L2/L3 Other	-0.012	[-0.905, 0.880]	0.979
ASA Morbidity	**1** **2** **3**	**Reference** **0.582** **1.159**	 **[0.318, 0.847]** **[0.633, 1.686]**	 **<0.001** **<0.001**
Smoker	**No** **Yes**	**Reference** **0.613**	 **[0.319, 0.908]**	 **<0.001**
Baseline COMI score		**0.378**	**[0.309, 0.446]**	**<0.001**

Table 5.28: Coefficient estimates, 95% confidence intervals and p-values of mixed-effect model including random intercepts, slopes and non-linear time terms. Data is from a subset of patients that had smoking status available.

On the subset of patients with BMI available, none of the BMI sub-categories had significant p-values. This indicates that BMI is not associated with treatment outcome. The inclusion of smoking status however, indicated that smokers had higher COMI scores after intervention than non-smokers. The inclusion led to an increase of the R_z^2 to 0.759 (fixed effect component 0.131, random effect component 0.628). It should therefore be included in patient forms further on.

5.7.6 Limitations

The mixed-effects approach has shown better model fit than linear and logistic regression, but it has its limitations. Many patients did not have more than one or two measurements after surgery, which can produce misleading linearity in plots and good fits that do not accurately reflect real behaviour when measured more frequently. This is why further model specifications with more random effects were not explored. Another limitation is in prediction modelling. The estimates for random effects for individual patients improve the fit significantly, but these estimates cannot be computed for new patients for which there are no measurements yet. Mixed-effects models are more suitable for predicting a later outcome value when prior values (past surgery) are available and the random effects of patients are already estimated. For new patients without prior measurements, only fixed effects can be utilized.

5.8 Joint modelling approach

In the source material of the Spine Tango data set, there was another variable that was considered as helpful regarding modelling approaches. In forms from follow-up visits it was noted, if there was a further follow-up planned or not. However, this was this question was only part of the 2011 version of the forms. Whether or not further follow-ups are scheduled could be an indicator that the

treatment was successful and the patient can be considered as “healed”. In modelling terms, this can be interpreted as endpoint.

Joint modelling is a statistical approach that combines two modelling approaches. It is used to analyse data sets that have both longitudinal and time-to event data. Like the mixed-effects approach, it allows for the incorporation of both individual-level and population-level information in the analysis, resulting in a more comprehensive understanding of the data. For subsequent analysis, the R-package “joineRML” was used (Hickey et al., 2018). The standard joint model in this package is based on formulas (4).

$$\begin{cases} h_i(t) = h_0(t)\exp\{\gamma^T w_i + \alpha z_i^T(t)u_i\}, \\ y_i(t) = m_i(t) + \varepsilon_i \\ \qquad = x_i^T(t)\beta + z_i^T(t)u_i + \varepsilon_i \ , \end{cases}$$

(4)

The second part of the equation system is the same model as the longitudinal mixed-effects model from the previous section. The sum of fixed and random effects of each individual are integrated in combination with a scaling parameter α into the hazard function of a Cox-model (common time-to-event model). Parameters γ and w_i are vectors of coefficients and patient covariates (Abd ElHafeez et al., 2021). These do not necessarily need to be the same as in the longitudinal formula.

Regarding the data set that is provided, a joint model could utilise the information of “further follow-up scheduled” or “no further follow-up scheduled” as endpoint of a survival model, in addition to the longitudinal data of COMI scores.

In a survival analysis, censored data refers to observations for which the time of an event of interest has not yet occurred. For example, in a study of the survival of patients with a particular disease, censorship might occur when a patient is lost to follow-up, withdraws from the study, or is still alive at the end of the study period. Censored data can introduce bias into the analysis if not properly accounted for. In a survival model, censored data is typically handled using survival analysis, which uses statistical methods to model the probability of an event occurring at a given time. There are two main types of censoring: right censoring and left censoring. Right censoring occurs when the event of interest has not yet occurred at the time of observation, while left censoring occurs when the event of interest occurred prior to the start of the study. Therefore, the available data set from the Spine Tango registry was considered right censored.

The analysis will be designed in a such a way that the model fit can be compared to the model fit of a similar mixed-model. Therefore, patients who had both longitudinal COMI scores up to 2 years as well as the variable of "further follow-up scheduled" available were considered. Measurements at later time points were excluded. The number of patients that could be utilised for this analysis approach was 2,773, with 5,233 longitudinal observations. Since COMI is a patient-reported outcome, there were cases for which there were COMI scores available, even though no further follow-up was scheduled. It is important to note that while some cases were accurately labeled as having 'no further follow-up,' there were instances where participants did undergo further follow-up despite the indication. For the application of the "`joineRML`"-package, these values were removed, since the package considers this as an end-point, regardless of whether further follow-up may have occurred in some instances. By doing so, the aim was to maintain consistency in the analysis and interpretation of results.. This reduced the number of observations from 5,233 to 3,866 and the number of patients from 2,773 to 2,320.

There are a few specifications that need to be considered before applying the "`mjoint()`"–function to fit the model). One can specify the included patient variables for the longitudinal model, as well for the survival model. The number of possible model specification is therefore very large and the AIC model selection algorithm that was used in prior methods does not work on joint model objects. Therefore, the variables that were associated with treatment outcome in the mixed-model approach were considered. Analysis of fit was assessed regarding the RMSE and loglikelihood values and residual standard errors.

5.8.1 Results

The dataset of 3,866 measurements from 2,320 patients was first used to fit a mixed-effects model and a Cox-model separately to identify predictive patient covariates. The separate Cox-model identified sex, surgeon type, country ID, age and previous treatment as associated with survival. On this dataset, the mixed-effects model identified $f(t)$, age, surgeon type, country ID, previous treatment, ASA morbidity and baseline COMI scores as associated with COMI outcomes scores. These covariates were then used to fit the joint model using the `mjoint()`-function, using the same set of patients.

The RMSE was calculated for 2-year post-surgery measurements, but not all patients had this available. A subset was selected and 2-year outcomes were predicted using the mixed-model and joint model. The RMSE values were 1.289 and 1.290, respectively, for the joint and mixed-effects models. This similarity is due to the mixed-effects model being based on the same formula as the longitudinal

part of the joint model (4). The loglikelihoods were -12,241 and -9,102 and the residual standard errors were 1.701 and 1.707 for the joint and mixed-effects models, respectively.

The details of the mixed-effects model are provided in Table 5.29.

Variable		Estimate	95% confidence interval	p-value
Age		**-0.018**	[-0.027, -0.010]	**<0.001**
Surgeon type	BC-N	Reference		
	SSS	**-0.345**	**[-0.610, -0.079]**	**0.011**
	N-t	0.124	[-0.262, 0.510]	0.529
	BC-O	-0.101	[-0.701, 0.499]	0.742
	O-t	-0.114	[-1.053, 0.826]	0.813
	Other	0.034	[-1.029, 1.098]	0.950
Country ID	A	Reference		
	B	**-0.504**	**[-0.773, -0.236]**	**<0.001**
	D	0.138	[-0.557, 0.833]	0.697
	E	-0.309	[-0.765, 0.147]	0.185
	F	**-0.746**	**[-1.298, -0.195]**	**0.008**
	G	-0.842	[-1.872, 0.188]	0.109
	H	**-1.730**	**[-2.917, -0.543]**	**0.004**
	Other	-0.020	[-1.452, 1.413]	0.978
Previous treatment	none	Reference		
	<6 mon. cons.	**0.554**	**[0.175, 0.934]**	**0.005**
	>6 mon. cons.	0.261	[-0.045, 0.567]	0.079
	Surgical	-0.009	[-0.237, 0.180]	0.761
ASA	1	Reference		
	2	**0.387**	**[0.154, 0.621]**	**0.002**
	3	**0.677**	**[0.212, 1.143]**	**0.005**
Baseline COMI score		**0.432**	**[0.369, 0.495]**	**<0.001**
* $\boldsymbol{f(t)}$		**7.636**	**[6.137, 9.135]**	**<0.001**

Table 5.29: Coefficient estimates, 95%-confidence intervals and p-values for the variables in the mixed-effects model. * Function $f(t)$ is defined as previously, with t denoting time in weeks past surgery.

Details of the longitudinal sub-model of the joint model are shown in Table 5.30.

Variable		Estimate	95% confidence interval	p-value
Age		**-0.018**		**<0.001**
Surgeon type	BC-N	Reference		
	SSS	**-0.336**	**[-0.604, -0.068]**	**0.014**
	N-t	0.120	[-0.261, 0.501]	0.536
	BC-O	-0.113	[-0.763, 0.538]	0.734
	O-t	-0.126	[-1.011, 0.759]	0.780
	Other	0.017	[-1.086, 1.119]	0.976
Country ID	A	Reference		
	B	**-0.568**	**[-0.840, -0.296]**	**<0.001**
	D	-0.044	[-0.743, 0.656]	0.903
	E	-0.301	[-0.788, 0.185]	0.225
	F	**-0.811**	**[-0.374, -0.248]**	**0.005**
	G	-0.941	[-1.917, 0.034]	0.059
	H	**-1.742**	**[-2.890, -0.595]**	**0.003**
	Other	-0.125	[-1.828, 1.578]	0.885
Previous treatment	none	Reference		
	<6 mon. cons.	**0.553**	**[0.171, 0.935]**	**0.005**
	>6 mon. cons.	0.275	[-0.032, 0.582]	0.079
	Surgical	-0.032	[-0.241, 0.177]	0.761
ASA	1	Reference		
	2	**0.385**	[0.148, 0.623]	**0.002**
	3	**0.676**	[0.205, 1.146]	**0.005**
Baseline COMI score		**0.426**	**[0.357, 0.496]**	**<0.001**
* $\boldsymbol{f(t)}$		**9.252**	**[7.607, 10.896]**	**<0.001**

Table 5.30: Coefficient estimates, 95%-confidence intervals and p-values for the variables in the longitudinal sub-model of the joint model. * Function $f(t)$ is defined as previously, with t denoting time in weeks past surgery.

Coefficients and p-values are very similar between those two models. Notably, procedures by specialised spinal surgeons were associated with lower COMI outcome scores compared to board-certified orthopedic surgeons and previous treatment of less than 6 months was associated with higher COMI outcome scores compared to no previous treatment.

5.8.2 Conclusion

Both methods performed very similar, the only difference being their loglikelihood value. Although the longitudinal sub-model of the joint model is based on the same formula as the mixed model, the coefficients are not necessarily the same, however very similar. This is because the model fit of a joint model is done in combination with the Cox-model.

Overall, using a joint model did not improve prediction accuracy in terms of RMSE values. This indicates that there is no strong link between COMI outcomes and the practitioner's decision for "no further follow-up". One reason for this might be, that COMI is a patient reported outcome, whereas the practitioners' decision is made during visits. Some patients supplied COMI measurements after being discharged. For this analysis, such measurements were removed, as the software does not support longitudinal measurements recorded after the event of interest.

5.9 Prediction modelling with complications as outcome

Predictions regarding the quality of life questionnaire COMI scores, although model fit could be improved by using mixed effect models, had low accuracy. Substantial variation in treatment outcomes was left unexplained by the measured patient characteristics. However, factors that are associated with treatment outcome regarding quality of life could be identified. Instead of using quality of life one can also explore models that use complications during surgery as outcome. Since the occurrence of complications is a binary variable, logistic regression is a logical approach. This could identify factors that are connected with risks during surgery and help decide which treatment option might be optimal for a given patient.

All sciatica patients from surgery forms were considered (17,252). Of those 17,252 patients, 236 that had a missing value for surgery complication were excluded. Considering the rare incidence of complications (4.37%) the following patient baseline characteristics were excluded from this analysis: BMI, smoking status, Baseline COMI scores. The following characteristics were considered as potentially associated with the occurrence of complications: age, sex, level of spinal disc surgery, surgeon credentials, previous treatment, country, ASA Morbidity status. Missing values of covariates were imputed using the 'mice' package in R. Covariates that were inconsistently collected over the years and have substantial amount of missingness were analysed with the same methods in separate case distinctions.

Categories of complications were grouped into durotomies, serious complications (cauda equina damage, nerve root damage, vascular injury and bleeding inside spinal canal) and other minor complications (bleeding outside spinal canal, wound infection and other). The few cases of surgery at

wrong level (3) were excluded as well, resulting in a total sample size of 17,013 patients. For each case binary variables for the occurrence of complications and their subcategories were created. Logistic regression was then performed for each of the cases to identify risk factors.

The prevalence of the complications is displayed in Table 5.31.

Complication	N = 17,013
Any complication	743 (4.37%)
Durotomy	659 (3.87%)
Serious complications	57 (0.34%)
Other minor complications	37 (0.22%)

Table 5.31: Prevalence of complication categories in total numbers and percentages.

Goodness of model fit was, as previously, assessed using area under ROC-curves, which are displayed in Table 5.32.

Outcome	Area under ROC-curve	95% Confidence Interval
Any complication	0.676	[0.651, 0.688]
Durotomy	0.682	[0.660, 0.698]
Serious complications	0.712	[0.640, 0.769]
Other minor complications	0.833	[0.764, 0.892]

Table 5.32: Areas under ROC-curves and their confidence intervals for logistic regression models for adverse event categories.

Due to the prevalence shown in Table 5.31, AUROC values for serious complications and other minor complications should be interpreted with caution.

For each of these models, the estimates and confidence intervals of variables are displayed in Table 5.33.

		Any complication OR (95% CI), p-value	Durotomy OR (95% CI), p-value	Serious complications OR (95% CI), p-value	Other/minor complications OR (95% CI), p-value
Intercept		**-3.635 [-3.951, -3.319], p<0.001**	**-3.885 [-4.197, -3.511], p<0.001**	**-5.746 [-6.865, -4.627], p<0.001**	**-6.641 [-8.075, -5.207], p<0.001**
Surgeon	BC-N	Reference			
	SSS	0.157 [-0.014, 0.328], p=0.076	0.157 [-0.023, 0.337], p=0.095	-0.312 [-0.973, 0.349], p=0.348	0.672 [-0.075, 1.418], p=0.076
	N-t	0.089 [-0.159, 0.336], p=0.482	0.096 [-0.161, 0.353], p=0.464	0.151 [-0.321, 0.623], p=0.731	0.074 [-0.436, 0.584], p=0.925
	BC-O	-0.189 [-0.591, 0.214], p=0.357	-0.116 [-0.539, 0.307], p=0.590	-0.371 [-1.394, 0.652], p=0.591	0.049 [-1.584, 1.683], p=0.953
	O-t	-0.699 [-1.600, 0.201], p=0.127	-1.112 [-2.266, 0.042], p=0.058	1.34 [-0.020, 2.561], p=0.076	*NA in this outcome category*
	Other	-0.283 [-0.991, 0.426], p=0.543	-0.173 [-0.798, 0.452], p=0.710	*NA in this outcome category*	*NA in this outcome category*
Prev. treat.	None	Reference	Reference	Reference	
	<6 mon. conservative	-0.045 [-0.247, 0.157], p=0.665	-0.015 [-0.228, 0.198], p=0.893	-0.435 [-1.188, 0.318], p=0.249	-0.401 [-1.003, 0.201], p=0.436
	>6 mon. conservative	0.148 [-0.061, 0.357], p=0.165	0.143 [-0.079, 0.365], p=0.205	0.06 [-0.671, 0.791], p=0.872	-0.006 [-1.034, 1.022], p=0.991
	Prior surgery	**0.935 [0.641, 1.228], p<0.001**	**0.923 [0.613, 1.234], p<0.001**	**1.175 [0.231, 2.119], p=0.009**	*NA in this outcome category*
Sex	Female	Reference	Reference		Reference
	Male	**-0.280 [-0.429, -0.131], p<0.001**	-0.011 [-0.260, 0.238], p=0.929	-0.018 [-0.558, 0.522], p=0.945	**-0.747 [-1.384, -0.110], p=0.031**
Country ID	A	Reference	Reference		Reference
	B	**-0.347 [-0.548, -0.146], p=0.001**	**-0.519 [-0.737, -0.302], p<0.001**	0.255 [-0.434, 0.944], p=0.458	**1.998 [1.037, 2.958], p<0.001**
	C	**-1.98 [-2.543, -1.417], p<0.001**	**-2.084 [-2.697, -1.472], p<0.001**	-0.366 [-1.613, 0.880], p=0.570	*NA in this outcome category*
	D	**-1.561 [-2.119, -1.003], p<0.001**	**-1.686 [-2.309, -1.062], p<0.001**	*NA in this outcome category*	0.957 [-1.588, 3.501], p=0.228
	E	**0.512 [0.174, 0.850], p=0.003**	**0.48 [0.127, 0.832], p=0.008**	0.585 [-0.756, 1.926], p=0.396	1.322 [-1.369, 4.013], p= 0.128
	F	0.296 [-0.150, 0.742], p=0.191	0.184 [-0.297, 0.665], p=0.454	-0.048 [-2.076, 1.980], p=0.963	**2.213 [0.769, 3.657], p=0.003**
	G	-0.201 [-0.889, 0.486], p=0.568	-0.134 [-0.827, 0.558], p=0.702	*NA in this outcome category*	*NA in this outcome category*
	H	0.071 [-0.711, 0.853], p=0.857	-0.033 [-0.873, 0.807], p=0.938	-15.7 [-6264.897, 6233.497], p=0.995	**2.506 [0.323, 4.689], p=0.024**
	Other	0.167 [-0.300, 0.634], p=0.478	-0.232 [-0.792, 0.328], p=0.418	**1.425 [0.572, 2.278], p=0.008**	**2.345 [0.893, 3.798], p=0.002**
Age		**0.006 [0.001, 0.011], p=0.038**	**0.007 [-0.002, 0.016], p=0.152**	0.001 [-0.021, 0.024], p=0.896	-0.015 [-0.042, 0.012], p=0.268
Lvl. Spine	L5 / S1	Reference	Reference		
	L4 / L5	**0.507 [0.336, 0.678], p<0.001**	**0.577 [0.396, 0.758], p<0.001**	0.06 [-0.542, 0.663], p=0.838	-0.055 [-0.799, 0.689], p=0.886

	L3 / L4	**0.519 [0.231, 0.808], p<0.001**	**0.552 [0.246, 0.858], p<0.001**	-0.08 [-1.207, 1.047], p=0.888	0.467 [-0.718, 1.652], p=0.438
	L2 / L3	**0.703 [0.283, 1.123], p<0.001**	**0.769 [0.324, 1.214], p<0.001**	0.261 [-1.269, 1.790], p=0.738	0.46 [-1.187, 2.107], p=0.581
	L1 / L2	**1.804 [0.638, 2.970], 0=0.002**	0.346 [-0.769, 1.461], p=0.745	**2.611 [-0.807, 6.028], p=0.019**	**3.248 [1.630, 4.866], p<0.001**
	Other	**0.572 [0.093, 1.051], p=0.019**	**0.727 [0.245, 1.210], p=0.003**	*NA in this outcome category*	*NA in this outcome category*
Morbidity	ASA 1	Reference	Reference		
	ASA 2	**0.196 [0.032, 0.361], p=0.021**	**0.216 [0.041, 0.390], p=0.016**	0.026 [-0.554, 0.606], p=0.931	0.418 [-0.378, 1.215], p=0.285
	ASA 3	**0.383 [0.078, 0.688], p=0.014**	**0.347 [0.020, 0.675], p=0.038**	0.295 [-0.818, 1.408], p=0.615	0.976 [-0.332, 2.284], p=0.114
	ASA 4	**1.396 [0.292, 2.500], p=0.013**	**0.732 [-0.748, 2.213], p=0.330**	**2.985 [0.781, 5.189], p=0.008**	**3.01 [0.688, 5.332], p=0.011**

Table 5.33: Odds ratios (OR), 95% confidence intervals and p-values of variables.

To analyse patients with answered smoking status or BMI, similar analyses were conducted on subsets containing complete data for these columns. However, neither smoking status nor BMI showed significant p-values in any of the adverse category models. Given the reduction in sample size when fitting models with available BMI or smoking status data, coupled with the rare occurrence of complications, these results were considered unreliable and were not reported.

Adverse events that are included in the other/minor category appears to be rather due to different guidelines between countries and therefore a reporting issue, rather than a difference in the patient demographic.

Logistic regression models with complications as outcome variable could identify a few risk factors. Similar to the studies done by Sobottke et al. and Zehnder et al., ASA morbidity status and prior surgery could be identified as risk factor for any adverse events (Sobottke et al., 2017, Zehnder et al., 2021). Prior surgery was the only risk factor for serious adverse events such as nerve root damage during surgery. Risk factors for any other complications such as durotomies, bleeding, wound infections are disc levels other than L5/S1, ASA morbidity status of 3 or 4 and prior surgery. Durotomies were more common for surgery on women. 350 (52.95%) of all durotomies were female (only 46.58% of all patients were female). Additionally, the occurrence rates of complications differed slightly between countries.

5.10 Discussion

The large number of patients and outcomes in the Spine Tango registry allowed for multiple model approaches to be explored and is therefore a powerful source of insight into real-world healthcare. The amount of missing data however, especially concerning patient-reported outcome measures was challenging for the model approaches and reduced the sample size significantly. Additionally, changes in surgery forms caused high missingness in other potentially significant covariates, such as smoking status, BMI or duration of symptoms. More consistency would help improving model fits by increasing sample size.

Linear and logistic regression approaches had low model fits and poor prediction accuracy for COMI scores. Model fit could be improved significantly, when including all available data points from patients, especially when allowing for random intercepts and non-linear functions of time past surgery. Models that include random effects, allow each individual to have a unique estimate of the regarding variable. After exploring multiple model approaches, the mixed-effects model including

random intercepts and a random non-linear time term of the form $f(t) = {}^{1}/_{t}$, where t is time past surgery, fit the data the best and could identify factors associated with improvement of QoL. For most patients QoL improves over time after surgery, although the progression can be different for each individual. The model was capable of identifying subgroups of patients regarding their longitudinal outcome progression. Mixed effects models were superior over linear and logistic regression in terms of model fit and prediction accuracy, measured by RMSE. However, there are limitations when used as prognostic model. Estimates of random effects cannot be computed for patients for who there are no past-surgery measurements available. Prognosis of treatment outcome for new patients is therefore only possible using fixed effects.

Patient characteristics that were correlated with treatment outcome were consistent throughout most modelling approaches and included sex, age, ASA morbidity, previous treatment and baseline COMI scores. Similar to the results of Sobottke et al. prior surgery and higher COMI scores at baseline were with worse outcomes post-surgery (Sobottke et al., 2017). Differences between countries were detected in all models and confirms previous findings by Aghayev et al., the reasons of which needs to be discussed further (Aghayev et al., 2020). Smoking status, which has been included since surgery forms of 2011, was also consistently considered as correlated to treatment outcome (Sobottke et al., 2012, Zehnder et al., 2021).

Even though model fit could be improved significantly by including random effects, there still is significant variation in treatment outcome, that is left unexplained. This could be due to the subjectivity of quality of life, but there could also be factors that are associated with treatment outcome, such as size of disc, intensity of protrusion or MRI scans, which are not routinely collected. It needs to be investigated, if there are measures that are not routinely collected, but associated with treatment outcome, and if they could potentially be integrated in a registry. The non-existence of a core outcome set leads to an inconsistency in measurements in this disease area. This means that analysis of specific characteristics, as seen for smoking status, needs to be done on subsets and therefore a significant reduction of the sample size.

A joint model approach utilising the time to event variable “further follow-up scheduled” was considered to improve the prediction of COMI outcomes. To compare it to the prior mixed-model, this was done on the same subset of patients. The prediction accuracy regarding RMSE values however, did not indicate a significant improvement.

While this study has provided valuable insights into the modelling of the COMI outcomes using the, it's important to acknowledge several limitations in the approach undertaken. Notably, one limitation lies in the assumption of linear effects for continuous covariates. The decision to model continuous

covariates with linear effects might not fully capture complex relationships that could exist between these variables and the outcomes. Other functional forms, such as fractional polynomials or splines, could potentially better represent these associations, allowing for more accurate modelling and prediction. By only considering linear effects, the study might have missed nuanced and non-linear relationships that could be present in the data. Exploring alternative functional forms in future analyses could provide a more comprehensive understanding of the relationships between continuous covariates and the outcomes of interest.

Logistic regression models with complications as outcome variable could identify a few risk factors. Similar to the studies done by Sobottke et al. and Zehnder et al. ASA morbidity status and prior surgery could be identified as risk factor for any adverse events (Sobottke et al., 2012, Zehnder et al., 2021). Furthermore, level of spine other than S1/L5 and age were detected as risk factor as well. Prior surgery was the only risk factor for serious adverse events such as nerve root damage during surgery. The occurrence rates of complications differed slightly between countries, which could partly be caused by reporting policies, but it's possible that the quality of procedures is different across countries too. When dealing with logistic regression models and rare binary outcomes (where one category is significantly less frequent than the other), it is important to interpret estimates cautiously. In the future, more advanced methods can be explored for further analysis.

To summarize, outcome modelling using the patient-reported outcomes questionnaire COMI had poor predictability. A mixed-effects model approach had a good model fit, but outcome prediction for new patients is challenging. Some factors that are associated with QoL improvement could be identified, such as prior surgery, country, baseline QoL scores and time after surgery. However, there still seemed to be a lot of unexplained variation in outcome. Logistic regression approaches of outcome modelling with adverse events showed slightly better prediction accuracy and could identify a few risk factors, depending on the category of adverse events, that can help clinicians decide if surgery is appropriate.

Chapter 6: Conclusions and further work

6.1 Literature review and descriptive analysis

The purpose behind comparing the two data sources was to analyse both observational studies and randomized controlled trials (RCTs) within the context of sciatica-affected patients who underwent microdiscectomy. This analysis had two primary objectives: firstly, to assess the utilization of registries in these studies, and secondly, to examine the alignment in methods and outcomes between RCTs and observational studies. This alignment would emphasize the potential of routinely collected data for the use in RCTs.

The findings from the literature review clearly point to significant alignment between these two types of publications, particularly in terms of the gathered outcomes. Even though a standardized core outcome set was lacking for the studied patient population, key metrics such as ODI, visual or analogue pain scales, and the SF-36 were consistently captured across both RCTs and observational studies. This suggests that observational studies can provide valuable complementary data to RCTs by offering information on real-world practices and potentially more accurate estimates through larger sample sizes. Registries can be used in both RCTs and observational studies for patient identification, tracking progress, and data collection. However, only one RCT and 22 observational studies in the reviewed publications used a registry. While setting up registries can be costly, integrating existing registries in RCTs can save money and even enable certain studies that would not have been feasible otherwise. Registry-based RCTs are a modern trial design that combines the benefits of large-scale registries with randomization, allowing for larger patient populations and cost-effective long-term follow-up (James et al., 2015).

Shifting to the analysis of available data in this study, the NERVES trial and the Spine Tango registry were closely examined to assess similarities. It becomes evident that the NERVES trial and the Spine Tango registry diverge in terms of captured outcomes, baseline covariates, data collection timeframes, and the presence of missing data. The NERVES trial primarily focused on ODI scores as its main outcome at 18 weeks, although data collection extended up to 54 weeks post-randomization. In contrast, the Spine Tango registry represents an ongoing data collection initiative without a predefined study protocol or specific goal, tracking outcomes for up to 3 years post-surgery. It's important to note that both sources incorporated the COMI questionnaire, albeit in varying versions.

Further discrepancies emerged in terms of patient characteristics and surgery details gathered from the two sources. Details like prior treatment, ASA morbidity, and surgeon credentials were documented in the Spine Tango registry but not in the RCT. These differences may stem from pre-specified regulations that eliminate the need for notation (e.g., excluding patients with prior surgery).

The Spine Tango registry faced notable issues with missing data, primarily due to changes in forms over the years that introduced inconsistency in collected variables. Particularly, patient-reported outcomes were inconsistently obtained, significantly reducing the sample size for subsequent analysis. This underscores the necessity for guidelines on handling missing patient-reported outcome data, including establishing a threshold for allowable missing items before disregarding a questionnaire.

While outcomes are often comparable in studies, measured patient covariates differed. This underscores the importance of a core covariate set for this patient population to enhance comparability and deepen the understanding of the indication and procedure. This becomes particularly crucial when establishing a registry for RCTs; incorporating standardized core measurement sets specific to the studied patient population would facilitate uniform data collection across trials and registries.

One of the main future research projects in this field is therefore, to establish a standardized core outcome and covariate sets specific to the patient population undergoing microdiscectomy for sciatica. This would enhance comparability across different studies, including randomized controlled trials (RCTs) and observational studies, leading to more robust and consistent insights into treatment outcomes and patient characteristics.

Enhancing the consistency of data collection in routinely collected data could lead to more comprehensive datasets, aligning closely with data collected in RCTs. This makes insights from registries more comparable and can aid in the planning of RCTs, potentially integrating registries in the form of patient identification and data collection. Given the issues encountered with missing data in the Spine Tango registry, future research could also explore strategies to minimize missing patient-reported outcome data. Developing guidelines for handling missing data and motivating patients to complete questionnaires at key timepoints can help maintain the integrity of large sample sizes and improve the reliability of registry-based analyses.

Noteworthy examples, including the TASTE-trial, the DETO2X-AMI trial, the SORT OUT trials, and the SAFE-PCI for women trials, illustrate how integrating infrastructure for routinely collected data can significantly enhance the quality of RCTs. To establish a registry tailored to RCTs, the inclusion of

standardized core measurement sets specific to the studied patient population is pivotal. Such an approach ensures uniform data collection across registries and trials. To enhance data accuracy and comprehensiveness, these registries should seamlessly integrate with various data sources, including electronic health records and administrative databases. Comprehensive patient data collection becomes crucial when cross-referencing between registries and other sources is unfeasible. Furthermore, registries must possess the capability to track participants over time and gather follow-up data for a thorough assessment of long-term intervention outcomes. However, this study design has shown effectiveness in other clinical areas and could potentially be applied in studies regarding surgical interventions for herniated discs as well. However, despite its advantages, there are several limitations and essential prerequisites for effectively integrating registries in RCTs. Given the potential cost-effectiveness and advantages of using existing registries in RCTs, future research should delve into methodologies and best practices for integrating registry data into clinical trial designs. The above-mentioned studies provide valuable examples of how registry infrastructure can enhance the quality of RCTs across various clinical domains and it is important to increase awareness of this option so that future eligible trials consider it. However, this approach is only effective if a registry routinely collects all the outcomes that a trial plans to collect, or several registries can be cross-referenced for individuals. Additionally, it would be important to identify all the registries that meet these criteria, so that clinical trial planners can check if a suitable registry could be used for their study. Awareness should also be raised about which parts of a clinical trial can be integrated with a registry, such as site and patient identification, data collection, and endpoint detection.

Further methodological work is needed to establish guidelines for registry implementation, including data protection, informatics guidelines for registry programming, data quality, and integration of data from multiple sources to improve completeness and accuracy. Regulatory guidelines that support this trial design should also be established to ensure patient privacy, data protection, and ethical considerations are met (Good Clinical Practice). Guidelines such as this, in combination of motivating payment tariffs such as the spinal best practice tariff on compliance with the British Spine Registry, can impact the general quality of observational registry data (Habeebullah et al., 2021).

6.2 Addressing Missing Data in Patient-Reported Outcome Measures

As mentioned, missing data in COMI outcomes reduced the sample size of the Spine Tango dataset significantly. The conducted simulation study in Chapter 4 explored how to handle missing data, either on the item-level or the score level. This study builds upon research by Marshall et al., who looked at missing data in covariates, by extending the investigation to patient-reported outcome

questionnaires, in particular for the COMI questionnaire, a domain where missing data is a frequent challenge. Previous work by Eekhout et al. (Eekhout et al., 2014) suggested using item-level multiple imputation methods in a simulation study with the Pain Coping Inventory (PCI) for low-back pain.

Unlike typical simulation studies that use artificial data, the simulation study in this thesis used actual patient data from the Spine Tango registry, a less common approach. The aim was to confirm Eekhout's results for the COMI questionnaire, indicating generalizability, or to identify differences that should be considered in future studies using imputation for questionnaire data.

Overall, the findings of the simulation study emphasize the effectiveness of item-based imputation over score-based imputation. This confirmation of the results by Eekhout et al. might therefore extend beyond the study's scope and could apply to similar datasets across various clinical fields. Item-based imputation proved robust in handling missing data for specific questionnaire items, particularly when precision in imputation matters. It's worth noting that these results were obtained using the PMM method exclusively. Further investigations into whether similar results hold when employing alternative imputation methods could be a subject of future research.

When only the entire questionnaire is missing data but not specific items, both imputation methods offer similar results. In such cases, computational efficiency could guide the choice, favouring score-based imputation due to lower computational burden. However, when missing data is isolated to certain items within the questionnaire, item-based imputation is superior. It consistently yields lower root mean squared error (RMSE) values and restores population means (COMI scores) more accurately. The results hold for most scenarios due to ample sample size, but as missingness increases, errors rise, especially under missing not at random (MNAR) conditions. In score-based imputation, selecting high cut-off points becomes essential to prevent errors. For the COMI questionnaire, sensitive cut-off points should be avoided.

While the study's focus is specific to its dataset, the principles governing imputation efficacy could likely extend to similar datasets and clinical contexts. Further validation and insight can be gained from future simulation studies, which should be consulted before applying imputation techniques. However, conducting simulation studies requires significant time and effort, often exceeding project scopes. While their insights can be widely relevant, applying them directly to different questionnaires or datasets requires careful consideration.

In essence, this research offers a comprehensive view of missing data imputation strategies by connecting insights from various studies. The simulation study directly comparing item-level and total

score-level imputation methods enriches our understanding of their effectiveness in real-world scenarios.

While the thesis emphasizes the relevance of its findings to the COMI questionnaire, future research should explore the transferability of these insights to other patient-reported outcome questionnaires with distinct characteristics. Conducting similar simulation studies for different questionnaires can help researchers identify nuances and patterns specific to each instrument, leading to more informed decisions when choosing imputation methods. To further enhance the practical utility of imputation techniques, future research projects should undertake benchmarking exercises that validate the effectiveness of different imputation strategies using real-world datasets with known missing data patterns. This validation process would provide a reference for researchers and practitioners when selecting the most suitable imputation approach for their specific datasets.

6.3 Prognostic Modelling for Enhanced Decision Making

Several approaches of prognostic modelling were considered in this work regarding multiple outcomes. The goal was to develop a model that could identify risk factors for non-significant change in COMI scores and complications during surgery and thereby aid in decision making. Of all the variables that are available in the source data, the following patient characteristics were considered as potentially connected to treatment outcome: sex, age, surgeon credentials, country ID, level of spine, ASA morbidity status, BMI, smoking status, previous treatment, and baseline COMI scores.

Simple models such as linear and logistic regression using COMI scores could identify which patient characteristics were consistently connected to outcomes, which included country ID, previous treatment, ASA morbidity and baseline COMI scores. However, these models exhibited poor fit.

Hence, a more complex model approach was considered by fitting a longitudinal mixed effects model using COMI scores as outcome. More data points could be included in contrast to linear and logistic regression, where only one time point must be chosen as primary outcome. Risk factors that were identified in previous models could be confirmed and including random effects also allowed for adjusting the model to individual patients and therefore achieving a better model fit. However, for patients that were not used to train the model and for which first measurements are not available, estimates cannot be calculated without further assumptions and the use of advanced Bayesian models (Fong et al., 2010).

A joint model approach was considered, utilising the variable “further follow-up scheduled” from follow-up surgery forms as time to event variable. Patient data up of COMI scores up to 2 years were

included in the fit of this model and a mixed model for comparison. However, prediction accuracy regarding RMSE values did not improve.

Again, it needs to be reminded that one main limitation is the data quality. Longitudinal models (both mixed-model and joint model) work best when each individual has numerous outcome measurements, but most patients have only one or two measurements after surgery, which can lead to misleading linearity in plots and good fits that do not accurately reflect real behaviour when measured more frequently. Additionally, the data set needed to be reduced to an appropriate format for the "`mjoint()`" R-function, which requires there to be no further measurements after "no further follow-up scheduled". This shows that the chosen endpoint "no further follow-up" was not a consistent indicator for successful treatment. Patients that were not scheduled for further follow-up still had COMI outcomes after. Reason for this could be that COMI questionnaires are patient-reported, and surgery forms are from practitioners.

Using complications during surgery was also considered as treatment outcome and logistic regression models were fitted to identify risk factors for different types of complications. These risk factors predominantly included prior surgery, ASA morbidity, sex, age, and level of spine. This finding parallels the studies conducted by Sobottke et al. and Zehnder et al., where ASA morbidity status and prior surgery were identified as risk factors for adverse events (Sobottke et al., 2017, Zehnder et al., 2021). However, due to the rare occurrence of complications, estimates of subgroups of patient covariates should be interpreted with caution.

Models that used quality-of-life outcomes had unexplained variation in treatment outcomes that need further examination with medical experts and practitioners. Clinicians should explore ways to improve this variability, such as educating patients about outcome responses and emphasizing the importance of honest and accurate responses. Other techniques for measuring quality of life and pain would also be helpful to reduce reliance on subjective patient responses. Additional measurements, such as the size of the disc and information from MRI scans and estimated disc prolapse, could potentially improve the precision of the models. Incorporating additional clinical measures, such as disc size, MRI scan information, and estimated disc prolapse, could significantly enhance the precision of the prognostic models. Future research should explore ways to integrate these relevant clinical factors into the modelling process, which may lead to a more comprehensive understanding of treatment outcomes and risk factors.

Data quality emerged as a concern, with missing data present in both form changes and on an individual level. Longitudinal models, which benefit from multiple measurements over time, perform better when multiple measurements are available for each individual. However, many patients in the

study had only one or two measurements post-surgery. As the Spine Tango registry is an observational database, improving individual-level missing data can be challenging, relying on each participating research unit. Despite the data limitations resulting in smaller subsets and less accurate model fits, risk factors remained consistently identified across various model approaches, which can aid healthcare providers in decision-making. Future research could explore strategies to encourage and facilitate longitudinal data collection, potentially utilizing technology to enable patients to submit outcome measurements more frequently, thereby creating a more comprehensive dataset aligning with longitudinal modelling requirements.

Sunderland et al. investigated lumbar decompression surgery success using COMI score improvement and identified ASA morbidity status, age, lateral stenosis, revision surgery, and surgeon training as prognostic factors (Sunderland et al., 2021). The Spine Tango registry has been a valuable resource for studies on patients with spinal conditions. Studies that used this data source regarding similar patient populations, like that by Sobottke et al., employed COMI scores as outcomes to identify predictors for quality of life improvement after open decompression surgery for lumbar spinal canal stenosis, revealing the influence of baseline COMI scores, number of prior surgeries, patient comorbidity, and stabilization techniques (Sobottke et al., 2017). Staub et al. developed predictive models for 1-year clinical outcomes after decompression surgery using data from the Spine Tango registry, pointing out considerable uncertainty on individual level (Staub et al., 2020). Aghayev et al. identified high baseline, department-level, and potentially country-level factors associated with negative global treatment outcomes for patients undergoing surgical treatment for lumbar spinal stenosis (Aghayev et al., 2020).

In conclusion, this study's findings demonstrate the challenges and potential of prognostic modelling in the context of spinal surgery outcomes. Despite limitations in data quality and model fitting, consistent risk factors have been identified across various modelling approaches that complement the results found in similar patient populations, providing valuable insights for healthcare decision-making. Future research can explore strategies to enhance data collection, incorporate additional clinical measures, and overcome the limitations of observational data to yield more accurate and informative prognostic models.

References

ABD ELHAFEEZ, S., D'ARRIGO, G., LEONARDIS, D., FUSARO, M., TRIPEPI, G. & ROUMELIOTIS, S. 2021. Methods to Analyze Time-to-Event Data: The Cox Regression Analysis. *Oxid Med Cell Longev,* 2021**,** 1302811.

AGHAYEV, E., MANNION, A. F., FEKETE, T. F., JANSSEN, S., GOODWIN, K., ZWAHLEN, M., BERLEMANN, U. & LORENZ, T. 2020. Risk Factors for Negative Global Treatment Outcomes in Lumbar Spinal Stenosis Surgery: A Mixed Effects Model Analysis of Data from an International Spine Registry. *World Neurosurg,* 136**,** e270-e283.

AHN, Y., LEE, U., KIM, W.-K. & KEUM, H. J. 2018. Five-year outcomes and predictive factors of transforaminal full-endoscopic lumbar discectomy. *Medicine,* 97**,** e13454-e13454.

ANESTHESIOLOGISTS, A. S. O. 2020. *ASA Physical Status Classification System* [Online]. Available: https://www.asahq.org/standards-and-guidelines/asa-physical-status-classification-system [Accessed Jan 27th, 2023].

ARNAB, R. 2017. Chapter 15 - Nonsampling Errors. *In:* ARNAB, R. (ed.) *Survey Sampling Theory and Applications.* Academic Press.

ATLAS, S. J., DEYO, R. A., KELLER, R. B., CHAPIN, A. M., PATRICK, D. L., LONG, J. M. & SINGER, D. E. 1996. The Maine Lumbar Spine Study, Part III. 1-year outcomes of surgical and nonsurgical management of lumbar spinal stenosis. *Spine (Phila Pa 1976),* 21**,** 1787-94; discussion 1794-5.

AUSTIN, P. C., WHITE, I. R., LEE, D. S. & VAN BUUREN, S. 2021. Missing Data in Clinical Research: A Tutorial on Multiple Imputation. *Can J Cardiol,* 37**,** 1322-1331.

AZUR, M. J., STUART, E. A., FRANGAKIS, C. & LEAF, P. J. 2011. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res,* 20**,** 40-9.

BÄCK, M., LEOSDOTTIR, M., HAGSTRÖM, E., NORHAMMAR, A., HAG, E., JERNBERG, T., WALLENTIN, L., LINDAHL, B. & HAMBRAEUS, K. 2021. The SWEDEHEART secondary prevention and cardiac rehabilitation registry (SWEDEHEART CR registry). *Eur Heart J Qual Care Clin Outcomes,* 7**,** 431-437.

BAILEY, B. E., ANDRIDGE, R. & SHOBEN, A. B. 2020. Multiple imputation by predictive mean matching in cluster-randomized trials. *BMC Med Res Methodol,* 20**,** 72.

BENSON, K. & HARTZ, A. J. 2000. A comparison of observational studies and randomized, controlled trials. *N Engl J Med,* 342**,** 1878-86.

BURHOLT, V. & NASH, P. 2011. Short Form 36 (SF-36) Health Survey Questionnaire: normative data for Wales. *J Public Health (Oxf),* 33**,** 587-603.

BURNS, R. A., BUTTERWORTH, P., KIELY, K. M., BIELAK, A. A., LUSZCZ, M. A., MITCHELL, P., CHRISTENSEN, H., VON SANDEN, C. & ANSTEY, K. J. 2011. Multiple imputation was an efficient method for harmonizing the Mini-Mental State Examination with missing item-level data. *J Clin Epidemiol,* 64**,** 787-93.

BURTON, A., ALTMAN, D. G., ROYSTON, P. & HOLDER, R. L. 2006. The design of simulation studies in medical statistics. *Stat Med,* 25**,** 4279-92.

BUTTERMANN, G. R. 2004. Treatment of lumbar disc herniation: epidural steroid injection compared with discectomy. A prospective, randomized study. *J Bone Joint Surg Am,* 86**,** 670-9.

BUUREN, S. V. 2010. Item Imputation Without Specifying Scale Structure. *Methodology: European Journal of Research Methods for The Behavioral and Social Sciences,* 6**,** 31-36.

CHEN, B.-L., GUO, J.-B., ZHANG, H.-W., ZHANG, Y.-J., ZHU, Y., ZHANG, J., HU, H.-Y., ZHENG, Y.-L. & WANG, X.-Q. 2018. Surgical versus non-operative treatment for lumbar disc herniation: a systematic review and meta-analysis. *Clinical Rehabilitation,* 32**,** 146-160.

CHIAROTTO, A., BOERS, M., DEYO, R. A., BUCHBINDER, R., CORBIN, T. P., COSTA, L. O., FOSTER, N. E., GROTLE, M., KOES, B. W. & KOVACS, F. M. 2018. Core outcome measurement instruments for clinical trials in nonspecific low back pain. *Pain,* 159**,** 481.

COLDITZ, G. 2010. Overview of the Epidemiology Methods and Applications: Strengths and Limitations of Observational Study Designs. *Critical reviews in food science and nutrition,* 50 Suppl 1**,** 10-2.

COLLINS, R., BOWMAN, L., LANDRAY, M. & PETO, R. 2020. The Magic of Randomization versus the Myth of Real-World Evidence. Mass Medical Soc.

COLLINS, R., REITH, C., EMBERSON, J., ARMITAGE, J., BAIGENT, C., BLACKWELL, L., BLUMENTHAL, R., DANESH, J., SMITH, G. D. & DEMETS, D. 2016. Interpretation of the evidence for the efficacy and safety of statin therapy. *The Lancet,* 388**,** 2532-2561.

CONCATO, J., LAWLER, E. V., LEW, R. A., GAZIANO, J. M., ASLAN, M. & HUANG, G. D. 2010. Observational methods in comparative effectiveness research. *Am J Med,* 123**,** e16-23.

CONCATO, J., SHAH, N. & HORWITZ, R. I. 2000. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med,* 342**,** 1887-92.

DANIELS, M. J. & ZHAO, Y. D. 2003. Modelling the random effects covariance matrix in longitudinal data. *Statistics in medicine,* 22**,** 1631-1647.

DEYO, R. A., BATTIE, M., BEURSKENS, A. J. H. M., BOMBARDIER, C., CROFT, P., KOES, B., MALMIVAARA, A., ROLAND, M., VON KORFF, M. & WADDELL, G. 1998. Outcome Measures for Low Back Pain Research: A Proposal for Standardized Use. *Spine,* 23**,** 2003-2013.

DOMBKOWSKI, K. J., COSTELLO, L. E., HARRINGTON, L. B., DONG, S., KOLASA, M. & CLARK, S. J. 2014. Age-specific strategies for immunization reminders and recalls: a registry-based randomized trial. *Am J Prev Med,* 47**,** 1-8.

EEKHOUT, I., DE BOER, R. M., TWISK, J. W., DE VET, H. C. & HEYMANS, M. W. 2012. Missing data: a systematic review of how they are reported and handled. *Epidemiology,* 23**,** 729-32.

EEKHOUT, I., DE VET, H. C., TWISK, J. W., BRAND, J. P., DE BOER, M. R. & HEYMANS, M. W. 2014. Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *J Clin Epidemiol,* 67**,** 335-42.

EUROSPINE. 2022a. *Forms* [Online]. Available: https://www.eurospine.org/forms.htm [Accessed January 3rd 2023].

EUROSPINE. 2022b. *Spine Tango Overview* [Online]. Available: https://www.eurospine.org/spine-tango.htm [Accessed Jan 27th 2023].

FAIRBANK, J. C. & PYNSENT, P. B. 2000. The Oswestry Disability Index. *Spine (Phila Pa 1976),* 25**,** 2940-52; discussion 2952.

FARAONI, D. & SCHAEFER, S. T. 2016. Randomized controlled trials vs. observational studies: why not just live together? *BMC anesthesiology,* 16**,** 102-102.

FIGTREE, G. A., VERNON, S. T., HADZIOSMANOVIC, N., SUNDSTRÖM, J., ALFREDSSON, J., NICHOLLS, S. J., CHOW, C. K., PSALTIS, P., RØSJØ, H., LEÓSDÓTTIR, M. & HAGSTRÖM, E. 2022. Mortality and Cardiovascular Outcomes in Patients Presenting With Non-ST Elevation Myocardial Infarction Despite No Standard Modifiable Risk Factors: Results From the SWEDEHEART Registry. *J Am Heart Assoc,* 11**,** e024818.

FONG, Y., RUE, H. & WAKEFIELD, J. 2010. Bayesian inference for generalized linear mixed models. *Biostatistics,* 11**,** 397-412.

FRITZELL, P., KNUTSSON, B., SANDEN, B., STRÖMQVIST, B. & HÄGG, O. 2015. Recurrent Versus Primary Lumbar Disc Herniation Surgery: Patient-reported Outcomes in the Swedish Spine Register Swespine. *Clin Orthop Relat Res,* 473**,** 1978-84.

GILMARTIN-THOMAS, J. F., LIEW, D. & HOPPER, I. 2018. Observational studies and their utility for practice. *Aust Prescr,* 41**,** 82-85.

GROOTENDORST, D. C., JAGER, K. J., ZOCCALI, C. & DEKKER, F. W. 2010. Observational studies are complementary to randomized controlled trials. *Nephron Clin Pract,* 114**,** c173-7.

GRØVLE, L., HAUGEN, A. J., KELLER, A., NATVIG, B., BROX, J. I. & GROTLE, M. 2010. The bothersomeness of sciatica: patients' self-report of paresthesia, weakness and leg pain. *Eur Spine J,* 19**,** 263-9.

HABEEBULLAH, A., RAJGOR, H. D., GARDNER, A. & JONES, M. 2021. The impact of a spinal best practice tariff on compliance with the British Spine Registry. *Bone Jt Open,* 2**,** 198-201.

HARITON, E. & LOCASCIO, J. J. 2018. Randomised controlled trials - the gold standard for effectiveness research: Study design: randomised controlled trials. *Bjog,* 125**,** 1716.

HAUCK, W. W. & ANDERSON, S. 1984. A Survey regarding the Reporting of Simulation Studies. *The American Statistician,* 38**,** 214-216.

HAWTHORNE, G. & ELLIOTT, P. 2005. Imputing cross-sectional missing data: comparison of common techniques. *Aust N Z J Psychiatry,* 39**,** 583-90.

HESS, C. N., RAO, S. V., KONG, D. F., ABERLE, L. H., ANSTROM, K. J., GIBSON, C. M., GILCHRIST, I. C., JACOBS, A. K., JOLLY, S. S., MEHRAN, R., MESSENGER, J. C., NEWBY, L. K., WAKSMAN, R. & KRUCOFF, M. W. 2013. Embedding a randomized clinical trial into an ongoing registry infrastructure: unique opportunities for efficiency in design of the Study of Access site For Enhancement of Percutaneous Coronary Intervention for Women (SAFE-PCI for Women). *Am Heart J,* 166**,** 421-8.

HOAGLIN, D. C. & ANDREWS, D. F. 1975. The Reporting of Computation-Based Results in Statistics. *The American Statistician,* 29**,** 122-126.

HOFMANN, R., JAMES, S. K., JERNBERG, T., LINDAHL, B., ERLINGE, D., WITT, N., AREFALK, G., FRICK, M., ALFREDSSON, J., NILSSON, L., RAVN-FISCHER, A., OMEROVIC, E., KELLERTH, T., SPARV, D., EKELUND, U., LINDER, R., EKSTRÖM, M., LAUERMANN, J., HAAGA, U., PERNOW, J., ÖSTLUND, O., HERLITZ, J. & SVENSSON, L. 2017. Oxygen Therapy in Suspected Acute Myocardial Infarction. *N Engl J Med,* 377**,** 1240-1249.

HOOFF, M. L. V., JACOBS, W. C. H., WILLEMS, P. C., WOUTERS, M. W. J. M., KLEUVER, M. D., PEUL, W. C., OSTELO, R. W. J. G. & FRITZELL, P. 2015. Evidence and practice in spine registries. *Acta Orthopaedica,* 86**,** 534-544.

INFANTE-RIVARD, C. & CUSSON, A. 2018. Reflection on modern methods: selection bias-a review of recent developments. *Int J Epidemiol,* 47**,** 1714-1722.

JAKOBSEN, J. C., GLUUD, C., WETTERSLEV, J. & WINKEL, P. 2017. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC Medical Research Methodology,* 17**,** 162.

JAMES, S., RAO, S. V. & GRANGER, C. B. 2015. Registry-based randomized clinical trials--a new clinical trial paradigm. *Nat Rev Cardiol,* 12**,** 312-6.

JERNBERG, T., ATTEBRING, M. F., HAMBRAEUS, K., IVERT, T., JAMES, S., JEPPSSON, A., LAGERQVIST, B., LINDAHL, B., STENESTRAND, U. & WALLENTIN, L. 2010. The Swedish Web-system for enhancement and development of evidence-based care in heart disease evaluated according to recommended therapies (SWEDEHEART). *Heart,* 96**,** 1617-21.

JOSÉ, M. & EDELMAN, E. R. 2017. From nonclinical research to clinical trials and patient-registries: challenges and opportunities in biomedical research. *Revista Española de Cardiología (English Edition),* 70**,** 1121-1133.

KARANATSIOS, B., PRANG, K.-H., VERBUNT, E., YEUNG, J. M., KELAHER, M. & GIBBS, P. 2020. Defining key design elements of registry-based randomised controlled trials: a scoping review. *Trials,* 21**,** 552.

KIM, M., GUILFOYLE, M. R., SEELEY, H. M. & LAING, R. J. 2010. A modified Roland-Morris disability scale for the assessment of sciatica. *Acta Neurochir (Wien),* 152**,** 1549-53; discussion 1553.

KOES, B. W., VAN TULDER, M. W. & PEUL, W. C. 2007. Diagnosis and treatment of sciatica. *BMJ (Clinical research ed.),* 334**,** 1313-1317.

LAGERQVIST, B., FRÖBERT, O., OLIVECRONA, G. K., GUDNASON, T., MAENG, M., ALSTRÖM, P., ANDERSSON, J., CALAIS, F., CARLSSON, J., COLLSTE, O., GÖTBERG, M., HÅRDHAMMAR, P., IOANES, D., KALLRYD, A., LINDER, R., LUNDIN, A., ODENSTEDT, J., OMEROVIC, E., PUSKAR, V., TÖDT, T., ZELLEROTH, E., ÖSTLUND, O. & JAMES, S. K. 2014. Outcomes 1 year after thrombus aspiration for myocardial infarction. *N Engl J Med,* 371**,** 1111-20.

LAUER, M. S. & D'AGOSTINO, R. B. 2013. The Randomized Registry Trial — The Next Disruptive Technology in Clinical Research? *New England Journal of Medicine,* 369**,** 1579-1581.

LEE, J., SHIN, K., PARK, S., LEE, G., LEE, C., KIM, D., KIM, D. & YANG, H. 2018. Comparison of Clinical Efficacy Between Transforaminal and Interlaminar Epidural Injections in Lumbosacral Disc Herniation: A Systematic Review and Meta-Analysis. *Pain physician,* 21**,** 433.

LI, G., SAJOBI, T. T., MENON, B. K., KORNGUT, L., LOWERISON, M., JAMES, M., WILTON, S. B., WILLIAMSON, T., GILL, S., DROGOS, L. L., SMITH, E. E., VOHRA, S., HILL, M. D. & THABANE, L. 2016. Registry-based randomized controlled trials- what are the advantages, challenges, and areas for future research? *J Clin Epidemiol,* 80**,** 16-24.

MANNION, A. F., VILA-CASADEMUNT, A., DOMINGO-SÀBAT, M., WUNDERLIN, S., PELLISÉ, F., BAGO, J., ACAROGLU, E., ALANAY, A., PÉREZ-GRUESO, F. S., OBEID, I. & KLEINSTÜCK, F. S. 2016. The Core Outcome Measures Index (COMI) is a responsive instrument for assessing the outcome of treatment for adult spinal deformity. *Eur Spine J,* 25**,** 2638-48.

MANSON, N. A., MCKEON, M. D. & ABRAHAM, E. P. 2013. Transforaminal epidural steroid injections prevent the need for surgery in patients with sciatica secondary to lumbar disc herniation: a retrospective case series. *Canadian Journal of Surgery,* 56**,** 89.

MARS, K., WALLERT, J., HELD, C., HUMPHRIES, S., PINGEL, R., JERNBERG, T., OLSSON, E. M. G. & HOFMANN, R. 2021. Association between β-blocker dose and cardiovascular outcomes after myocardial infarction: insights from the SWEDEHEART registry. *Eur Heart J Acute Cardiovasc Care,* 10**,** 372-379.

MARSHALL, A., ALTMAN, D. G., ROYSTON, P. & HOLDER, R. L. 2010. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Med Res Methodol,* 10**,** 7.

MEHROTRA, D. V., LIU, F. & PERMUTT, T. 2017. Missing data in clinical trials: control-based mean imputation and sensitivity analysis. *Pharm Stat,* 16**,** 378-392.

MÖLLER, M., WOLF, O., BERGDAHL, C., MUKKA, S., RYDBERG, E. M., HAILER, N. P., EKELUND, J. & WENNERGREN, D. 2022. The Swedish Fracture Register – ten years of experience and 600,000 fractures collected in a National Quality Register. *BMC Musculoskeletal Disorders,* 23**,** 141.

MORRIS, T. P., WHITE, I. R. & CROWTHER, M. J. 2019. Using simulation studies to evaluate statistical methods. *Stat Med,* 38**,** 2074-2102.

MORRIS, T. P., WHITE, I. R. & ROYSTON, P. 2014. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology,* 14**,** 75.

MOUSSA, I., HERMANN, A., MESSENGER, J. C., DEHMER, G. J., WEAVER, W. D., RUMSFELD, J. S. & MASOUDI, F. A. 2013. The NCDR CathPCI Registry: a US national perspective on care and outcomes for percutaneous coronary intervention. *Heart,* 99**,** 297-303.

NIELSEN, S. F., NORDESTGAARD, B. G. & BOJESEN, S. E. 2012. Statin use and reduced cancer-related mortality. *New England Journal of Medicine,* 367**,** 1792-1802.

NØRGAARD, M., EHRENSTEIN, V. & VANDENBROUCKE, J. P. 2017. Confounding in observational studies based on large health care databases: problems and potential solutions - a primer for the clinician. *Clinical epidemiology,* 9**,** 185-193.

OSTERMAN, H., SEITSALO, S., KARPPINEN, J. & MALMIVAARA, A. 2006. Effectiveness of microdiscectomy for lumbar disc herniation: a randomized controlled trial with 2 years of follow-up. *Spine (Phila Pa 1976),* 31**,** 2409-14.

PEDERSEN, A. B., MIKKELSEN, E. M., CRONIN-FENTON, D., KRISTENSEN, N. R., PHAM, T. M., PEDERSEN, L. & PETERSEN, I. 2017. Missing data and multiple imputation in clinical epidemiological research. *Clin Epidemiol,* 9**,** 157-166.

PINHEIRO, J., BATES, D., DEBROY, S. S. & SARKAR, D. 2013. Nlme: Linear and Nonlinear Mixed Effects Models. *R package version 31-110,* 3**,** 1-113.

PRICE, D. D., MCGRATH, P. A., RAFII, A. & BUCKINGHAM, B. 1983. The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. *PAIN,* 17**,** 45-56.

RANGANATHAN, P. & AGGARWAL, R. 2018. Study designs: Part 1 - An overview and classification. *Perspectives in clinical research,* 9**,** 184-186.

RAO, S. V., HESS, C. N., BARHAM, B., ABERLE, L. H., ANSTROM, K. J., PATEL, T. B., JORGENSEN, J. P., MAZZAFERRI, E. L., JR., JOLLY, S. S., JACOBS, A., NEWBY, L. K., GIBSON, C. M., KONG, D. F., MEHRAN, R., WAKSMAN, R., GILCHRIST, I. C., MCCOURT, B. J., MESSENGER, J. C., PETERSON, E. D., HARRINGTON, R. A. & KRUCOFF, M. W. 2014. A registry-based randomized trial comparing radial and femoral approaches in women undergoing percutaneous coronary intervention: the SAFE-PCI for Women (Study of Access Site for Enhancement of PCI for Women) trial. *JACC Cardiovasc Interv,* 7**,** 857-67.

RILEY, R. D., SNELL, K. I., ENSOR, J., BURKE, D. L., HARRELL, F. E., JR., MOONS, K. G. & COLLINS, G. S. 2019. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med,* 38**,** 1276-1296.

ROLAND, M. & MORRIS, R. 1983. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. *Spine (Phila Pa 1976),* 8**,** 141-4.

ROPPER, A. H. & ZAFONTE, R. D. 2015. Sciatica. *N Engl J Med,* 372**,** 1240-8.

ROTH, P. L., SWITZER, F. S. & SWITZER, D. M. 1999. Missing Data in Multiple Item Scales: A Monte Carlo Analysis of Missing Data Techniques. *Organizational Research Methods,* 2**,** 211 - 232.

SCHOUTEN, R. M., LUGTIG, P. & VINK, G. 2018. Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation,* 88**,** 2909-2930.

SIGMUNDSSON, F. G., JÖNSSON, B. & STRÖMQVIST, B. 2013. Impact of pain on function and health related quality of life in lumbar spinal stenosis. A register study of 14,821 patients. *Spine (Phila Pa 1976),* 38**,** E937-45.

SOBOTTKE, R., AGHAYEV, E., RÖDER, C., EYSEL, P., DELANK, S. K. & ZWEIG, T. 2012. Predictors of surgical, general and follow-up complications in lumbar spinal stenosis relative to patient age as emerged from the Spine Tango Registry. *Eur Spine J,* 21**,** 411-7.

SOBOTTKE, R., HERREN, C., SIEWE, J., MANNION, A. F., RÖDER, C. & AGHAYEV, E. 2017. Predictors of improvement in quality of life and pain relief in lumbar spinal stenosis relative to patient age: a study based on the Spine Tango registry. *Eur Spine J,* 26**,** 462-472.

SPIETH, P. M., KUBASCH, A. S., PENZLIN, A. I., ILLIGENS, B. M.-W., BARLINN, K. & SIEPMANN, T. 2016. Randomized controlled trials - a matter of design. *Neuropsychiatric disease and treatment,* 12**,** 1341-1349.

STATISTICSGLOBE. 2022. *Predictive mean matching imputation method* [Online]. Available: https://statisticsglobe.com/predictive-mean-matching-imputation-method/ [Accessed Dec 18th 2022].

STAUB, L. P., AGHAYEV, E., SKRIVANKOVA, V., LORD, S. J., HASCHTMANN, D. & MANNION, A. F. 2020. Development and temporal validation of a prognostic model for 1-year clinical outcome after decompression surgery for lumbar disc herniation. *Eur Spine J,* 29**,** 1742-1751.

SUNDERLAND, G., FOSTER, M., DHEERENDRA, S. & PILLAY, R. 2021. Patient-Reported Outcomes Following Lumbar Decompression Surgery: A Review of 2699 Cases. *Global Spine Journal,* 11**,** 172-179.

THUESEN, L., JENSEN, L. O., TILSTED, H. H., MÆNG, M., TERKELSEN, C., THAYSSEN, P., RAVKILDE, J., CHRISTIANSEN, E. H., BØTKER, H. E., MADSEN, M. & LASSEN, J. F. 2013. Event detection using population-based health care databases in randomized clinical trials: a novel research tool in interventional cardiology. *Clin Epidemiol,* 5**,** 357-61.

TRIPEPI, G., CHESNAYE, N. C., DEKKER, F. W., ZOCCALI, C. & JAGER, K. J. 2020. Intention to treat and per protocol analysis in clinical trials. *Nephrology (Carlton),* 25**,** 513-517.

U.S. DEPARTMENT OF LABOR. *Worker's Compensation* [Online]. Available: https://www.dol.gov/general/topic/workcomp [Accessed 08 July 2023].

VAN BUUREN, S. & GROOTHUIS-OUDSHOORN, K. 2011. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software,* 45**,** 1 - 67.

VINK, G., FRANK, L., PANNEKOEK, J. & BUUREN, S. 2014. Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica,* 68.

WANG, J. C., LIN, E., BRODKE, D. S. & YOUSSEF, J. A. 2002. Epidural injections for the treatment of symptomatic lumbar herniated discs. *J Spinal Disord Tech,* 15**,** 269-72.
WEINSTEIN, J. N., LURIE, J. D., TOSTESON, T. D., TOSTESON, A. N., BLOOD, E., ABDU, W. A., HERKOWITZ, H., HILIBRAND, A., ALBERT, T. & FISCHGRUND, J. 2008. Surgical versus non-operative treatment for lumbar disc herniation: four-year results for the Spine Patient Outcomes Research Trial (SPORT). *Spine,* 33**,** 2789.
WILBY, M. J., BEST, A., WOOD, E., BURNSIDE, G., BEDSON, E., SHORT, H., WHEATLEY, D., HILL-MCMANUS, D., SHARMA, M., CLARK, S., BARANIDHARAN, G., PRICE, C., MANNION, R., HUTCHINSON, P. J., HUGHES, D. A., MARSON, A. & WILLIAMSON, P. R. 2021. Surgical microdiscectomy versus transforaminal epidural steroid injection in patients with sciatica secondary to herniated lumbar disc (NERVES): a phase 3, multicentre, open-label, randomised controlled trial and economic evaluation. *Lancet Rheumatol,* 3**,** e347-e356.
YUAN, Y. 2005. Multiple Imputation for Missing Data: Concepts and New Development.
ZEHNDER, P., HELD, U., PIGOTT, T., LUCA, A., LOIBL, M., REITMEIR, R., FEKETE, T., HASCHTMANN, D. & MANNION, A. F. 2021. Development of a model to predict the probability of incurring a complication during spine surgery. *Eur Spine J,* 30**,** 1337-1354.
ZHANG, D. 2020. *Coefficients of Determination for Mixed-Effects Models*.
ZHANG, Z. 2016. Missing data imputation: focusing on single imputation. *Ann Transl Med,* 4**,** 9.

Appendix

Appendix A: COMI version in the Spine Tango registry

Question 1: Which of the following problems troubles you the most? Please tick ONE BOX only.

( ) back pain

( ) leg / buttock pain

( ) sensory disturbances in the back/legs/buttocks, e.g. tingling 'pins and needles', numbness

( ) none of the above

Question 2: For the following 2 questions (2a and 2b) we would like you to indicate the severity of your pain, by ticking the appropriate box (where "0" = no pain, "10" = worst pain you can imagine). There are separate questions for back pain and for leg pain (sciatica)/buttock pain.

Question 2a: How severe was your back pain in the last week?

No pain 0 ( ) 1 ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 ( ) 8 ( ) 9 ( ) 10 ( ) worst pain

Question 2b: How severe was your leg pain in the last week?

No pain 0 ( ) 1 ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 ( ) 8 ( ) 9 ( ) 10 ( ) worst pain

Question 3: During the past week, how much did your back problem interfere with your normal work (including both work outside the home and housework)?

( ) not at all

( ) a little bit

( ) moderately

( ) quite a bit

( ) extremely

Question 4: If you had to spend the rest of your life with the symptoms you have right now, how would you feel about it?

( ) very satisfied

( ) somewhat satisfied

( ) neither satisfied nor dissatisfied

( ) somewhat dissatisfied

( ) very dissatisfied

Question 5: Please reflect on the last week. How would you rate your quality of life?

( ) very good

( ) good

( ) moderate

( ) bad

( ) very bad

Question 6: During the past 4 weeks, how many days did you cut down on the things you usually do (work, housework, school, recreational activities) because of your back problem?

( ) none

( ) between 1 and 7 days

( ) between 8 and 14 days

( ) between 15 and 21 days

( ) more than 21 days

Question 7: During the past 4 weeks, how many days did your back problem keep you from going to work (job, school, housework)?

( ) none

( ) between 1 and 7 days

( ) between 8 and 14 days

( ) between 15 and 21 days

( ) more than 21 days

Appendix B: COMI version in NERVES trial

In addition to optional follow-up sections and private patient details, the following items are included throughout the examination interval.Question 1: During the past week, how bothersome have each of the following symptoms been?

Question 1a: Low back pain

( ) Not at all bothersome

( ) Slightly bothersome

( ) Moderately bothersome

( ) Very bothersome

( ) Extremely bothersome

Question 1b: Leg pain (sciatica)

( ) Not at all bothersome

( ) Slightly bothersome

( ) Moderately bothersome

( ) Very bothersome

( ) Extremely bothersome

Question 2: During the past week, how much did pain interfere with your normal work (including both work outside the home and housework)?

( ) Not at all

( ) A little bit

( ) Moderately

( ) Quite a bit

( ) Extremely

Question 3: If you had to spend the rest of your life with the symptoms you have right now, how would you feel about it?

( ) Very satisfied

( ) Somewhat satisfied

( ) Neither satisfied nor dissatisfied

( ) Somewhat dissatisfied

( ) Very dissatisfied

Question 4: During the past 4 weeks, about how many days did you cut down on the things you usually do for more than half the day because of back pain or leg pain (sciatica)?

___ (number of days)

Question 5: During the past 4 weeks, how many days did low back pain or leg pain (sciatica) keep you from going to work or school?

___ (number of days)

Appendix C: List of papers of observational studies included in literature review in Chapter 2

#	Pubmed ID	Authors	Title	Journal
1	34815704	Ma, Cheng et al.	"Percutaneous Endoscopic Lumbar Discectomy for Huge Lumbar Disc Herniation with Complete Dural Sac Stenosis via an Interlaminar Approach: An Observational Retrospective Cohort Study.Approach: An Observational Retrospective Cohort Study"	International journal of general medicine vol. 14 8317-8324. 16 Nov. 2021, doi:10.2147/IJGM.S341309
2	34238046	Evaniew, Nathan et al.	"Minimally Invasive Tubular Lumbar Discectomy Versus Conventional Open Lumbar Discectomy: An Observational Study From the Canadian Spine Outcomes and Research Network."	Global spine journal, 21925682211029863. 9 Jul. 2021, doi:10.1177/21925682211029863
3	33981932	Wakaizumi, Kenta et al.	"Momentary pain assessments reveal benefits of endoscopic discectomy: a prospective cohort study."	Pain reports vol. 6,1 e906. 17 Mar. 2021, doi:10.1097/PR9.0000000000000906
4	33185390	Ren, Chunpeng et al.	"Microendoscopic Discectomy Combined with Annular Suture Versus Percutaneous Transforaminal Endoscopic Discectomy for Lumbar Disc Herniation: A Prospective Observational Study."	Pain physician vol. 23,6 (2020): E713-E721.
5	33181867	Kang, Suk-Hyung et al.	"A Prospective Observational Study of Return to Work after Single Level Lumbar Discectomy."	Journal of Korean Neurosurgical Society vol. 63,6 (2020): 806-813. doi:10.3340/jkns.2020.0227
6	31414288	Wagner, Arthur et al.	"Psychological predictors of quality of life and functional outcome in patients undergoing elective surgery for degenerative lumbar spine disease."	European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society vol. 29,2 (2020): 349-359. doi:10.1007/s00586-019-06106-x
7	29942349	Vinas-Rios, Juan Manuel et al.	"Incidence of early postoperative complications requiring surgical revision for recurrent lumbar disc herniation after spinal surgery: a retrospective observational study of 9,310 patients from the German Spine Register."	Patient safety in surgery vol. 12 9. 21 May. 2018, doi:10.1186/s13037-018-0157-1
8	29523985	Elkan, P et al.	"Response rate does not affect patient-reported outcome after lumbar discectomy."	European spine journal: official publication of the European Spine Society, the European Spinal Deformity Society,

				and the European Section of the Cervical Spine Research Society vol. 27,7 (2018): 1538-1546. doi:10.1007/s00586-018-5541-0
9	29201144	Segura-Trepichio, Manuel et al.	"Length of stay, costs, and complications in lumbar disc herniation surgery by standard PLIF versus a new dynamic interspinous stabilization technique."	Patient safety in surgery vol. 11 26. 23 Nov. 2017, doi:10.1186/s13037-017-0141-1
10	27653010	Wang, Wenjun et al.	"Application of Laparoscopic Lumbar Discectomy and Artificial Disc Replacement: At Least Two Years of Follow-Up."	Spine vol. 41 Suppl 19 (2016): B38-B43. doi:10.1097/BRS.0000000000001820
11	26451868	Schiavolin, Silvia et al.	"Change in quality of life, disability, and well-being after decompressive surgery: results from a longitudinal study."	International journal of rehabilitation research. Internationale Zeitschrift fur Rehabilitationsforschung. Revue internationale de recherches de readaptation vol. 38,4 (2015): 357-63. doi:10.1097/MRR.0000000000000136
12	24598904	Löbner, Margrit et al.	"Inpatient or outpatient rehabilitation after herniated disc surgery? - Setting-specific preferences, participation and outcome of rehabilitation."	PloS one vol. 9,3 e89200. 5 Mar. 2014, doi:10.1371/journal.pone.0089200
13	24561397	el Barzouhi, Abdelilah et al.	"Reliability of gadolinium-enhanced magnetic resonance imaging findings and their correlation with clinical outcome in patients with sciatica."	The spine journal : official journal of the North American Spine Society vol. 14,11 (2014): 2598-607. doi:10.1016/j.spinee.2014.02.028
14	22999108	Haugen, Anne Julsrud et al.	"Prognostic factors for non-success in patients with sciatica and disc herniation."	BMC musculoskeletal disorders vol. 13 183. 22 Sep. 2012, doi:10.1186/1471-2474-13-183
15	22943189	Konnopka, Alexander et al.	"Psychiatric comorbidity as predictor of costs in back pain patients undergoing disc surgery: a longitudinal observational study."	BMC musculoskeletal disorders vol. 13 165. 3 Sep. 2012, doi:10.1186/1471-2474-13-165
16	22453894	Lee, Kong Hwee et al.	"Clinical and radiological outcomes of open versus minimally invasive transforaminal lumbar interbody fusion."	European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society vol. 21,11 (2012): 2265-70. doi:10.1007/s00586-012-2281-4
17	25729660	Farzanegan, Gholamreza et al.	"Quality-of-Life Evaluation of Patients Undergoing Lumbar Discectomy Using Short Form 36."	Anesthesiology and pain medicine vol. 1,2 (2011): 73-6. doi:10.5812/kowsar.22287523.1998
18	25729651	Farzanegan, Gholamreza et al.	"Effects of lumbar discectomy on disability and depression in patients with chronic low back pain."	Anesthesiology and pain medicine vol. 1,1 (2011): 20-4. doi:10.5812/kowsar.22287523.1529
19	20689982	Zieger, Margrit et al.	"The impact of psychiatric comorbidity on the return to work in patients undergoing herniated disc surgery."	Journal of occupational rehabilitation vol. 21,1 (2011): 54-65. doi:10.1007/s10926-010-9257-1
20	20135333	Cobo Soriano, Javier et al.	"Predictors of outcome after decompressive lumbar surgery and instrumented posterolateral fusion."	European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society vol. 19,11 (2010): 1841-8. doi:10.1007/s00586-010-1284-2

21	19301042	Schluessmann, E et al.	"SWISSspine: a nationwide registry for health technology assessment of lumbar disc prostheses."	European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society vol. 18,6 (2009): 851-61. doi:10.1007/s00586-009-0934-8
22	17849238	Heider, Dirk et al.	"Health-related quality of life in patients after lumbar disc surgery: a longitudinal observational study."	Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation vol. 16,9 (2007): 1453-60. doi:10.1007/s11136-007-9255-8
23	16924216	Slover, James et al.	"The impact of comorbidities on the change in short-form 36 and oswestry scores following lumbar spine surgery."	Spine vol. 31,17 (2006): 1974-80. doi:10.1097/01.brs.0000229252.30903.b9
24	16540869	Carragee, Eugene J et al.	"A prospective controlled study of limited versus subtotal posterior discectomy: short-term outcomes in patients with herniated lumbar intervertebral discs and large posterior anular defect."	Spine vol. 31,6 (2006): 653-7. doi:10.1097/01.brs.0000203714.76250.68
25	14589465	Bogduk, Nikolai, and Michael Karasek.	"Two-year follow-up of a controlled trial of intradiscal electrothermal anuloplasty for chronic low back pain resulting from internal disc disruption."	The spine journal : official journal of the North American Spine Society vol. 2,5 (2002): 343-50. doi:10.1016/s1529-9430(02)00409-6
26	12533579	Carragee, Eugene J et al.	"Clinical outcomes after lumbar discectomy for sciatica: the effects of fragment type and anular competence."	The Journal of bone and joint surgery. American volume vol. 85,1 (2003): 102-8.
27	9253102	Carragee, E J, and D H Kim.	"A prospective analysis of magnetic resonance imaging findings in patients with sciatica and lumbar disc herniation. Correlation of outcomes with disc fragment and canal morphology."	Spine vol. 22,14 (1997): 1650-60. doi:10.1097/00007632-199707150-00025
28	8153823	Pople, I K, and H B Griffith.	"Prediction of an extruded fragment in lumbar disc patients from clinical presentations."	Spine vol. 19,2 (1994): 156-8. doi:10.1097/00007632-199401001-00007
29	35040807	Thomé, Claudius et al.	"Motor Recovery Depends on Timing of Surgery in Patients With Lumbar Disk Herniation."	Neurosurgery vol. 90,3 (2022): 347-353. doi:10.1227/NEU.0000000000001825
30	34622848	Ahn, Yong et al.	"The irony of the transforaminal approach: A comparative cohort study of transforaminal endoscopic lumbar discectomy for foraminal versus paramedian lumbar disc herniation."	Medicine vol. 100,40 (2021): e27412. doi:10.1097/MD.0000000000027412
31	34213872	Song, Sung Kyu et al.	"Comparison of the Outcomes of Percutaneous Endoscopic Interlaminar Lumbar Discectomy and Open Lumbar Microdiscectomy at the L5-S1 Level."	Pain physician vol. 24,4 (2021): E467-E475.
32	34106586	Ji, Qing-Hui et al.	"Study on the effect of percutaneous intervertebral foraminoscopic discectomy in the treatment of lumbar disc herniation."	Medicine vol. 100,19 (2021): e25345. doi:10.1097/MD.0000000000025345

33	33988950	An, Gang et al.	"Pathomechanism of Lower-level Discogenic Groin Pain and Clinical Outcomes of Percutaneous Endoscopic Discectomy for the Treatment of Discogenic Groin Pain."	Pain physician vol. 24,3 (2021): E289-E297.
34	33573617	Takahashi, Hiroshi et al.	"Characteristics of relief and residual low back pain after discectomy in patients with lumbar disc herniation: analysis using a detailed visual analog scale."	BMC musculoskeletal disorders vol. 22,1 167. 11 Feb. 2021, doi:10.1186/s12891-021-04015-z
35	33515794	Chen, Chao et al.	"Full Endoscopic Lumbar Foraminoplasty with Periendoscopic Visualized Trephine Technique for Lumbar Disc Herniation with Migration and/or Foraminal or Lateral Recess Stenosis."	World neurosurgery vol. 148 (2021): e658-e666. doi:10.1016/j.wneu.2021.01.062
36	33371065	Son, Seong et al.	"Outcomes of epiduroscopic laser ablation in patients with lumbar disc herniation."	Medicine vol. 99,51 (2020): e23337. doi:10.1097/MD.0000000000023337
37	33290379	Chen, Xiaolong et al.	"Do Markers of Inflammation and/or Muscle Regeneration in Lumbar Multifidus Muscle and Fat Differ Between Individuals with Good or Poor Outcome Following Microdiscectomy for Lumbar Disc Herniation?."	Spine vol. 46,10 (2021): 678-686. doi:10.1097/BRS.0000000000003863
38	33230085	Holmberg, Siril T et al.	"Pain During Sex Before and After Surgery for Lumbar Disc Herniation: A Multicenter Observational Study."	Spine vol. 45,24 (2020): 1751-1757. doi:10.1097/BRS.0000000000003675
39	32285191	Polak, Samuel B et al.	"Surgery for extraforaminal lumbar disc herniation: a single center comparative observational study."	Acta neurochirurgica vol. 162,6 (2020): 1409-1415. doi:10.1007/s00701-020-04313-w
40	32115977	Mlaka, J et al.	"Endoscopic discectomy as an effective treatment of a herniated intervertebral disc."	Bratislavske lekarske listy vol. 121,3 (2020): 199-205. doi:10.4149/BLL_2020_030
41	32081830	Vangen-Lønne, Vetle et al.	"Microdiscectomy for Lumbar Disc Herniation: A Single-Center Observational Study."	World neurosurgery vol. 137 (2020): e577-e583. doi:10.1016/j.wneu.2020.02.056
42	32065508	Beyaz, Serbülent Gökhan et al.	"A Novel Combination Technique: Three Points of Epiduroscopic Laser Neural Decompression and Percutaneous Laser Disc Decompression With the Ho:YAG Laser in an MSU Classification 3AB Herniated Disc."	Pain practice : the official journal of World Institute of Pain vol. 20,5 (2020): 501-509.
43	31899404	Aghayev, Emin et al.	"Risk Factors for Negative Global Treatment Outcomes in Lumbar Spinal Stenosis Surgery: A Mixed Effects Model Analysis of Data from an International Spine Registry."	World neurosurgery vol. 136 (2020): e270-e283. doi:10.1016/j.wneu.2019.12.147
44	31804313	Cao, Jian et al.	"Percutaneous endoscopic lumbar discectomy for lumbar disc herniation as day surgery - short-term clinical results of 235 consecutive cases."	Medicine vol. 98,49 (2019): e18064. doi:10.1097/MD.0000000000018064

45	31786988	Lagerbäck, Tobias et al.	“Lumbar disc herniation surgery in adolescents and young adults: a long-term outcome comparison.”	The bone & joint journal vol. 101-B,12 (2019): 1534-1541. doi:10.1302/0301-620X.101B12.BJJ-2019-0621.R1
46	31725682	Kim, Jang Hun et al.	“Efficacy of automated percutaneous lumbar discectomy for lumbar disc herniation in young male soldiers.”	Medicine vol. 98,46 (2019): e18044. doi:10.1097/MD.0000000000018044
47	31679049	Papanastasiou, Evangelos I et al.	“Association between MRI findings and clinical outcomes in a period of 5 years after lumbar spine microdiscectomy.”	European journal of orthopaedic surgery & traumatology : orthopedie traumatologie vol. 30,3 (2020): 441-446. doi:10.1007/s00590-019-02588-z
48	31573063	Koksal, Vaner, and Rahmi Kemal Koc.	“Microsurgery versus Medical Treatment for Neuropathic Pain Caused by Foraminal Extraforaminal Lumbar Disc Herniation: An Observational Study.”	Turkish neurosurgery vol. 29,6 (2019): 915-926. doi:10.5137/1019-5149.JTN.26988-19.1
49	31151337	Ahn, Yong et al.	“Transforaminal Endoscopic Lumbar Discectomy Versus Open Lumbar Microdiscectomy: A Comparative Cohort Study with a 5-Year Follow-Up.”	Pain physician vol. 22,3 (2019): 295-304.
50	30929479	Fjeld, O R et al.	“Complications, reoperations, readmissions, and length of hospital stay in 34 639 surgical cases of lumbar disc herniation.”	The bone & joint journal vol. 101-B,4 (2019): 470-477. doi:10.1302/0301-620X.101B4.BJJ-2018-1184.R1
51	30689304	Akuthota, Venu et al.	“Clinical Course of Motor Deficits from Lumbosacral Radiculopathy Due to Disk Herniation.”	PM & R : the journal of injury, function, and rehabilitation vol. 11,8 (2019): 807-814. doi:10.1002/pmrj.12082
52	30508967	Hua, Wenbin et al.	“Outcomes of discectomy by using full-endoscopic visualization technique via the interlaminar and transforaminal approaches in the treatment of L5-S1 disc herniation: An observational study.”	Medicine vol. 97,48 (2018): e13456. doi:10.1097/MD.0000000000013456
53	30269234	Lagerbäck, Tobias et al.	“Effectiveness of surgery for sciatica with disc herniation is not substantially affected by differences in surgical incidences among three countries: results from the Danish, Swedish and Norwegian spine registries.”	European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society vol. 28,11 (2019): 2562-2571. doi:10.1007/s00586-018-5768-9
54	29753167	Klessinger, Stephan.	“The frequency of re-surgery after lumbar disc Nucleoplasty in a ten-year period.”	Clinical neurology and neurosurgery vol. 170 (2018): 79-83. doi:10.1016/j.clineuro.2018.05.004
55	29703053	Hua, Wenbin et al.	“Full-endoscopic discectomy via the interlaminar approach for disc herniation at L4-L5 and L5-S1: An observational study.”	Medicine vol. 97,17 (2018): e0585. doi:10.1097/MD.0000000000010585
56	29475577	Segura-Trepichio, Manuel et al.	“Lumbar disc herniation surgery with microdiscectomy plus interspinous stabilization: Good clinical results, but failure to lower the incidence of re-operation.”	Journal of clinical neuroscience : official journal of the Neurosurgical Society of Australasia vol. 51 (2018): 29-34. doi:10.1016/j.jocn.2018.02.010
57	29454134	Owens, R Kirk 2nd et al.	“Back pain improves significantly following discectomy for lumbar disc herniation.”	The spine journal : official journal of the North American Spine Society vol. 18,9 (2018): 1632-1636. doi:10.1016/j.spinee.2018.02.014

58	29223520	Madsbu, Mattis A et al.	“Lumbar Microdiscectomy in Obese Patients: A Multicenter Observational Study.”	World neurosurgery vol. 110 (2018): e1004-e1010. doi:10.1016/j.wneu.2017.11.156
59	29045852	Madsbu, Mattis A et al.	“Surgery for Herniated Lumbar Disc in Daily Tobacco Smokers: A Multicenter Observational Study.”	World neurosurgery vol. 109 (2018): e581-e587. doi:10.1016/j.wneu.2017.10.024
60	28837531	O'Donnell, Jeffrey A et al.	“Preoperative Opioid Use is a Predictor of Poor Return to Work in Workers' Compensation Patients After Lumbar Diskectomy.”	Spine vol. 43,8 (2018): 594-602. doi:10.1097/BRS.0000000000002385
61	28735120	Debono, Bertrand et al.	“Outpatient Lumbar Microdiscectomy in France: From an Economic Imperative to a Clinical Standard-An Observational Study of 201 Cases.”	World neurosurgery vol. 106 (2017): 891-897. doi:10.1016/j.wneu.2017.07.065
62	28392347	Oba, Hiroki et al.	“Predictors of improvement in low back pain after lumbar decompression surgery: Prospective study of 140 patients.”	Journal of orthopaedic science : official journal of the Japanese Orthopaedic Association vol. 22,4 (2017): 641-646. doi:10.1016/j.jos.2017.03.011
63	28339437	Xie, Tian-Hang et al.	“Complications of Lumbar Disc Herniation Following Full-endoscopic Interlaminar Lumbar Discectomy: A Large, Single-Center, Retrospective Study.”	Pain physician vol. 20,3 (2017): E379-E387.
64	28241227	Madsbu, Mattis A et al.	“Surgery for Herniated Lumbar Disk in Individuals 65 Years of Age or Older: A Multicenter Observational Study.”	JAMA surgery vol. 152,5 (2017): 503-506. doi:10.1001/jamasurg.2016.5557
65	28091818	Gulati, Sasha et al.	“Lumbar microdiscectomy for sciatica in adolescents: a multicentre observational registry-based study.”	Acta neurochirurgica vol. 159,3 (2017): 509-516. doi:10.1007/s00701-017-3077-4
66	28072800	Li, Zhen-Zhou et al.	“Modified Percutaneous Lumbar Foraminoplasty and Percutaneous Endoscopic Lumbar Discectomy: Instrument Design, Technique Notes, and 5 Years Follow-up.”	Pain physician vol. 20,1 (2017): E85-E98.
67	28072799	Liu, Chao et al.	“Percutaneous Endoscopic Lumbar Discectomy for Highly Migrated Lumbar Disc Herniation.”	Pain physician vol. 20,1 (2017): E75-E84.
68	28004244	Tschugg, Anja et al.	“Gender differences after lumbar sequestrectomy: a prospective clinical trial using quantitative sensory testing.”	European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society vol. 26,3 (2017): 857-864.
69	27759631	Zhu, Ru-Sen et al.	“Does local lavage influence functional recovery during lumber discectomy of disc herniation?: One year's systematic follow-up of 410 patients.”	Medicine vol. 95,42 (2016): e5022. doi:10.1097/MD.0000000000005022
70	27495101	Qi, Lei et al.	“The clinical application of "jetting suture" technique in annular repair under microendoscopic discectomy: A	Medicine vol. 95,31 (2016): e4503. doi:10.1097/MD.0000000000004503

			prospective single-cohort observational study."	
71	27243810	Dorow, Marie et al.	"The Course of Pain Intensity in Patients Undergoing Herniated Disc Surgery: A 5-Year Longitudinal Observational Study."	PloS one vol. 11,5 e0156647. 31 May. 2016, doi:10.1371/journal.pone.0156647
72	27225576	Papić, Monika et al.	"Return to Work After Lumbar Microdiscectomy - Personalizing Approach Through Predictive Modelling."	Studies in health technology and informatics vol. 224 (2016): 181-3.
73	26882505	Shin, Joon-Shik et al.	"Long-Term Course of Alternative and Integrative Therapy for Lumbar Disc Herniation and Risk Factors for Surgery: A Prospective Observational 5-Year Follow-Up Study."	Spine vol. 41,16 (2016): E955-E963. doi:10.1097/BRS.0000000000001494
74	26849140	Elkan, P et al.	"Similar result after non-elective and elective surgery for lumbar disc herniation: an observational study based on the SweSpine register."	European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society vol. 25,5 (2016): 1460-1466. doi:10.1007/s00586-016-4419-2
75	26820687	Stienen, Martin N et al.	"Short- and long-term effects of smoking on pain and health-related quality of life after non-instrumented lumbar spine surgery."	Clinical neurology and neurosurgery vol. 142 (2016): 87-92. doi:10.1016/j.clineuro.2016.01.024
76	26815256	Choi, Kyung Chul et al.	"Percutaneous Endoscopic Lumbar Discectomy as an Alternative to Open Lumbar Microdiscectomy for Large Lumbar Disc Herniation."	Pain physician vol. 19,2 (2016): E291-300.
77	26512594	Ni, Jianqiang et al.	"Anterior Lumbar Interbody Fusion for Degenerative Discogenic Low Back Pain: Evaluation of L4-S1 Fusion."	Medicine vol. 94,43 (2015): e1851. doi:10.1097/MD.0000000000001851
78	26211851	Stienen, Martin N et al.	"Surgical Resident Education in Noninstrumented Lumbar Spine Surgery: A Prospective Observational Study with a 4.5-Year Follow-Up."	World neurosurgery vol. 84,6 (2015): 1589-97. doi:10.1016/j.wneu.2015.07.030
79	26143123	Pochon, L et al.	"Influence of gender on patient-oriented outcomes in spine surgery."	European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society vol. 25,1 (2016): 235-246. doi:10.1007/s00586-015-4062-3
80	26140406	Whitmore, Robert G et al.	"Predictive value of 3-month lumbar discectomy outcomes in the NeuroPoint-SD Registry."	Journal of neurosurgery. Spine vol. 23,4 (2015): 459-66. doi:10.3171/2015.1.SPINE14890
81	25962814	Elkan, P et al.	"Markers of inflammation and fibrinolysis in relation to outcome after surgery for lumbar disc herniation. A prospective study on 177 patients."	European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society vol. 25,1 (2016): 186-191. doi:10.1007/s00586-015-3998-7
82	25701544	Lagerbäck, Tobias et al.	"An observational study on the outcome after surgery for lumbar disc herniation in adolescents	The spine journal : official journal of the North American Spine Society vol. 15,6 (2015): 1241-7. doi:10.1016/j.spinee.2015.02.024

			compared with adults based on the Swedish Spine Register."	
83	25504102	Cheng, Jiwei et al.	"Posterolateral transforaminal selective endoscopic diskectomy with thermal annuloplasty for discogenic low back pain: a prospective observational study."	Spine vol. 39,26 Spec No. (2014): B60-5. doi:10.1097/BRS.0000000000000495
84	25240564	Crockett, M T et al.	"Ozone-augmented percutaneous discectomy: a novel treatment option for refractory discogenic sciatica."	Clinical radiology vol. 69,12 (2014): 1280-6. doi:10.1016/j.crad.2014.08.008
185	24063775	Majeed, Shiju A et al.	"Comparison of outcomes between conventional lumbar fenestration discectomy and minimally invasive lumbar discectomy: an observational study with a minimum 2-year follow-up."	Journal of orthopaedic surgery and research vol. 8 34. 24 Sep. 2013, doi:10.1186/1749-799X-8-34
86	28392347	Oba, Hiroki et al.	"Predictors of improvement in low back pain after lumbar decompression surgery: Prospective study of 140 patients."	Journal of orthopaedic science : official journal of the Japanese Orthopaedic Association vol. 22,4 (2017): 641-646. doi:10.1016/j.jos.2017.03.011
87	24010898	Ghogawala, Zoher et al.	"The efficacy of lumbar discectomy and single-level fusion for spondylolisthesis: results from the NeuroPoint-SD registry: clinical article."	Journal of neurosurgery. Spine vol. 19,5 (2013): 555-63. doi:10.3171/2013.7.SPINE1362
88	20515353	Bakhsh, Ahmed.	"Long-term outcome of lumbar disc surgery: an experience from Pakistan."	Journal of neurosurgery. Spine vol. 12,6 (2010): 666-70. doi:10.3171/2009.10.SPINE09142

Reference List 1: List of papers of observational studies included in Chapter 3.

Appendix D: List of papers of RCTs included in literature review in Chapter 2

#	Pubmed ID	Authors	Title	Journal
1	6857385	Weber, H.	"Lumbar disc herniation. A controlled, prospective study with ten years of observation."	Spine vol. 8,2 (1983): 131-40.
2	6339137	Ejeskär, A et al.	"Surgery versus chemonucleolysis for herniated lumbar discs. A prospective study with random assignment."	Clinical orthopaedics and related research ,174 (1983): 236-42.
3	1579871	Muralikuttan, K P et al.	"A prospective randomized trial of chemonucleolysis and conventional disc surgery in single level lumbar disc herniation."	Spine vol. 17,4 (1992): 381-7. doi:10.1097/00007632-199204000-00001
4	8434309	Revel, M et al.	"Automated percutaneous lumbar discectomy versus chemonucleolysis in the treatment of sciatica. A randomized multicenter trial."	Spine vol. 18,1 (1993): 1-7. doi:10.1097/00007632-199301000-00001
5	8434321	Tullberg, T et al.	"Does microscopic removal of lumbar disc herniation lead to better results than the standard procedure? Results of a one-year randomized study."	Spine vol. 18,1 (1993): 24-7. doi:10.1097/00007632-199301000-00005
6	8367786	Zdeblick, T A.	"A prospective, randomized study of lumbar fusion. Preliminary results."	Spine vol. 18,8 (1993): 983-91. doi:10.1097/00007632-199306150-00006
7	7604351	Chatterjee, S et al.	"Report of a controlled clinical trial comparing automated percutaneous lumbar discectomy and microdiscectomy in the treatment of contained lumbar disc herniation."	Spine vol. 20,6 (1995): 734-8. doi:10.1097/00007632-199503150-00016
8	8799541	Henriksen, L et al.	"A controlled study of microsurgical versus standard lumbar discectomy."	British journal of neurosurgery vol. 10,3 (1996): 289-93. doi:10.1080/02688699650040160
9	10428127	Hermantin, F U et al.	"A prospective, randomized study comparing the results of open discectomy with those of video-assisted arthroscopic microdiscectomy."	The Journal of bone and joint surgery. American volume vol. 81,7 (1999): 958-65. doi:10.2106/00004623-199907000-00008
10	10990391	Krugluger, J, and K Knahr.	"Chemonucleolysis and automated percutaneous discectomy--a prospective randomized comparison."	International orthopaedics vol. 24,3 (2000): 167-9. doi:10.1007/s002640000139

11	11550835	Bernsmann, K et al.	"Lumbar micro disc surgery with and without autologous fat graft. A prospective randomized trial evaluated with reference to clinical and social factors."	Archives of orthopaedic and trauma surgery vol. 121,8 (2001): 476-80. doi:10.1007/s004020100277
12	11795752	McAfee, Paul C et al.	"Anterior BAK instrumentation and fusion: complete versus partial discectomy."	Clinical orthopaedics and related research ,394 (2002): 55-63. doi:10.1097/00003086-200201000-00007
13	12163718	Gibson, Suzy et al.	"Allograft versus autograft in instrumented posterolateral lumbar spinal fusion: a randomized control trial."	Spine vol. 27,15 (2002): 1599-603. doi:10.1097/00007632-200208010-00002
14	12217670	Haines, Stephen J et al.	"Discectomy strategies for lumbar disc herniation: results of the LAPDOG trial."	Journal of clinical neuroscience : official journal of the Neurosurgical Society of Australasia vol. 9,4 (2002): 411-7. doi:10.1054/jocn.2002.1120
15	12438990	Burkus, J Kenneth et al.	"Clinical and radiographic outcomes of anterior lumbar interbody fusion using recombinant human bone morphogenetic protein-2."	Spine vol. 27,21 (2002): 2396-408. doi:10.1097/00007632-200211010-00015
16	12461392	Boden, Scott D et al.	"Use of recombinant human bone morphogenetic protein-2 to achieve posterolateral lumbar spine fusion in humans: a prospective, randomized clinical pilot trial: 2002 Volvo Award in clinical studies."	Spine vol. 27,23 (2002): 2662-73. doi:10.1097/00007632-200212010-00005
17	12592544	Hägg, O et al.	"Predictors of outcome in fusion surgery for chronic low back pain. A report from the Swedish Lumbar Spine Study."	European spine journal: official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society vol. 12,1 (2003): 22-33. doi:10.1007/s00586-002-0465-z
18	12902960	McAfee, Paul C et al.	"SB Charité disc replacement: report of 60 prospective randomized cases in a US center."	Journal of spinal disorders & techniques vol. 16,4 (2003): 424-33. doi:10.1097/00024720-200308000-00016
19	14560188	Delamarter, Rick B et al.	"ProDisc artificial total lumbar disc replacement: introduction and early results from the United States clinical trial."	Spine vol. 28,20 (2003): S167-75. doi:10.1097/01.BRS.0000092220.66650.2B
20	14722400	Sasso, Rick C et al.	"A prospective, randomized controlled clinical trial of anterior lumbar	Spine vol. 29,2 (2004): 113-22; discussion 121-2. doi:10.1097/01.BRS.0000107007.31714.77

			interbody fusion using a titanium cylindrical threaded fusion device."	
21	15069129	Buttermann, Glenn R.	"Treatment of lumbar disc herniation: epidural steroid injection compared with discectomy. A prospective, randomized study."	The Journal of bone and joint surgery. American volume vol. 86,4 (2004): 670-9.
22	15830983	Reverberi, C et al.	"Disc coablation and epidural injection of steroids: a comparison of strategies in the treatment of mechanical spinal discogenic pain."	Acta neurochirurgica. Supplement vol. 92 (2005): 127-8. doi:10.1007/3-211-27458-8_27
23	16025024	Blumenthal, Scott et al.	"A prospective, randomized, multicenter Food and Drug Administration investigational device exemptions study of lumbar total disc replacement with the CHARITE artificial disc versus lumbar fusion: part I: evaluation of clinical outcomes."	Spine vol. 30,14 (2005): 1565-75; discussion E387-91. doi:10.1097/01.brs.0000170587.32676.0e
24	16826006	Katayama, Yoshito et al.	"Comparison of surgical outcomes between macro discectomy and micro discectomy for lumbar disc herniation: a prospective randomized study with surgery performed by the same spine surgeon."	Journal of spinal disorders & techniques vol. 19,5 (2006): 344-7. doi:10.1097/01.bsd.0000211201.93125.1c
25	17023847	Osterman, Heikki et al.	"Effectiveness of microdiscectomy for lumbar disc herniation: a randomized controlled trial with 2 years of follow-up."	Spine vol. 31,21 (2006): 2409-14. doi:10.1097/01.brs.0000239178.08796.52
26	17108817	Hoogland, Thomas et al.	"Transforaminal posterolateral endoscopic discectomy with or without the combination of a low-dose chymopapain: a prospective randomized study in 280 consecutive cases."	Spine vol. 31,24 (2006): E890-7. doi:10.1097/01.brs.0000245955.22358.3a
27	17119140	Weinstein, James N et al.	"Surgical vs nonoperative treatment for lumbar disk herniation: the Spine Patient Outcomes Research Trial (SPORT): a randomized trial."	JAMA vol. 296,20 (2006): 2441-50. doi:10.1001/jama.296.20.2441
28	17538084	Peul, Wilco C et al.	"Surgery versus prolonged conservative treatment for sciatica."	The New England journal of medicine vol. 356,22 (2007): 2245-56. doi:10.1056/NEJMoa064039
29	17545903	Hallett, Alison et al.	"Foraminal stenosis and single-level degenerative disc disease: a	Spine vol. 32,13 (2007): 1375-80. doi:10.1097/BRS.0b013e318064520f

			randomized controlled trial comparing decompression with decompression and instrumented fusion."	
30	17621025	Xie, Jing-cheng, and R John Hurlbert.	"Discectomy versus discectomy with fusion versus discectomy with fusion and instrumentation: a prospective randomized study."	Neurosurgery vol. 61,1 (2007): 107-16; discussion 116-7. doi:10.1227/01.neu.0000279730.44016.da
31	17881967	Righesso, Orlando et al.	"Comparison of open discectomy with microendoscopic discectomy in lumbar disc herniations: results of a randomized controlled trial."	Neurosurgery vol. 61,3 (2007): 545-9; discussion 549. doi:10.1227/01.NEU.0000290901.00320.F5
32	18300905	Ryang, Yu-Mi et al.	"Standard open microdiscectomy versus minimal access trocar microdiscectomy: results of a prospective randomized study."	Neurosurgery vol. 62,1 (2008): 174-81; discussion 181-2. doi:10.1227/01.NEU.0000311075.56486.C5
33	18427312	Ruetten, Sebastian et al.	"Full-endoscopic interlaminar and transforaminal lumbar discectomy versus conventional microsurgical technique: a prospective, randomized, controlled study."	Spine vol. 33,9 (2008): 931-9. doi:10.1097/BRS.0b013e31816c8af7
34	18502911	Peul, Wilco C et al.	"Prolonged conservative care versus early surgery in patients with sciatica caused by lumbar disc herniation: two year results of a randomised controlled trial."	BMJ (Clinical research ed.) vol. 336,7657 (2008): 1355-8. doi:10.1136/bmj.a143
35	19360440	Franke, Jörg et al.	"Comparison of a minimally invasive procedure versus standard microscopic discotomy: a prospective randomised controlled clinical trial."	European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society vol. 18,7 (2009): 992-1000. doi:10.1007/s00586-009-0964-2
36	19506919	Berg, Svante et al.	"Total disc replacement compared to lumbar fusion: a randomised controlled trial with 2-year follow-up."	European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society vol. 18,10 (2009): 1512-9. doi:10.1007/s00586-009-1047-0
37	19584344	Arts, Mark P et al.	"Tubular diskectomy vs conventional microdiskectomy for sciatica: a randomized controlled trial."	JAMA vol. 302,2 (2009): 149-58. doi:10.1001/jama.2009.972
38	19819762	Berg, Svante et al.	"Sex life and sexual function in men and women before and after total disc	The spine journal : official journal of the North American Spine Society vol. 9,12

			replacement compared with posterior lumbar fusion."	(2009): 987-94. doi:10.1016/j.spinee.2009.08.454
39	20556439	Arts, Mark et al.	"Does minimally invasive lumbar disc surgery result in less muscle injury than conventional surgery? A randomized controlled trial."	European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society vol. 20,1 (2011): 51-7. doi:10.1007/s00586-010-1482-y
40	21036279	McMorland, Gordon et al.	"Manipulation or microdiskectomy for sciatica? A prospective randomized clinical study."	Journal of manipulative and physiological therapeutics vol. 33,8 (2010): 576-84. doi:10.1016/j.jmpt.2010.08.013
41	21519072	Garg, Bhavuk et al.	"Microendoscopic versus open discectomy for lumbar disc herniation: a prospective randomised study."	Journal of orthopaedic surgery (Hong Kong) vol. 19,1 (2011): 30-4. doi:10.1177/230949901101900107
42	21613439	Erginousakis, Dimitrios et al.	"Comparative prospective randomized study comparing conservative treatment and percutaneous disk decompression for treatment of intervertebral disk herniation."	Radiology vol. 260,2 (2011): 487-93. doi:10.1148/radiol.11101094
43	23609203	Wardlaw, Douglas et al.	"Prospective randomized trial of chemonucleolysis compared with surgery for soft disc herniation with 1-year, intermediate, and long-term outcome: part I: the clinical outcome."	Spine vol. 38,17 (2013): E1051-7. doi:10.1097/BRS.0b013e31829729b3
44	23764765	Rodríguez-Vela, Javier et al.	"Clinical outcomes of minimally invasive versus open approach for one-level transforaminal lumbar interbody fusion at the 3- to 4-year follow-up."	European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society vol. 22,12 (2013): 2857-63. doi:10.1007/s00586-013-2853-y
45	24136677	Gempt, J et al.	"Long-term follow-up of standard microdiscectomy versus minimal access surgery for lumbar disc herniations."	Acta neurochirurgica vol. 155,12 (2013): 2333-8. doi:10.1007/s00701-013-1901-z
46	24153171	Lurie, Jon D et al.	"Surgical versus nonoperative treatment for lumbar disc herniation: eight-year results for the spine patient outcomes research trial."	Spine vol. 39,1 (2014): 3-16. doi:10.1097/BRS.0000000000000088
47	24736930	Hussein, Mohamed et al.	"Surgical technique and effectiveness of microendoscopic discectomy for large uncontained lumbar disc herniations: a prospective,	European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine

			randomized, controlled study with 8 years of follow-up."	Research Society vol. 23,9 (2014): 1992-9. doi:10.1007/s00586-014-3296-9
48	25614151	Brouwer, Patrick A et al.	"Percutaneous laser disc decompression versus conventional microdiscectomy in sciatica: a randomized controlled trial."	The spine journal : official journal of the North American Spine Society vol. 15,5 (2015): 857-65. doi:10.1016/j.spinee.2015.01.020
49	26851686	Nikoobakht, Mehdi et al.	"Plasma disc decompression compared to physiotherapy for symptomatic contained lumbar disc herniation: A prospective randomized controlled trial."	Neurologia i neurochirurgia polska vol. 50,1 (2016): 24-30. doi:10.1016/j.pjnns.2015.11.001
50	26887645	Pan, Zhimin et al.	"Efficacy of Transforaminal Endoscopic Spine System (TESSYS) Technique in Treating Lumbar Disc Herniation."	Medical science monitor : international medical journal of experimental and clinical research vol. 22 530-9. 18 Feb. 2016, doi:10.12659/msm.894870
51	26898494	Belykh, Evgenii et al.	"Prospective Comparison of Microsurgical, Tubular-Based Endoscopic, and Endoscopically Assisted Diskectomies: Clinical Effectiveness and Complications in Railway Workers."	World neurosurgery vol. 90 (2016): 273-280. doi:10.1016/j.wneu.2016.02.047
52	27276397	Cristante, Alexandre Fogaça et al.	"Randomized clinical trial comparing lumbar percutaneous hydrodiscectomy with lumbar open microdiscectomy for the treatment of lumbar disc protrusions and herniations."	Clinics (Sao Paulo, Brazil) vol. 71,5 (2016): 276-80. doi:10.6061/clinics/2016(05)06
53	27704286	Krappel, Ferdinand et al.	"Herniectomy versus herniectomy with the DIAM spinal stabilization system in patients with sciatica and concomitant low back pain: results of a prospective randomized controlled multicenter trial."	European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society vol. 26,3 (2017): 865-876. doi:10.1007/s00586-016-4796-6
54	27673378	Hussein, Mohamed.	"Minimal Incision, Multifidus-sparing Microendoscopic Diskectomy Versus Conventional Microdiskectomy for Highly Migrated Intracanal Lumbar Disk Herniations."	The Journal of the American Academy of Orthopaedic Surgeons vol. 24,11 (2016): 805-813. doi:10.5435/JAAOS-D-15-00588
55	27885470	Gibson, J N Alaistair et al.	"A randomised controlled trial of transforaminal endoscopic discectomy vs microdiscectomy."	European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine

				Research Society vol. 26,3 (2017): 847-856. doi:10.1007/s00586-016-4885-6
56	28454511	Brouwer, Patrick A et al.	"Percutaneous laser disc decompression versus conventional microdiscectomy for patients with sciatica: Two-year results of a randomised controlled trial."	Interventional neuroradiology : journal of peritherapeutic neuroradiology, surgical procedures and related neurosciences vol. 23,3 (2017): 313-324. doi:10.1177/1591019917699981
57	28550071	Overdevest, Gijsbert M et al.	"Tubular discectomy versus conventional microdiscectomy for the treatment of lumbar disc herniation: long-term results of a randomised controlled trial."	Journal of neurology, neurosurgery, and psychiatry vol. 88,12 (2017): 1008-1016. doi:10.1136/jnnp-2016-315306
58	30423641	Gu, Honglin et al.	"Efficacy of the Wallis interspinous implant for primary lumbar disc herniation : a prospective randomised controlled trial."	Acta orthopaedica Belgica vol. 83,3 (2017): 405-415.
59	29149145	Park, Chang Hong, and Sang Ho Lee.	"Endoscopic Epidural Laser Decompression Versus Transforaminal Epiduroscopic Laser Annuloplasty for Lumbar Disc Herniation: A Prospective, Randomized Trial."	Pain physician vol. 20,7 (2017): 663-670.
60	30076437	Kong, Lei et al.	"Percutaneous endoscopic lumbar discectomy and microsurgical laminotomy : A prospective, randomized controlled trial of patients with lumbar disc herniation and lateral recess stenosis."	Der Orthopade vol. 48,2 (2019): 157-164. doi:10.1007/s00132-018-3610-z
61	31703056	Chen, Zihao et al.	"Percutaneous Transforaminal Endoscopic Discectomy Versus Microendoscopic Discectomy for Lumbar Disc Herniation: Two-Year Results of a Randomized Controlled Trial."	Spine vol. 45,8 (2020): 493-503. doi:10.1097/BRS.0000000000003314
62	31852061	Yadav, Ram Ishwar et al.	"Comparison of the effectiveness and outcome of microendoscopic and open discectomy in patients suffering from lumbar disc herniation."	Medicine vol. 98,50 (2019): e16627. doi:10.1097/MD.0000000000016627
63	32187469	Bailey, Chris S et al.	"Surgery versus Conservative Care for Persistent Sciatica Lasting 4 to 12 Months."	The New England journal of medicine vol. 382,12 (2020): 1093-1102. doi:10.1056/NEJMoa1912658

64	32215135	Gao, Xiang et al.	"Efficacy Analysis of Percutaneous Endoscopic Lumbar Discectomy Combined with PEEK Rods for Giant Lumbar Disc Herniation: A Randomized Controlled Study."	Pain research & management vol. 2020 3401605. 10 Mar. 2020, doi:10.1155/2020/3401605
65	32539752	Hamawandi, Sherwan A et al.	"Open fenestration discectomy versus microscopic fenestration discectomy for lumbar disc herniation: a randomized controlled trial."	BMC musculoskeletal disorders vol. 21,1 384. 15 Jun. 2020, doi:10.1186/s12891-020-03396-x
66	33474962	Li, Jizheng et al.	"A novel full endoscopic annular repair technique combined with autologous conditioned plasma intradiscal injection: a new safe serial therapeutic model for the treatment of lumbar disc herniation."	Annals of palliative medicine vol. 10,1 (2021): 292-301. doi:10.21037/apm-20-2257
67	33871219	Hadžić, Ermin et al.	"Comparison of early and delayed lumbar disc herniation surgery and the treatment outcome."	Medicinski glasnik : official publication of the Medical Association of Zenica-Doboj Canton, Bosnia and Herzegovina vol. 18,2 (2021): 456-462. doi:10.17392/1343-21
68	34488362	Zhou, Fei et al.	"Clinical effect of TESSYS technique under spinal endoscopy combined with drug therapy in patients with lumbar disc herniation and its effect on quality of life and serum inflammatory factors: results of a randomized trial."	Annals of palliative medicine vol. 10,8 (2021): 8728-8736. doi:10.21037/apm-21-1282
69	34866340	Hamawandi, Sherwan A et al.	"Effect of Duration of Symptoms on the Clinical and Functional Outcomes of Lumbar Microdiscectomy: A Randomized Controlled Trial."	Orthopaedic surgery vol. 14,1 (2022): 157-168. doi:10.1111/os.13114
70	34896609	Kelekis, Alexis et al.	"Intradiscal oxygen-ozone chemonucleolysis versus microdiscectomy for lumbar disc herniation radiculopathy: a non-inferiority randomized control trial."	The spine journal : official journal of the North American Spine Society vol. 22,6 (2022): 895-909. doi:10.1016/j.spinee.2021.11.017
71	34929381	Fasoli, Fabrizio et al.	"Minimally-invasive percutaneous treatments for low back pain and leg pain: a randomized controlled study of thermal disc decompression versus mechanical percutaneous disc decompression."	The spine journal : official journal of the North American Spine Society vol. 22,5 (2022): 709-715.

72	35190388	Gadjradj, Pravesh S et al.	"Full endoscopic versus open discectomy for sciatica: randomised controlled non-inferiority trial."	BMJ (Clinical research ed.) vol. 376 e065846. 21 Feb. 2022, doi:10.1136/bmj-2021-065846
73	12973134	Brox, Jens Ivar et al.	"Randomized clinical trial of lumbar instrumented fusion and cognitive intervention and exercises in patients with chronic low back pain and disc degeneration."	Spine vol. 28,17 (2003): 1913-21. doi:10.1097/01.BRS.0000083234.62751.7A
74	33969319	Wilby, Martin John et al.	"Surgical microdiscectomy versus transforaminal epidural steroid injection in patients with sciatica secondary to herniated lumbar disc (NERVES): a phase 3, multicentre, open-label, randomised controlled trial and economic evaluation."	The Lancet. Rheumatology vol. 3,5 e347-e356. 18 Mar. 2021, doi:10.1016/S2665-9913(21)00036-9

Reference List 2: List of Papers of RCTs included in Chapter3.

Appendix E: Associations between patient covariates in the Spine Tango data set

1) Sex vs other patient covariates

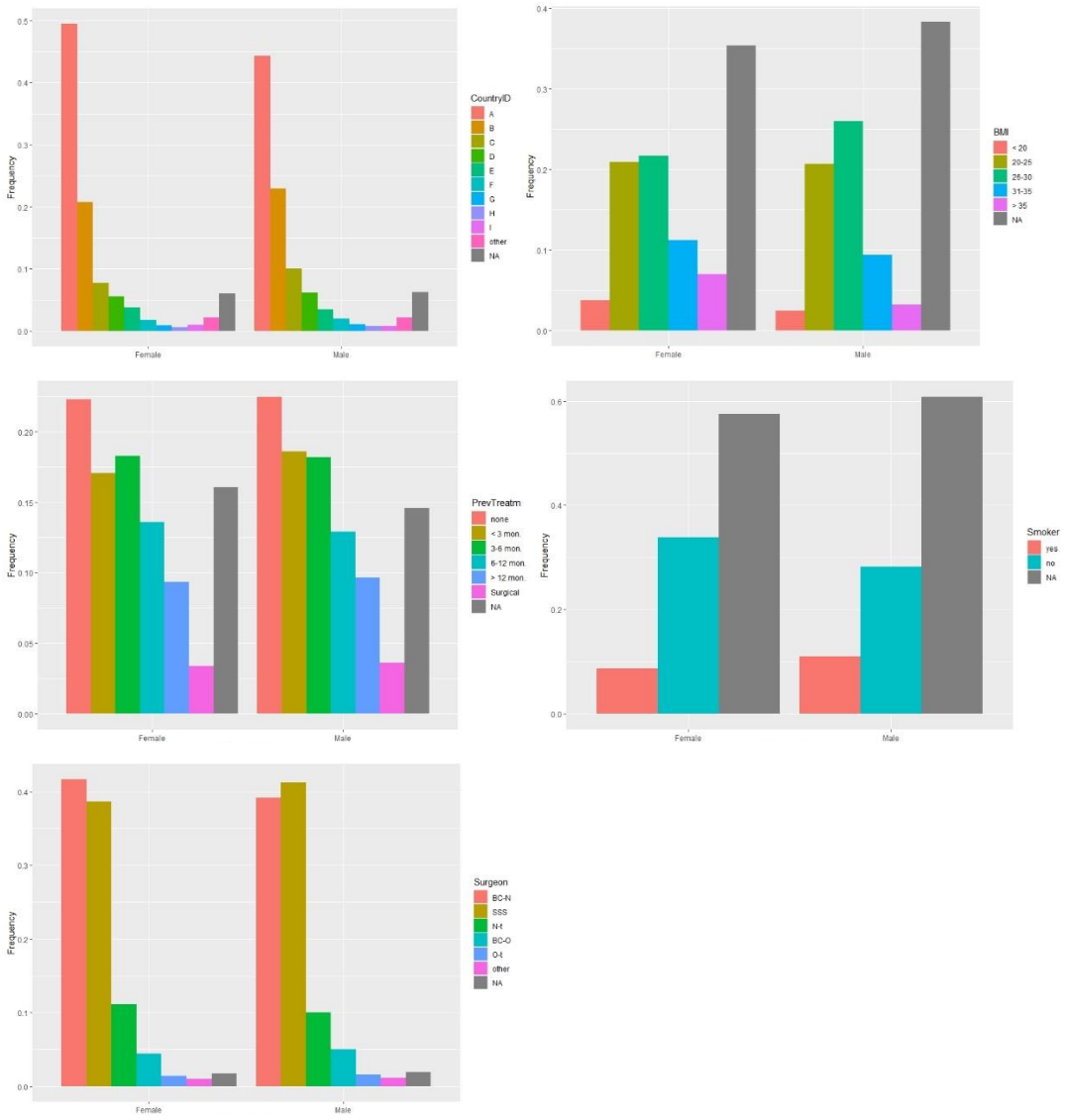
Combination of categorical and continuous variables (KS-test)

Sex - Age: KS-test: 0.331

Combinations of two categorical variables (Chi-Square test)

Second variable	Chi-square test p-value
Surgeon Type	**0.019**
Country	**0.002**
Level of Spine	0.316
Previous Treatment	**0.041**
Smoking Status	**0.024**
BMI	**0.013**
Morbidity	0.076
Complication	0.294
COMI available	0.316

Plots of frequencies for the significant Chi-square results



Although there were significant p-values for the Chi-square tests of these variable combinations, these plots of frequencies of the sub-categories vs the sub-categories of the other variable did not show systematic dependencies. Notably, it seems that a BMI of 25-30 and a smoking status "yes" was slightly more frequent in males than in females.

2) Age vs other patient covariates

Combination of categorical and continuous variables (KS-test)

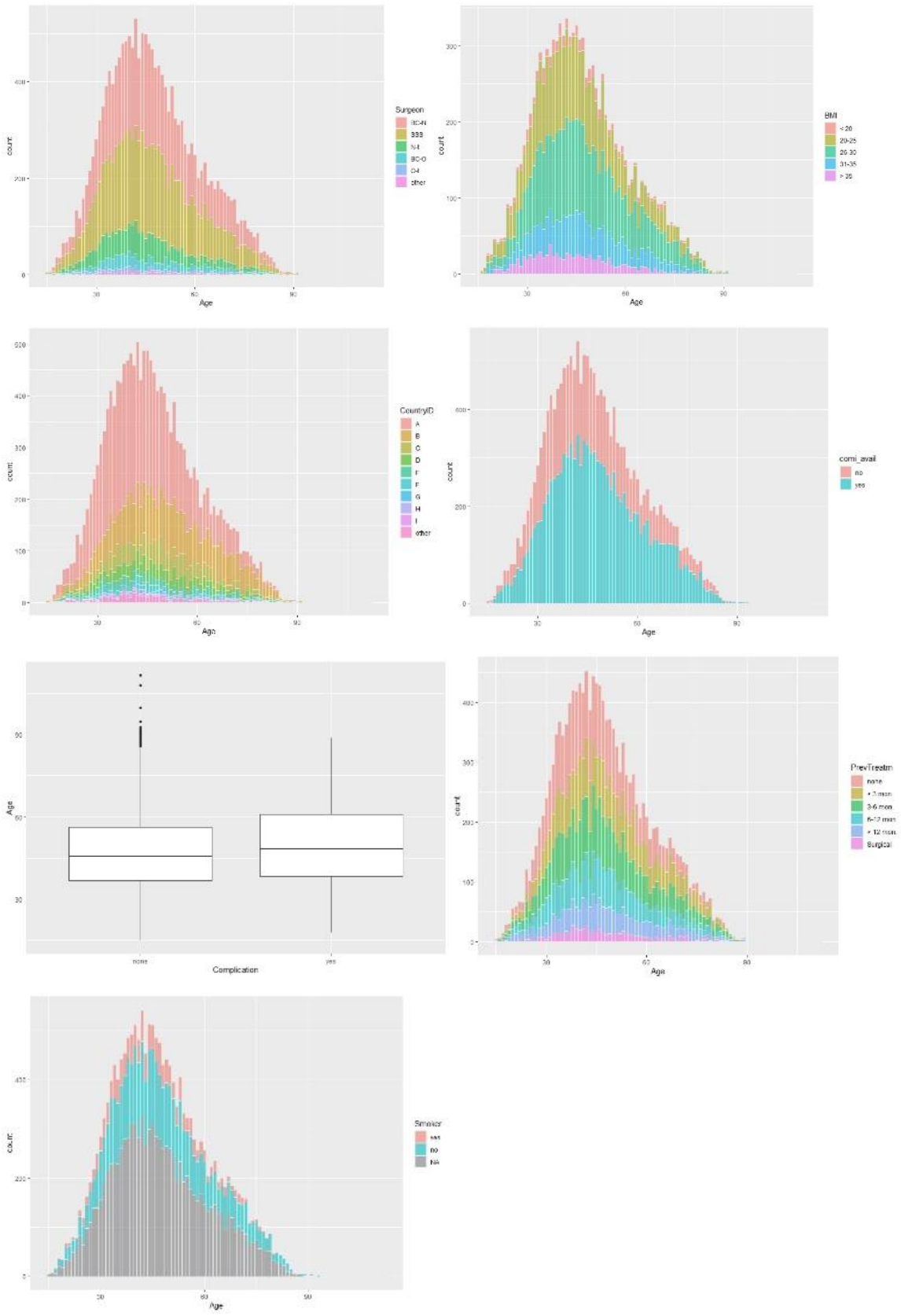
Second variable		KS-test test p-value
Surgeon Credentials	BC-N - SSS	0.940
	BC-N - N-t	0.215
	BC-N - BC-O	0.006
	BC-N - O-t	0.052
	BC-N – Other	<0.001
	SSS - N-t	0.940
	SSS - BC-O	0.052

	SSS - O-t	0.101
	SSS – Other	<0.001
	N-t - BC-O	0.795
	N-t - O-t	0.001
	N-t – Other	0.293
	BC-O - O-t	0.049
	BC-O – Other	0.027
	O-t – Other	0.810
Country ID	A - B	0.002
	A - C	0.027
	A - D	0.024
	A - E	<0.001
	A - F	<0.001
	A - G	0.007
	A - H	0.036
	A - I	0.811
	A - Other	0.022
	B - C	0.374
	B - D	0.585
	B - E	0.046
	B - F	0.004
	B - G	0.001
	B - H	0.009
	B - I	0.007
	B – Other	<0.001
	C - D	<0.001
	C - E	0.140
	C - F	0.001
	C - G	<0.001
	C - H	0.618
	C - I	0.033
	C – Other	0.038
	D - E	0.001
	D - F	0.201
	D - G	0.001
	D - H	0.342
	D - I	0.117
	D – Other	0.431
	E - F	0.001
	E - G	0.006
	E - H	0.048
	E - I	0.304
	E – Other	0.771
	F - G	0.002
	F - H	<0.001
	F - I	0.882
	F – Other	0.129
	G - H	0.250
	G - I	0.034
	G – Other	0.198
	H - I	<0.001

	H – Other	0.3246
	I - Other	0.7611
Previous Treatment	None - <6mon.	0.197
	None - >6mon.	0.164
	None – Surgical	0.105
	<6mon. - >6mon.	0.618
	<6mon. - Surgical	0.078
	>6mon. - Surgical	0.381
Level of Spine	L5/S1 – L4/L5	<0.001
	L5/S1 - L3/L4	<0.001
	L5/S1 - L2/L3	<0.001
	L5/S1 – L1/L2	<0.001
	L5/S1 – Other	<0.001
	L4/L5 - L3/L4	<0.001
	L4/L5 - L2/L3	<0.001
	L4/L5 – L1/L2	0.0339
	L4/L5 – Other	0.1222
	L3/L4 - L2/L3	<0.001
	L3/L4 – L1/L2	0.6753
	L3/L4 – Other	<0.001
	L2/L3 – L1/L2	0.02293
	L2/L3 – Other	<0.001
	L1/L2 - Other	0.102
Smoker	No - Yes	0.013
BMI	<20 vs – 20-25	0.1778
	<20 vs – 25-30	<0.001
	<20 vs – 30-35	<0.001
	<20 vs – >35	0.614
	20-25 vs – 25-30	<0.001
	20-25 vs – 30-35	<0.001
	20-25 vs – >35	0.005
	25-30 vs – 30-35	0.519
	35-30 vs – >35	<0.001
	35-30 vs >35	<0.001
COMI available	Yes – No	0.076
Complication	No - Yes	<0.001

Although there are many significant p-values detected, the distribution of age in the sub-categories did not show systematic dependencies between age and the respective paired variable.

Histograms for age distribution colored by subgroups

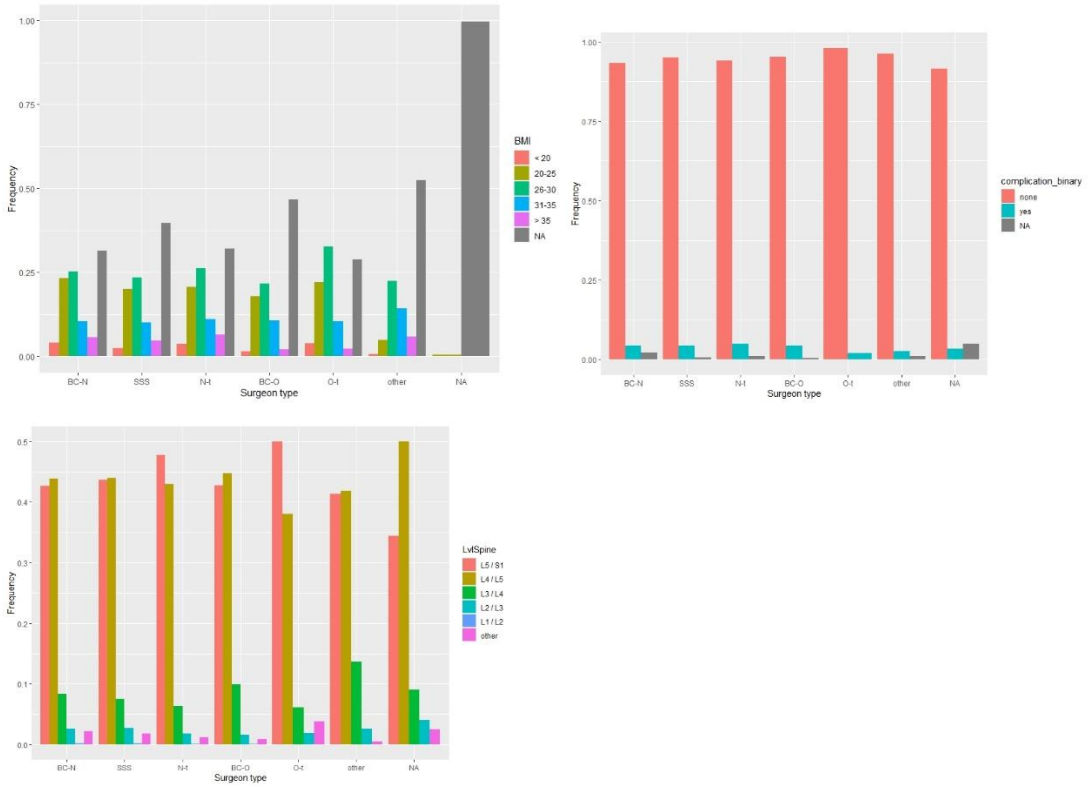


3) Surgeon credentials vs other variables:

<u>Chi-square results</u>

Second variable	Chi-square test p-value
Level of Spine	0.041
Smoking Status	0.248
BMI	0.173
Morbidity	0.094
Complication	0.037
COMI available	0.316

Plots of frequencies for the significant Chi-square results



Although there were significant p-values for the Chi-square tests of these variable combinations, these plots of frequencies of the sub-categories vs the sub-categories of the other variable did not show systematic dependencies. Notably, L4/L5 was more frequent than L5/S1 for the category of surgeon type "NA" and less frequent in the category of orthopaedic surgeons in training.

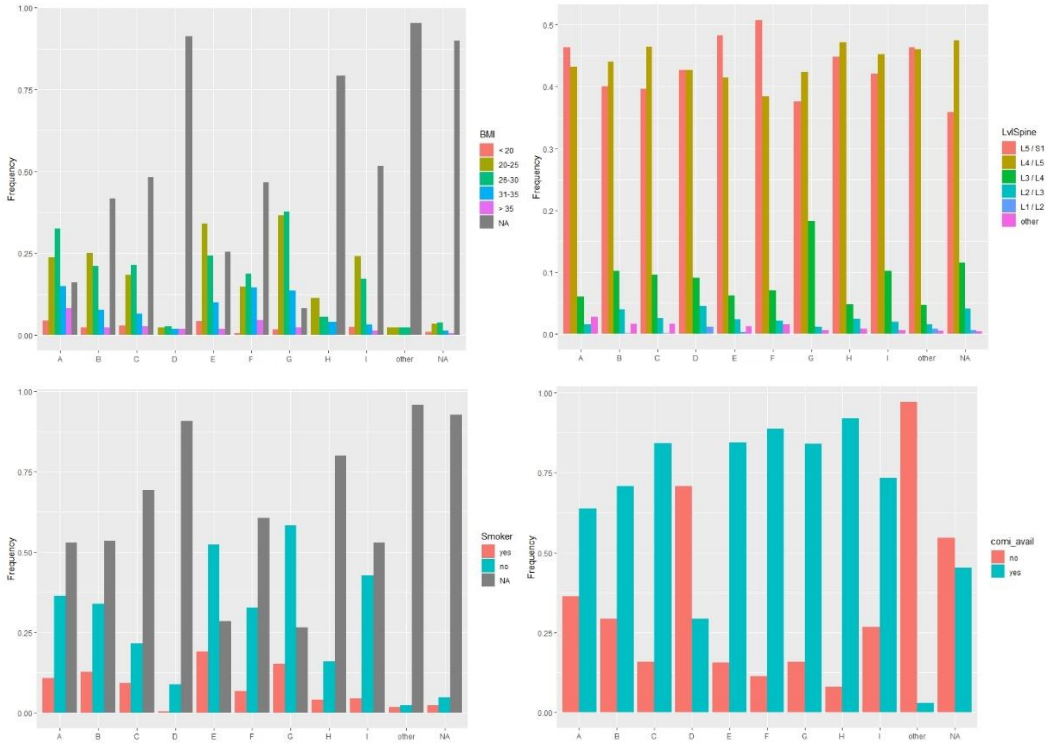
4) Country ID vs other variables

Chi-square results

Second variable	Chi-square test p-value
Level of Spine	<0.001

Smoking Status	<0.001
BMI	<0.001
Complication	0.081
COMI available	<0.001

Plots of frequencies for the significant Chi-square results



Although there were significant p-values for the Chi-square tests of these variable combinations, these plots of frequencies of the sub-categories vs the sub-categories of the other variable did not show systematic dependencies. Noteworthy were country D, G and "other", for which these frequencies could differ. However, these were the categories with low count of patients.

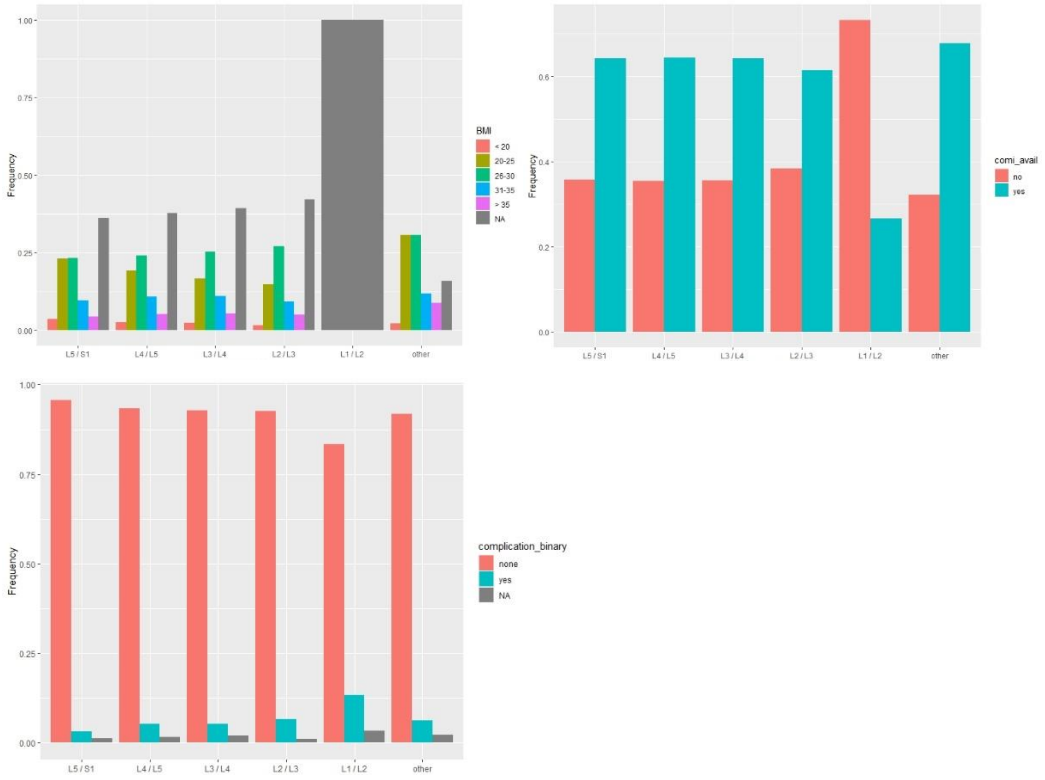
5) Level of Spine vs other variables

Chi-square results

Second variable	Chi-square test p-value
Previous Treatment	0.083
Smoking Status	0.263
BMI	<0.001
Morbidity	0.067

Complication	<0.001
COMI available	<0.001

Plots of frequencies for the significant Chi-square results



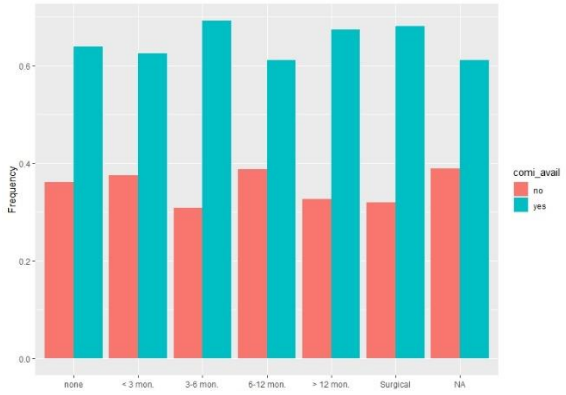
Although there were significant p-values for the Chi-square tests of these variable combinations, these plots of frequencies of the sub-categories vs the sub-categories of the other variable did not show systematic dependencies. Notably, patients that had surgery at the L1/L2 level had less often COMI available and a slightly higher rate for complications.

6) Previous Treatment vs other variables

Chi-square results

Second variable	Chi-square test p-value
Smoking Status	0.171
BMI	0.106
Morbidity	
COMI available	<0.001

Plots of frequencies for the significant Chi-square results



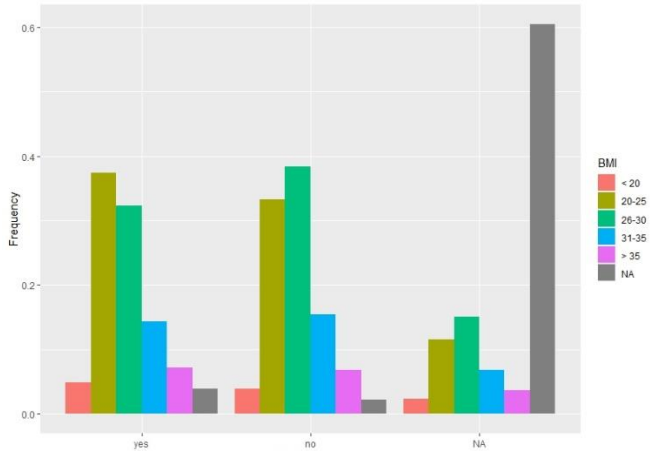
Although there were significant p-values for the Chi-square tests of these variable combinations, these plots of frequencies of the sub-categories vs the sub-categories of the other variable did not show systematic dependencies.

7) Smoking status vs other variables

Chi-square results

Second variable	Chi-square test p-value
BMI	
Morbidity	0.174
Complication	0.318
COMI available	0.291

Plots of frequencies for the significant Chi-square results



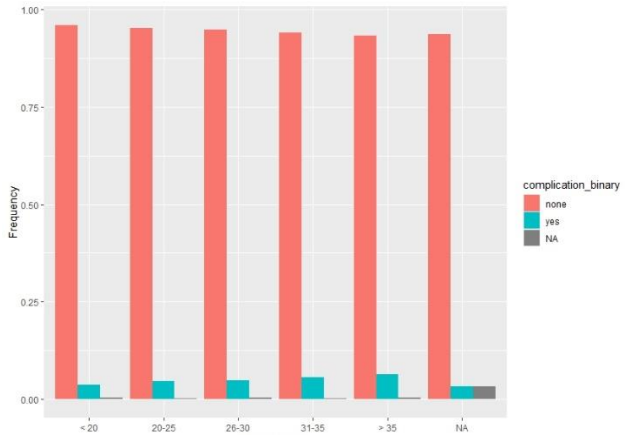
Although there were significant p-values for the Chi-square tests of these variable combinations, these plots of frequencies of the sub-categories vs the sub-categories of the other variable did not show systematic dependencies.

8) BMI vs other variables

Chi-square results

Second variable	Chi-square test p-value
Complication	0.028
COMI available	0.431

Plots of frequencies for the significant Chi-square results



Although there were significant p-values for the Chi-square tests of these variable combinations, these plots of frequencies of the sub-categories vs the sub-categories of the other variable did not show systematic dependencies.

9) ASA Morbidity status vs other variables

Chi-square results

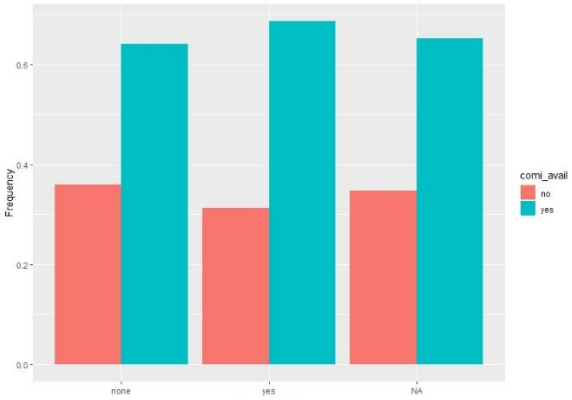
Second variable	Chi-square test p-value
Complication	0.068
COMI available	0.080

10) Complication vs COMI availability

Chi-square results

Second variable	Chi-square test p-value
COMI available	0.010

Plots of frequencies for the significant Chi-square results



Although there were significant p-values for the Chi-square tests of these variable combinations, these plots of frequencies of the sub-categories vs the sub-categories of the other variable did not show systematic dependencies.

Appendix F: Results of simulations with dataset of patients with complete data

All simulations in this Appendix part are programmed the same way as the simulations in Chapter 4, the only difference being the sample size of patients. The dataset used here, is the complete case analysis (CCA) of patients that had no missingness in any covariates, which were 4,312. This was done as a sensitivity analysis, to support the results from Chapter 4, that used a larger sample of patients, but with imputed covariates.

Questionnaires at baseline for missing data underlying MCAR mechanism

The results regarding RMSE are shown in Appendix Table 1 and 2. Each table shows each combination of scenarios for the RMSE using item-based imputation and score-based imputation.

Probability / Cut-off point	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	0.41	0.58	0.71	0.82	0.92	1.01	1.08	1.16	1.23
2	0.40	0.58	0.72	0.82	0.92	1.01	1.08	1.16	1.24
3	0.41	0.59	0.71	0.82	0.92	1.00	1.08	1.17	1.23
4	0.42	0.57	0.71	0.81	0.91	1.00	1.09	1.16	1.22
5	0.42	0.58	0.71	0.81	0.91	1.01	1.09	1.16	1.23
6	0.41	0.58	0.71	0.81	0.92	1.01	1.10	1.16	1.24
7	0.42	0.58	0.72	0.82	0.91	0.99	1.09	1.16	1.22
8	0.40	0.58	0.72	0.82	0.92	1.01	1.09	1.16	1.23

Appendix Table 1: Mean of RMSE of each combination of probability of amputed missingness and cut-off point, using item-based imputation method. RMSE was averaged over number of simulations (N=50).

Probability / Cut-off point	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	0.76	1.07	1.31	1.51	1.69	1.86	2.01	2.15	2.30
2	0.70	1.01	1.24	1.44	1.61	1.78	1.92	2.07	2.20
3	0.65	0.91	1.14	1.30	1.46	1.60	1.74	1.89	2.00
4	0.55	0.78	0.98	1.12	1.26	1.39	1.51	1.62	1.73
5	0.49	0.68	0.84	0.97	1.09	1.21	1.33	1.42	1.53
6	0.45	0.65	0.81	0.95	1.05	1.16	1.26	1.35	1.45
7	0.48	0.66	0.81	0.94	1.05	1.15	1.25	1.34	1.44
8	0.45	0.65	0.80	0.92	1.04	1.15	1.25	1.34	1.44

Appendix Table 2: Mean of RMSE of each combination of probability of amputed missingness and cut-off point, using score-based imputation method. RMSE was averaged over number of simulations (N=50).

Comparing these results with the results in Chapter 4, it can be observed that the values in the tables are nearly identical, only differing by a maximum margin of 0.02. Specifically, the root mean square errors (RMSEs) ranged from 0.40 to 1.24 for item-wise imputation, and from 0.45 to 2.30 for score-wise imputation in this simulation utilizing CCA. Notably, these ranges closely mirror the ranges

observed in Chapter 4, namely 0.41 to 1.25 for item-wise imputation, and 0.45 to 2.30 for score-wise imputation.

For each combination the population mean was calculated and presented in Appendix Table 3 and 4.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	7.67	7.67	7.67	7.67	7.66	7.67	7.66	7.68	7.67
2	7.67	7.67	7.66	7.67	7.67	7.68	7.67	7.67	7.67
3	7.66	7.66	7.67	7.67	7.66	7.67	7.68	7.67	7.68
4	7.67	7.67	7.67	7.67	7.66	7.67	7.68	7.68	7.68
5	7.67	7.67	7.67	7.67	7.67	7.67	7.66	7.67	7.68
6	7.66	7.67	7.67	7.67	7.67	7.67	7.68	7.68	7.67
7	7.67	7.67	7.67	7.68	7.67	7.67	7.66	7.67	7.67
8	7.67	7.67	7.67	7.67	7.67	7.67	7.67	7.68	7.67

Appendix Table 3: Estimated population mean of COMI baseline scores for each scenario, using item-based imputation.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	7.67	7.66	7.66	7.67	7.63	7.65	7.63	7.65	7.61
2	7.67	7.67	7.66	7.67	7.66	7.65	7.66	7.65	7.67
3	7.66	7.66	7.66	7.66	7.67	7.66	7.67	7.64	7.66
4	7.66	7.67	7.66	7.66	7.66	7.66	7.65	7.66	7.67
5	7.67	7.67	7.67	7.67	7.67	7.67	7.65	7.67	7.67
6	7.67	7.67	7.67	7.66	7.67	7.67	7.67	7.67	7.66
7	7.66	7.66	7.67	7.67	7.66	7.67	7.67	7.67	7.66
8	7.66	7.67	7.67	7.67	7.67	7.67	7.67	7.67	7.66

Appendix Table 4: Estimated population mean of COMI baseline scores for each scenario, using score-based imputation.

It appears that the population mean of COMI baseline scores was systematically estimated smaller than in the simulation done in Chapter 4 (by a small margin), however, on this sample size, the true population mean was smaller as well, with a mean and standard deviation of 7.66 (s.d. 1.74).

Questionnaires at baseline for missing data underlying MAR mechanism

The results regarding RMSE are shown in Appendix Table 5 and 6. Each table shows each combination of scenarios for the RMSE using item-based imputation and score-based imputation.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	0.41	0.57	0.71	0.81	0.94	0.99	1.08	1.18	1.25
2	0.41	0.57	0.69	0.82	0.92	0.99	1.09	1.15	1.23
3	0.42	0.57	0.72	0.82	0.92	1.00	1.09	1.15	1.23
4	0.40	0.57	0.71	0.83	0.90	1.01	1.07	1.17	1.24
5	0.42	0.58	0.71	0.82	0.91	1.00	1.08	1.16	1.22
6	0.40	0.57	0.71	0.83	0.93	1.01	1.09	1.17	1.22
7	0.39	0.57	0.71	0.83	0.93	1.02	1.11	1.17	1.23
8	0.41	0.58	0.70	0.82	0.91	0.99	1.09	1.17	1.24

Appendix Table 5: Mean of RMSE of each combination of probability of amputed missingness and cut-off point, using item-based imputation method. RMSE was averaged over number of simulations (N=50).

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	0.72	1.06	1.30	1.53	1.71	1.90	2.08	2.22	2.34
2	0.70	1.01	1.26	1.44	1.62	1.78	1.95	2.06	2.22
3	0.63	0.90	1.12	1.28	1.44	1.60	1.74	1.87	2.00
4	0.54	0.75	0.93	1.11	1.24	1.37	1.48	1.63	1.75
5	0.47	0.65	0.81	0.94	1.07	1.19	1.32	1.40	1.51
6	0.44	0.63	0.78	0.90	1.02	1.13	1.23	1.35	1.44
7	0.42	0.61	0.77	0.90	1.03	1.15	1.25	1.33	1.42
8	0.42	0.63	0.77	0.89	1.01	1.14	1.24	1.35	1.42

Appendix Table 6: Mean of RMSE of each combination of probability of amputed missingness and cut-off point, using score-based imputation method. RMSE was averaged over number of simulations (N=50).

The results for the MAR mechanism align with those observed for the MCAR mechanism. In Chapter 4, the RMSE ranges for item-wise imputation were 0.40 to 1.25, while in the CCA simulation, they were 0.39 to 1.25. For score-wise imputation, the ranges were 0.43 to 2.41 in Chapter 4 and 0.42 to 2.34 in the CCA simulation. Notably, the score-wise imputation exhibited slightly smaller RMSEs.

For each combination the population mean was calculated and presented in Appendix Table 7 and 8.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	7.67	7.66	7.67	7.67	7.66	7.67	7.68	7.67	7.67
2	7.66	7.67	7.67	7.67	7.67	7.66	7.66	7.66	7.68
3	7.66	7.67	7.66	7.67	7.66	7.65	7.67	7.68	7.67
4	7.66	7.67	7.67	7.67	7.66	7.66	7.67	7.66	7.66
5	7.67	7.67	7.66	7.67	7.68	7.67	7.67	7.67	7.67
6	7.66	7.66	7.67	7.68	7.66	7.67	7.67	7.67	7.69
7	7.66	7.66	7.66	7.67	7.66	7.66	7.68	7.67	7.66
8	7.67	7.67	7.67	7.66	7.68	7.67	7.67	7.66	7.67

Appendix Table 7: Estimated population mean of COMI baseline scores for each scenario, using item-based imputation.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	7.64	7.56	7.54	7.48	7.45	7.34	7.30	7.30	7.29
2	7.63	7.59	7.53	7.54	7.51	7.46	7.44	7.49	7.53
3	7.65	7.62	7.60	7.58	7.57	7.56	7.57	7.56	7.61
4	7.66	7.65	7.65	7.64	7.64	7.65	7.63	7.62	7.64
5	7.67	7.67	7.68	7.68	7.69	7.68	7.68	7.69	7.67
6	7.67	7.67	7.69	7.70	7.71	7.69	7.69	7.69	7.68
7	7.67	7.68	7.69	7.68	7.69	7.68	7.69	7.69	7.68
8	7.68	7.68	7.69	7.69	7.69	7.69	7.69	7.67	7.68

Appendix Table 8: Estimated population mean of COMI baseline scores for each scenario, using score-based imputation.

Again, considering the true population mean of this smaller sample size (7.66) the results exhibit a high degree of similarity, even mirroring the under-estimation of mean population COMI scores for high missingness and sensitive cut-off points, when imputing score-wise.

Questionnaires at baseline for missing data underlying MNAR mechanism

The results regarding RMSE are shown in Appendix Table 9 and 10. Each table shows each combination of scenarios for the RMSE using item-based imputation and score-based imputation.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	0.39	0.55	0.70	0.80	0.89	0.97	1.08	1.14	1.22
2	0.38	0.54	0.67	0.80	0.90	0.98	1.06	1.15	1.22
3	0.37	0.55	0.68	0.79	0.88	0.98	1.06	1.14	1.23
4	0.38	0.54	0.68	0.79	0.89	0.98	1.05	1.13	1.23
5	0.39	0.55	0.69	0.79	0.89	0.98	1.06	1.13	1.22
6	0.39	0.56	0.68	0.80	0.90	0.98	1.06	1.15	1.23
7	0.37	0.55	0.68	0.78	0.89	0.97	1.05	1.14	1.22
8	0.39	0.55	0.68	0.80	0.88	0.98	1.05	1.14	1.22

Appendix Table 9: Mean of RMSE of each combination of probability of amputed missingness and cut-off point, using item-based imputation method. RMSE was averaged over number of simulations (N=50).

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	0.71	1.06	1.30	1.56	1.76	1.96	2.21	2.37	2.60
2	0.67	0.99	1.25	1.45	1.68	1.84	2.01	2.18	2.33
3	0.62	0.89	1.09	1.32	1.47	1.63	1.76	1.92	2.03
4	0.53	0.75	0.92	1.09	1.21	1.36	1.48	1.59	1.72
5	0.43	0.63	0.79	0.94	1.06	1.16	1.27	1.38	1.50
6	0.40	0.59	0.75	0.86	0.97	1.10	1.19	1.30	1.40
7	0.40	0.58	0.73	0.85	0.98	1.09	1.19	1.30	1.40
8	0.42	0.58	0.73	0.86	0.97	1.10	1.18	1.30	1.40

Appendix Table 10: Mean of RMSE of each combination of probability of amputed missingness and cut-off point, using score-based imputation method. RMSE was averaged over number of simulations (N=50).

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	7.63	7.60	7.56	7.55	7.54	7.53	7.54	7.56	7.60
2	7.63	7.60	7.57	7.54	7.54	7.53	7.54	7.56	7.61
3	7.63	7.60	7.56	7.54	7.54	7.54	7.54	7.56	7.60
4	7.63	7.60	7.57	7.54	7.54	7.53	7.55	7.57	7.60
5	7.63	7.60	7.57	7.54	7.54	7.53	7.54	7.57	7.60
6	7.63	7.60	7.57	7.54	7.54	7.53	7.55	7.55	7.60
7	7.63	7.59	7.57	7.55	7.53	7.53	7.54	7.56	7.60
8	7.63	7.59	7.57	7.54	7.54	7.53	7.54	7.57	7.60

Appendix Table 11: Estimated population mean of COMI baseline scores for each scenario, using item-based imputation.

Cut-off point \ Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	7.58	7.47	7.39	7.25	7.17	7.05	6.87	6.83	6.69
2	7.59	7.50	7.41	7.32	7.23	7.16	7.10	7.10	7.17
3	7.60	7.52	7.47	7.39	7.36	7.33	7.32	7.32	7.47
4	7.62	7.57	7.53	7.49	7.48	7.46	7.48	7.51	7.57
5	7.63	7.60	7.58	7.55	7.54	7.54	7.55	7.58	7.61
6	7.64	7.62	7.59	7.58	7.58	7.58	7.58	7.60	7.63
7	7.64	7.61	7.60	7.59	7.57	7.57	7.58	7.60	7.63
8	7.64	7.62	7.60	7.58	7.59	7.58	7.59	7.61	7.64

Appendix Table 12: Estimated population mean of COMI baseline scores for each scenario, using score-based imputation.

Questionnaires at baseline for missing data underlying MNAR mechanism

In Appendix Table 13 the RMSEs are summarised for each mechanism of missingness and method of imputation.

Imputation type \ Probability		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
MCAR	item-based	0.77	1.09	1.33	1.52	1.71	1.89	1.99	2.18	2.27
	score-based	0.78	1.07	1.33	1.54	1.72	1.88	2.04	2.19	2.31
MAR	item-based	0.79	1.12	1.34	1.55	1.74	1.91	2.05	2.13	2.25
	score-based	0.79	1.11	1.32	1.53	1.72	1.87	2.02	2.19	2.26
MNAR	item-based	0.72	1.11	1.36	1.60	1.84	2.09	2.31	2.56	2.81
	score-based	0.74	1.10	1.35	1.61	1.86	2.10	2.29	2.58	2.96

Appendix Table 13: RMSEs of both imputation methods for all mechanisms of missingness. Columns are ordered regarding the probability of missingness (complete questionnaire missingness).

In Appendix Table 16 the estimated population means are summarised for each mechanism of missingness and method of imputation.

Imputation type \ Probability		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
MCAR	item-based	7.68	7.67	7.65	7.70	7.65	7.65	7.64	7.70	7.61
	score-based	7.67	7.67	7.66	7.66	7.65	7.63	7.58	7.64	7.57
MAR	item-based	7.67	7.65	7.67	7.68	7.67	7.68	7.64	7.66	7.56
	score-based	7.67	7.65	7.67	7.65	7.68	7.69	7.71	7.57	7.56
MNAR	item-based	7.56	7.43	7.30	7.17	7.00	6.83	6.64	6.43	6.21
	score-based	7.55	7.43	7.31	7.16	7.00	6.83	6.69	6.44	6.04

Appendix Table 14: Estimated population means of baseline COMI scores for both imputation methods and all mechanisms of missingness. Columns are ordered regarding the probability of missingness (complete questionnaire missingness).

Overall, the same patterns can be observed for this study on the sample size of patients with complete covariate data. This speaks for the robustness of the results in Chapter 4. A complete case analysis of

the missing data in outcomes at 3 months past surgery was not done, since this sub-chapter was experimental by nature and imputing questionnaire scores or items raises ethical concerns.