UiT The Arctic University of Norway

[Department of Neurosurgery]

[Development of a prognostic model for unfavorable outcome after lumbar microdiscectomy]

[David Werner]

A dissertation for the degree of Philosophiae Doctor [November 2020]

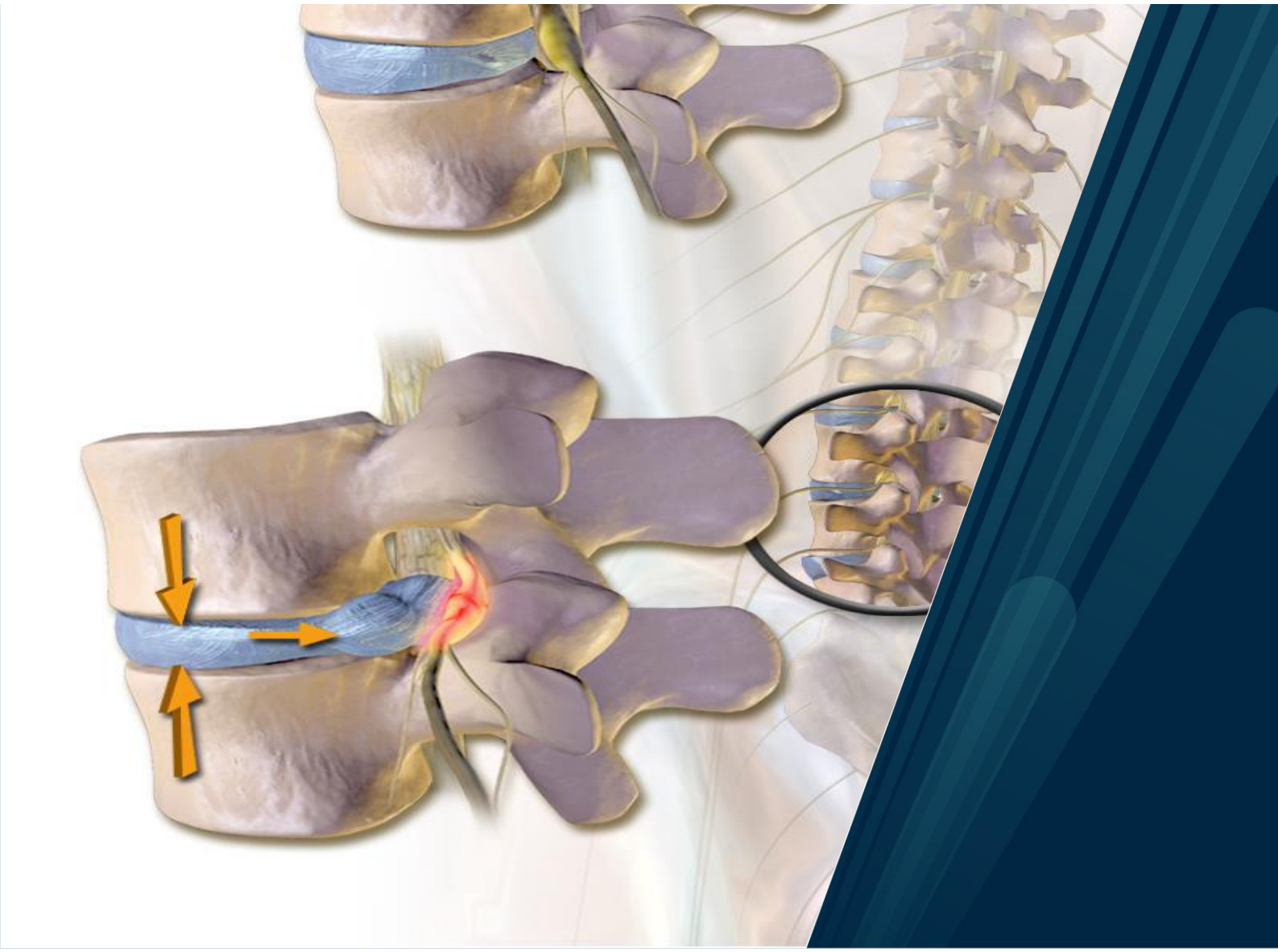


Table of contents

1 Tables 6

2 Figures 6

3 Preface 7

4 Acknowledgements 8

5 Funding 10

6 List of papers 11

7 What is this thesis about? 13

7.1 Overall aim 13

7.2 Outcome definition 13

7.3 Outcome prediction 13

7.4 Structure 13

8 Introduction 15

8.1 Lumbar disc herniation 15

8.1.1 Anatomy 15

8.1.2 Pathophysiology 15

8.1.3 Epidemiology 16

8.1.4 Diagnosis 16

8.1.5 Treatment 17

8.2 Clinical registry 21
8.2.1 Definition 21
8.2.2 Purpose 21
8.2.3 Design 21
8.3 Outcome interpretation 22
8.3.1 Patient Reported Outcome Measures (PROMs) 22
8.3.2 Minimal Clinical Important Change (MCIC) 24
8.3.3 Substantial clinical change 25
8.4 Outcome prediction 26
8.5 Quality of care 27
9 Materials and Methods 28
9.1 Design 28
9.2 Data source 28
9.3 Study population 29
9.4 Data collection 29
9.5 Analyses 30
9.6 Statistics 30
9.7 Ethical considerations 31
10 Results 32

10.1 Outcome definition (Paper I and II) 32
10.2 Outcome prediction (Paper III) 36
11 Discussion 40
11.1 Main finding 40
11.2 Outcome definition 40
11.2.1 Advantages and disadvantages of PROMs 40
11.2.2 Choosing the right anchor 41
11.2.3 Failure and worsening 42
11.2.4 ODI superiority and final score versus change score 44
11.2.5 The impact of baseline disability 45
11.2.6 Limitations of the minimal clinical important difference 45
11.3 Outcome prediction 47
11.3.1 Creating a prognostic model 47
11.3.2 Choice of risk factors 50
11.4 Handling of missing data 51
11.5 Model application 52
12 Future Perspectives 53
13 Conclusion 54
14 Works cited 55

15 Papers 71
15.1 Paper I. 72
15.2 Paper II. 82
15.3 Paper III. 90
16 Appendix 120
16.1 NORspine questionnaires (in Norwegian) 120
16.2 Supplementary appendix to paper I 126
16.3 Supplementary appendix to paper II 142
16.4 Supplementary appendix to paper III 152

1 Tables

Table 1. PROM cut-offs for failure and worsening, for the entire study population. 34

Table 2. Baseline dependent cut-offs for success. 35

2 Figures

Figure 1 Model validation 37

Figure 2. Prediction model for failure or worsening..... 38

Figure 3. Analysis of variance (ANOVA) 43

Figure 4. The Minimal Clinical Important Difference (MCIC) versus the cut-offs for failure and worsening on the final ODI raw score 46

3 Preface

Since the Norwegian Registry for Spine Surgery (NORSpine) was started in 2007, more than 50000 patients operated for lumbar spinal degenerative disorders have been included. The registry was started in Tromsø by my mentor Dr. Solberg, and has since spread out to all public and private clinics in the country. Similar registries have been developed in Europe[1], and the United States[2], collecting large amounts of data. However, use of this data at the hospitals is scarce, and while positive effects of quality registries have been shown in some medical disciplines[3], so far there is little evidence for spine registries having an impact on clinical practice. This thesis is aimed at bridging this gap by developing a decision support tool that conveys information from the NORspine about those previously operated back to patients and physicians, so that they can make better and more informed decisions about treatments for future patients.

4 Acknowledgements

This thesis is the brainchild of my mentor Tore Solberg, one of the founding fathers of the NORspine. Tore, with your positive and encouraging attitude it was easy for you to win me over for this project, and with your bottomless supply of coffee and patience you kept me going until this point. Thanks for your big brain, humor, the writing sessions during late night on calls, the meetings and conferences, and the occasional bottle of wine. Without you, this would have never happened.

I also owe my gratitude to Margreth and Milada, who in countless sessions have with the utmost patience managed to hammer basic statistics into a surgeon's brain. Without you, these papers would never have materialized. Thanks for the highly productive meetings in your small office in Oslo Margreth, and for the midnight answering service in regards to statistical technicalities Milada. Thank you, Sasha for your unwavering positive attitude, no matter what manuscripts I threw at you, and for the good brainstorming during meetings and conferences. Øystein, thanks for the golden nuggets of tips and insight, and for staying awake during your long Nepal flights to review my manuscripts. Ivar, thanks for taking the time to contribute to this work, whilst you very much where busy with your own PhD. Thanks also to Øyvind, who solved some very important methodological problems in the final steps of this work.

To all my colleagues at the Neurosurgical Department in Tromsø, thanks for supporting me when I needed to take time off clinical work to write papers, and for tolerating my constant bickering about this work. Lars, my roomie for more than six years, we have had all the ups and downs young neurosurgeons can experience during their training. Lasse, thanks for being my clinical supervisor and close friends during all this time. Kay, Jens, Andreas, and Tore, thanks for your valuable insights

into spine surgery, and for the countless hours in and out of theater discussing treatment approaches and surgical technique. Lasse, Roar, Jørgen, and Kristin thanks for training me in intracranial surgery, and being able to consult with you not matter what time of day. Roar and Kristin, thanks so much for giving med the administrative leeway to make this project possible, and for redirecting clinical workflow in order to get me all the time I needed to cross the finish line. To all my fellow residents, Helene, Maria, Kaja, Fanny, Daniel, and all the others over the last six years thanks for being flexible so that I could take the occasional short notice meeting, on call rotation or project-related trip.

Needless to say, without the unwavering support from the two most important persons in my life, I would have never made it this far, neither with this project, nor in my job. Thanks to Kristina and Philip, you mean the world to me.

5 Funding

I am very grateful for the financial support for this project received through research grants from both the Regional Health Authority of Northern Norway (Helse Nord) and the Norwegian Medical Association (Legeforeningen). Their support made it possible for me to take time off intermittently from my clinical work, in order to pursue this research.

6 List of papers

I. Werner DAT, Grotle M, Gulati S, Austevoll IM, Lønne G, Nygaard ØP, Solberg TK (2016) Criteria for failure and worsening after surgery for lumbar disc herniation: a multicenter observational study based on data from the Norwegian Registry for Spine Surgery. Eur Spine J [Internet]. 2017; Available from: http://link.springer.com/10.1007/s00586-017-5185-

II. Werner DAT, Grotle M, Gulati S, Austevoll IM, Madsbu MA, Lønne G, Solberg TK (2019) Can a successful outcome after surgery for lumbar disc herniation be defined by the Oswestry disability index raw score? Glob Spine J [Internet]. 2019;219256821985148. Available from: http://journals.sagepub.com/doi/10.1177/2192568219851480

III. Werner DAT, Grotle M, Gulati S, Salvesen Ø, Nygaard ØP, Ingebrigtsen T, Solberg TK (2020) A prognostic model for failure and worsening one year after lumbar microdiscectomy. A multicenter observational study based on the Norwegian Registry for Spine Surgery (NORSpine). Current under review at Acta Neurochirurgica (https://www.springer.com/journal/701/)

Abbreviations

AI	Artificial Intelligence
ANCOVA	Analysis of Covariance
ANOVA	Analysis of Variance
AUC	Area Under the Curve
CI	Confidence Interval
EQ-5D	EuroQol 5 Dimensions
GPE	Global Perceived Effectiveness
L1-5	Lumbar level 1-5
MCIC	Minimal Clinical Important Change
MRI	Magnetic Resonance Imaging
NORspine	Norwegian Registry for Spine Surgery
NRS	Numerical Rating Scale
ODI	Oswestry Disability Index
PASS	Patient Acceptable Symptom State
PROM	Patient Reported Outcome Measure
RCT	Randomized Controlled Trial
ROC	Receiver Operating Curve
SD	Standard Deviation
SDC	Smallest Detectable Change
SEM	Standard Error of Measurement
VAS	Visual Analogue Scale

7 What is this thesis about?

7.1 Overall aim

The overall aim of this research work was to develop a clinical tool which would be used by both surgeons and patients to predict outcome 12 months after lumbar microdiscectomy. With an outcome prediction the patient and surgeon would then be able to make an evidence-based informed decision about the question whether to operate or not. It is important to note that this thesis does not concern non-operative treatment of lumbar disc herniation.

7.2 Outcome definition

In order to be able to predict an outcome, first it must be clearly defined. This was the aim of papers I and II, where we defined criteria for success, failure and worsening 12 months after microdiscectomy for lumbar disc herniation, based on different Patient Reported Outcome Measures (PROMs). Notably we chose to define these criteria based on much larger magnitude of PROM changes, than the previously defined Minimal Clinical Important Change (MCIC).

7.3 Outcome prediction

With established outcome criteria, we developed a predictive tool by utilizing known risk factors and patient characteristics in paper III. The resulting model was implemented into a risk matrix, with an algorithm allowing us to calculate the probability of a negative outcome after surgery.

7.4 Structure

In the introduction, I will outline the clinical entity of lumbar disc herniation. Further, I will introduce clinical quality registries, and on this background introduce the metric of a PROM and how treatment outcomes are assessed with this tool. I will then briefly discuss the imperative of

quality assessment in modern medicine, and the of role prognostic research in its context. Since the papers are closely related, the methodological section, results, discussion and conclusion will comprise all three studies together. Finally, I will outline some future perspectives based on my research.

8 Introduction

8.1 Lumbar disc herniation

8.1.1 Anatomy

The spine or vertebral column is made up of bony building blocks (vertebrae) which are connected by intervertebral discs made up of a collagenous perimeter (annulus fibrosus) containing a liquid rich mucoprotein gel (nucleus pulposus), and by facet joints and ligaments. Behind the lumbar vertebrae and the intervertebral disc runs a bundle of nerve roots, covered by a layer of connective tissue (dura mater). A bony lamina is attached to each side of the vertebral body by the pedicles. This bony arch, the facet joints, and the yellow ligament (ligamentum flavum) protect the spinal cord from posterior. Thereby the spinal cord is run through a protective bony "tunnel" giving off one nerve root on each side at each vertebra of the spine. The lumbar spine denotes the last five vertebrae (L1-L5) making up the lower back, before the tail bone (sacrum and coccygeus). Approximately at the level of L1 the spinal cord ends. Below the dural sac contains peripheral nerves, i.e. the L1-L5, as well as the sacral 1-5 nerve roots, and collectively termed the "cauda equina" due to its resemblance of the tail of a horse[4]. The nerve roots exit the spinal canal by the foramen, defined by the pedicle above, the intervertebral disc medially, and the facet joint and isthmus laterally and below.

8.1.2 Pathophysiology

The degenerative process of the spine (spondylosis) increases with age, and starts in the intervertebral disc. Weakening or rupture of the annulus fibrosus can lead to herniation of the nucleus pulposus and impingement of nerve root(s) against the wall bony walls of the spinal canal or foramen[5]. Mechanical compression and inflammation can lead to pain and neurological

deficits, such as loss of both sensory and motor function according to the innervation of the affected nerve root. This radiculopathy manifests as radiating pain down the leg, and potentially numbness on the thigh, calf and areas of the foot[6–11]. In the case of lumbar disc herniation, the 4^{th} and 5^{th} disc are most commonly affected. The sum of these symptoms can be highly invalidating for the patient[12,13]. While the cause of a disc herniation is not entirely clear, both age, environmental, and genetic factors are suspected[14–17].

8.1.3 Epidemiology

The lifetime prevalence of lumbosacral radiculopathy is estimated to be between 12-27%. While the symptoms clear with the spontaneous resorption of the disc herniation in the majority of patients, surgery for lumbar disc herniation is the most common spinal surgical intervention[18–20], and whilst incidence and prevalence rates are constant, surgery rates are sharply increasing[21,22].

8.1.4 Diagnosis

Clinical

Lumbar disc herniation causes radiculopathy and leg pain. In addition, back pain is often be present. However, leg pain worse than back pain carries a high sensitivity for lumbar disc herniation[12]. The pain is often mechanical, i.e. increasing upon coughing, sneezing or lifting. Sensory loss for light touch, pain, and temperature can be present in the area known to be innervated by a given nerve root. Physical examination can show a mechanically irritable nerve root by maneuvers stretching the femoral or sciatic nerve (ipsilateral straight leg or inverted leg raising

test). Motor testing can reveal a paresis in muscle groups innervated by the given nerve root, as well as impaired reflex arcs innervated by the given root. In late stages of the disorder muscle wasting can be seen[12,23].

Imaging

Magnetic Resonance Imaging (MRI) is the gold standard for diagnosing intervertebral disc herniations[24]. A magnetic field is used to excite water molecules from their natural state and then to measure signals given off by these molecules upon returning to their resting state. This type of imaging is well suited to depict water rich anatomical structures, such as the nucleus pulposus and the contents of the dural sac. MRI for lumbar disc herniation has a sensitivity and specificity of 81% and 77%[25].

It is important to note that not all disc herniations with nerve impingement necessary lead to symptoms. Disc herniations can be found in 30-40% of asymptomatic individuals, increasing by advancing age[26,27].

8.1.5 Treatment

Non-surgical

In 70% of cases a herniated disc will dry out and shrink spontaneously within 3 - 12months, leading to a spontaneous improvement in symptoms[28]. Conservative regimens usually include rest, non-steroidal anti-inflammatory agents, and in some cases opiates and/or neuropathic pain medications. There is no clear evidence as to the benefit of surgical treatment over conservative approaches 12-24 months after onset of symptoms. Still, patients undergoing initial conservative

treatment will experience longer duration of pain, physical impairment, and sick leave than patients undergoing surgery within the first 6-12 weeks[19,29].

Surgical

In Norway and Europe microscope assisted discectomy (microdiscectomy) is the gold standard for surgical treatment of lumbar disc herniation. The procedure is performed under general anesthesia, with the patient in prone or knee/elbow position[30,31]. A 3-4cm incision is made between the two spinous processes above and below the affected disc. The thoracolumbar fascia is opened near the midline, and the underlying rectus muscle is dissected away from the lamina of the two vertebrae in a subperiosteal fashion. A retractor instrument is then placed so that the interspace between the two laminae is visualized. With the visual assistance of an operating microscope or loupes, the ligamentum flavum is opened and the underlying thecal sac and the affected nerve root are identified. Depending on the location of the disc herniation, the root is medialized or lateralized and the underlying posterior longitudinal ligament might be opened and the disc material is extracted.

Open discectomy was the most commonly used surgical method before the general advent of microscopes and it is rarely used nowadays. The procedure requires a larger incision and may lead to more soft tissue trauma, and may require more removal of more bone to improve visualization[32]. In contrast, use of the microscope allowed for minimal incision size with improved lightning and visualization.

In a minimal invasive discectomy, in a small tube is placed into the interlaminar space through av 1-2 cm incision. By use of an endoscope the thecal sac and nerve root are visualized, and the

herniated disc material is then removed in a similar fashion as in a microdiscectomy. While this procedure requires an even smaller skin incision, it requires additional instruments, yet the evidence is not clear on whether this procedure leads to superior outcome[33,34].

Nonsurgical invasive methods such as chemonucleosis[35], thermal nucleotomy[36] and epidural steroid injections[7,37] may be used as an alternative to surgical management. Evaluation of these methods is beyond the scope of this thesis.

Treatment decision

Lumbar disc herniations can in rare cases cause severe symptoms, also known as cauda equina or conus syndrome[38]. In these cases several nerve-roots are affected leading to sensation loss in the perineum and loss of bladder and external anal sphincter control, as well as paresis in the lower extremities. More commonly, nerve-root compression can cause loss of motor function, leaving the patient with a limp. Progressive neural deficits due to a lumbar disc herniation require an urgent decompression of the nerve-root, and thus they are considered absolute indications for surgery[39–41]. Radicular pain alone or in conjunction with back pain due to a disc herniation is a relative indication for surgery. This applies to cases were the nerve root compression does not naturally resolve, or where the pain is so invalidating that non-surgical treatment approaches yield no acceptable quality of life for the patient[19,29]. Since the indication for surgery is relative in most cases, it is important that the possibilities for both favorable and unfavorable outcome are discussed between patient and surgeon. In the Norwegian Registry for Spine Surgery (NORspine) 20% of cases are operated for paresis, and 1.3% for cauda equina. Thus about 80% are operated for

pain alone. It is therefore imperative that the treatment decision is based on the best available evidence, in order to provide quality care[41–43].

8.2 Clinical registry

8.2.1 Definition

A clinical registry is defined as an organized system for the collection, storage, retrieval, analysis, and dissemination of information on individual persons who have a particular disease […][44].

8.2.2 Purpose

Registry based research is aimed at improving the quality of health care in daily clinical practice[45]. The goal is to evaluate how treatments work in everyday clinical practice, when surgeons and patients have chosen a given type of treatment according to preferences. In contrast, randomized controlled trials (RCTs), intend to evaluate if treatment can work in idealized "homogenous" conditions[46]. While RCTs have high internal validity, they lack external validity, i.e. how does a given treatment perform in the "heterogenous" real life world of medical practice. In the latter, personal preferences of both patients and physicians, heterogenous comorbidities and lifestyle factors, as well as shortcuts in treatments and non-compliance introduce factors influencing the outcome[46,47].

8.2.3 Design

Clinical registries are designed a priori, collecting data based on a predefined purpose, i.e. quality in assessment and research. Unlike clinical trials with predefined patient management protocols, clinical registries "shadow" patient evaluation, treatment, and follow-up without influencing the course of these steps. This also means that data collection is prospective according to the general purpose of the registry at predefined time points, as opposed to data being collected retrospectively from other data sources such as the patient record[44].

8.3 Outcome interpretation

Modern day healthcare is based on scientific evidence. This evidence should weigh the patient's perspective on treatment and outcomes, as well as taking into account costs[48]. In order to measure outcome after interventions for multifactorial pain conditions such as degenerative spinal disorders, patient centered outcome measures have gained popularity and are now considered to be a gold standard[49,50].

8.3.1 Patient Reported Outcome Measures (PROMs)

In 1978 Lee et.al.[51] noted in a paper about surgery for spinal stenosis, that objective clinical findings did not reflect the patient's functional outcome. They thus proposed a self-rated questionnaire allowing the patient to score his/her functional abilities for several domains of daily living, laying the fundament for PROMs in spine care[52]. Since then, PROMs have become the gold standard outcome measure in spine care, and their use has increased exponentially[53,54].

PROMs can be defined as "any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else"[55]. PROMs can be measured in absolute terms, or as a change from a previous measure[56]. They let patients themselves report symptom intensity and functional impairment by answering specific questions, such as what type of chair they can sit in, or by grading symptom level, e.g. by a Visual Analogue Scale (VAS) for pain intensity. Answers to these questions are subsequently converted into a point score depicting the patient's functional level and quality of life as dictated by symptom burden. PROMs intend to answer questions regarding quality of care, such as "does the given treatment work?" and "does the patient perceive the treatment effect as expected by the

caregiver?"[57]. These constructs became popular quickly, because they allowed for a patient centered evaluation of treatment outcomes, shifting away from metrics like imaging diagnostics or biased caregiver opinion[53].

The most commonly used PROM in spine care, aside of the general VAS, is the Oswestry Disability Index (ODI) which asks the patient 10 questions regarding the limitations of daily life activities. Each answer is translated into a 0 to 5-point score, and consequently transferred into a percentage score with a range of 0 (no disability) to 100 (maximum pain related disability). The ODI is a specific PROM used in degenerative spinal conditions. The current version, employed in this study, is the ODI v 2.1a[58].

A generic health-related quality of life assessment is the EuroQol-5D (EQ5D 3L), evaluating the five dimensions of mobility, self-care, pain, anxiety, and activities of daily living. The degree of problems the patient has in each dimension is rated as either none, mild-to-moderate, or severe. These answers are translated into a score range from -0.59 to 1, where 1 corresponds to perfect health and 0 to death. Notably, negative values are considered to be worse than death. The EQ-5D has been validated for spine care[59].

Pain-ratings can either be recorded by the VAS, or more simply by a Numeric Rating Scale (NRS) ranging from 0 to 10, where 0 equals no pain and 10 the worst pain imaginable[60].

The Norwegian version of the PROMs measured in the NORspine and used for this study can be found in the appendix.

8.3.2 Minimal Clinical Important Change (MCIC)

With the advent of clinical registries and the increasing popularity of PROMs, a new problem became apparent. Because clinical registries allow for collection and evaluation of much larger data sample than traditional RCTs, small differences in measured data points could potentially be statistically significant, yet meaningless in the clinical context[57]. In order to overcome this issue, the MCIC, or minimal important difference (MID), or minimal important change (MIC) was introduced in 1989[61], and defined as the smallest change in outcome that a patient perceives as clinically important[62]. In the current body of literature, the a clinical important change is mainly determined based on two approaches, namely an anchor based method or a distribution based method, the smallest detectable change (SDC)[57,62,63].

<u>Anchor based</u>

As the name implies, in the anchor-based approach a PROM cut-off for the MCIC is determined by comparing (anchoring) PROM scores to an overall rating of the treatment effect by either the patient or physician. The Global Perceived Effectiveness (GPE) scale is such a rating tool, asking the patient to score the perceived treatment effect on a scale from 1 to 7, where 1 is the best possible outcome, and 7 the worst[64]. Average PROM changes between each category of this anchor can then be calculated and serve as descriptions of treatment effect. The GPE has good test retest capabilities and has been adapted for use in Norway[64,65].

Distribution based

The SDC is a statistical approach to measuring the smallest relevant change that can be detected mathematically through the noise (distribution) of the data. The SDC is calculated based on standard error of measurement (SEM), or the standard deviation (SD)[57,62]. In short if the magnitude of the PROM change reaches a certain level of baseline SD, for example ½ the SD, this change is determined to carry a clinical significance. The SDC should ideally not be greater than the MCIC.

8.3.3 Substantial clinical change

It is important to distinguish the minimal perceived change of the MCIC from substantial clinical effects. While a treatment effect in the magnitude of the MCIC might be perceivable by the patient, it might not be the patients or practitioners' goal of the intervention. A substantial clinical change based on PROMs, is one that exceeds the threshold of the MCIC by a good margin. This type of treatment effect is not merely perceived as clinically meaningful (e.g. feeling a little better), but rather perceived as a significant change in clinical status (e.g. feeling much better, or feeling completely recovered)[66,67]. While changes in the dimension of the MCIC are useful e.g. for sample size calculation in RCTs, substantial clinical changes are making the biggest difference for both the patient and practitioner in real life clinical practice, and thus need to be aspired to in order to improve quality of care[67–69]. As part of this dissertation I will propose outcome criteria of substantial magnitude 12 months after lumbar discectomy, termed "success", "failure", and "worsening".

8.4 Outcome prediction

A physician essentially has three tasks, namely to diagnose a condition and its etiology, foresee the prognosis, and to treat based on the best current medical standards. While etiological and therapeutic research is receiving widespread attention, the field of prognostic research has been the most limited of the three. Prognosis does not simply inform about the expected course of an illness or a treatment. Prognostic research also intends to estimate the risk of future outcomes in individuals based on their clinical and socio-demographic characteristics[70]. Clinical registries open new possibilities for prognostic studies using a multivariable approach to predict outcome probabilities based on numerous patient and disease specific parameters. These studies result in so called predictive or prognostic models, estimating an absolute risk or probability for a given outcome[71]. A prominent example is the Framingham study for predicting the 10-year mortality due to a cardiovascular event, based on given risk factors[72].

It is important to distinguish between prognostic modelling and associative modelling. While a prognostic model aims solely at predicting a future outcome with the highest possible accuracy, an associative model aims to identify independent risk factors for an outcome, while adjusting for other possible causal factors[71].

Development of a predictive model should optimally be performed in the setting of a prospective cohort study containing generalizable data from patients with heterogenic risk profiles, as well as a long period of follow-up, as opposed to a RCT with small sample sizes, and very comparable patients[71,73]. Thus, the NORspine is a well-suited environment for conducting outcome and prognosis research.

8.5 Quality of care

The Institute of Medicine defines quality in healthcare as the “degree to which health services [..] increase the likelihood of desired outcomes and are consistent with current professional knowledge”[74]. Contributing to quality are the informed participation of the patient, attention to the scientific basis of treatment and the efficient use of resources. With the advent of PROMs, the results of care can be evaluated with greater scientific accuracy, and according to the patient’s own experience. At the same time indications, risk factors, complication- and outcome measurements are collected in large clinical registries, such as the NORspine. These data allow assessment of quality of care for the given collecting hospital, or region, and make benchmarking and comparison between different medical facilities possible[75]. Consequently, medical professionals now are compelled to utilize this evidence in order to improve and maintain the quality of care, yet so far spine registries have had limited impact on the quality of spine care[76].

In the field of surgery for lumbar disc herniation the indication for surgery is often relative, yet patient expectations exceed those in other fields of surgery for degenerative conditions[77]. Treatment decisions must be derived from the balance between possible benefits, risks and also costs[48]. In a setting of increasing number of spinal surgical interventions, avoiding inefficient surgeries might have a larger impact on treatment outcomes, than improving the surgical technique itself[22,78,79]. Just as Thomas Mroz stated in his note[48] on the advent of value based spine care, we need to start focusing on the “why”, instead of the “how”.

9 Materials and Methods

9.1 Design

This work is a prospective multicenter observational cohort study of patients operated with lumbar microdiscectomy over the period January 1st, 2007 and August 2nd 2015.

9.2 Data source

The NORspine collects data on patients operated for degenerative disorders of the spine associated to spondylosis and spondylarthrosis, such as lumbar disc herniation, lumbar spinal stenosis, degenerative or isthmic spondylolisthesis, and degenerative scoliosis and segmental back pain. Both emergency and elective surgeries are recorded[80].

The registry does not include patients fulfilling the following criteria:

- Patients unable to give informed consent due to cognitive deficits or reduced consciousness
- Children < 16 years
- Patients with serious drug abuse or severe psychiatric disorders
- Patients with fractures, primary infections or malignant conditions in the spine
- Patients unable to respond to the declaration of consent and/or the questionnaires due to language barriers

Data collection is done at admission for surgery, and both three and twelve months after the operation.

9.3 Study population

All patients were recruited from the NORspine over the period of 2007 to 2015. During the study period the registry had a coverage (proportion of spinal centers reporting to the registry) of 95% of all public and private institutions, and a completeness (proportion of operated patients reported to the registry) of 65%[43].

For the purpose of this study, we excluded all patients operated on for any other conditions than lumbar disc herniation and/or patients who underwent fusion procedures. After subsequent exclusion of cases lost to follow-up, as well as patients diagnosed with spondylolisthesis, a total of 11081 cases were used in the analyses. In paper I and II a smaller patient sample (6840 cases) was created following these steps. In paper III the material was split at random into a training sample (70% of cases) and validation sample (30% of cases). This was done to allow for building the prognostic model in the training sample, and validate it in the smaller validation sample[81].

9.4 Data collection

Patients included in the registry filled out a questionnaire collecting baseline data on demographics, lifestyle issues, and PROMs at admission for surgery (baseline). During the hospital stay the surgeon recorded data concerning diagnosis, treatment, and comorbidities on a standard registration form. Twelve months after surgery a questionnaire identical to that used at baseline was distributed by regular mail. It was completed at home by the patients and returned to the central registry unit without involvement of the treating hospitals. One reminder with a new copy of the questionnaire was sent to those who did not respond. A copy of both questionnaires is attached in the appendix.

9.5 Analyses

Baseline characteristics were compared between responders and non-responders at 12 months follow-up. Outcome differences were investigated between elective and emergent lumbar microdiscectomies. We then investigated the correlation of four PROMs, namely the ODI, EQ-5D, NRS back pain and NRS leg pain with the GPE, 12 months after surgery. Furthermore, we assessed variation in postoperative scores for these PROMs between the individual GPE categories. We consequently defined two different outcome types, “failure” and “worsening”, based on GPE categories. We then calculated cut-off points for both the change score, the percentage change score (except for the EQ-5D), and the final raw score for each PROM against the two outcome types 12 months after surgery. The PROM showing the highest accuracy was selected and cut-off values were entered as the dependent variable in regression models, with patient baseline characteristics as possible independent predictors. Based on the regression models we finally created a risk matrix calculating the risk for a given outcome in percent.

9.6 Statistics

All statistical analyses were performed with either SPSS (IBM, Version 23.0), or R (Version 2.13.1). In paper I we assessed the variance of PROM scores against the seven categories of the GPE with and without adjustment for the baseline ODI, by means of analysis of variance (ANOVA) and analysis of covariance (ANCOVA). For nominally distributed data we assessed correlation between the PROMs and the GPE by Spearman rank correlation, and for non-nominal data (ODI raw score) Pearson correlation was used. In papers I and III differences in baseline characteristics for patients lost to follow-up versus completed follow-up were investigated by independent sample t-tests (continuous variables) and chi square tests (categorical variables). Furthermore, we compared outcome after 12 months between emergency and elective cases, by an independent sample t-test

for the PROMs and by Mann-Whitney U test for the GPE. In both papers I and II we calculated cut-off values for the respective PROMs (in paper II only the ODI was used), by using the coordinates of Receiver Operating Curves (ROC) that showed the highest sensitivity and specificity for classifying a given outcome. Overall classification rates of the PROM cut-offs against the actual outcomes were identified by confusion matrices[82]. In paper III we identified potential risk factors for both failure and worsening by univariate binary logistic regression. Significant variables were consequently included in a binary logistic multiple regression model, and removed in a backwards manual fashion based on their level of significance. Goodness of predictions were analyzed by plotting the observed proportion of outcome against the average predicted proportion. Chi square test was used to assess if there were significant differences between the predicted and observed coordinates on the graph (calibration)[83]. The discriminative ability (discrimination) was determined by running ROC analyses of the risk values against the predicted outcome, where the area under the curve (AUC) served as an estimate for the accuracy (C-criterion)[84].

9.7 Ethical considerations

This study is based on data collected from clinical cases. No animals, drugs, human tissue or other live tissue samples were part of this investigation. The study protocol was submitted to the regional ethical committee for medical research which categorized it as a clinical audit study, not in need of their formal approval[85]. Participation in the registry is neither mandatory, nor required to receive treatment. Except for data registration, no differences in treatment decision and hospital protocol are done for patients participating in the registry or those who opt out. All patients are offered an outpatient follow-up 12 weeks after the surgery.

10 Results

10.1 Outcome definition (Paper I and II)

ANOVA showed that both the mean ODI, EQ-5D, NRS back-pain, and NRS leg-pain scores were significantly different between GPE groups 1-3 and 4, as well as GPE groups 4,5 and 6,7. We defined "failure" as a patient rated outcome of GPE 4 – 7 (no change, somewhat worse, much worse, worse than ever), and "worsening" as a patient rated outcome of GPE 6-7 (much worse, worse than ever). All PROMs correlated significantly with the GPE. For none of the PROMs were floor or ceiling effects found. The ODI percentage change, as well as the 12-months ODI raw score, were the most robust in defining failure and worsening. Initially we identified cut-offs for the whole study population (table 1). The overall correct classification rates were highest for the ODI raw and the ODI percentage change scores, however only the ODI raw score 12 months after surgery showed acceptable accuracy when defining failure or worsening.

During the analyses we noticed that the cut-offs are dependent on the level of baseline disability and we additionally calculated failure/worsening cut-offs on the ODI score for three baseline ODI groups, namely patients with low baseline disability (ODI <33, <25th percentile), moderate baseline disability (ODI 33-58, 25th-75th percentile), and high baseline disability (ODI >58, >75th percentile).

We also identified an ODI raw cut-off for success (GPE groups 1 and 2), for all three baseline ODI groups (table 2). Again, the ODI raw and ODI percentage change scores were the most accurate for defining the outcome.

Based on our cut-offs, 63-65% of patients had a successful outcome 12 months after microdiscectomy. Furthermore, 23-27% scored as failure, and 7-8% as worsening.

Table 1. PROM cut-offs for failure and worsening, for the entire study population.

	Failure					Worsening			
	Cut-off	**Sens/Spec**	**AUC (95% CI)**	**Corr. Class %**		**Cut-off**	**Sens/Spec**	**AUC (95% CI)**	**Corr. Class %**
ODI									
Change score	**13**	0.82, 0.82	0.89 (0.88 - 0.91)	82					
Percentage change score	**33**	0.86, 0.86	0.93 (0.92 - 0.94)	86					
12 month raw score	**25**	0.89, 0.81	0.92 (0.91 - 0.93)	86		**48**	0.70, 0.70	0.76 (0.72 - 0.80)	69
NRS leg-pain									
Change score	**1.5**	0.81, 0.76	0.87 (0.86 - 0.88)	84					
Percentage change score	**39**	0.86, 0.81	0.89 (0.88 - 0.90)	84					
12 month raw score	**4.5**	0.91, 0.85	0.90 (0.88 - 0.91)	84		**7.5**	0.64, 0.68	0.70 (0.66 - 0.75)	67
NRS back-pain									
Change score	**1.5**	0.74, 0.86	0.85 (0.84 - 0.86)	76					
Percentage change score	**24**	0.85, 0.81	0.87 (0.86 - 0.88)	86					
12 month raw score	**5.5**	0.81, 0.87	0.92 (0.91 - 0.93)	86		**7.5**	0.78, 0.64	0.77 (0.73 - 0.81)	68
EQ-5D[1]									
Change score	**0.10**	0.76, 0.83	0.85 (0.84 - 0.87)	82					
12 month raw score	**0.63**	0.81, 0.85	0.91 (0.90 - 0.92)	85		**0.09**	0.76, 0.60	0.71 (0.67 - 0.75)	65

Cut-offs were calculated for the four different PROMs against the GPE by means of ROC analyses. All cut-off values with corresponding sensitivity and specificity, area under the curve (95% confidence interval) and percentage of correctly classified. For worsening, only the 12-month raw scores were used, all other cut-offs had an AUC<0.70. [1] not possible to calculate % change score for EQ-5D.

Table 2. Baseline dependent cut-offs for success.

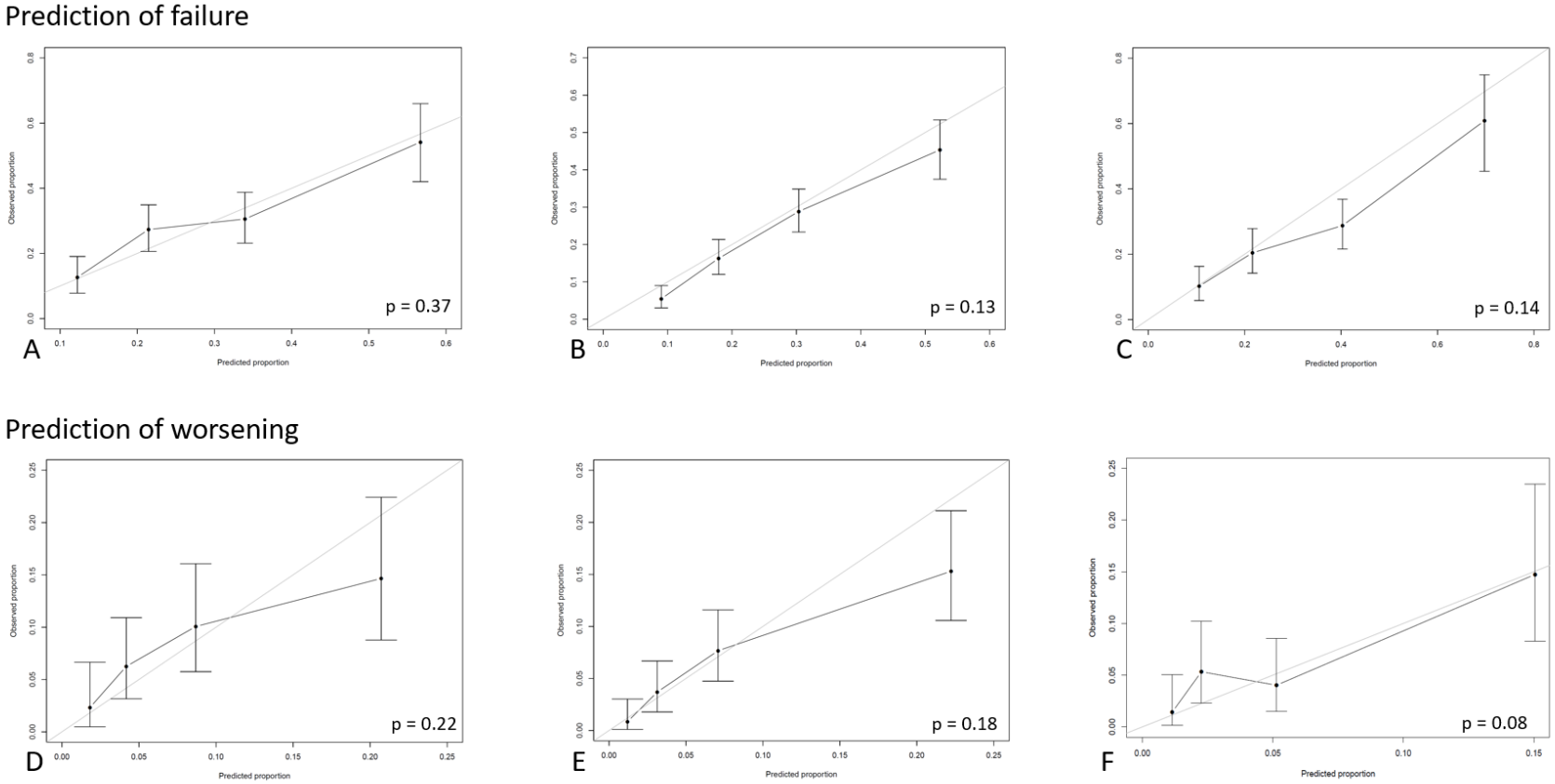
	AUC	**95% CI**	**Cut-off**	**sens/spec**	**Accuracy (%)**
	12-months ODI raw score				
ODI Prescore					
<25th percentile	0.92	0.90-0.93	**13**	0.81/0.88	83
25-75th percentile	0.95	0.94-0.95	**21**	0.85/0.89	86
>75th percentile	0.94	0.93-0.96	**28**	0.89/0.85	88
	ODI change				
ODI Prescore	**AUC**	**CI**	**Cut-off**	**sens/spec**	**Accuracy (%)**
<25th percentile	0.89	0.88-0.91	**9**	0.77/0.84	79
25-75th percentile	0.92	0.91-0.93	**24**	0.83/0.84	83
>75th percentile	0.92	0.91-0.94	**48**	0.85/0.84	85
	ODI % change				
ODI Prescore	**AUC**	**CI**	**Cut-off**	**sens/spec**	**Accuracy (%)**
<25th percentile	0.91	0.90-0.93	**39**	0.82/0.84	83
25-75th percentile	0.94	0.94-0.95	**53**	0.86/0.88	86
>75th percentile	0.94	0.93-0.96	**66**	0.85/0.88	88

Cut-offs were calculated for three ODI metrics, the change score, the % change score, and the 12-months ODI raw score, by means of ROC analyses. AUC = Area Under the Curve, CI = 95% Confidence Interval, sens = sensitivity, spec = specificity. Overall accuracy was determined by a confusion matrix.

10.2 Outcome prediction (Paper III)

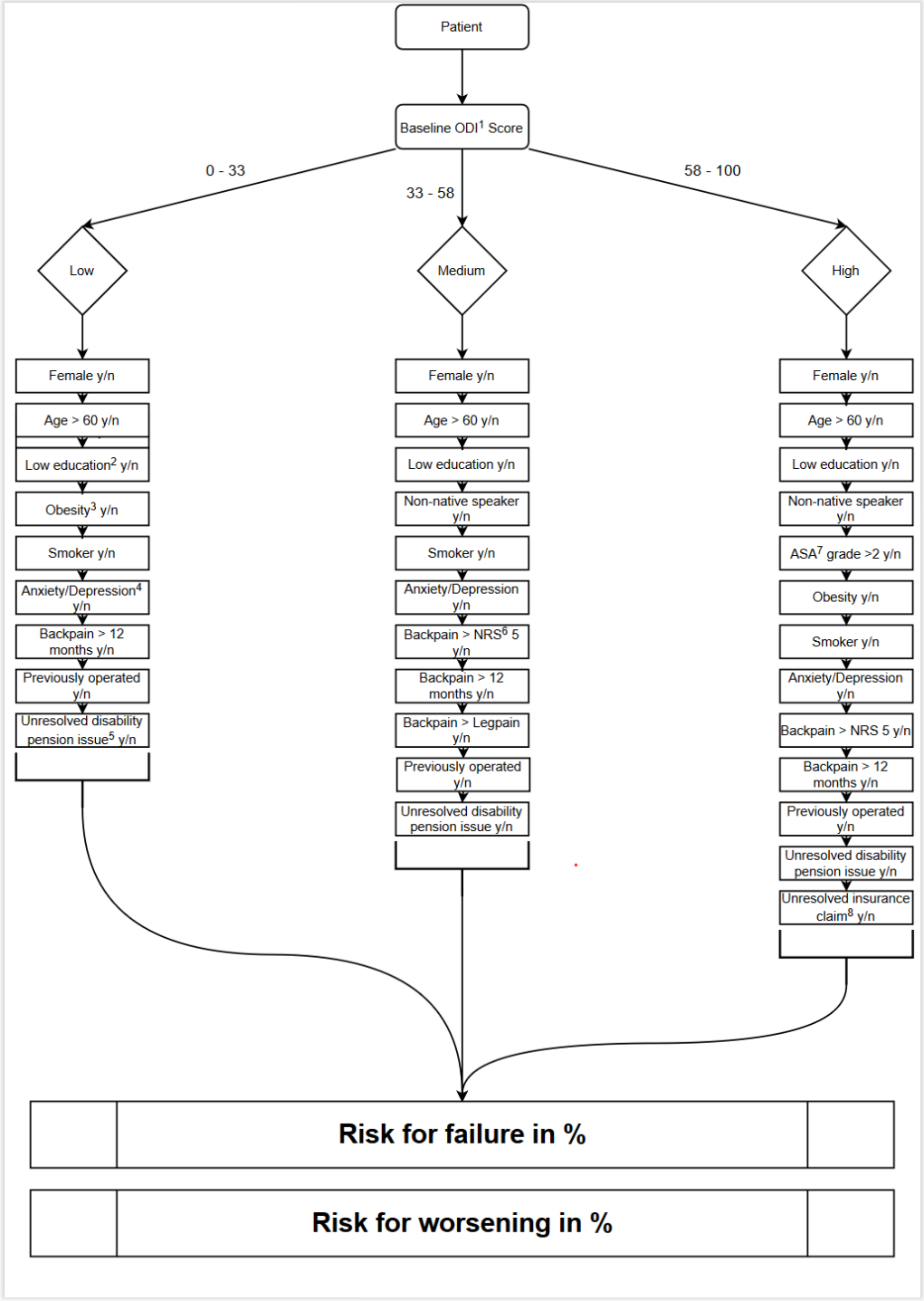
Based on the results in papers I and II we chose the ODI raw score at 12-months as the PROM which cut-offs should be predicted for both failure and worsening. Depending on the level of baseline disability (preoperative ODI) we split the study population into three groups, namely those with a baseline ODI below the 25th percentile, above the 75th percentile, and in between the 25th and 75th percentile. We built one model for both failure and worsening in each group. Based on the results from the uni- and multivariate regression analyses, each model resulted in three risk matrices, with 7-11 different covariates. Smoking, an educational level with less than four years of college or university education, and the presence of more than 12 months of back pain prior to surgery were significant risk factors common to all six matrices. Discriminative ability of the model was acceptable, but calibration testing showed that the matrix predicting worsening in the high ODI baseline group (ODI >58) deviated significantly ($p<0.1$) from the optimal prediction line (Fig 1), suggesting possible underestimation of the outcome. The final model is shown in figure 2.

Figure 1 Model validation



Observed proportion of the outcome (with confidence interval) on the vertical axis against average predicted probability of the outcome on the horizontal axis. Each coordinate with whiskers represents one quartile of estimated probability and its 95% confidence interval, compared to the observed proportion of the predicted outcome. The p-value from the chi square test for the coordinates vs the optimal prediction line is indicated in the lower right corner. A p-value < 0.1 indicates significant deviation from the average predicted probability. A-C show prediction of failure for the three baseline invalidity groups (A: Baseline ODI <25th percentile, B 25-75th percentile, C >75th percentile). D-F show prediction of worsening for the three baseline invalidity groups (D: Baseline ODI <25th percentile, E: 25th – 75th percentile, F: >75th percentile).

Figure 2. Prediction model for failure or worsening.



Model algorithm for the three ODI baseline groups. Based on the preoperative ODI the patient will be classified via one of the three pathways, calculating an overall risk for either failure or worsening. Risk is calculated from the odds of each risk factor. The risk factors are listed in random order, and their place in the sequence does not reflect their odds. [1]Range: 0-100 (no-maximal disability). The ODI score was <33, 33-58 and >58 in the subgroups with low, medium high baseline disability, respectively.[2]Less than four years of college/university education. [3]Body Mass Index ≥30. [4]EQ-5D 3L questionnaire; 5th item, moderate to severe problems. [5]Pending medical claim/ litigation the Norwegian public welfare agency fund concerning disability pension. [6]Numeric Rating Scale (0-10). [7]American Society of Anesthesiologists grade [8]Pending

medical compensation claim/litigation against private insurance companies or the public Norwegian System of Compensation to Patients.

11 Discussion

11.1 Main finding

The main finding of this thesis is that we were able to develop a prognostic model for failure and worsening 12 months after lumbar microdiscectomy, based on data from the Norwegian registry for spine surgery. We also found that unfavorable outcome can readily be defined by cut-offs on the ODI, NRS backpain, NRS legpain, and EQ-5D. The ODI percentage change, and the final ODI score 12 months after surgery were the most accurate PROMs for this purpose. The final ODI score after 12 months was also able to define a favorable outcome after surgery with high accuracy. Furthermore, cut-offs for all metrics, were depending on the amount of preoperative baseline disability.

11.2 Outcome definition

11.2.1 Advantages and disadvantages of PROMs

Since their inception, PROMs have gained in popularity due to several advantages. They allow for measuring the impact of chronic pain conditions such as disability, symptom burden, and quality of life from the patients' own perspective, whereby eliminating observer bias (no surgeons rating of the outcome). Moreover, they facilitate communication and shared decision making. Since they assess domains important to the patient, they also increase self-awareness[52,54,56,86].

While PROMs offer significant advantages, they also bear some inherent problems. Since they are based on the patient's own assessment, and often are measured prior to and after a given intervention, they are susceptible to the lapsing memory (recall bias) and change in the patients' value construct (response shift). Recall bias simply implies that a patient does not remember his or her rating on a given dimension of the PROM, e.g. how badly the symptom

intensity was 12 months ago[87]. Response shift basically implies a change in the patient's perspective of the PROM items. Response shift can be further classified into recalibration (change in internal standard of the patients assessment of his/her wellbeing), reprioritization (the same items of health related quality of life do not carry the same weight in the patient's own perception of quality of life), and reconceptualization (redefinition of the concept)[88]. In addition, PROMs can exhibit floor or ceiling effects, where the potential disability could extend beyond the scale leading to a grouping of patients who might consider their disability levels differently[89,90].

11.2.2 Choosing the right anchor

When choosing an anchor for establishing cut-offs of clinical significance against the PROMs, this anchor should be intuitively meaningful, able to inform on the change over time, as well as reflect the PROM`s concept[91]. Ideally it should also be objective, easy to measure, and applicable in all kinds of clinical settings. Yet no such anchor exists.

The GPE is based on the patient's ability to recall hers/his symptom state 12 months earlier, and compare it to the symptom state at the time of the assessment. Both assessments, the previous and the current, are potentially biased in the same way as PROMs, and as explained in the previous chapter. The ability to remember the level of symptoms and disability varies from patient to patient, and some might not recall accurately how they felt before the surgery (recall bias)[65]. Furthermore, when assessing the symptom state at the time of follow-up, other factors than pain and disability might influence the patients rating of his or her overall health, and thereby potentially influencing the rating of the outcome on the GPE scale (response shift)[57,92]. Patient expectations and their discrepancy to the actual outcome can also influence the overall rating of surgery[77,93]. Other measurements have been suggested, such as the clinicians rating of outcome, which has been proven to differ from the patients

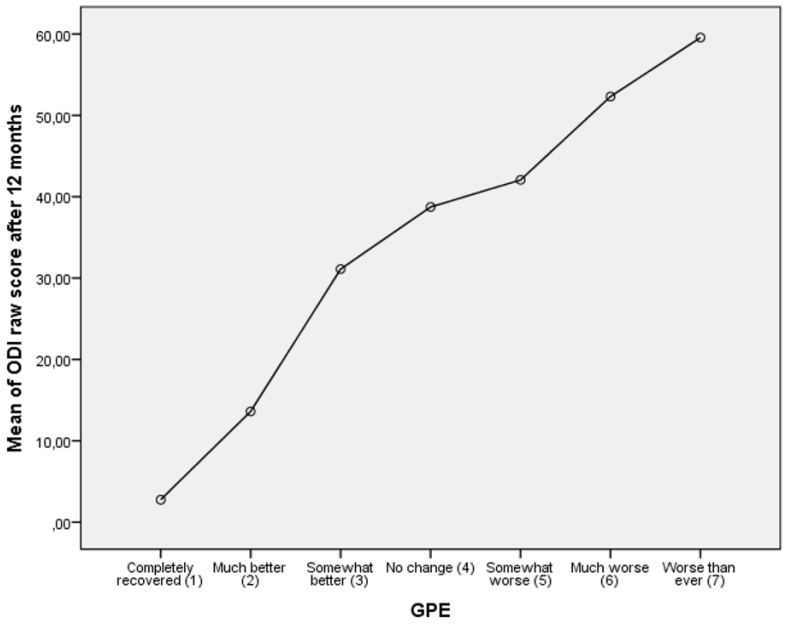
perspective [52,94]. More objective measures such as return to work, painkiller use, or other group specific metrics exclude subgroups of patients and are more difficult to measure.

To the best of my knowledge, the GPE is currently the most optimal approximation to a gold standard anchor. This is also reflected in the recommendations of the Food and Drug Administration (FDA) and IMMPACT consensus group[49,50].

11.2.3 Failure and worsening

In chronic pain conditions, any surgery resulting in no improvement or even worse symptoms after the surgery, can hardly be viewed as beneficial. Outcome constructs defining positive results after surgical interventions on the spine have previously been evaluated[67], and clearance of all symptoms naturally leads to an outcome being rated as successful. We aimed at defining the negative spectrum of outcome after lumbar microdiscectomy, and we chose two categories, namely failure and worsening. In paper I ANOVA analyses of the ODI against the GPE showed that GPE categories 1-3 (completely recovered, much better, somewhat better) were significantly different from category 4 (no change) (fig. 3). Furthermore, categories 4-5 (no change, slightly worse) and 6-7 (much worse, worse than ever) were clearly distinguishable. We therefore decided to define one outcome class as failure, where the patient reported no change or a worse status 12 months after surgery. We also defined a category where the patient reported at least a much worse outcome after surgery, termed worsening. A large proportion (24%) of patients classified themselves as somewhat better, unchanged, or somewhat worse after surgery, based on the GPE. However, those in the somewhat better group showed a mean improvement on the ODI over 15 points, which crosses the threshold of the MCIC[95]. Thus, these patients should neither be classified as success, nor as failure. We termed this group non-success.

Figure 3. Analysis of variance (ANOVA)



Analysis of variance (ANOVA) of the mean final ODI raw score 12 months after lumbar microdiscectomy against the Global Perceived Effectiveness scale (GPE) ranging from 1 – completely recovered, to 7 – worse than ever.

11.2.4 ODI superiority and final score versus change score

In paper I ROC analyses showed that the ODI was superior to the NRS back-pain, NRS leg-pain, and the EQ-5D when determining a cut-off for both failure and worsening. This is not surprising since ODI is both a disease specific metric, as it takes in account more than one dimension of pain. Thus, it is also from a clinical perspective more suitable, than e.g. the NRS leg-pain[67,96]. It has been previously validated for the Norwegian population[97], and in our studies we could not identify floor or ceiling effects.

In both papers I and II we could see that the ODI change score in points clearly had inferior measurement properties than the ODI raw score, or the ODI percentage score for defining outcomes cut-offs. This has also been shown in a large medical registry study in the US[98], as well as for a lumbar spinal stenosis study in the NORspine[99], and makes sense as the change in points does not reflect the underlying magnitude of improvement or worsening (i.e. a patient improving 30 points with a baseline ODI of 40 experienced a much larger improvement than a patient with a baseline ODI of 70). Our results also confirm the notion that the final ODI score is of importance when the patient rates his or her outcome after 12 months[65]. Symptoms may well have improved from baseline, yet the patient might consider the outcome as failed, or even worsened. Figure 4 shows how patients who actually experienced improvements from their baseline ODI, report outcome scores indicating that they feel unchanged or even worse after the surgery. This illustrates the importance of a disability score as the entity defining failure or worsening (or positive outcomes on the other end of the scale), versus simply using the patient rated outcome, or even the surgeon's own opinion. Furthermore, this implies that change alone is not the sole arbiter of a substantial benefit to the patient, and that the final ODI score plays a role in the patients' perceived benefit of the surgery. This is also a finding of other studies, identifying a Patient Acceptable Symptom State (PASS) on the ODI[69,100].

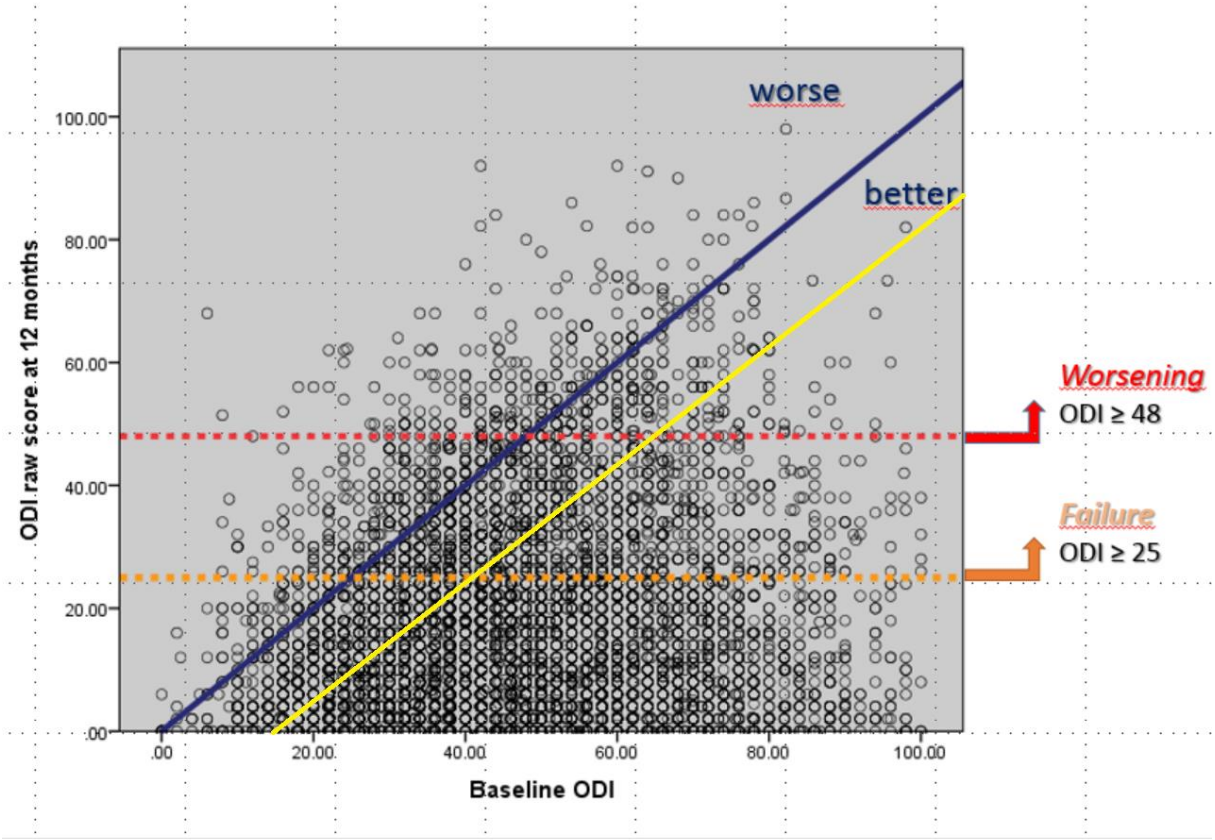
11.2.5 The impact of baseline disability

ROC analyses in papers I and II show that all cut-offs for the ODI, independent of metric, differ based on the preoperative ODI score. Patients with a larger amount of disability need to perceive a larger amount of improvement, not only in points but also in percent, in order to rate the surgery not as failure or worsening. This is in accordance with a similar study[98]. Consequently, the baseline ODI needs to be controlled for, when developing outcome criteria and prognostic models. One simply cannot apply the same criteria for a patient with a rather low baseline disability, vs a patient on the high end of the spectrum.

11.2.6 Limitations of the minimal clinical important difference

The MCIC has previously been recommended as an outcome criterion for success after spine surgery[62]. This is somewhat problematic as the MCIC is a fluid construct[101], proven to be shifting in magnitude based on the amount of baseline disability a patient experiences before surgery, as well as the time passed since the surgery[52]. Many patients might experience change corresponding to the MCIC, and yet rate their outcome negatively. This is illustrated in figure 4, where the yellow diagonal line represents a change of 15 points in the ODI between baseline and 12 months after lumbar microdiscectomy. This line delineates the generally accepted MCIC for the ODI[95]. All points to the right of that line have achieved a postoperative improvement larger than the MCIC, yet many patients score as failed or worsened.

Figure 4. *The Minimal Clinical Important Difference (MCIC) versus the cut-offs for failure and worsening on the final ODI raw score*



Y axis: ODI raw score at 12 months vs X axis: ODI at baseline. The blue diagonal line represents no change. The yellow diagonal line represents the MCIC of 15 points. Coordinates to the right of the yellow line represent patients who have achieved the MCIC 12 months after microdiscectomy. The red line indicates the 12-month ODI raw cut-off, above which patients consider themselves as worse, irrespective of the change experienced. The orange line represents the ODI raw cut-off for failure.

We suggest that stronger criteria, such as success or failure/worsening are to be used when one wishes to improve quality of care, instead of minimal changes such as the MCIC. Rather than drawing conclusions in regard to outcome and their implications in terms of clinical significance, the MCIC can be used when comparing outcome across groups or interventions.

Aside of questions around the clinical implication, neither the MCIC or metrics of larger amplitude such as the substantial clinical change take into account the cost of treatment. This might not matter to the patient, but very much to the legislator, administrator, politician, and society, who in the future might want to see value for their money in terms of clinical effect achieved per unit currency spent[47,101].

11.3 Outcome prediction

In paper III we developed a prognostic model resulting in six risk matrices predicting negative outcome (failure/worsening) 12 months after surgery for lumbar disc herniation. Each matrix is applicable to a baseline ODI range (<25th, 25th-75th, or >75th baseline ODI percentile). It is important to note that the model was built based on data from a population of patients who were all referred to surgery, and had undergone lumbar microdiscectomy. Thus, the model might not be applicable for patients who are evaluated in general practice and who might benefit from noninvasive treatment options. Furthermore, the model was built based on patient data from the NORspine, and thus usability and feasibility in other spine registries needs to be assessed.

11.3.1 Creating a prognostic model

In prognostic modelling, especially in the field of medicine, two main methods are used. The traditional approach is multivariable analysis, while the more novel approach is based on artificial neural networks[102]. The discussion of the latter is beyond the scope of this dissertation. Multivariable analysis determines contributions of various factors to a single outcome. It's a powerful tool which can be utilized for different purposes, mainly to either shed

light on the importance of each individual factor in regards to the outcome (used in epidemiological, associative studies), or to predict a given outcome based on the presence or absence of risk factors and possibly unknown secondary factors (confounders). The latter method is used for prognostic modelling and is the method of choice for this dissertation.

Due to the nature of the majority of variables collected in the NORspine, we chose logistic regression where the included covariates are dichotomous (yes or no). This allows for the calculation of odds ratios (OR), from which probabilities can be calculated. The advantage of this is that the concept of a probability for a given outcome is easy to understand for both patient and clinical caretaker, as opposed to coefficient values from linear logistic regression models. However, dichotomization also bears disadvantages. Information from continuous observations is lost, and patients are pooled into categories leading to the same outcome prediction, albeit having potentially different risk values. Dichotomization also hampers comparability with other studies on the same subject, using different cut-off points on the linear scales[103].

While associative models are sensitive for confounders, prognostic models make no assumptions in their regard[104]. Thus, based on our analyses we cannot make an assumption on the causal relationship between smoking and the outcome after lumbar microdiscectomy. Our model shows that smoking increases the risk for failure and worsening as an outcome after lumbar disc surgery. While smoking might directly have an impact on the outcome, its effect might very well be mediated by a known or unknown confounder.

When building a multiple regression model, one can choose between an automatic or manual approach, and in case of the latter between a forward, backward, or subset method. Automatic methods act non-discretionally based solely on mathematical reasoning. While this approach is criticized due to issues with confounding in associative modelling, it is also not optimal for predictive models. This is due to the fact that in some cases clinically important variables are

excluded in favor of other variables just because of a minor difference in mathematical statistical significance. Manual models have the advantage of clinical discretion and better transparency. For our purpose we chose manual backward regression. We first made a preselection of variables by univariate regression, assessing the predictive power of each covariate on its own against the outcome. Significant variables were consequently entered in the model simultaneously, and then the weakest one was excluded until only statistically significant variables remained[84]. Based on clinical discretion we also included age and gender, irrespective of statistical significance. Because previous findings indicated that the rating of the outcome 12 months after surgery is strongly influenced by the amount of baseline disability based on the ODI score, we chose to create subset models for three baseline strata, resulting in the six risk matrices.

Once a model is built and risk matrices calculate the probability for a given outcome, the question is as to the accuracy of said predicion. In the case of logistic regression, a recommended method is to compare the proportion of predicted risk to observed outcome in groups of patients, i.e. in a group of 100 patients averaging a 30% predicted probability of a given outcome, optimally 30 patients should achieve this outcome[84]. Our results illustrate this assessment in figure 2. The reader may note that the 95% CIs are larger for the three matrices predicting worsening, indicating a smaller sample size. This represents a weakness in the models, resulting from a rather low incidence of worsening as an outcome 12 months after surgery. Nevertheless, aside of the matrix predicting worsening in those with a baseline ODI above the 75th percentile, observed proportions of outcome did not deviate significantly from the average predicted probability. It is important to note though, that this might only hold true for our study population and that the model`s reliability might be insufficient when evaluated in other patient populations, for example in other clinical registries[71,84,105].

11.3.2 Choice of risk factors

When developing a prognostic model, one has to make a choice in regards to which risk factors to include in order to predict the given outcome. Simply including any factor available in the dataset would lead to the best accuracy for the given model applied in the data set it was developed from, and while that is well within the purview of a prognostic regression model, this approach would hamper its generalizability to other clinical registries and its applicability in clinical practice. In order to develop a both clinically meaningful and generalizable prediction model, risk factors included should be readily available, simple to measure, and at the same time carry a high predictive value. Based on these criteria, we chose covariates which have previously been identified, such as intensity of low back pain and leg pain, BMI, educational level, previous back surgery, smoking, and unresolved issues with disability funds or medical insurances[106,107,116–119,108–115]. While all operated patients had an MRI confirmed lumbar disc herniation, the registry does not collect data on prolapse morphology. This might be a weakness in the model, however the contribution of image findings to prognosis is not clearly established[120,121].

11.4 Handling of missing data

We report a lost to follow-up rate at 12 months of 31-32% in our papers. Among non-respondents we found a statistically significant higher proportion of risk factors for a negative outcome in patients that smoke, have a lower educational level, have had previous lumbar disc surgery, and those receiving sickness or disability payments. At the same time these cases also showed a higher proportion of positive risk factors, as in they were younger, more likely to be men, suffering from less comorbidities, and had less severe limb paresis. Still, loss to follow-up could represent a selection bias, if these cases would show a significant difference in outcome against our study population[122,123]. This issue has been addressed not only in the Norwegian, but also the Swedish and Danish spine registries, where patients lost to follow-up were traced and interviewed. The studies found the same baseline differences when comparing responders to non-responders, yet no differences in outcome between the two groups up to 24 months after surgery[124–126].

Aside of missing outcome data for those lost to follow-up, we reported low percentage of missing data for baseline values in all PROMs. The largest proportion of missing data in all three studies was found in paper III, where the BMI as a possible risk factor had approximately 10% of datapoints missing. Based on the results from the studies mentioned above, we deleted missing data in a pairwise fashion. In longitudinal studies one has the option to estimate values of missing data by different methods of imputation, namely cross sectional or longitudinal imputation. In regards of outcome data lost to follow-up, a popular method is imputation by carrying the last known observation forward to the end-point. Carrying forward 3-month follow-up values of the ODI to estimate 12-month outcomes is not advised[127,128]. The NORspine does not register outcome values between these two timepoints, thus more advanced longitudinal methods cannot be applied here. When handling missing baseline values, cross

sectional imputation could be applied by taking the mean of either all available values for a given variable (mean of series imputation), or the mean of a random set of cases with similar baseline characteristics for a given variable. Another, more sophisticated, method is building a prognostic model predicting a given baseline variable, based on the presence of other baseline variables. The main author considered the first two methods to be guessing at best. The last method was not considered, as predicting BMI based on other socioeconomic variables potentially carried significant bias in itself, and goes far beyond the scope of this study. For this reason, cross sectional regression is only advised for missing outcome variables but should not be applied to estimate predictor variables. It has also been shown that imputation of missing data has no significant impact on the final models with 10% missing values at baseline. Furthermore, imputation of missing data of larger proportion led to weakening of the regression models[129].

Instead of trying to estimate missing values, we rather recommend to use data from the whole study population in order to calibrate the prognostic models, and subsequently assess applicability and discriminative ability in another study population, such as in another spine registry.

11.5 Model application

The prediction model developed in paper III is based on data from patients who already have undergone lumbar microdiscectomy. Thus, these patients have prior been selected as suitable candidates for surgery by either a neurosurgeon, or an orthopedic surgeon. The model is aimed at aiding the surgeon and the patient in the shared decision-making process, especially when the grounds for a surgical indication are weak or uncertain. Furthermore, the predicted outcome can be helpful in setting expectation levels prior to surgery.

12 Future Perspectives

Spine care is an expensive business, and regulators are voicing the need for a value-based approach[130,131]. Spine care patients are a heterogenous population, with varying levels of expectations and means to cope with pain. There is a need to make personalized informed decisions about a surgical treatment the outcome of which cannot be measured by means of survival rates, but is so very important for the quality of life[132]. This work represents the first step in predicting outcome after a common surgical intervention for a chronic pain condition in a heterogenous patient group, on a large scale. I hope that our model can be externally validated in another spine registry population, and consequently be the platform for a risk calculator to be tested in clinical practice. A possible scenario would be to compare differences in practice and outcome between two units, where one is applying the calculator during the surgeon/patient shared decision-making process. If proven useful, similar concepts can be applied in decision processes for surgical treatment of other degenerative disorders.

Novel, artificial intelligence (AI) driven prediction of outcomes, is increasingly gaining ground also in the field of neurosurgery[133,134]. We expect these techniques to take a central part in medical decision making in the future, also in the field of surgical indication judgement. It will be interesting to see if an AI can handle the complex interaction of risk factors better than a human brain, in order to improve quality of care.

13 Conclusion

In this dissertation we have analyzed data on patients operated for lumbar disc herniation by means of lumbar microdiscectomy. Patient participation in the studies, by partaking in the NORspine registry, has not led to additional examinations or deviation of treatment protocol from regular clinical practice. We defined cut-off values on validated PROMs to classify outcomes as “success”, “failure”, and “worsening”. The ODI percentage change, and the final ODI score 12-months after lumbar microdiscectomy were the most accurate in defining these outcomes. As the result of our analyses, we propose a prognostic model for two outcomes, namely failure and worsening, one year after surgery for lumbar disc herniation. The model is built based on previously identified risk factors, and outcome criteria identified based on the patient’s own rating of outcome compared to self-reported disability scores.

14 Works cited

1 Aebi M, Grob D. SSE Spine Tango: a European Spine Registry promoted by the Spine Society of Europe (SSE). *EurSpine J* 2004;**13**:661–2.

2 McGirt MJ, Speroff T, Dittus RS, *et al.* The National Neurosurgery Quality and Outcomes Database (N2QOD): general overview and pilot-year project description. *Neurosurg Focus* 2013;**34**:E6. doi:10.3171/2012.10.FOCUS12297

3 Hoque DME, Kumari V, Hoque M, *et al.* Impact of clinical registries on quality of patient care and clinical outcomes: A systematic review. *PLoS One* 2017;**12**:1–20. doi:10.1371/journal.pone.0183667

4 Sattar MH, Guthrie ST. *Anatomy, Back, Sacral Vertebrae*. 2020. http://www.ncbi.nlm.nih.gov/pubmed/31869117

5 Haefeli M, Kalberer F, Saegesser D, *et al.* The Course of Macroscopic Degeneration in the Human Lumbar Intervertebral Disc. *Spine (Phila Pa 1976)* 2006;**31**:1522–31. doi:10.1097/01.brs.0000222032.52336.8e

6 Valat J-P, Genevay S, Marty M, *et al.* Sciatica. *Best Pract Res Clin Rheumatol* 2010;**24**:241–52. doi:10.1016/j.berh.2009.11.005

7 Yin M, Mo W, Wu H, *et al.* Efficacy of Caudal Epidural Steroid Injection with Targeted Indwelling Catheter and Manipulation in Managing Patients with Lumbar Disk Herniation and Radiculopathy: A Prospective, Randomized, Single-Blind Controlled Trial. *World Neurosurg* 2018;**114**:e29–34. doi:10.1016/j.wneu.2018.01.162

8 Goupille P, Jayson MI, Valat JP, *et al.* The role of inflammation in disk herniation-associated radiculopathy. *Semin Arthritis Rheum* 1998;**28**:60–71. doi:10.1016/s0049-0172(98)80029-2

9 Mulleman D, Mammou S, Griffoul I, *et al.* Pathophysiology of disk-related sciatica. I.—Evidence supporting a chemical component. *Jt Bone Spine* 2006;**73**:151–8. doi:10.1016/j.jbspin.2005.03.003

10 Molinos M, Almeida CR, Caldeira J, *et al.* Inflammation in intervertebral disc degeneration and regeneration. *J R Soc Interface* 2015;**12**:20141191. doi:10.1098/rsif.2014.1191

11 Burke JG, Watson RWG, McCormack D, *et al.* Intervertebral discs which cause low back pain secrete high levels of proinflammatory mediators. *J Bone Joint Surg Br* 2002;**84**:196–201. doi:10.1302/0301-620x.84b2.12511

12 Deyo RA, Mirza SK. Herniated Lumbar Intervertebral Disk. *N Engl J Med* 2016;**374**:1763–72. doi:10.1056/NEJMcp1512658

13 Jensen LD, Frost P, Schiøttz-Christensen B, *et al.* Predictors of vocational prognosis after herniated lumbar disc: a two-year follow-up study of 2039 patients diagnosed at hospital. *Spine (Phila Pa 1976)* 2011;**36**:E791-7. doi:10.1097/BRS.0b013e3181ef6243

14 Sambrook PN, MacGregor AJ, Spector TD. Genetic influences on cervical and lumbar disc degeneration: A magnetic resonance imaging study in twins. *Arthritis Rheum* 1999;**42**:366–72. doi:10.1002/1529-0131(199902)42:2<366::AID-ANR20>3.0.CO;2-6

15 Belavy DL, Adams M, Brisby H, *et al.* Disc herniations in astronauts: What causes them, and what does it tell us about herniation on earth? *Eur Spine J* 2016;**25**:144–54. doi:10.1007/s00586-015-3917-y

16 Violante FS, Zompatori M, Lovreglio P, *et al.* Is age more than manual material handling associated with lumbar vertebral body and disc changes? A cross-sectional multicentre MRI study. *BMJ Open* 2019;**9**:e029657. doi:10.1136/bmjopen-2019-029657

17 Miyamoto H, Saura R, Doita M, *et al.* The role of cyclooxygenase-2 in lumbar disc herniation. *Spine (Phila Pa 1976)* 2002;**27**:2477–83. doi:10.1097/00007632-200211150-00011

18 Konstantinou K, Dunn KM. Sciatica: review of epidemiological studies and prevalence estimates. *Spine (Phila Pa 1976)* 2008;**33**:2464–72. doi:10.1097/BRS.0b013e318183a4a2

19 Peul WC, van Houwelingen HC, van den Hout WB, *et al.* Surgery versus prolonged conservative treatment for sciatica. *N Engl J Med* 2007;**356**:2245–56. doi:10.1056/NEJMoa064039

20 Bailey CS, Rasoulinejad P, Taylor D, *et al.* Surgery versus conservative care for persistent sciatica lasting 4 to 12 months. *N Engl J Med* 2020;**382**:1093–102. doi:10.1056/NEJMoa1912658

21 Grøvle L, Fjeld OR, Haugen AJ, *et al.* The Rates of LSS Surgery in Norwegian Public Hospitals: A Threefold Increase From 1999 to 2013. *Spine (Phila Pa 1976)* 2019;**44**:E372–8. doi:10.1097/BRS.0000000000002858

22 Grotle M, Småstuen MC, Fjeld O, *et al.* Lumbar spine surgery across 15 years: Trends, complications and reoperations in a longitudinal observational study from Norway. *BMJ Open* 2019;**9**:1–7. doi:10.1136/bmjopen-2018-028743

23 Van Der Windt DAWM, Simons E, Riphagen I, *et al.* Physical examination for lumbar radiculopathy due to disc herniation in patients with low-back pain. *Cochrane Database Syst Rev* Published Online First: 2008. doi:10.1002/14651858.CD007431

24 Bono CM, Ghiselli G, Gilbert TJ, *et al. An evidence-based clinical guideline for the diagnosis and treatment of cervical radiculopathy from degenerative disorders.* 2011. doi:10.1016/j.spinee.2010.10.023

25 Kim J-H, van Rijn RM, van Tulder MW, *et al.* Diagnostic accuracy of diagnostic imaging for lumbar disc herniation in adults with low back pain or sciatica is unknown; a systematic review. *Chiropr Man Therap* 2018;**26**:37. doi:10.1186/s12998-018-0207-x

26 Kim SJ, Lee TH, Lim SM. Prevalence of disc degeneration in asymptomatic Korean subjects. Part 1: Lumbar spine. *J Korean Neurosurg Soc* 2013;**53**:31–8. doi:10.3340/jkns.2013.53.1.31

27 Brinjikji W, Luetmer PH, Comstock B, *et al.* Systematic literature review of imaging features of spinal degeneration in asymptomatic populations. *AJNR Am J Neuroradiol* 2015;**36**:811–6. doi:10.3174/ajnr.A4173

28 Weber H, Holme I, Amlie E. The natural course of acute sciatica with nerve root symptoms in a double-blind placebo-controlled trial evaluating the effect of piroxicam. 1993. doi:10.1097/00007632-199318110-00006

29 Jacobs WCH, van Tulder M, Arts M, *et al.* Surgery versus conservative management of sciatica due to a lumbar herniated disc: a systematic review. *Eur Spine J* 2011;**20**:513–22. doi:10.1007/s00586-010-1603-7

30 Atenello F, Hsieh P. Lumbar Microdiskectomy: Midline Open and Far-Lateral Techniques. In: *Surgical Anatomy & Techniques to the Spine 2nd Edition*. 2013. 404–11.

31 Truumees E, Geck M, Stokes JK, *et al.* Lumbar Microdiscectomy. *JBJS Essent Surg Tech* 2016;**6**:e3. doi:10.2106/JBJS.ST.N.00093

32 Zhang Y, Chong F, Feng C, *et al.* Comparison of Endoscope-Assisted and Microscope-Assisted Tubular Surgery for Lumbar Laminectomies and Discectomies: Minimum 2-Year Follow-Up Results. *Biomed Res Int* 2019;**2019**:1–7. doi:10.1155/2019/5321580

33 Gibson JN, Waddell G. Surgical interventions for lumbar disc prolapse.

CochraneDatabaseSystRev 2007;:CD001350.

34 Kamper SJ, Ostelo RWJG, Rubinstein SM, *et al.* Minimally invasive surgery for lumbar disc herniation: A systematic review and meta-analysis. *Eur Spine J* 2014;**23**:1021–43. doi:10.1007/s00586-013-3161-2

35 Couto JMC, Castilho EA de, Menezes PR. Chemonucleolysis in lumbar disc herniation: a meta-analysis. *Clinics (Sao Paulo)* 2007;**62**:175–80. doi:10.1590/s1807-59322007000200013

36 Goupille P, Mulleman D, Mammou S, *et al.* Percutaneous laser disc decompression for the treatment of lumbar disc herniation: a review. *Semin Arthritis Rheum* 2007;**37**:20–30. doi:10.1016/j.semarthrit.2007.01.006

37 Kennedy DJ, Zheng PZ, Smuck M, *et al.* A minimum of 5-year follow-up after lumbar transforaminal epidural steroid injections in patients with lumbar radicular pain due to intervertebral disc herniation. *Spine J* 2018;**18**:29–35. doi:10.1016/j.spinee.2017.08.264

38 Lavy C, James A, Wilson-MacDonald J, *et al.* Cauda equina syndrome. *BMJ* 2009;**338**:b936–b936. doi:10.1136/bmj.b936

39 Kapetanakis S, Chaniotakis C, Kazakos C, *et al.* Cauda Equina Syndrome Due to Lumbar Disc Herniation: a Review of Literature. *Folia Med (Plovdiv)* 2017;**59**:377–86. doi:10.1515/folmed-2017-0038

40 Sharma H, Lee SWJ, Cole AA. The management of weakness caused by lumbar and lumbosacral nerve root compression. *J Bone Joint Surg Br* 2012;**94**:1442–7. doi:10.1302/0301-620X.94B11.29148

41 Lønne G, Solberg TK, Sjaavik K, *et al.* Recovery of muscle strength after microdiscectomy for lumbar disc herniation: A prospective cohort study with 1-year follow-up. *Eur Spine J*

2012;**21**:655–9. doi:10.1007/s00586-011-2122-x

42 Lurie JD, Tosteson TD, Tosteson ANA, *et al.* Surgical Versus Nonoperative Treatment for Lumbar Disc Herniation. *Spine (Phila Pa 1976)* 2014;**39**:3–16. doi:10.1097/BRS.0000000000000088

43 Solberg TK, Olsen LR. NORspine annual report 2015 [Nasjonalt kvalitetsregister for ryggkirurgi (NKR): Årsrapport for 2015 med plan for forbedringstiltak 2016]. 2016.

44 Gliklich RE, Dreyer NA. Registries for Evaluating Patient Outcomes, 3rd edition. 2014. Chapter 3, section 10.https://www.ncbi.nlm.nih.gov/books/NBK208616/

45 van Hooff ML, Jacobs WCH, Willems PC, *et al.* Evidence and practice in spine registries. *Acta Orthop* 2015;**86**:1–11. doi:10.3109/17453674.2015.1043174

46 McGirt MJ, Parker SL, Asher AL, *et al.* Role of prospective registries in defining the value and effectiveness of spine care. *Spine (Phila Pa 1976)* 2014;**39**:S117-28. doi:10.1097/BRS.0000000000000552

47 Asher AL, Devin CJ, Mroz T, *et al.* Clinical Registries and Evidence-Based Care Pathways. *Spine (Phila Pa 1976)* 2014;**39**:S136–8. doi:10.1097/BRS.0000000000000543

48 Mroz TE, McGirt M, Chapman JR, *et al.* More “Why” and Less “How”. *Spine (Phila Pa 1976)* 2014;**39**:S7–8. doi:10.1097/BRS.0000000000000539

49 Dworkin RH, Turk DC, Wyrwich KW, *et al.* Interpreting the Clinical Importance of Treatment Outcomes in Chronic Pain Clinical Trials: IMMPACT Recommendations. *J Pain* 2008;**9**:105–21. doi:10.1016/j.jpain.2007.09.005

50 McLeod LD, Coon CD, Martin SA, *et al.* Interpreting patient-reported outcome results: US FDA guidance and emerging methods. *Expert Rev Pharmacoecon Outcomes Res* 2011;**11**:163–9.

doi:10.1586/erp.11.12

51 Lee CK, Hansen HT, Weiss AB. Developmental Lumbar Spinal Stenosis. *Spine (Phila Pa 1976)* 1978;**3**:246–55. doi:10.1097/00007632-197809000-00010

52 Finkelstein JA, Schwartz CE. Patient-reported outcomes in spine surgery: Past, current, and future directions. *J Neurosurg Spine* 2019;**31**:155–64. doi:10.3171/2019.1.SPINE18770

53 Guzman JZ, Cutler HS, Connolly J, *et al.* Patient-Reported Outcome Instruments in Spine Surgery. *Spine (Phila Pa 1976)* 2016;**41**:429–37. doi:10.1097/BRS.0000000000001211

54 Nilsson E, Orwelius L, Kristenson M. Patient-reported outcomes in the Swedish National Quality Registers. *J Intern Med* 2016;**279**:141–53. doi:10.1111/joim.12409

55 FDA. Patient-Reported Outcomes (PROs) in Medical Device Decision Making. https://www.fda.gov/about-fda/cdrh-patient-engagement/patient-reported-outcomes-pros-medical-device-decision-making

56 Johnston BC, Patrick DL, Busse JW, *et al.* Patient-reported outcomes in meta-analyses - Part 1: Assessing risk of bias and combining outcomes. *Health Qual Life Outcomes* 2013;**11**:1–10. doi:10.1186/1477-7525-11-109

57 Wyrwich KW, Norquist JM, Lenderking WR, *et al.* Methods for interpreting change over time in patient-reported outcome measures. *Qual Life Res* 2013;**22**:475–83. doi:10.1007/s11136-012-0175-x

58 Fairbank JCT. Why are there different versions of the Oswestry Disability Index? *J Neurosurg Spine* 2014;**20**:83–6. doi:10.3171/2013.9.SPINE13344

59 Solberg TK, Olsen JA, Ingebrigtsen T, *et al.* Health-related quality of life assessment by the EuroQol-5D can provide cost-utility data in the field of low-back surgery. *EurSpine J*

2005;**14**:1000–7.

60 MP J, Karoly P. Self-report Scales and Procedures for Assessing Pain in Adults. In: DC T, Melzack R, eds. *Handbook of Pain Assessment*. New York: : The Guilford Press 1992. 135–51.

61 Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control ClinTrials* 1989;**10**:407–15.

62 Copay AG, Subach BR, Glassman SD, *et al.* Understanding the minimum clinically important difference: a review of concepts and methods. *Spine J* 2007;**7**:541–6.

63 Wright A, Hannon J, Hegedus EJ, *et al.* Clinimetrics corner: a closer look at the minimal clinically important difference (MCID). *J Man Manip Ther* 2012;**20**:160–6. doi:10.1179/2042618612Y.0000000001

64 Kamper SJ, Ostelo RW, Knol DL, *et al.* Global Perceived Effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status. *JClinEpidemiol* 2010;**63**:760–6.

65 Grøvle L, Haugen AJ, Hasvik E, *et al.* Patients' ratings of global perceived change during 2 years were strongly influenced by the current health status. *J Clin Epidemiol* 2014;**67**:508–15. doi:10.1016/j.jclinepi.2013.12.001

66 Glassman SD, Copay AG, Berven SH, *et al.* Defining substantial clinical benefit following lumbar spine arthrodesis. *JBone Jt SurgAm* 2008;**90**:1839–47.

67 Solberg T, Johnsen LG, Nygaard OP, *et al.* Can we define success criteria for lumbar disc surgery? Estimates for a substantial amount of improvement in core outcome measures. *Acta Orthop* 2013;**84**:196–201.

68 Tubach F, Dougados M, Falissard B, *et al.* Feeling good rather than feeling better matters

more to patients. *Arthritis Care Res* 2006;**55**:526–30. doi:10.1002/art.22110

69 van Hooff ML, Mannion AF, Staub LP, *et al.* Determination of the Oswestry Disability Index score equivalent to a 'satisfactory symptom state' in patients undergoing surgery for degenerative disorders of the lumbar spine-a Spine Tango registry-based study. *Spine J* 2016;**16**:1221–30. doi:10.1016/j.spinee.2016.06.010

70 Hemingway H, Croft P, Hayden JA, *et al.* Prognosis research strategy (PROGRESS) 1 : A. 2013;**5595**:1–11. doi:10.1136/bmj.e5595

71 Moons KG, Royston P, Vergouwe Y, *et al.* Prognosis and prognostic research: what, why, and how? *BMJ* 2009;**338:b375.**:b375.

72 Mahmood SS, Levy D, Vasan RS, *et al.* The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet (London, England)* 2014;**383**:999–1008. doi:10.1016/S0140-6736(13)61752-3

73 Riley RD, Hayden JA, Steyerberg EW, *et al.* Prognosis Research Strategy (PROGRESS) 2 : Prognostic Factor Research. 2013;**10**. doi:10.1136/bmj.e5595

74 Blumenthal D. Part 1: Quality of care--what is it? N. Engl. J. Med. 1996;**335**:891–4. doi:10.1056/NEJM199609193351213

75 Lønne G, Schoenfeld AJ, Cha TD, *et al.* Variation in selection criteria and approaches to surgery for Lumbar Spinal Stenosis among patients treated in Boston and Norway. *Clin Neurol Neurosurg* 2017;**156**:77–82. doi:10.1016/j.clineuro.2017.03.008

76 van Hooff ML, Jacobs WCH, Willems PC, *et al.* Evidence and practice in spine registries: A systematic review, and recommendations for future design of registries. *Acta Orthop* 2015;**86**:1–11. doi:10.3109/17453674.2015.1043174

77 Mancuso CA, Duculan R, Stal M, *et al.* Patients expectations of lumbar spine surgery. *Eur Spine J* 2014;**24**:2362–9. doi:10.1007/s00586-014-3597-z

78 Deyo RA, Mirza SK. The case for restraint in spinal surgery: Does quality management have a role to play? *EurSpine J* 2009;**18**:331–7. doi:10.1007/s00586-009-0908-x

79 Fjeld OR, Grøvle L, Helgeland J, *et al.* Complications, reoperations, readmissions, and length of hospital stay in 34 639 surgical cases of lumbar disc herniation. *Bone Joint J* 2019;**101**-**B**:470–7. doi:10.1302/0301-620X.101B4.BJJ-2018-1184.R1

80 Solberg TK, Olsen LR. Nasjonalt kvalitetsregister for ryggkirurgi (NKR): Årsrapport for 2015 med plan for forbedringstiltak 2016. 2016.

81 Dahl FA, Grotle M, Šaltyte Benth J, *et al.* Data splitting as a countermeasure against hypothesis fishing: With a case study of predictors for low back pain. *Eur J Epidemiol* 2008;**23**:237–42. doi:10.1007/s10654-008-9230-x

82 Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett* 2006;**27**:861–74. doi:10.1016/j.patrec.2005.10.010

83 Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med* 2014;**33**:517–35. doi:10.1002/sim.5941

84 Royston P, Moons KGM, Altman DG, *et al.* Prognosis and prognostic research: Developing a prognostic model. *BMJ* 2009;**338**:b604. doi:10.1136/bmj.b604

85 Ruyter KW. REK sør-øst Knut W. Ruyter 22845518 24.06.2015. 2015.

86 Jaeschke R, Singer J, Guyatt GH. Measurement of health status. *Control Clin Trials* 1989;**10**:407–15. doi:10.1016/0197-2456(89)90005-6

87 Hassan ES. Recall Bias can be a Threat to Retrospective and Prospective Research Designs. *Internet J Epidemiol* 2005;**3**:1–7.

88 Campbell H, Rivero-Arias O, Johnston K, *et al.* Responsiveness of objective, disease-specific, and generic outcome measures in patients with chronic low back pain: an assessment for improving, stable, and deteriorating patients. *Spine (Phila Pa 1976)* 2006;**31**:815–22. doi:10.1097/01.brs.0000207257.64215.03

89 de Vet HC, Ostelo RW, Terwee CB, *et al.* Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. *QualLife Res* 2007;**16**:131–42.

90 Terwee CB, Bot SD, De Boer MR, *et al.* Quality criteria were proposed for measurement properties of health status questionnaires. *JClinEpidemiol* 2007;**60**:34–42.

91 Devji T, Carrasco-Labra A, Qasim A, *et al.* Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. *BMJ* 2020;:m1714. doi:10.1136/bmj.m1714

92 Wilson IB. Clinical understanding and clinical implications of response shift. *Soc Sci Med* 1999;**48**:1577–88. doi:10.1016/S0277-9536(99)00050-7

93 Witiw CD, Mansouri A, Mathieu F, *et al.* Exploring the expectation-actuality discrepancy: a systematic review of the impact of preoperative expectations on satisfaction and patient reported outcomes in spinal surgery. *Neurosurg Rev* 2018;**41**:19–30. doi:10.1007/s10143-016-0720-0

94 Kleinstuck FS, Grob D, Lattig F, *et al.* The influence of preoperative back pain on the outcome of lumbar decompression surgery. *Spine (Phila Pa 1976)* 2009;**34**:1198–203.

95 Ostelo RWJG, Deyo RA, Stratford P, *et al.* Interpreting Change Scores for Pain and Functional Status in Low Back Pain. *Spine (Phila Pa 1976)* 2008;**33**:90–4. doi:10.1097/BRS.0b013e31815e3a10

96 Copay AG, Martin MM, Subach BR, *et al.* Assessment of spine surgery outcomes: inconsistency of change amongst outcome measurements. *Spine J* 2010;**10**:291–6. doi:10.1016/j.spinee.2009.12.027

97 Grotle M, Brox JI, Vøllestad NK. Cross-cultural adaptation of the Norwegian versions of the Roland-Morris Disability Questionnaire and the Oswestry Disability Index. *J Rehabil Med* 2003;**35**:241–7.http://www.ncbi.nlm.nih.gov/pubmed/14582557

98 Asher AM, Oleisky ER, Pennings JS, *et al.* Measuring clinically relevant improvement after lumbar spine surgery: is it time for something new? *Spine J* 2020;**20**:847–56. doi:10.1016/j.spinee.2020.01.010

99 Austevoll IM, Gjestad R, Grotle M, *et al.* Follow-up score, change score or percentage change score for determining clinical important outcome following surgery? An observational study from the Norwegian registry for Spine surgery evaluating patient reported outcome measures in lumbar spinal steno. *BMC Musculoskelet Disord* 2019;**20**:31. doi:10.1186/s12891-018-2386-y

100 Tubach F, Dougados M, Falissard B, *et al.* Feeling Good Rather Than Feeling Better Matters More to Patients. 2006;**55**:526–30. doi:10.1002/art.22110

101 Draak THP, de Greef BTA, Faber CG, *et al.* The minimum clinically important difference: which direction to take. *Eur J Neurol* 2019;**26**:850–5. doi:10.1111/ene.13941

102 Harbaugh RE. Artificial neural networks for neurosurgical diagnosis, prognosis, and management. *JNS* 2018;**45**:20–1. doi:10.3171/2018.8.FOCUS17773.

103 Naggara O, Raymond J, Guilbert F, *et al.* Analysis by categorizing or dichotomizing continuous variables is inadvisable: An example from the natural history of unruptured aneurysms. *Am J Neuroradiol* 2011;**32**:437–40. doi:10.3174/ajnr.A2425

104 Moons KGM, Royston P, Vergouwe Y, *et al.* Prognosis and prognostic research: What, why, and how? *BMJ* 2009;**338**:1317–20. doi:10.1136/bmj.b375

105 Katz MH. Multivariable Analysis: A Primer for Readers of Medical Research. *Ann Intern Med* 2003;**138**:644. doi:10.7326/0003-4819-138-8-200304150-00012

106 Pieber K, Salomon N, Inschlag S, *et al.* Predictors of an unfavorable outcome 1 . 5 and 12 years after a first , uncomplicated lumbar disc surgery. 2016;:3520–7. doi:10.1007/s00586-016-4700-4

107 Hebert JJ, Fritz JM, Koppenhaver SL, *et al.* Predictors of clinical outcome following lumbar disc surgery: the value of historical, physical examination, and muscle function variables. *Eur Spine J* 2016;**25**:310–7. doi:10.1007/s00586-015-3916-z

108 Koerner JD, Glaser J. Which Variables Are Associated With Patient-reported Outcomes After Discectomy ? Review of SPORT Disc Herniation Studies. *Clin Orthop Relat Res* 2015;:2000–6. doi:10.1007/s11999-014-3671-1

109 Fjeld O, Grotle M, Siewers V, *et al.* Prognostic Factors for Persistent Leg-Pain. 2017;**42**. doi:10.1097/BRS.0000000000001773

110 Madsbu MA, Salvesen Ø, Werner DAT, *et al.* Surgery for Herniated Lumbar Disc in Daily Tobacco Smokers: A Multicenter Observational Study. *World Neurosurg* 2018;**109**:e581–7. doi:10.1016/j.wneu.2017.10.024

111 Zehnder P, Aghayev E, Fekete TF, *et al.* Influence of previous surgery on patient-rated

outcome after surgery for degenerative disorders of the lumbar spine. *Eur Spine J* 2016;**25**:2553–62. doi:10.1007/s00586-016-4383-x

112 Haugen AJ, Brox JI, Grovle L, *et al.* Prognostic factors for non-success in patients with sciatica and disc herniation. *BMCMusculoskeletDisord* 2012;**13:183. do**:113–83.

113 Kleinstueck FS, Fekete T, Jeszenszky D, *et al.* The outcome of decompression surgery for lumbar herniated disc is influenced by the level of concomitant preoperative low back pain. *Eur Spine J* 2011;**20**:1166–73. doi:10.1007/s00586-010-1670-9

114 Strömqvist F, Strömqvist B, Jönsson B, *et al.* The outcome of lumbar disc herniation surgery is worse in old adults than in young adults. *Acta Orthop* 2016;**87**:516–21. doi:10.1080/17453674.2016.1205173

115 Strömqvist F, Strömqvist B, Jönsson B, *et al.* Outcome of surgical treatment of lumbar disc herniation in young individuals. *Bone Jt J* 2015;**97B**:1675–82. doi:10.1302/0301-620X.97B12.36258

116 McKillop AB, Carroll LJ, Battié MC. Depression as a prognostic factor of lumbar spinal stenosis: a systematic review. *Spine J* 2014;**14**:837–46. doi:10.1016/j.spinee.2013.09.052

117 Ablin JN, Berman M, Aloush V, *et al.* Effect of Fibromyalgia Symptoms on Outcome of Spinal Surgery. *Pain Med* 2016;:pnw232. doi:10.1093/pm/pnw232

118 Mannion AF, Elfering A. Predictors of surgical outcome and their assessment. *EurSpine J* 2006;**15 Suppl 1**:S93-108. doi:10.1007/s00586-005-1045-9

119 Strömqvist F, Strömqvist B, Jönsson B, *et al.* Inferior Outcome of Lumbar Disc Surgery in Women Due to Inferior Preoperative Status. *Spine (Phila Pa 1976)* 2016;**41**:1247–52. doi:10.1097/BRS.0000000000001492

120 Masui T, Yukawa Y, Nakamura S, *et al.* Natural history of patients with lumbar disc herniation observed by magnetic resonance imaging for minimum 7 years. *J Spinal Disord Tech* 2005;**18**:121–6. doi:10.1097/01.bsd.0000154452.13579.b2

121 Vroomen PCAJ, Wilmink JT, De KMCTFM. Prognostic value of MRI findings in sciatica. *Neuroradiology* 2002;**44**:59–63. doi:10.1007/s002340100650

122 Ware JH, Ph D, Harrington D, *et al.* Missing Data. *N Engl J Med* 2012;:1353–4. doi:10.1056/NEJMsm1210043

123 Nunan D, Aronson J, Bankhead C. Catalogue of bias : attrition bias. *BMJ* 2018;**23**:21–2.

124 Hojmark K, Stottrup C, Carreon L, *et al.* Patient-reported outcome measures unbiased by loss of follow-up. Single-center study based on DaneSpine, the Danish spine surgery registry. *Eur spine J Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc* 2015;:282–6. doi:10.1007/s00586-015-4127-3

125 Solberg TK, Sorlie A, Sjaavik K, *et al.* Would loss to follow-up bias the outcome evaluation of patients operated for degenerative disorders of the lumbar spine?: A study of responding and non-responding cohort participants from a clinical spine surgery registry. *Acta Orthop* 2011;**82**:56–63. doi:10.3109/17453674.2010.548024

126 Elkan P, Möller TLH, Gerdhem P. Response rate does not affect patient - reported outcome after lumbar discectomy. *Eur Spine J* 2018;**27**:1538–46. doi:10.1007/s00586-018-5541-0

127 Asher AL, Chotai S, Devin CJ, *et al.* Inadequacy of 3-month oswestry disability index outcome for assessing individual longer-term patient experience after lumbar spine surgery. *J Neurosurg Spine* 2016;**25**:170–80. doi:10.3171/2015.11.SPINE15872

128 Parker SL, Asher AL, Godil SS, *et al.* Patient-reported outcomes 3 months after spine surgery:

is it an accurate predictor of 12-month outcome in real-world registry platforms? *Neurosurg Focus* 2015;**39**:E17. doi:10.3171/2015.9.FOCUS15356

129 Twisk J, De Vente W. Attrition in longitudinal studies: How to deal with missing data. *J Clin Epidemiol* 2002;**55**:329–37. doi:10.1016/S0895-4356(01)00476-0

130 McGirt MJ, Resnick D, Edwards N, *et al.* Background to understanding value-based surgical spine care. *Spine (Phila Pa 1976)* 2014;**39**:S51–2. doi:10.1097/BRS.0000000000000544

131 Resnick DK, Tosteson AN a., Groman RF, *et al.* Setting the Equation. *Spine (Phila Pa 1976)* 2014;**39**:S43–50. doi:10.1097/BRS.0000000000000581

132 Gutacker N, Street A. Use of large-scale HRQoL datasets to generate individualised predictions and inform patients about the likely benefit of surgery. *Qual Life Res* 2017;**26**:2497–505. doi:10.1007/s11136-017-1599-0

133 Perez-Breva L, Shin JH. Artificial Intelligence in Neurosurgery: A Comment on the Possibilities. *Neurospine* 2019;**16**:640–2. doi:10.14245/ns.1938404.202

134 Panesar SS, Kliot M, Parrish R, *et al.* Promises and Perils of Artificial Intelligence in Neurosurgery. *Neurosurgery* 2020;**87**:33–44. doi:10.1093/neuros/nyz471

15 Papers

15.1 Paper I.

Eur Spine J
DOI 10.1007/s00586-017-5185-5

ORIGINAL ARTICLE

Criteria for failure and worsening after surgery for lumbar disc herniation: a multicenter observational study based on data from the Norwegian Registry for Spine Surgery

David A. T. Werner[1,2] · Margreth Grotle[3,4] · Sasha Gulati[5] · Ivar M. Austevoll[6] · Greger Lønne[7] · Øystein P. Nygaard[5,8] · Tore K. Solberg[1,2,8]

Received: 6 November 2016 / Revised: 10 May 2017 / Accepted: 7 June 2017

Abstract

Purpose In clinical decision-making, it is crucial to discuss the probability of adverse outcomes with the patient. A large proportion of the outcomes are difficult to classify as either failure or success. Consequently, cutoff values in patient-reported outcome measures (PROMs) for "failure" and "worsening" are likely to be different from those of "non-success". The aim of this study was to identify dichotomous cutoffs for failure and worsening, 12 months after surgical treatment for lumbar disc herniation, in a large registry cohort.

Methods A total of 6840 patients with lumbar disc herniation were operated and followed for 12 months, according to the standard protocol of the Norwegian Registry for Spine Surgery (NORspine). Patients reporting to be unchanged or worse on the Global Perceived Effectiveness (GPE) scale at 12-month follow-up were classified as "failure", and those considering themselves "worse" or "worse than ever" after surgery were classified as "worsening". These two dichotomous outcomes were used as anchors in analyses of receiver operating characteristics (ROC) to define cutoffs for failure and worsening on commonly used PROMs, namely, the Oswestry Disability Index (ODI), the EuroQuol 5D (EQ-5D), and Numerical Rating Scales (NRS) for back pain and leg pain.

Results "Failure" after 12 months for each PROM, as an insufficient improvement from baseline, was (sensitivity and specificity): ODI change <13 (0.82, 0.82), ODI% change <33% (0.86, 0.86), ODI final raw score >25 (0.89, 0.81), NRS back-pain change <1.5 (0.74, 0.86), NRS back-pain % change <24 (0.85, 0.81), NRS back-pain final raw score >5.5 (0.81, 0.87), NRS leg-pain change <1.5 (0.81, 0.76), NRS leg-pain % change <39 (0.86, 0.81), NRS leg-pain final raw score >4.5 (0.91, 0.85), EQ-5D change <0.10 (0.76, 0.83), and EQ-5D final raw score >0.63 (0.81, 0.85). Both a final raw score >48 for the ODI and an NRS >7.5 were indicators for "worsening" after 12 months, with acceptable accuracy.

Conclusion The criteria with the highest accuracy for defining failure and worsening after surgery for lumbar disc herniation were an ODI percentage change score <33% for failure and a 12-month ODI raw score >48. These cutoffs can facilitate shared decision-making among doctors and patients, and improve quality assessment and comparison

Electronic supplementary material The online version of this article (doi:10.1007/s00586-017-5185-5) contains supplementary material, which is available to authorized users.

✉ David A. T. Werner
Mail@david-werner.com

[1] Department of Neurosurgery, University Hospital of Northern Norway, Tromsø, Norway

[2] Department of Clinical Medicine, University of Tromsø, Tromsø, Norway

[3] Institute of Physiotherapy, Faculty of Health Sciences, Oslo and Akershus University College of Applied Sciences, Oslo, Norway

[4] Communication Unit for Musculoskeletal Disorders (FORMI), Oslo University Hospital, Oslo, Norway

[5] Department of Neurosurgery, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway

[6] Kysthospitalet in Hagevik, Orthopedic Clinic, Haukeland University Hospital, Bergen, Norway

[7] Department of Orthopedic Surgery, Innlandet Hospital Trust, Lillehammer, Norway

[8] The Norwegian Registry for Spine Surgery (NORspine), Northern Norway Regional Health Authority, Tromsø, Norway

Published online: 14 June 2017

of clinical outcomes across surgical units. In addition to clinically relevant improvements, we propose that rates of failure and worsening should be included in reporting from clinical trials.

Keywords Lumbar disc surgery outcome · Failure · Worsening · Spine registry · Patient-reported outcome measures

Introduction

In spine surgery, several well-validated patient-reported outcome measures (PROMs) have been recommended, such as the Oswestry Disability Index (ODI) [1], Numerical Rating Scale (NRS) for leg pain and back pain [2], and the EuroQol 5D (EQ-5D) [3]. Still, clinicians are often unfamiliar with their interpretation. In large cohorts, even small and clinically irrelevant PROM changes tend to reach statistical significance [4]. To provide cutoffs on PROM changes that are perceived as meaningful and important by the patients, the "minimal important change" (MIC) has been defined by various methods [5–7]. A recent review proposed an MIC cutoff for the ODI of ten points, or 30% improvement from baseline [8]. Several studies have identified MIC cutoffs for the NRS back pain and leg pain from 2 to 2.5 [8, 9]. In addition, cutoffs for substantial clinical improvements, such as "success" after lumbar disc surgery, have been reported both for the ODI (20), NRS back pain (2.5), NRS leg pain (3.5), and EQ-5D (0.3) [9–11]. A large proportion of the patients are difficult to classify as either improved, unchanged, or worse after surgery [12]. Consequently, cutoffs on the PROMs for deterioration and "failure" may be different from those of "non-success". Previously, authors have used various methods and different concepts for defining cutoffs for clinical meaningful improvements [10, 12], resulting in a diversity of recommended threshold values [8, 21, 22]. This makes it even more difficult to disentangle "failure" from constructs developed to identify improvements. There is clearly a grey zone between "failure" and "non-success" [13], "minimal meaningful improvements", or a "satisfactory symptom state" [14]. Using an external anchor method to define "failure" more accurately could provide more robust definitions of this outcome category [11]. It is, therefore, important to differentiate between "failure" and "non-success".

The indication for operative treatment of lumbar disc herniation is relative, and the decision to operate must be based on a trade-off between possible benefits, risks, and costs [15]. In clinical trials, focus is generally placed on improvements such as "success rates". To enhance quality assessment and shared decision-making, it is crucial to consider the other end of the scale and to discuss the possibility of adverse outcomes with the patients. Avoiding inefficient operations may have a greater impact on treatment outcomes, than improving surgical technique [16]. The first step would be to try to define cutoffs for "failure" and "worsening" on the PROMs. When informing the patient about possible outcomes, we think that it is important to differentiate between being unchanged after surgery, which might be an acceptable risk, and actually getting worse, which might be harmful. Previous studies show that larger cohorts are needed to clearly define clinically meaningful thresholds for such outcomes, especially for worsening [17, 18].

The Norwegian Registry for Spine Surgery (NorSpine) collects clinical data (PROMs) on the majority of patients operated for lumbar disc herniation in Norway. Its purpose is to evaluate treatment outcomes from the "real life" of daily clinical practice and use this information to improve the quality of the health services [19, 20]. The aim of this study was to estimate the most accurate cutoffs for both failure and worsening after surgical treatment of lumbar disc herniation, using data from the large registry cohort of the NORspine. Such benchmark criteria could be used for calculating sample size in research and facilitate shared decision-making among doctors and patients, clinical audit, and comparisons of outcomes across surgical units.

Methods

Patient population and data collection

6840 patients operated for lumbar disc herniation between January 1st, 2007 and February 28th, 2014 were followed for 12 months, according to the standard NORspine protocol. The NORspine is a comprehensive clinical registry for quality control and research. Both emergency and elective cases are registered. We included all patients who were treated for lumbar disc herniation with lumbar discectomy and/or herniectomy. Fusion procedures or laminectomy with removal of midline structures were not included. Table 1 describes the exclusion criteria in the current study. This study comprises 38 of 40 (95%) Norwegian private and public centers, performing surgery for degenerative spinal disorders. The inclusion rate for lumbar disc herniation is currently about 65% in the NORspine.

Informed consent was obtained from all patients and participation was neither mandatory, nor required to gain access to healthcare. According to Norwegian legislation, patients over the age of 15 can independently consent to participation in the registry. The registry protocol has been approved by the Data Inspectorate of Norway. This study was submitted to the regional ethical committee for

Table 1 NORspine exclusion criteria

• Patients unable to give informed consent due to cognitive deficits or reduced consciousness
• Children <16 years
• Patients with serious drug abuse or severe psychiatric disorders
• Patients with fractures, primary infections or malignant conditions in the spine
• Patients unable to respond to the declaration of consent and/or the questionnaires due to language barriers

medical research which categorized it as a clinical audit study, not in need of their formal approval [21].

At admission for surgery, the patients completed a baseline questionnaire on demographics, lifestyle issues, and PROMs. During the hospital stay, the surgeon recorded data concerning diagnosis, treatment, and comorbidity on a standard registration form. Twelve months after surgery, a questionnaire was distributed by regular post, completed at home by the patients, and returned to the central registry unit without involvement of the treating hospitals. One reminder with a new copy of the questionnaire was sent to those who did not respond.

Patient-reported outcome measures

The PROM questionnaires were identical at baseline and follow-up. The ODI version 2.0 was used to assess pain-related disability. It contains ten questions on limitations of activities of daily living. Each item is rated 0–5 and then transferred into a percentage score ranging from 0 (none) to 100 (maximum pain-related disability) [1].

Pain was reported on the numerical rating scale of 0–10 for both back pain (NRS back pain) and leg pain (NRS leg pain), where 0 = no pain and 10 = worst conceivable pain [2].

Generic health-related quality of life was assessed by the EQ-5D [22], which has been validated for a similar patient population [23]. It evaluates five dimensions: mobility, self-care, activities of daily living, pain, and anxiety and/or depression. For each dimension, the patient describes three possible levels of problems (none, mild-to-moderate, and severe). This descriptive system, therefore, contains $3^5 = 243$ combinations or index values for health status. The total score ranges from −0.59 to 1, where 1 corresponds to perfect health and 0 to death. Negative values are considered to be worse than death.

The patient-rated benefit of the operation was rated on a Global Perceived Effect scale (GPE) at follow-up [24]. The response alternatives were: 1 = "completely recovered", 2 = "much better", 3 = "somewhat better", 4 = "no change", 5 = "somewhat worse", 6 = "much worse", and 7 = "worse than ever".

Definition of failure and worsening

Patients reporting to be unchanged or worse (categories 4–7) on the GPE scale at 12-month follow-up were classified as "failure", and those considering themselves worse or worse than ever (GPE 6–7) were classified as "worsening".

Statistics

All statistical analyses were performed with the Statistical Package for the Social Sciences (SPSS, IBM Version 23.0).

We excluded all patients who did not respond at 12 months, and compared baseline characteristics of both respondents and non-respondents. This strategy was based on a study on a comparable patient population from NORspine and a recent and similar Danish study [25, 26].

For all PROMs, the mean change, mean % change (except for EQ-5D), and mean final raw score were assessed against the GPE by one-way analyses of variance (ANOVA) with post hoc analysis (Tukey, $\alpha = 0.05$) and by analyses of co-variance (ANCOVA, generalized linear model) with adjustment for baseline scores. Correlation analyses between PROMs and the GPE were done by Spearman rank correlation for all measures, except for the final raw scores in which Pearson was used.

Cutoffs for all scores were estimated by Receiver Operating Characteristic (ROC) curves. When analyzing criteria for "failure", cases with failure were defined as those who reported to be unchanged or worse (categories 4–7) on the GPE scale at 12 months. All other categories on the GPE scale (1–3) were defined as "no failure". When comparing patients, reporting being considerably worse (GPE 6–7), with those who reported an unchanged status (GPE 4–5), those reporting improvement (GPE 1–3) were excluded from these analyses. To determine the cutoff with the highest sensitivity and specificity for both failure and worsening, the closest point to the upper left corner of the ROC curve was calculated from the coordinates of the curve. Area under the curve (AUC) calculations were performed to determine how well the instruments differentiated between the outcome groups. An AUC value of >0.70 was considered acceptable [27]. The overall

accuracy for each cutoff was calculated with a confusion matrix. In the presentation of the results, we included AUC and cutoff values only for variables with an AUC value above 0.70. Results for PROMS with poorer accuracy can be provided on request.

To investigate whether the optimal cutoffs differed between important subgroups in the registry sample, sensitivity analyses were performed between first time vs reoperation and between macroscopic ("open") vs microscope or loupe-assisted discectomy. To evaluate the impact of different baseline scores on the cutoffs, cutoff calculations were also carried out on those with low- and high baseline disability.

Differences between elective and emergency cases at 12-month follow-up were calculated for all PROMs by Student's *t* test and for the GPE by Mann–Whitney *U* test.

Floor and ceiling effects were assessed by calculating the frequency of the highest and lowest possible scores at baseline. If 15% of patients had a minimal or maximal score value at baseline, these were considered as floor or ceiling effects [27, 28].

Results

6840 out of 9930 (69%) patients had 12-month follow-up data. Among those lost to follow-up were more smokers, a higher number of sickness benefits recipients, and more patients who had been operated previously (Table 2). Furthermore, they had a lower level of education, and fewer were operated on for paresis. Except for back pain, there was no statistical significant difference in PROMs at baseline. Patients who did not respond to the follow-up scored slightly higher for back pain than those who responded.

During surgery, an operating microscope or loupes were used in 5936 of 6840 (87%) cases. A total of 885 (13%) had a reoperation on the same level, 466 (7%) on a different level, and 66 (1%) on both the same and a different level between L1 and S1. The perioperative complication rate was 169 (3%) with 115 (2%) dural tears, 21 (0.3%) nerve root injuries, 24 (0.4%) hematomas requiring transfusion or reoperation, and 9 (0.1%) cardiorespiratory complications.

Few data points were missing for the baseline PROMs: ODI (13, 0.2%), EQ-5D (252, 3.7%), NRS back pain (170, 2.5%), and NRS leg pain (159, 2.3%). At 12-month follow-up, 40 (0.6%) were missing data on GPE, 11 (0.2%) on ODI, 520 (7.6%) on EQ-5D, 47 (0.7%) on NRS back pain, and 66 (1%) on NRS leg pain. GPE scores for the entire population are shown in Table 3. Mean improvement (95% CI) for each PROM from baseline to 12-month follow-up for the total sample was 28.7 (28.2–29.2) for the ODI, 0.45 (0.44–0.46) for EQ-5D, 3.2 (3.1–3.3) for back pain, and 4.4 (4.3–4.5) for leg pain, $p < 0.001$.

The Spearman rank correlation coefficients between the GPE and the change scores of the instruments were high for mean % changes with 0.8 for the ODI, 0.7 for NRS back pain and leg pain, and moderate for mean changes with 0.6 (ODI), 0.5 (NRS back pain), 0.6 (NRS leg pain), and 0.5 (EQ-5D). The Pearson correlation coefficients were high for all the final raw scores with 0.8 (ODI), 0.7 (NRS leg pain), 0.8 (NRS back pain), and 0.7 (EQ-5D). All correlation coefficients were statistically significant ($p < 0.001$).

ANOVA with post hoc analysis (Tukey, $\alpha = 0.05$) indicated that the mean changes of all of the PROMs were

Table 2 Baseline patient characteristics for respondents vs non-respondents

Characteristic	Respondents	Non-respondents	*p* value
Receiving sickness or disability payment, *n* (%)	4180 (61)	2026 (66)	<0.001
Smokers, *n* (%)	1936 (29)	1222 (40)	<0.001
BMI, mean (SD)	26.6 (4.2)	27.0 (4.7)	<0.001
University or college education, *n* (%)	2561 (37)	962 (31)	<0.001
Operated for paresis, *n* (%)	1321 (19)	530 (17)	0.01
Emergency surgery, *n* (%)	653 (9)	291 (9)	0.84
Previous lumbar disc surgery, *n* (%)	1417 (21)	745 (24)	<0.001
ASA, mean (SD)	1.5 (0.6)	1.5 (0.6)	0.10
Comorbidity, *n* (%)	1664 (28)	674 (26)	0.014
Mean ODI (SD)	45.99 (18.9)	45.69 (18.4)	0.46
Mean EQ-5d (SD)	0.27 (0.35)	0.26 (0.36)	0.18
Mean NRS back pain (SD)	6.23 (2.5)	6.36 (2.4)	0.02
Mean NRS leg pain (SD)	6.9 (2.2)	6.9 (2.12)	0.87

SD standard deviation

Table 3 Baseline adjusted mean of the change score, % change score, and final raw score for all PROMS (95% of CI) according to the global perceived effect scale at 1-year follow-up

GPE	Completely recovered (1)	Much better (2)	Somewhat better (3)	No change (4)	Somewhat worse (5)	Much worse (6)	Worse than ever (7)
N (%)	1659 (24)	3265 (48)	1093 (16)	358 (5)	216 (3)	153 (2)	66 (1)
ODI							
Mean change	43.1 (42.7–43.6)	32.4 (32.1–32.7)	15.1 (14.5–15.7)	6.6 (5.6–7.6)	3.7 (2.5–5.0)	−5.6 (−7.1 to −4.1)	−12.9 (−15.2 to −1.5)
Mean % change	93.4 (91.9–94.9)	66.7 (65.6–67.8)	26.7 (24.9–28.6)	1.3 (−1.9 to 4.5)	−6.7 (−10.8 to −2.5)	−13.7 (−18.6 to −8.7)	−30.7 (−38.3 to −23.0)
12 month raw score	2.9 (2.4–3.3)	13.6 (13.3–14)	30.9 (30.3–31.5)	39.4 (38.4–40.4)	42.3 (41–43.6)	51.6 (50–53.1)	58.9 (56.5–61.2)
NRS back pain							
Mean change	5.7 (5.6–5.7)	3.6 (3.6–3.7)	1.0 (0.9–1.1)	−0.1 (−0.3 to 0.05)	−0.5 (−0.7 to −0.3)	−1.7 (−2.0 to −1.4)	−2.0 (−2.4 to −1.6)
Mean % change	91.4 (89.2–93.5)	51.3 (49.8–52.8)	6.3 (3.6–9.0)	−11.1 (−15.8 to −6.4)	−17.8 (−23.8 to −11.8)	−24.0 (−31.1 to −16.9)	−59.8 (−70.7 to −49.0)
12 month raw score	0.6 (0.5 –0.7)	2.6 (2.5–2.7)	5.2 (5.1–5.3)	6.4 (6.2–6.5)	6.7 (6.5–7.0)	7.9 (7.7–8.2)	8.2 (7.8–8.6)
NRS leg pain							
Mean change	6.5 (6.4–6.6)	4.9 (4.8–5.0)	2.5 (2.4–2.6)	1.0 (0.8–1.2)	0.5 (0.2–0.7)	−0.5 (−0.8 to −0.2)	−1.0 (−1.5 to −0.6)
Mean % change	92.6 (90.5–94.6)	67.8 (66.3–69.2)	29.6 (27.1–32.1)	3.5 (−0.82 to 7.9)	−2.6 (−8.2 to 3.0)	−9.2 (−15.9 to 2.6)	−19.2 (−30.0 to −8.8)
12 month raw score	0.4 (0.3–0.5)	2.0 (1.9–2.1)	4.4 (4.3–4.5)	5.9 (5.7–6.1)	6.4 (6.2–6.7)	7.4 (7.1–7.7)	7.9 (7.5–8.4)
EQ-5D[a]							
Mean change	0.68 (0.67–0.69)	0.50 (0.50–0.51)	0.28 (0.27–0.29)	0.07 (0.06–0.09)	0.03 (−0.00 to 0.05)	−0.15 (−0.18 to −0.12)	−0.24 (−0.29 to −0.19)
12 month raw score	0.95 (0.95–0.96)	0.77 (0.77–0.78)	0.55 (0.54–0.56)	0.34 (0.32–0.36)	0.30 (0.27–0.32)	0.12 (0.09–0.15)	0.03 (−0.02 to 0.08)

Negative prefix = worsening of the score

[a] Not possible to calculate % change score for EQ-5D

significantly different between GPE categories 1–3 and 4. The mean of the final raw scores for all of the PROMs, as well as the mean change in ODI, EQ-5D, and NRS leg pain, and the mean ODI% change score at 12 months were able to differentiate between "no change" (4) and "much worse" (6) with statistical significance. Mean changes in NRS back pain, as well as mean % change in NRS back- and leg pain were not statistically significant different between those "unchanged" (4) and those reporting to be "much worse" (6).

After evaluating the mean score differences of all PROMs across the categories of the GPE, the study group concluded that the definition of a score range of 4–7 for "failure" and 6–7 for "worsening" was appropriate (Table 3). Figures illustrating these differences are shown in the appendix (Figs. 1x–4x).

For each GPE outcome group, the baseline adjusted mean scores of the PROMs (ANCOVA) after 12 months are shown in Table 3.

Cutoff values

For differentiation between "failure" vs no failure in the whole cohort, all PROMs had an acceptable AUC of >0.70 (Table 4). The PROM with the highest accuracy was the mean ODI% change score with an AUC of 0.93 and a correct classification rate of 86% (Fig. 1).

For differentiation between "worsening" vs unchanged and slightly worse, the AUCs were poor (<0.70) for score changes of all outcome measures. The final raw scores of all four PROMs showed acceptable AUCs. The PROM with the highest accuracy was the ODI raw score with an AUC of 0.76 and a correct classification rate of 69% (Fig. 2). The ROCs for all of the PROMs are illustrated in the appendix (Figs. 5x–9x).

Based on these cutoff values, the ODI change classified 26%, the ODI% change score 23%, and the ODI raw score at 12 months 27% of lumbar disc surgeries as failure. Failure rates assessed by cutoffs of the less accurate PROMs are shown in the appendix (Table 4x).

The percentages of patients classified as worsening by the cutoffs on the final PROM raw scores were 7% for ODI, 8% for EQ-5D, 7% for NRS leg pain, and 8% for NRS back pain.

Sensitivity analysis

When comparing patients operated for the first time with those who had been operated previously, values for cutoff, sensitivity, and specificity were similar (Tables 2x and 3x in appendix). When investigating the effect of low and high baseline disability (based on the 25th and 75th percentile of the baseline score for ODI), the cutoffs for

Table 4 All cutoff values with corresponding sensitivity and specificity, area under the curve (95% confidence interval), and percentage of correctly classified

	Failure				Worsening			
	Cutoff	Sens/spec	AUC (95% CI)	Corr. class %	Cutoff	Sens/spec	AUC (95% CI)	Corr. class %
ODI								
Mean change	13	0.82, 0.82	0.89 (0.88–0.91)	82				
Mean % change	33	0.86, 0.86	0.93 (0.92–0.94)	86				
12 month raw	25	0.89, 0.81	0.92 (0.91–0.93)	86	48	0.70, 0.70	0.76 (0.72–0.80)	69
NRS leg pain								
Mean change	1.5	0.81, 0.76	0.87 (0.86–0.88)	84				
Mean % change	39	0.86, 0.81	0.89 (0.88–0.90)	84				
12 month raw	4.5	0.91, 0.85	0.90 (0.88–0.91)	84	7.5	0.64, 0.68	0.70 (0.66–0.75)	67
NRS back pain								
Mean change	1.5	0.74, 0.86	0.85 (0.84–0.86)	76				
Mean % change	24	0.85, 0.81	0.87 (0.86–0.88)	86				
12 month raw	5.5	0.81, 0.87	0.92 (0.91–0.93)	86	7.5	0.78, 0.64	0.77 (0.73–0.81)	68
EQ-5D								
Mean change	0.1	0.76, 0.83	0.85 (0.84–0.87)	82				
12 month raw	0.6	0.81, 0.85	0.91 (0.90–0.92)	85	0.1	0.76, 0.60	0.71 (0.67–0.75)	65

For worsening, only the 12-month raw scores were used, and all the other cutoffs had an AUC < 0.70

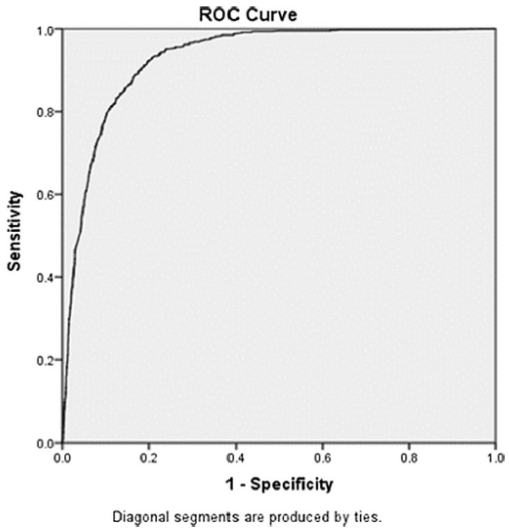


Fig. 1 ODI% change vs external anchor, GPE 4–7 vs 1–3 (AUC 0.893) at 12-month follow-up

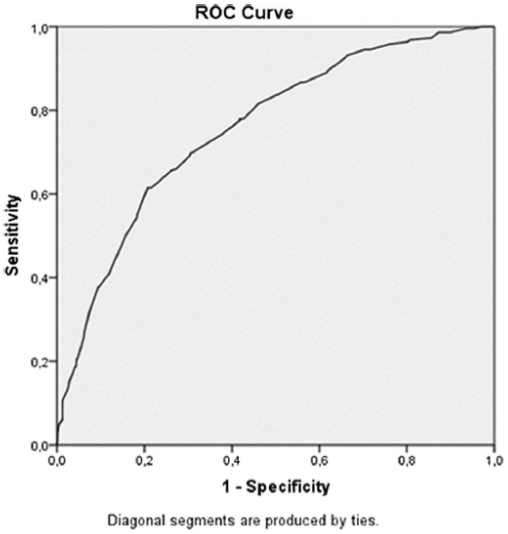


Fig. 2 ODI 12-month raw vs external anchor, GPE 4–5 vs 6 + 7 (AUC 0.758) at 12-month follow-up

"failure" and "worsening" in the PROMs varied considerably, both for change scores, % change scores, and the final raw score (Table 1x, appendix). For example, in the group with high disability at baseline, the failure cutoff for the mean % change in ODI was 30% higher than in the low disability group.

Compared to elective surgery, emergency cases had statistically significant worse baseline PROM scores and experienced a greater score improvement at 12 months. Accordingly, no statistically difference in any of the 12-month PROM raw scores was found between these two

groups. Furthermore, they reported the same GPE after 12 months, with a median score of 2 (Table 5x, appendix).

Floor and ceiling effects

No floor or ceiling effects were detected. Only 9 (0.1%) patients scored 0 and 7 (0.1%) patients scored 100 on the baseline ODI. Furthermore, 107 (1.6%) scored 0 and 590 (8.8%) scored 10 in the NRS back-pain scale. For the NRS leg pain, scale numbers were 55 (0.8%) for 0 and 728 (10.9%) for 10. In the EQ-5D, only 12 (0.2%) patients scored the minimum and 20 (0.3%) the maximum at baseline.

Discussion

We estimated the optimal cutoff values for failure and worsening 12 months after surgery for lumbar disc herniation, using four recommended PROMs. An ODI% improvement of less than 33% was the most accurate measure for identifying patients for whom the surgery had failed. Back pain, both the mean % change, and the final raw score at 12 months, also showed high accuracy for identifying failure. We found no significant difference in outcome scores among patient groups who considered themselves as "unchanged" or "slightly worse", which is in accordance with a previous study [12]. A final ODI raw score of more than 48 at 12-month follow-up had the highest accuracy for identifying patients reporting worsening, followed by a final raw score of 7.5 for NRS back pain. A potential explanation for this finding might be that those with a final ODI over a threshold value of 48 will tend to consider themselves as worse, irrespective of the amount of change. These patients are exhausted after more than a year with unresolved severe pain and disability, not compatible with a normal life (Fig. 10x, appendix). One previous study also found a high correspondence between the final raw score and the GPE scale as an external anchor [17].

Compared to the GPE, all cutoffs categorized a higher proportion of the outcomes as "failure" or "worsening". Since the individual PROMs represent different concepts, the variation between the individual outcome measures and GPE scale is to be expected [10, 11]. For instance, even the disease-specific ten item ODI could fail to address issues important to patients. Individuals might also weigh each item differently according to their preferences.

We chose to classify all patients who scored unchanged or worse (GPE > 3), as "failure" and those scoring much worse or worse than ever (GPE 6–7) as "worsening". These definitions are supported by our data, i.e., differences in mean PROMs between the GPE groups in ANOVA and ANCOVA analyses, as shown in Table 3 and Figs. 1x–5x (appendix). A large group of patients ($n = 1676$, 24%) classified themselves as "slightly better", "unchanged", or "slightly worse" on the GPE, and would be the most susceptible of being misclassified [12]. While it was not possible to separate the "unchanged" from the "slightly worse" based on PROMs, patients defining themselves as "slightly better" (16%) had a mean improvement in the ODI score of 15.1, more than the previously defined cutoff for the Minimal Clinical Important Change (MCIC) [8]. Hence, it is reasonable not to include them in the failure group. While non-success implies a degree of improvement, failure does not, which might be of importance for litigation issues. The distinction between these two concepts could also be used in the development of predictive models in value-based health care [29].

The mean PROM improvements in this study were in line with results from other clinical trials [30–33]. Failure and success rates, however, are highly dependent on where the cutoff levels are set to classify outcomes, and types of PROMs used [11]. Mean change in NRS back pain showed the highest failure rate (31%) and mean change in NRS leg pain the lowest (20%). Back-pain intensity is not the primary indication for lumbar discectomy without fusion. It could therefore be expected that, for instance, the NRS leg pain classified a lower failure rate [34]. Our findings indicate that patients reporting failure and worsening tend to be concerned about back pain, even though leg pain may have improved. An explanation may be that a large proportion of patients operated for lumbar disc disease will expect a substantial improvement in back pain [35].

Methodological challenges

The global perceived effect is a frequently used external anchor to define cutoffs on PROMs. Still, it has several weaknesses related to recall bias [17], lack of objectivity [36], and for not taking into account the measurement precision [6]. More objective criteria, such as return to work or use of pain killers, have been proposed [36]. However, they tend to be subgroup specific (e.g., only considering the working population) and may also be susceptible to confounding [37]. Some authors argue that the criteria should be defined prior to treatment by letting the patients quantify, e.g., on a pain scale, how great a satisfying improvement should be [38]. To the best of our knowledge, no such alternative and well-validated external anchors for self-reported questionnaires exist. Unlike the European Spine Tango registry, the NORspine does not collect data on the surgeon's overall assessment of outcome [39]. Lack of "expert opinion" might represent a weakness. However, surgeons and patients agree only in 50% of cases when assessing outcomes, and surgeons tend

to rate the end result over-optimistically [40]. Another weakness related to anchor-based methods is misclassification. In our population, the ODI% cutoff of 33% improvement at 12 months (AUC 0.93, sensitivity/specificity 86%) gave a false-positive rate of 14% and a false-negative rate of 15%.

Importantly, we found that the cutoffs also were highly depending on the baseline PROM score. For instance, severely disabled patients will require disproportionally greater improvements than the less disabled, not to consider the surgery as failed. This is in accordance with findings of other studies and illustrates the importance of taking into account the baseline score while interpreting PROM change scores, regardless of using absolute or percentage change scores [18, 41]. Consequently, one should adjust for the baseline score when using such outcome criteria in clinical trials and risk factor analyses. A possible cause might be higher expectations towards improvements among patients with high baseline pain and disability [42]. Fulfillment of expectations has also been identified as a major predictor for positive patient-rated positive outcome after surgery [35]. Similar to findings by Elkan et al., emergency cases presented with more severe symptoms and had a greater amount of change on the PROM scores, thus reported the same improvement on the GPE scale [43].

Limitations and strengths of this study

Loss to follow-up at 12 months was 31.1%. Two Scandinavian registry studies found that a loss to follow-up of 12–22% did not bias conclusions about treatment effects [25, 26]. Even if baseline PROMs were similar between respondents and non-respondents in our study, several baseline characteristics of non-respondents have been associated with poorer outcomes [44]. This could represent a selection bias, especially when measuring the exact failure and worsening rate, but less so when defining PROM cutoffs over a large range of outcomes. Follow-up was only 12 months, but previous studies have shown mean outcome values to be stable from 1 up to 8 years [26, 45].

An advantage of this study is the large sample size and high external validity due to patient recruitment from everyday practice. In a smaller single-center study from 2013, Gum et al. tried to define clinically important deterioration among patients operated with lumbar fusion for various diagnoses, but found it difficult to define cutoffs. They concluded that a larger patient population was needed to identify accurate cutoffs, since worsening is a relatively rare event [41]. We have used a much larger and more condition-specific cohort.

Future perspectives

Both clinicians and administrators have questioned whether quality registries can improve clinical practice and feedback comprehendible information to patients and clinicians [3]. An advantage of dichotomous outcomes is the possibility to provide risk estimates in terms of probability. In clinical decision-making, percentwise probability would be easier to understand than estimates based on continuous outcome data (e.g., linear regressing coefficients). More research is needed to identify risk factors for adverse outcomes and to learn how such new knowledge can be conveyed efficiently to patients and health care providers.

Conclusion

We have defined cutoff values with acceptable sensitivity and specificity on validated PROMs to classify outcomes as "failure" and "worsening" 12 months after lumbar disc surgery.

Implication

These criteria could facilitate shared decision-making among physicians and patients, quality assessment, and comparison of clinical outcomes across surgical units. In addition to clinically relevant improvements, we propose that rates of failure and worsening should be included in reporting from clinical trials.

Acknowledgements This project receives funding from The Norwegian Medical Association—Foundation for quality improvement and patient safety, and The Regional Health Authority of Northern Norway.

Compliance with ethical standards

Conflict of interest None of the authors has any potential conflict of interest.

References

1. Baker DJ, PPB, FCT (1990) The Oswestry Disability Index revisited: its reliability, repeatability and validity, and a

comparison with the St Thomas's Disability Index. Back pain. New approaches to rehabilitation and education. 174–186
2. Mp J, Karoly P (1992) Self-report scales and procedures for assessing pain in adults. In: DC T, Melzack R (eds) Handbook of pain assessment. The Guilford Press, New York, pp 135–151
3. van Hooff ML, Jacobs WCH, Willems PC et al (2015) Evidence and practice in spine registries: a systematic review, and recommendations for future design of registries. Acta Orthop 86:1–11. doi:10.3109/17453674.2015.1043174
4. Wyrwich KW, Norquist JM, Lenderking WR, Acaster S (2013) Methods for interpreting change over time in patient-reported outcome measures. Qual Life Res 22:475–483. doi:10.1007/s11136-012-0175-x
5. Wright A, Hannon J, Hegedus EJ, Kavchak AE (2012) Clinimetrics corner: a closer look at the minimal clinically important difference (MCID). J Man Manip Ther 20:160–166. doi:10.1179/2042618612Y.0000000001
6. Copay AG, Subach BR, Glassman SD et al (2007) Understanding the minimum clinically important difference: a review of concepts and methods. Spine J 7:541–546
7. van der Roer N, Ostelo RWJG, Bekkering GE et al (2006) Minimal clinically important change for pain intensity, functional status, and general health status in patients with nonspecific low back pain. Spine (Phila Pa 1976) 31:578. doi:10.1097/01.brs.0000201293.57439.47
8. Ostelo RWJG, Deyo RA, Stratford P et al (2008) Interpreting Change Scores for Pain and Functional Status in Low Back Pain. Spine (Phila Pa 1976) 33:90–94. doi:10.1097/BRS.0b013e31815e3a10
9. Glassman SD, Copay AG, Berven SH et al (2008) Defining substantial clinical benefit following lumbar spine arthrodesis. J Bone Jt Surg Am 90:1839–1847
10. Copay AG, Martin MM, Subach BR et al (2010) Assessment of spine surgery outcomes: inconsistency of change amongst outcome measurements. Spine J 10:291–296
11. Solberg T, Johnsen LG, Nygaard OP, Grotle M (2013) Can we define success criteria for lumbar disc surgery? Estimates for a substantial amount of improvement in core outcome measures. Acta Orthop 84:196–201
12. Copay AG, Glassman SD, Subach BR et al (2008) Minimum clinically important difference in lumbar spine surgery patients: a choice of methods using the Oswestry Disability Index, Medical Outcomes Study questionnaire Short Form 36, and Pain Scales. Spine J 8:968–974. doi:10.1016/j.spinee.2007.11.006
13. van Kampen DA, Willems WJJ, van Beers LWAH et al (2013) Determination and comparison of the smallest detectable change (SDC) and the minimal important change (MIC) of four-shoulder patient-reported outcome measures (PROMs). J Orthop Surg Res 8:40. doi:10.1186/1749-799X-8-40
14. Fekete TF, Haschtmann D, Kleinstück FS et al (2016) What level of pain are patients happy to live with after surgery for lumbar degenerative disorders? Spine J 16:S12–S18. doi:10.1016/j.spinee.2016.01.180
15. Mroz TE, McGirt M, Chapman JR et al (2014) More "Why" and Less "How". Spine (Phila Pa 1976) 39:S7–S8. doi:10.1097/BRS.0000000000000539
16. Deyo RA, Mirza SK (2009) The case for restraint in spinal surgery: does quality management have a role to play? EurSpine J 18(Suppl 3):331–337
17. Grøvle L, Haugen AJ, Hasvik E et al (2014) Patients' ratings of global perceived change during 2 years were strongly influenced by the current health status. J Clin Epidemiol 67:508–515. doi:10.1016/j.jclinepi.2013.12.001
18. Mannion AF, Porchet F, Kleinstück FS et al (2009) The quality of spine surgery from the patient's perspective: part 2. Minimal clinically important difference for improvement and deterioration as measured with the Core Outcome Measures Index. Eur Spine J 18:374–379. doi:10.1007/s00586-009-0931-y
19. McGirt MJ, Parker SL, Asher AL et al (2014) Role of prospective registries in defining the value and effectiveness of spine care. Spine (Phila Pa 1976) 39:S117–S128. doi:10.1097/BRS.0000000000000552
20. Larsson S, Lawyer P, Silverstein MB (2010) From concept to reality. Aging (Albany NY). doi:10.1140/epjcd/s2004-03-1694-8
21. Ruyter KW (2015) REK sør-øst Knut W. Ruyter 22845518 24.06.2015
22. Dolan P, Gudex C, Kind P, Williams A (1996) The time trade-off method: results from a general population study. Heal Econ 5:141–154
23. Solberg TK, Olsen JA, Ingebrigtsen T et al (2005) Health-related quality of life assessment by the EuroQol-5D can provide cost-utility data in the field of low-back surgery. Eur Spine J 14:1000–1007
24. Kamper SJ, Ostelo RW, Knol DL et al (2010) Global Perceived Effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status. J Clin Epidemiol 63:760–766
25. Hojmark K, Stottrup C, Carreon L, Andersen MO (2015) Patient-reported outcome measures unbiased by loss of follow-up. Single-center study based on DaneSpine, the Danish spine surgery registry. Eur spine J Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc. doi:10.1007/s00586-015-4127-3
26. Solberg TK, Sorlie A, Sjaavik K et al (2011) Would loss to follow-up bias the outcome evaluation of patients operated for degenerative disorders of the lumbar spine?: a study of responding and non-responding cohort participants from a clinical spine surgery registry. Acta Orthop 82:56–63. doi:10.3109/17453674.2010.548024
27. de Vet HC, Ostelo RW, Terwee CB et al (2007) Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. Qual Life Res 16:131–142
28. Terwee CB, Bot SD, De Boer MR et al (2007) Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 60:34–42
29. Resnick DK, Tosteson ANA, Groman RF, Ghogawala Z (2014) Setting the equation. Spine (Phila Pa 1976) 39:S43–S50. doi:10.1097/BRS.0000000000000581
30. Weinstein JN, Tosteson TD, Lurie JD et al (2006) Surgical vs nonoperative treatment for lumbar disk herniation: the Spine Patient Outcomes Research Trial (SPORT): a randomized trial. JAMA 296:2441–2450
31. Stromqvist B, Fritzell P, Hagg O et al (2013) Swespine: the Swedish spine register: the 2012 report. Eur Spine J 22:953–974
32. Whitmore RG, Curran JN, Ali ZS et al (2015) Predictive value of 3-month lumbar discectomy outcomes in the NeuroPoint-SD Registry. J Neurosurg Spine 23:1–8. doi:10.3171/2015.1.SPINE14890
33. Porchet F, Bartanusz V, Kleinstueck FS et al (2009) Microdiscectomy compared with standard discectomy: an old problem revisited with new outcome measures within the framework of a spine surgical registry. Eur Spine J 18(Suppl 3):360–366
34. Sørile A, Moholdt V, Kvistad KA et al (2012) Modic type i changes and recovery of back pain after lumbar microdiscectomy. Eur Spine J 21:2252–2258. doi:10.1007/s00586-012-2419-4
35. Mannion AF, Junge A, Elfering A et al (2009) Great expectations: really the novel predictor of outcome after spinal surgery? Spine (Phila Pa 1976) 34:1590–1599
36. Gatchel RJ, Mayer TG (2010) Testing minimal clinically important difference: additional comments and scientific reality testing. Spine J 10:330–332. doi:10.1016/j.spinee.2010.01.019

37. Glassman SD, Carreon LY (2010) Thresholds for health-related quality of life measures: reality testing. Spine J 10:328–329. doi:10.1016/j.spinee.2009.12.026
38. Ferreira ML, Herbert RD, Ferreira PH et al (2012) A critical review of methods used to determine the smallest worthwhile effect of interventions for low back pain. J Clin Epidemiol 65:253–261
39. Roder C, Chavanne A, Mannion AF et al (2005) SSE Spine Tango–content, workflow, set-up. www.eurospine.org-Spine Tango. Eur Spine J 14:920–924
40. Kleinstuck FS, Grob D, Lattig F et al (2009) The influence of preoperative back pain on the outcome of lumbar decompression surgery. Spine (Phila Pa 1976) 34:1198–1203
41. Gum JL, Glassman SD, Carreon LY (2013) Clinically important deterioration in patients undergoing lumbar spine surgery: a choice of evaluation methods using the Oswestry Disability Index, 36-Item Short Form Health Survey, and pain scales. J Neurosurg Spine 19:564–568. doi:10.3171/2013.8.SPINE12804
42. Mancuso CA, Duculan R, Stal M, Girardi FP (2014) Patients expectations of lumbar spine surgery. Eur Spine J 24:2362–2369. doi:10.1007/s00586-014-3597-z
43. Elkan P (2016) Similar result after non-elective and elective surgery for lumbar disc herniation : an observational study based on the SweSpine register. Eur Spine J. doi:10.1007/s00586-016-4419-2
44. Mannion AF, Elfering A (2006) Predictors of surgical outcome and their assessment. Eur Spine J 15(Suppl 1):S93–108
45. Lurie JD, Tosteson TD, Tosteson ANA et al (2014) Surgical versus nonoperative treatment for lumbar disc herniation. Spine (Phila Pa 1976) 39:3–16. doi:10.1097/BRS.0000000000000088

15.2 Paper II.

Original Article

AOSPINE

Global Spine Journal
1-8

DOI: 10.1177/2192568219851480
journals.sagepub.com/home/gsj
SAGE

Can a Successful Outcome After Surgery for Lumbar Disc Herniation Be Defined by the Oswestry Disability Index Raw Score?

David A. T. Werner, MD[1,2], Margreth Grotle, MD, PhD[3,4], Sasha Gulati, MD, PhD[5,6], Ivar M. Austevoll, MD[7], Mattis A. Madsbu, MD[5], Greger Lønne, MD, PhD[6,8,9], and Tore K. Solberg, MD, PhD[1,2,9]

Abstract

Study Design: Prospective multicenter cohort study.

Objective: To investigate (1) the discriminative ability and cutoff estimates for success 12 months after surgery for lumbar disc herniation on the Oswestry Disability Index (ODI) raw score compared with a change and a percentage change score and (2) to what extent these clinical outcomes depend on the baseline disability.

Methods: A total of 6840 patients operated for lumbar disc herniation from the Norwegian Registry for Spine Surgery (NORspine) were included. In receiver operating characteristic (ROC) curve analyses, a global perceived effect (GPE) scale (1-7) was used an external anchor. Success was defined as categories 1-2, "completely recovered" and "much better." Cutoffs for success for subgroups with different preoperative disability were also estimated.

Results: When defining success after surgery for lumbar disc herniation, the accuracy (sensitivity, specificity, area under the curve, 95% CI) for the ODI raw score (0.83, 0.87, 0.930, 0.924-0.937) was comparable to the ODI percentage change score (0.85, 0.85, 0.925, 0.918-0.931), and higher than the ODI change score (0.79, 0.73, 0.838, 0.830-0.852). The cutoff for success was highly dependent on the amount of baseline disability (low-high), with cutoffs ranging from 13 to 28 for the ODI raw score and 39% to 66% for ODI percentage change. The ODI change score (points) was not as accurate.

Conclusion: The 12-month ODI raw score, like the ODI percentage change score, can define a successful outcome with excellent accuracy. Adjustment for the baseline ODI score should be performed when comparing outcomes across groups, and one should consider using cutoffs according to preoperative disability (low, medium, high ODI scores).

Keywords

PASS, ODI, success criteria, lumbar disc surgery, PROM, lumbar disc herniation

[1] University Hospital of Northern Norway, Tromsø, Norway
[2] University of Tromsø, Institute of Clinical Medicine, Tromsø, Norway
[3] Oslo Metropolitan University, Oslo, Norway
[4] Oslo University Hospital, Oslo, Norway
[5] St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway
[6] Norwegian University of Science and Technology, Trondheim, Norway
[7] Kysthospitalet in Hagevik, Haukeland University Hospital, Bergen, Norway
[8] Innlandet Hospital Trust, Lillehammer, Norway
[9] Norwegian Registry for Spine Surgery, Northern Norway Regional Health Authority, Bodø, Norway

Corresponding Author:
David Werner, Department of Neurosurgery, University Hospital of Northern Norway, Sykehusveien 38, 9019 Tromsø, Norway.
Email: mail@david-werner.com

Introduction

In Norway, operative treatment of lumbar disc herniation is the most frequently performed spine surgery procedure in patients younger than 50 years.[1] The indication for surgery is most often relative, that is, reducing pain-related disability.[2] To compare treatment effects across interventions and institutions, changes in patient-reported outcome measures (PROMs) are frequently used, but their interpretation is complex. Previous studies have used score changes of the Oswestry Disability Index (ODI) to calculate clinically meaningful improvements, such as cutoffs for a "successful outcome."[3-6] However, the amount of change needed for success is highly dependent on the baseline scores of the PROMs.[7]

Studies from other medical fields, such as rheumatology, have used a Patient Acceptable Symptom State (PASS)[8-10] in order to define a cutoff for a successful outcome on a PROM. The PASS could be viewed as a separate entity to the underlying change score.[8,9,11] We have previously defined cutoffs for success based on PROM change scores for patients operated for lumbar disc herniation, by either open- or micro-discectomy.[6] In a recent study, we found cutoffs on the 12-month ODI raw score that had the highest accuracy for identifying cases that could be classified as failed and worsened after lumbar disc surgery,[11] indicating that patients could be more focused on their current disability than on health changes when reporting clinical outcomes.

In the present study, we sought (1) to define the discriminative ability and cutoff estimates of success for a 12-month ODI raw score (current disability), an ODI change and ODI percentage change score and (2) to investigate if these clinical outcomes depend on the baseline disability, that is, the preoperative ODI score. We defined success by the patient's ratings of a substantial effect of surgery (Global Perceived Effect scale, GPE), that is, when the patient is feeling "completely recovered" (GPE = 1), or "much better" (GPE = 2) 12 months after the operation. Such information would aid in the classification and understanding and of successful outcome, facilitating reporting and comparisons of treatment results.

Materials and Methods

Patient Population and Data Collection

A total of 6840 patients operated for lumbar disc herniation at 38 different surgical units between January 1, 2007 and February 28, 2014 were followed for 12 months, according to the standard protocol of the Norwegian registry for spine surgery (NORspine). The NORspine is a comprehensive clinical registry for quality control and research. During the study period, the NORspine comprised 95% (38 of 40) Norwegian public and private centers performing lumbar disc surgery. Completeness, the proportion of patients operated on for lumbar disc herniation reported to the NORspine, was 65%.[1] The registry excluded patients unable to consent, children aged <16 years, patients with documented drug abuse or severe psychiatric disorders, and patients with traumatic, infectious or malignant conditions in the spine. In this study, we included all elective and emergency cases operated for lumbar disc herniation. Fusion procedures and/or procedures including laminectomy were not included.

Informed consent was obtained from all patients and participation was neither mandatory, nor required to gain access to healthcare. The registry protocol has been approved by the Data Inspectorate of Norway. The study protocol had been submitted to the regional ethical committee for medical research which categorized it as a clinical audit study, not in need of their formal approval.[12]

At admission for surgery, the patients completed a baseline questionnaire on demographics, lifestyle issues and PROMs (Figures 10x-13x, appendix). During the hospital stay, the surgeon recorded data concerning diagnosis, treatment, and comorbidity on a standard registration form (Figures 14x-15x, appendix). Twelve months after surgery a questionnaire was distributed by regular post, completed at home by the patients, and returned in prestamped envelopes to the central registry unit without involvement of the treating hospitals. One reminder with a new copy of the questionnaire was sent to those who did not respond.

Patient-Reported Outcome Measures

This study is based on the cohort used in a previous study by the authors.[11] The ODI version 2.1a was used to assess pain-related disability. It contains 10 questions on limitations of activities of daily living. Each item is rated 0 to 5 and then transferred into a percentage score ranging from 0 (none) to 100 (maximum pain-related disability).[13]

The patient-rated benefit of the operation was rated on the GPE at follow-up.[11,14] The response alternatives were as follows: 1 = "completely recovered," 2 = "much better," 3 = "somewhat better," 4 = "no change," 5 = "somewhat worse," 6 = "much worse," and 7 = "worse than ever."

Statistics

All statistical analyses were performed with the Statistical Package for the Social Sciences (SPSS, IBM Version 23.0).

We excluded all patients who did not respond at 12 months. This strategy was based on a study from the NORspine on a comparable patient population, and a recent and similar Danish registry study, both indicating that patients lost to follow-up could be handled as missing at random in the analyses.[15,16]

We assessed the mean 12-month ODI raw score, as well as the mean ODI percentage change score and the mean ODI change score after 12 months against the GPE by one-way analyses of variance (ANOVA) with post hoc analysis (Tukey, $\alpha = .05$) and by analyses of covariance (ANCOVA, generalized linear model) with adjustment for baseline scores. Correlation analyses between the different ODI tools and the GPE were done by Spearman rank correlation.

Cutoffs for all scores were estimated by receiver operating characteristic (ROC) curves. We calculated cutoffs for a

Table 1. Baseline Characteristics of the Study Population.

Characteristic	Mean (SD)
Age (years)	48.7 (13.6)
BMI (kg/m^2)	26.6 (4.2)
ASA	1.5 (0.6)
ODI	46 (18.9)
Backpain, NRS	6.2 (2.5)
Legpain, NRS	6.9 (2.2)
Characteristic	**n (%)**
Female	3952 (58)
Smoker	1936 (27)
Married	3827 (56)
Emergency surgery	653 (10)
Lower education	4279 (63)
Comorbidity	1664 (28)
Previously operated	1417 (21)
Sickness benefits	4180 (61)

Abbreviations: SD, standard deviation; BMI, body mass index; NRS, numeric rating scale; ASA, American Society of Anesthesiologists score; ODI, Oswestry Disability Index score.

substantial improvement from baseline (GPE 1-2 vs 3-7), termed success.

To determine the cutoff with the highest sensitivity and specificity, the closest point to the upper left corner of the ROC curve was calculated from the coordinates of the curve. The area under the curve (AUC) determined how well the instruments differentiated between the outcome groups. An AUC value of >0.70 was considered acceptable, >0.80 good, and >0.9 excellent. The overall accuracy for each cutoff was calculated with a confusion matrix.[17]

To be able to study the impact of low and high baseline disability on the outcome cutoffs (success criteria), we split the patient sample based on the baseline ODI score into low (<25th percentile), medium (25th-75th percentile) and high disability (>75th percentile) and calculated cutoffs for the 12-months ODI raw score, ODI percentage change and ODI change after 12 months, for each of these percentiles.

Floor and ceiling effects were assessed by calculating the frequency of the highest and lowest possible scores at baseline. If 15% of patients had a minimal or maximal score value at baseline, these were considered as floor or ceiling effects.[18,19]

Results

Baseline characteristics of both respondents and nonrespondents of this patient population have been shown and discussed in a previous study.[11] Characteristics of the study population are listed in Table 1. Follow-up data after 12 months were available for 6840 (69%) out of 9930 of patients. The sample was divided into low ODI baseline (n = 1617), medium ODI baseline (n = 3718), and high ODI baseline (n = 1505). Only 13 data points (0.2%) were missing for the baseline ODI. At 12-month follow-up, 40 values (0.6%) were missing for the GPE and 11 (0.2%) on the ODI. As shown in a previous article, the lost to

Table 2. Baseline Adjusted Mean Scores of the 12-Month Oswestry Disability Index (ODI) Raw Score, ODI Change Score, and ODI Percentage Change Score Analysis of Covariance (ANCOVA) by the Global Perceived Effectiveness (GPE) Scale.[a]

	GPE						
	Completely Recovered (1)	Much Better (2)	Somewhat Better (3)	No Change (4)	Somewhat Worse (5)	Much Worse (6)	Worse Than Ever (7)
n (%)	1659 (24)	3265 (48)	1093 (16)	358 (5)	216 (3)	153 (2)	66 (1)
Mean baseline ODI (SD)	45.4 (19.6)	46.1 (19.1)	47.3 (18.0)	42.0 (17.1)	44.7 (16.9)	50.7 (16.5)	50.9 (17.1)
12-month ODI raw score (95% CI)	2.9 (2.4-3.3)	13.6 (13.3-14.0)	30.9 (30.3-31.5)	39.4 (38.4-40.4)	42.3 (41-43.6)	51.6 (50-53.1)	58.9 (56.5-61.2)
ODI change (95% CI)	43.1 (42.7-43.6)	32.4 (32.1-32.7)	15.1 (14.5-15.7)	6.6 (5.6-7.6)	3.7 (2.5-5.0)	−5.6 (−7.1 to −4.1)	−12.9 (−15.2 to −1.5)
ODI % change (95% CI)	93.4 (91.9-94.9)	66.7 (65.6-67.8)	26.7 (24.9-28. 6)	1.3 (−1.9 to 4.5)	−6.7 (−10.8 to −2.5)	−13.7 (−18.6 to −8.7)	−30.7 (−38.3 to −23.0)

Abbreviation: CI, confidence interval.
[a]Negative prefix indicates a worsening of the ODI from baseline. The mean ODI score for the entire study population prior to surgery was 46.

Table 3. Cutoff for the 12-Month Oswestry Disability Index (ODI) Raw Score, the 12-Month ODI Change Score, and the 12-Month ODI Percentage Change Score, Classifying Success in the Whole Study Population (Receiver Operating Curve [ROC] Analyses) and Accuracy (Confusion Matrix).

	AUC	95% CI	Cutoff	Sens/ Spec	Accuracy (%)
12-month ODI raw score	0.93	0.92-0.94	19	0.83/0.87	84
ODI change score	0.84	0.83-0.85	19	0.79/0.73	78
ODI percentage change score	0.93	0.92-0.93	52	0.85/0.85	85

Abbreviations: AUC, area under the curve; 95% CI, 95% confidence interval; Sens, sensitivity; Spec, specificity.

Table 4. Cutoffs for the 12-Month Oswestry Disability Index (ODI) Raw Score, the 12-Month ODI Change Score, and the 12-Month ODI Percentage Change Score When Classifying Success in Each of the 3 ODI Baseline Subgroups.[a]

Baseline Subgroup	AUC	95% CI	Cutoff	Sens/ Spec	Accuracy (%)
12-month ODI raw score					
<25th perc	0.92	0.90-0.93	13	0.81/0.88	83
25th-75th perc	0.95	0.94-0.95	21	0.85/0.89	86
>75th perc	0.94	0.93-0.96	28	0.89/0.85	88
ODI change					
<25th perc	0.89	0.88-0.91	9	0.77/0.84	79
25th-75th perc	0.92	0.91-0.93	24	0.83/0.84	83
>75th perc	0.92	0.91-0.94	48	0.85/0.84	85
ODI % change					
<25th perc	0.91	0.90-0.93	39	0.82/0.84	83
25th-75th perc	0.94	0.94-0.95	53	0.86/0.88	86
>75th perc	0.94	0.93-0.96	66	0.85/0.88	88

Abbreviations: AUC, area under the curve; 95% CI, 95% confidence interval; Sens, sensitivity; Spec, specificity; perc, percentile.
[a]Analyses were done by receiver operating curve (ROC) analyses. Overall accuracy was determined by a confusion matrix.

follow-up group contained more smokers, fewer with higher education, more sickness benefits recipients, more previously operated patients and fewer cases operated for paresis.[11] The Spearman correlation coefficients were 0.6 for the 12-month ODI change ($P < .001$), 0.8 for the 12-month ODI percentage change ($P < .001$), and 0.8 for the 12-month ODI raw score ($P < .001$). ANOVA with post hoc analysis indicated that the 12-month ODI raw scores of all estimates were significantly different between GPE categories. For each outcome, baseline adjusted mean ODI scores (ANCOVA), are shown in Table 2.

Cutoffs for Success

The discriminative ability for success was significantly higher for the ODI percentage change and the 12-month ODI raw score in comparison with the ODI change score (Table 3). In the subgroup analyses, we found that the cutoffs for success

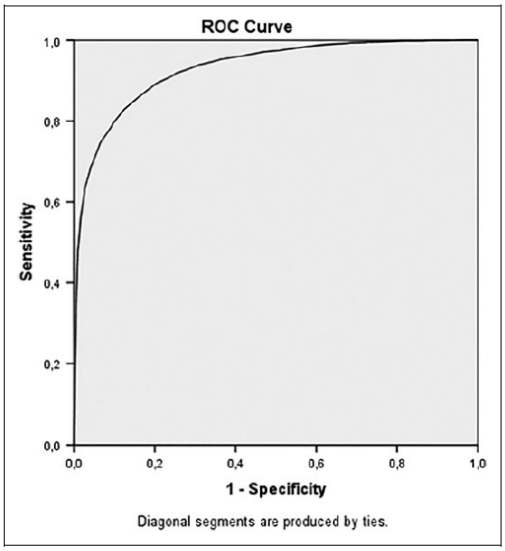


Figure 1. Receiver operating curve for the 12-month Oswestry Disability Index (ODI) score cutoff for "success". AUC (area under the curve) = 0.93 (0.92-0.94).

were dependent on the baseline ODI score. Patients with a low baseline ODI (<25th percentile, ODI score <32) had a cutoff on the 12-month ODI raw score (ODI % change) of 13 points (39%), those with medium baseline ODI (25th-75th percentile, ODI score 32-60) a cutoff of 21 points (55%), and those with high baseline ODI (>75th percentile, ODI score >60) a cutoff of 28 points (66%). The cutoffs for all ODI scores for all the different ODI baseline groups are listed in Table 4. Figure 4 shows that for the subgroups, the change cutoff (downward arrow) reaches the ODI raw score cutoff for success (horizontal line).

For the entire population, the cutoffs were 19 (ODI raw score), 19 (ODI change score), and 52% (ODI percentage change score) (Table 3). AUCs were high for all curves, ranging from 0.84 (ODI change score) to 0.93 (ODI raw score, ODI percentage score) (Figures 1–3).

Proportion of Success at 12 Months

For the entire population, the ODI percentage change score and the ODI raw scores corresponded better to a successful outcome (groups 1 and 2 on the GPE-scale) than the ODI change score (Table 3). Table 5 shows the proportion of cases classified as success 12 months after surgery. Table 1x (appendix) shows these proportions using separate cutoffs based on the different baseline ODI levels (percentiles). The ODI percentage change classified the highest proportions of success for the whole sample.

Floor and Ceiling Effects

No floor or ceiling effects were detected.

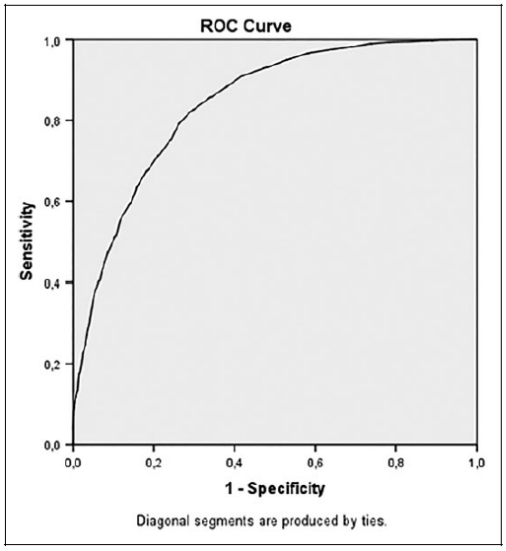


Figure 2. Receiver operating curve for the 12-month Oswestry Disability Index (ODI) change cutoff for "success". AUC (area under the curve) = 0.84 (0.83-0.85).

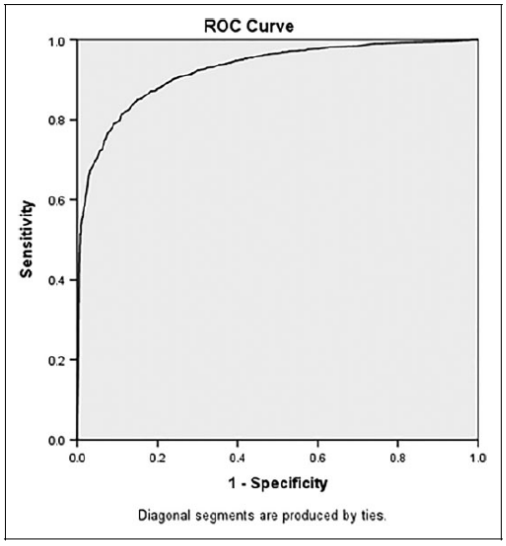


Figure 3. Receiver operating curve for the 12-month Oswestry Disability Index (ODI) percentage change cutoff for "success". AUC (area under the curve) = 0.93 (0.92-0.93).

Discussion

We found that success after surgery for lumbar disc herniation could as accurately be defined by the 12-month ODI raw score, as by the ODI percentage change score, and more accurately than by the ODI change scores from baseline. In a previous study we also found that the 12-month ODI raw score was more robust than the change scores for defining failure and worsening.[11] In the subgroup analyses we found that the cutoffs for success were dependent on the baseline ODI score. For those with low baseline disability the amount of improvement from baseline was considerably lower than for those with high baseline disability (Table 4). This dependency on the baseline score illustrates that patients perceive their postoperative improvements based on the amount of disability they experienced prior to surgery. Thus, in a patient sample with a low mean ODI other criteria for a positive outcome need to be applied, than in a patient sample with medium or high baseline ODI scores. This also implies that the previous recommendation to use a 30% change score cutoff for minimal clinical change[3] must be reconsidered for patients with medium and high baseline ODI scores. Our results confirm the importance of adjusting for baseline scores when comparing success rates between groups, for example, hospitals and surgical interventions.[11,20] When evaluating outcomes for individual patients or groups, one should consider using cutoffs according to baseline disability (low, medium, or high ODI scores). Moreover, statistical studies aimed at predicting outcome after surgery for lumbar disc herniation should be modeled with adjustment for preoperative ODI score, for example, by stratification.

The ODI change score had the lowest accuracy for defining success, especially among patients with high and low baseline disability. Therefore, we only recommend using the 12-month ODI raw score and the ODI percentage change score cutoffs.

Interestingly, the success rates among patients with low and high baseline scores were the same. This indicates that patients with low baseline disability may have higher demands for physical performance, and they may be more sensitive to smaller improvements which they would consider meaningful compared to those with high baseline disability.[21] Prior to surgery, these issues should be discussed with the patient. Differences in symptom tolerance before and after the operation may also reflect variation in patient expectations and coping strategies.

Table 5. Total Number (N) of Cases Classified as Success by Each Oswestry Disability Index (ODI) Outcome Tool, for the Entire Population.

	Total population, n (%)
12-month ODI score	
Success	4322 (63)
No success	2507 (37)
ODI change score	
Success	4391 (64)
No success	2425 (36)
ODI percentage change score	
Success	4446 (65)
No success	2361 (35)

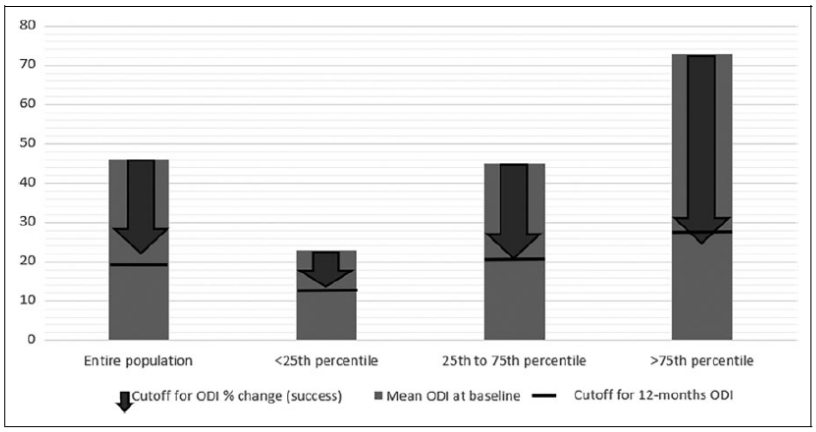


Figure 4. Mean baseline Oswestry Disability Index (ODI) (bar) with the 12-month ODI percentage change cutoff for success (arrow) and the 12-month ODI score cutoff for success (line).

Studies in rheumatology suggest that a treatment needs to reduce symptom intensity below a certain threshold to be perceived as successful by the patient. This threshold has been termed a "patient acceptable symptom state" (PASS).[8,9,22] As illustrated in Figure 4, it does not matter if a patient experiences, for instance, a 30% or 50% improvement of the baseline score, as long as he or she achieves the cutoff for the 12-month ODI raw score. Moreover, for the study population as a whole, the 12-month ODI raw score cutoff for success ($\leq$19) corresponds to what van Hooff et al[23] defined a cutoff for a patient acceptable symptom state (PASS = ODI $\leq$ 22 at follow-up).

Methodological Challenges

By collecting data from "real-world" clinical practice, studies from clinical registries not only have advantages such as large sample sizes and high external validity but also limitations such as lower follow-up rates compared with closely monitored clinical trials. Still, there is increasing evidence in the literature that observational studies conducted according to the STROBE check list report corresponding results similar to those found in randomized controlled trials.[24]

Loss to follow-up was 31%. In three previous studies from the Scandinavian spine registries (NORspine, SWEspine, and DANEspine), dropout cases (rates of 12%-38%) were traced and interviewed. These studies found the same differences in baseline characteristics that we found between patients who responded and those who did not, yet the same clinical outcomes at 1 and 2 years of follow-up.[15,16,25] Thus, we do not expect that loss to follow-up would bias our success rate estimates. Furthermore, the aim of the study was not a clinical effectiveness evaluation, but rather to define cutoffs for success over the wide range of different outcomes found in this large cohort. Generalizability of our findings beyond the Norwegian population is supported by previous comparative studies in Scandinavian countries and the United States, who report conceding results on baseline data and clinical outcomes (effect sizes).[26-28]

Using the GPE as an external anchor has been criticized since recall bias may exist. Moreover, the patients tend to be more focused on their current health state than health change when responding on a GPE scale, indicating a weakness of its construct validity.[11,29] The ideal anchor should objectively measure the patient's status before and after surgery with high reliability and validity. It should be easy to use, and universally applicable in different clinical settings. However, to the best of our knowledge no such anchor exists. In the search for such a tool, other variables have been evaluated by different research groups, such as return to work, use of painkillers, or surgeon-reported outcome. However, such measures also have limitations, namely, bias due to selection of certain subgroups, and subjective information based on surgeon's assessment of the clinical outcome.[30-33] Acknowledging these limitations, both the FDA (Food and Drug Administration) and the IMMPACT consensus group recommend a 7-point Likert-type scale, like the GPE, to be used as an external anchor.[34,35]

Conclusion

The ODI raw score can be used to define a successful outcome 12 months after surgery for lumbar disc herniation with high accuracy, similar that of the ODI percentage change. The ODI change score in points was not as accurate. Since these cutoffs are point estimates and vary depending on the baseline disability, adjustment for the baseline ODI should be performed when comparing success rates between hospitals or interventions. We recommend using ODI raw score or ODI percentage change (value in parentheses) cutoffs for success, according to their level of baseline disability, low= 13 points (39%), medium= 21 points (53%), or high= 28 points (66%).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by The Norwegian Medical Association–Foundation for Quality Improvement and Patient Safety, and The Regional Health Authority of Northern Norway.

ORCID iD

David A. T. Werner https://orcid.org/0000-0002-0865-7395

Supplemental Material

The supplemental material is available in the online version of the article.

References

1. Solberg TK, Olsen LR. Nasjonalt kvalitetsregister for ryggkirurgi (NKR): Årsrapport for 2015 med plan for forbedringstiltak 2016. In: Solberg TK, ed. *NORspine Annual Report*. Bodø, Norway: NORspine; 2016:46.
2. Deyo RA, Mirza SK. Clinical practice. Herniated lumbar intervertebral disk. *N Engl J Med*. 2016;374:1763-1772. doi:10.1056/NEJMcp1512658
3. Ostelo RWJG, Deyo RA, Stratford P, et al. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. *Spine (Phila Pa 1976)*. 2008;33:90-94. doi:10.1097/BRS.0b013e31815e3a10
4. Glassman SD, Copay AG, Berven SH, Polly DW, Subach BR, Carreon LY. Defining substantial clinical benefit following lumbar spine arthrodesis. *J Bone Joint Surg Am*. 2008;90:1839-1847.
5. Copay AG, Martin MM, Subach BR, et al. Assessment of spine surgery outcomes: inconsistency of change amongst outcome measurements. *Spine J*. 2010;10:291-296.
6. Solberg T, Johnsen LG, Nygaard ØP, Grotle M. Can we define success criteria for lumbar disc surgery? Estimates for a substantial amount of improvement in core outcome measures. *Acta Orthop*. 2013;84:196-201.
7. de Vet HCW, Foumani M, Scholten MA, et al. Minimally important change values of a measurement instrument depend more on baseline values than on the type of intervention. *J Clin Epidemiol*. 2015;68:518-524. doi:10.1016/j.jclinepi.2014.07.008
8. Escobar A, Riddle DL. Concordance between important change and acceptable symptom state following knee arthroplasty: the role of baseline scores. *Osteoarthritis Cartilage*. 2014;22:1107-1110. doi:10.1016/j.joca.2014.06.006
9. Tubach F, Ravaud P, Baron G, et al. Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement. *Ann Rheum Dis*. 2005;64:29-33. doi:10.1136/ard.2004.022905
10. Myles PS, Myles DB, Galagher W, et al. Measuring acute postoperative pain using the visual analog scale: the minimal clinically important difference and patient acceptable symptom state. *Br J Anaesth*. 2017;118:424-429. doi:10.1093/bja/aew466
11. Werner DAT, Grotle M, Gulati S, et al. Criteria for failure and worsening after surgery for lumbar disc herniation: a multicenter observational study based on data from the Norwegian Registry for Spine Surgery. *Eur Spine J*. 2017;26:2650-2659. doi:10.1007/s00586-017-5185-5
12. Ruyter KW. The Regional Ethical Committe for Medical Research, South-Eastern Norway Regional Health Trust: 22845518 24.06.2015. 2015.
13. Fairbank JCT. Why are there different versions of the Oswestry Disability Index? *J Neurosurg Spine*. 2014;20:83-86. doi:10.3171/2013.9.SPINE13344
14. Kamper SJ, Ostelo RW, Knol DL, Maher CG, de Vet HC, Hancock MJ. Global Perceived Effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status. *J Clin Epidemiol*. 2010;63:760-766.e1.
15. Hojmark K, Støttrup C, Carreon L, Andersen MO. Patient-reported outcome measures unbiased by loss of follow-up. Single-center study based on DaneSpine, the Danish spine surgery registry. *Eur Spine J*. 2015;25:282-286. doi:10.1007/s00586-015-4127-3
16. Solberg TK, Sørlie A, Sjaavik K, Nygaard ØP, Ingebrigtsen T. Would loss to follow-up bias the outcome evaluation of patients operated for degenerative disorders of the lumbar spine? *Acta Orthop*. 2011;82:56-63. doi:10.3109/17453674.2010.548024
17. Fawcett T. An introduction to ROC analysis. *Pattern Recogn Lett*. 2006;27:861-874. doi:10.1016/j.patrec.2005.10.010
18. de Vet HC, Ostelo RW, Terwee CB, et al. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. *Qual Life Res*. 2007;16:131-142.
19. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol*. 2007;60:34-42.
20. Fekete TF, Haschtmann D, Kleinstück FS, Porchet F, Jeszenszky D, Mannion AF. What level of pain are patients happy to live with after surgery for lumbar degenerative disorders? *Spine J*. 2016;16(4 suppl):S12-S18. doi:10.1016/j.spinee.2016.01.180
21. Mancuso CA, Duculan R, Stal M, Girardi FP. Patients expectations of lumbar spine surgery. *Eur Spine J*. 2014;24:2362-2369. doi:10.1007/s00586-014-3597-z
22. Tubach F, Dougados M, Falissard B, Baron G, Logeart I, Ravaud P. Feeling good rather than feeling better matters more to patients. *Arthritis Rheum*. 2006;55:526-530. doi:10.1002/art.22110
23. van Hooff ML, Mannion AF, Staub LP, Ostelo RWJG, Fairbank JCT. Determination of the Oswestry Disability Index score equivalent to a "satisfactory symptom state" in patients undergoing surgery for degenerative disorders of the lumbar spine—a Spine Tango registry-based study. *Spine J*. 2016;16:1221-1230. doi:10.1016/j.spinee.2016.06.010
24. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med*. 2000;342:1878-1886. doi:10.1056/NEJM200006223422506

25. Elkan P, Lagerbäck T, Möller TLH, Gerdhem P. Response rate does not affect patient-reported outcome after lumbar discectomy. *Eur Spine J*. 2018;27:1538-1546. doi:10.1007/s00586-018-5541-0
26. Weinstein J, Tosteson TD, Lurie JD, et al. Surgical vs nonoperative treatment for lumbar disk herniation: the Spine Patient Outcomes Research Trial (SPORT): a randomized trial. 2006;296: 2441-2450. doi:10.1001/jama.296.20.2441
27. Lønne G, Schoenfeld AJ, Cha TD, Nygaard ØP, Zwart JAH, Solberg T. Variation in selection criteria and approaches to surgery for lumbar spinal stenosis among patients treated in Boston and Norway. *Clin Neurol Neurosurg*. 2017;156:77-82. doi:10. 1016/j.clineuro.2017.03.008
28. Lagerbäck T, Fritzell P, Hägg O, Nordvall D, Lønne G. 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 [published online September 29, 2018]. *Eur Spine J*. doi:10.1007/s00586-018-5768-9
29. Grøvle L, Haugen AJ, Hasvik E, Natvig B, Brox JI, Grotle M. Patients' ratings of global perceived change during 2 years were strongly influenced by the current health status. *J Clin Epidemiol*. 2014;67:508-515. doi:10.1016/j.jclinepi.2013.12.001
30. Gatchel RJ, Mayer TG. Testing minimal clinically important difference: additional comments and scientific reality testing. *Spine J*. 2010;10:330-332. doi:10.1016/j.spinee.2010. 01.019
31. Glassman SD, Carreon LY. Thresholds for health-related quality of life measures: reality testing. *Spine J*. 2010;10:328-329. doi:10. 1016/j.spinee.2009.12.026
32. Ferreira ML, Herbert RD, Ferreira PH, et al. A critical review of methods used to determine the smallest worthwhile effect of interventions for low back pain. *J Clin Epidemiol*. 2012;65: 253-261.
33. Pochon L, Kleinstück FS, Porchet F, Mannion AF. Influence of gender on patient-oriented outcomes in spine surgery. *Eur Spine J*. 2016;25:235-246. doi:10.1007/s00586-015-4062-3
34. McLeod LD, Coon CD, Martin SA, Fehnel SE, Hays RD. Interpreting patient-reported outcome results: US FDA guidance and emerging methods. *Expert Rev Pharmacoecon Outcomes Res*. 2011;11:163-169. doi:10.1586/erp.11.12
35. Dworkin RH, Turk DC, Wyrwich KW, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *J Pain*. 2008;9:105-121. doi:10.1016/j.jpain.2007.09.005

15.3 Paper III.

A prognostic model for failure and worsening after lumbar microdiscectomy. A multicenter study from the Norwegian Registry for Spine Surgery (NORspine).

David A T Werner[1,2], Margreth Grotle[3,4], Milada Cvancarova Småstuen[3,4], Sasha Gulati[5,6], Øystein P Nygaard[5,6], Øyvind Salvesen[7], Tor Ingebrigtsen[1,2,5], and Tore K Solberg[1,2,5].

[1]Dept. of Neurosurgery, University Hospital of Northern Norway, Tromsø, Norway

[2]Department of Clinical Medicine, Faculty of Health Sciences, UiT the Arctic University of Norway, Tromsø, Norway

[3] Faculty of Health Sciences, Oslo Metropolitan University, Oslo, Norway

[4] Communication Unit for Musculoskeletal Disorders (FORMI), Oslo University Hospital, Oslo, Norway

[5]The Norwegian Registry for Spine Surgery (NORspine), University Hospital of Northern Norway, Tromsø, Norway

[6]Department of Neurosurgery, St. Olav University Hospital, Trondheim, Norway

[7]Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, Trondheim, Norway

Contact data of the corresponding author:
Name: David Werner
Institution: Dept. of Neurosurgery, University Hospital of Northern Norway, Tromsø, Norway
E-Mail address: Mail@david-werner.com
Telephone: +47 46951791

Element counts

Abstract word count: 242
Text word count: 3029
Number of references: 39
Number of tables: 4
Number of figures: 2

Abstract

Objective

To develop a prognostic model for failure and worsening one year after surgery for lumbar disc herniation.

Methods

This multicenter cohort study included 11081 patients operated with lumbar microdiscectomy, registered at the Norwegian Registry for Spine Surgery. Follow up was one year. Uni- and multivariate logistic regression analyses were used to assess potential prognostic factors for previously defined cut-offs for failure and worsening on the Oswestry Disability Index scores 12-months after surgery. Since the cut-offs for failure and worsening are different for patients with low, moderate and high baseline ODI scores, the multivariate analyses were run separately for these subgroups. Data were split into a training (70%) and a validation set (30%). The model was developed in the training set and tested in the validation set. A prediction (%) of an outcome was calculated for each patient in a risk matrix.

Results

The prognostic model produced six risk matrices based on three baseline ODI ranges (low, medium and high) and two outcomes (failure and worsening), each containing 7 to 11 prognostic factors. Model discrimination and calibration were acceptable. The estimated preoperative probabilities ranged from 3% to 94% for failure and from 1% to 72% for worsening in our validation cohort.

Conclusion

We developed a prognostic model for failure and worsening 12 months after surgery for lumbar disc herniation. The model showed acceptable calibration and discrimination and could be useful in assisting physicians and patients in clinical decision-making process prior to surgery.

Key Words: Microdiscectomy, outcome, PROM, quality, ODI, Lumbar disc surgery

BACKGROUND

Worldwide, low back pain is the leading cause for years lived with disability[14]. The most common indication for low back surgery is sciatica caused by lumbar disc herniation (LDH)[9]. The lifetime prevalence of sciatica in the general population has been reported between 12-27%[19]. If left untreated, most patients with LDH will have a favorable outcome. Surgery is typically offered to patients with persisting and/or intolerable leg pain with or without and low back pain, or with severe limb or bowel/bladder paresis (cauda equina syndrome)[3, 28]. The majority of the operations are preformed electively on relative indications.

Most clinical studies tend to focus on favorable outcomes after surgery based on mean improvements or success rates according to patient reported outcome measures (PROMs)[2, 3, 20, 28, 37], and predictive models for such outcomes have been developed[22, 24, 25]. An efficient strategy for improving the quality and safety of the health service is to increase the focus on unfavorable outcomes[8, 35]. Although the majority of patients experience substantial improvements, up to 30-40% report non-successful outcomes[2, 12, 23, 38], a large proportion of these cases cannot be classified as "failure"[6], indicating that non-success and failure are not interchangeable concepts.

The risk of a poor outcome is a frequent concern among patients being operated, especially the risk of getting worse, which indicates a harmful (adverse) treatment effect[32]. To enhance individualized risk prediction and prevention of unfavourable outcomes, we have previously defined benchmark criteria for both failure and worsening, based on frequently used PROMs[38]. A prediction model for unfavourable outcomes be can be further developed into a risk calculator, which could enhance shared clinical decision-making and improve selection of patients prior to lumbar disc surgery.

The aim of this study was to develop a prognostic model calculating individual risk (%) for failure and worsening after surgery for lumbar disc herniation, based on a large cohort from the Norwegian

registry for spine surgery (NORspine). Data from this large registry cohort, collected in daily surgical practice, would ensure high external validity, and thus clinical relevance.

MATERIAL AND METHODS

Design

Multicentre observational study following the recommendations for reporting in observational studies, STROBE criteria[36], and the methodological framework proposed by the PROGRESS group[34].

Study population and data collection

A total of 26427 patients operated for degenerative disorders of the lumbar spine reported to the NORspine registry between January 1st 2007 and August 2nd 2015 were screened for eligibility and followed for 12 months. The NORspine includes patients operated for degenerative disorders of the spinal column. It does not include patients with fractures, primary infections of the spine, or with spinal malignancies. Furthermore, it does not include children <16 years of age, as well as patients with known serious drug abuse or severe psychiatric disorders. For the purpose of this study, we included all patients who had a microscope or loupe assisted lumbar disc microdiscectomy for a magnetic resonance imaging (MRI) confirmed lumbar disc herniation. Both emergency and elective cases were registered. Patients diagnosed with lumbar spinal stenosis or spondylolisthesis, and those operated with more comprehensive decompression techniques including laminectomy, disc prosthesis or fusion procedures, were excluded.

The NORspine is a comprehensive clinical registry for quality control and research, covering 95% of public and private operating centers in Norway, with a completeness (proportion of operated patients reported to the registry) of 65% over the study period. It comprises a range of baseline data on known

and potential predictors for different outcomes[27]. Participation in NORspine is not required for a patient to gain access to the health care, or to receive payment/reimbursement for a provider.

At admission for surgery (baseline) the patients completed a questionnaire on demographics, lifestyle issues, and the PROMs. During the hospital stay the surgeon recorded data concerning diagnosis, treatment, comorbidity on a standard registration form. Twelve months after surgery a questionnaire identical to that used at baseline was distributed by regular mail. It was completed at home by the patients and returned to the central registry unit without involvement of the treating hospitals. One reminder with a new copy of the questionnaire was sent to those who did not respond. Informed consent was obtained from all patients.

The NORspine registry protocol has been approved by the Data Protection Authority of Norway. This study was submitted to the regional ethical committee for medical research which categorized it as a clinical audit study (2015/1829/REK South-East Regional Health Authority).

Outcomes

Failure and worsening, were defined according to validated cut-offs on the Oswestry Disability Index (ODI) version 2.1a, which showed the highest accuracy identifying these outcomes when evaluated against the Numeric Rating Scale for back-pain, leg-pain, and the EuroQol 5D (EQ-5D)[38]. The ODI contains ten questions about limitations of activities of daily living. Each item is rated from 0 to 5 and then transformed into a score ranging from 0 (none) to 100 (maximum pain-related disability)[4]. The ODI cut-offs have been determined according to an external anchor, the Global Perceived Effect scale (GPE, 1-7): 1 “fully recovered”, 2 “much better”, 3 “somewhat better”, 4 “unchanged”, 5 “somewhat worse”, 6 “much worse”, 7 “worse than ever”. Failure corresponds to GPE range 4-7, and worsening to GPE range 6-7[38, 39]. We have also shown that that both the ODI change score, as well as the final ODI score after 12 months are highly dependent on the preoperative ODI score [38, 39]. Therefore, we

stratified our model according to the preoperative ODI score (percentiles). Failure was defined as an ODI raw score 12 months after lumbar microdiscectomy ≥ 18 (low baseline ODI group, < 25 percentile), ≥ 29 (medium baseline ODI group, 25 to 75 percentile), and ≥ 34 (high baseline ODI group, > 75th percentile). Worsening was defined accordingly as an ODI raw score 12 months after lumbar discectomy ≥ 33 (low baseline ODI group), ≥ 47 (medium baseline ODI group), and ≥ 58 (high baseline ODI group)[38].

Possible Prognostic factors

We included prognostic factors, previously reported in the literature[10, 12, 15, 17, 18, 29]. Sociodemographic and anthropometric factors included were; gender, age > 60, obesity (body mass index, BMI ≥ 30), marital status (living alone yes/no), employment status (employed/unemployed) and low educational level (yes/no), i.e. less than four years of college/university education. Anxiety or depression was assessed by the item on the EuroQol-5D-3L questionnaire, (yes = "moderate" to "severe" problems, no = "no problems"). In Norway public health insurance is compulsory, thus no distinction was made between public or private insurance, or between public and private hospitals. A recent study has shown equivalent effectiveness of lumbar disc surgery between the public and private sector[21]. Patients were also asked if they had a pending or unresolved claim or litigation issue (yes/no) against, (1) the Norwegian public welfare agency fund concerning permanent disability pension, or (2) a compensation claim against private insurance companies or the public Norwegian System of Compensation to Patients. As shown in the tables we also assessed other clinical parameters, including the baseline PROM scores, smoking, duration of symptoms, previous lumbar spine surgery and use of analgesics[12, 15, 17, 18, 29].

Statistical analyses

All statistical analyses were performed with the Statistical Package for the Social Sciences (SPSS, IBM Version 23.0) and R (Version 2.13.1.) To assess potential sources of selection bias among patients, baseline differences between respondents and non-respondents at 12 months of follow-up were evaluated using the Students t-test for continuous variables or chi-square test for pairs of categorical variables. The proportions of missing data were small, <10% for all the analyzed variables. No imputation of missing values was performed.

Cases were selected for the training set (70%, n= 5741) and validation set (30%, n= 2218,) by the random sample function in SPSS (Figure 1)[7]. The models were built using the training set, and then the final models were assessed in the validation set. Since the ODI threshold values for failure and worsening after 12 months depend on the preoperative ODI baseline score, we stratified the prediction model into the three ODI percentiles of "low" ODI baseline scores (<33), "medium" (33 – 58), and "high" (>58) for each outcome[38, 39].

Training set

The outcomes failure versus no failure and worsening versus no worsening were modeled separately. Crude associations between each selected covariate and the outcome were assessed using univariate logistic regression. Variables that reached $p<0.1$ in these analyses were entered into the multivariate analyses (binary logistic regression model). In a next step, variables that were no longer statistically significant ($p<0.05$) were removed from the model using backward selection. We chose to include gender and age in all models, irrespectively of their statistical significance[31]. Continuous variables were dichotomized in order to be adapted into a risk matrix. Collinearity between possible predictors was assessed with Spearmans rho, with correlation coefficients (CC) >0.3 considered as weak, >0.5 as moderate, and >0.7 as strong. Associations between outcomes and prognostic factors were expressed as odd ratios (OR) with a 95% confidence interval (CI). Regression coefficients from the final models were

converted to probabilities for the risk matrix. Depending on the presence or absence of the risk factors, the matrix then calculated a probability for both failure and worsening for each patient.

Validation set

For each model, calibration was assessed by dividing the sample into four prediction groups (quartiles) with increasing probabilities for failure and worsening. We then plotted the observed proportion for these outcomes against the average predicted probability, using a logistic regression model with the observed binary outcome as dependent and the log odds of the validated regression model as independent. Chi square test was used to assess difference between coordinates the optimal prediction line. Significant deviation, indicating over- or underestimation, was defined as p-values <0.1. Discrimination was assessed by the c-criterion (C), calculated as the area under the curve (AUC) in a receiver operating analysis (ROC), plotting predicted probability against failure and worsening. C values >0.6 were considered acceptable[31].

RESULTS

Study population and data collection

We included 11081 patients in the analyses. Of these, 3621 (32.7%) were lost to follow-up 12 months after surgery (Figure1). Baseline characteristics for the entire study population are shown in table 1.

Mean age was 47.8 years (SD 13.61), and 42% of patients were females. Non-respondents at 12 months were younger, more likely to be men, had less severe comorbidity, and less severe limb paresis, but were more likely to be smokers, obese, anxious or depressed and previously operated. There were no clinically relevant differences in baseline pain and disability (PROMS) between respondents and non-respondents. The amount (n, %) of missing data for the prognostic factors was low for: age (6, 0.01),

gender (none), non-native Norwegian speaker (19, 0) living alone (43, 0.01), smoking (76, 0.01), having low education (52, 0.01), BMI (522, 11.2), American Association of Anesthesiologists (ASA) grade>2 (128, 1.8), unresolved disability pension issue (182, 3.4), unresolved insurance claim (171, 3.4), anxiety/depression (117, 1.6), duration of back pain >12 months (391, 5.6), back pain intensity (176, 2.4), and leg pain intensity (157, 2.2). Patient reported outcomes by baseline ODI (percentiles) subgroups in the training and the validation sets are shown in table 1x (supplementary appendix). For the entire study population, a total of 1779 cases (24.1%) were classified as failed and 469 (6.3%) as worsened.

Prognostic factors and outcomes

Tables 2x and 3x in the supplementary appendix show the results from the univariate analyses for all potential prognostic factors for failure and worsening, in both the training and validation sets. The results from the multivariate regression analyses for all three ODI baseline groups are shown in tables 2 (failure) and 3 (worsening). Duration of preoperative back pain was highly correlated (CC >0.7) with duration of preoperative leg pain. Duration of preoperative leg pain was consequently excluded from the model because of suspected multi-collinearity. Otherwise, all correlations between potential prognostic factors were low (CC ≤0.3).

The combination of the presence (yes) or absence (no) of each prognostic factor yield an overall probability for failure or worsening in each of the three ODI baseline groups. The matrices are shown as a flow chart in the supplementary appendix (Fig 7x). Table 4 illustrates three example cases from the risk matrices applied on the validation set. Each patient was allocated into 1 out of 6 matrices, based the baseline ODI (3 subgroups) and outcomes (2 subgroups). In the validation cohort, the individual predicted risk for failure ranged from 3% to 94% and from 1% to 72% for worsening.

The calibration plots showing agreement between the average predicted and observed proportion of failure and worsening (Figure 2) illustrate that the predicted and observed probabilities coincided well. There was no statistically significant deviation of the coordinates from the optimal prediction line, except for the model predicting worsening in the >75th percentile ODI baseline group. C-criterion values [95% CI] were; 0.66 [0.58 – 0.74], 0.74 [0.70 – 0.78], and 0.71 [0.66 – 0.76] for prediction of failure in the low, medium, and high baseline ODI groups, respectively, indicating acceptable discrimination. The corresponding c-criterion values for predicting worsening were similar: 0.68 [0.60 – 0.76], 0.74 [0.68 – 0.79], and 0.71 [0.61 – 0.81] All ROC curves for C calculations are shown in the supplementary appendix (Figs 1x-6x).

DISCUSSION

We have developed a prognostic model for unfavorable outcomes 12 months after surgery for lumbar disc herniation, based on validated and recommended PROMs[5]. Of all outcomes, 24% were classified as failure, and 6% as worsening. The estimated preoperative probabilities in our study population ranged from 3% to 94% for failure and from 1% to 72% for worsening, exemplified by three cases. This means that the model can identify patients with high and low baseline probability for unfavorable outcomes. It can be further developed into a calculator providing an absolute risk estimate. It should however be externally validated in other cohorts, and its feasibility should be confirmed by patients and clinicians before being implemented in regular clinical practice. Importantly, we have not assessed outcomes after non-operative treatment. Therefore, it is highly uncertain if the model could be useful in other settings, e.g. among patients seen in general practice.

The discriminative ability of risk the matrices was acceptable. Calibration assessment showed that patients with high baseline disability (>75th percentile of ODI) tended to underestimate the proportion of worsening, and the prediction of worsening among those cases was too inaccurate. A

reason could be the small sample size (type II error) this subgroup, or confounding due to unmeasured factors, such as widespread body pain and pain interference [1]. Confounding is the most likely source of bias in our study. We assessed anxiety and depression using one item of the EQ-5D 3L questionnaire, instead of a condition specific questionnaire which could be more sensitive. This may represent an information bias [12].

All cases of lumbar disc herniation were verified on MRI scans. evaluated by radiologists and surgeons, However, we did not have data on more specific morphological changes, e.g. contained versus uncontained herniation or additional Modic changes, which could influence the surgeons recommendation about surgery. This illustrates that statistical probabilities cannot be used as surrogate for clinical judgement, but rather as a supplementary decision support. We suggest that our model could be used in cases where the indication for surgery is uncertain. The model could be also helpful in calibrating surgeons' and patients' expectations about surgical outcomes.

To the best of our knowledge, this is the first registry study modeling unfavorable patient reported outcomes after lumbar disc surgery Three American studies have assessed patient populations operated for different degenerative spine disorders, including disc replacement and arthrodesis surgery[16, 24, 25]. The models were developed for predicting improvements, such as Minimal Clinically Important Change (MCIC), rather than unfavorable outcomes. Interestingly, 12 months of follow-up data from the latter paper by Khor et al. on a subgroup of 528 surgical patients showed that 222 of them reported an unsuccessful outcome (not reaching MCIC on the ODI scale)[16]. Of these, 86 (39%) reported to be unchanged or worse. The remaining 136 (61%) did not, hence representing a "grey zone" of patients with minor improvements. This supports our strategy of distinguishing failed from non-successful outcomes[38, 39].

It is important to acknowledge the conceptual differences between prognostic modelling and prognostic factors research. The prognostic model, developed in our study, aims at calculating the overall probability (individual absolute risk) for an outcome. Our study was not designed for prognostic factor research, which focuses on identifying independent prognostic (risk) factors[30, 34].

Still, our results can lend support to previously studies identifying as long duration low back pain and leg pain, anxiety and/or depression, previous back surgery, smoking, lower education, BMI, and unresolved disability pension or insurance issues as predictors for inferior outcomes[12, 15, 17, 18, 29].

Prediction models have to balance the need for accurate predictions against the risk of overfitting. Model overfitting implies lack of generalizability, i.e. it might work well for the population it was developed on, but not for others[26]. For instance, it is important not to include too many and/or too specific covariates. Our model appeared to be well balanced between an acceptable accuracy and a limited number of predictors, which are available in most clinical trials and regular clinical practice at the hospitals. We stratified our model by different levels of baseline disability (low ,medium and high ODI score), since the outcome score is highly dependent on the baseline score, and the actual cut offs for failure and worsening are different in these subgroups[16, 18, 38].

Registry-based studies collecting “real-life” data from daily clinical practice have advantages such as large sample sizes and high external validity, but also limitations such as lower follow-up rates[11]. Loss to follow-up at 12 months was 32.7%. Baseline characteristics-linked inferior outcomes seemed to be equally distributed between responders and non-responders. Still, loss to follow-up could represent a selection bias, especially when estimating exact failure and worsening rates. However, two Scandinavian registry studies on similar patient populations found that a loss to follow-up of did not bias conclusions about treatment effects[13, 33]. Moreover, the objective of or study was not effectiveness evaluations, but rather to develop a prediction model over a wide range of outcomes.

CONCLUSION

We have developed a prognostic model to identify patients at risk of unfavorable outcomes after lumbar microdiscectomy, which could assist physicians and patients in clinical decision-making prior to surgery.

Figure 1. Flow diagram of patient enrollment, exclusion and allocation.

Figure 2. Model validation. Observed proportion of the outcome (with confidence interval) on the vertical axis against average predicted probability of the outcome on the horizontal axis. Each coordinate with whiskers represents one quartile of estimated probability and its 95% confidence interval, compared to the observed proportion of the predicted outcome. The p-value from the chi square test for the coordinates vs the optimal prediction line is indicated in the lower right corner. A p-value < 0.1 indicates significant deviation from the average predicted probability. A-C show prediction of failure for the three baseline invalidity groups (A: Baseline ODI <25th percentile, B 25-75th percentile, C >75th percentile). D-F show prediction of worsening for the three baseline invalidity groups (D: Baseline ODI <25th percentile, E: 25th – 75th percentile, F: >75th percentile).

DISCLOSURES

Conflicts of interest

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Funding

The main author, David Werner, has received grants from the Regional Health Authority of Northen Norway, and the Norwegian Medical Association – Foundation for quality improvement and patient safety, for the purpose of this project.

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee (name of institute/committee) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.
This project was drafted with the regional ethical committee, which categorized it as a clinical audit study, not in need of their formal approval (REK 22845518.06.2015).

Informed consent

Informed consent was obtained from all individual participants of this study

REFERENCES

1. Ablin JN, Berman M, Aloush V, Regev G, Salame K, Buskila D, Lidar Z (2016) Effect of Fibromyalgia Symptoms on Outcome of Spinal Surgery. Pain Med pnw232
2. Atlas SJ, Keller RB, Wu YA, Deyo RA, Singer DE (2005) Long-term outcomes of surgical and nonsurgical management of lumbar spinal stenosis: 8 to 10 year results from the maine lumbar spine study. Spine (Phila Pa 1976) 30(1528–1159 (Electronic)):936–943
3. Bailey CS, Rasoulinejad P, Taylor D, et al (2020) Surgery versus conservative care for persistent sciatica lasting 4 to 12 months. N Engl J Med 382(12):1093–1102
4. Baker DJ PPB and FCT (1990) The Oswestry Disability Index revisited: its reliability, repeatability and validity, and a comparison with the St Thomas's Disability Index. Back PainNew approaches to Rehabil Educ 174–186
5. Clement RC, Welander A, Stowell C, et al (2015) A proposed set of metrics for standardized outcome reporting in the management of low back pain. Acta Orthop 86(5):523–533
6. Copay AG, Glassman SD, Subach BR, Berven S, Schuler TC, Carreon LY (2008) Minimum clinically important difference in lumbar spine surgery patients: a choice of methods using the Oswestry Disability Index, Medical Outcomes Study questionnaire Short Form 36, and pain scales. Spine J 8(1529-9430 (Print)):968–974
7. Dahl FA, Grotle M, Šaltyte Benth J, Natvig B (2008) Data splitting as a countermeasure against hypothesis fishing: With a case study of predictors for low back pain. Eur J Epidemiol 23(4):237–242
8. Deyo RA, Mirza SK (2009) The case for restraint in spinal surgery: Does quality management have a role to play? EurSpine J 18(1432–0932 (Electronic)):331–337
9. Deyo RA, Mirza SK (2016) Herniated Lumbar Intervertebral Disk. N Engl J Med 374(18):1763–1772

10. Fjeld O, Grotle M, Siewers V, Pedersen LM, Nilsen KB, Zwart J (2017) Prognostic Factors for Persistent Leg-Pain. doi: 10.1097/BRS.0000000000001773
11. Gliklich RE, Dreyer NA (2014) Registries for Evaluating Patient Outcomes, 3rd edition. , p Chapter 3, section 10
12. Haugen AJ, Brox JI, Grovle L, Keller A, Natvig B, Soldal D, Grotle M (2012) Prognostic factors for non-success in patients with sciatica and disc herniation. BMCMusculoskeletDisord 13:183. do:113–183
13. Hojmark K, Stottrup C, Carreon L, Andersen MO (2015) Patient-reported outcome measures unbiased by loss of follow-up. Single-center study based on DaneSpine, the Danish spine surgery registry. Eur spine J Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc 282–286
14. James SL, Abate D, Abate KH, et al (2018) Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 392(10159):1789–1858
15. Järvimäki V, Kautiainen H, Haanpää M, Koponen H (2016) Depressive symptoms are associated with poor outcome for lumbar spine surgery. 12:13–17
16. Khor S, Lavallee D, Cizik AM, et al (2018) Development and validation of a prediction model for pain and functional outcomes after lumbar spine surgery. JAMA Surg 153(7):634–642
17. Kleinstueck FS, Fekete T, Jeszenszky D, Mannion AF, Grob D, Lattig F, Mutter U, Porchet F (2011) The outcome of decompression surgery for lumbar herniated disc is influenced by the level of concomitant preoperative low back pain. Eur Spine J 20(7):1166–1173
18. Koerner JD, Glaser J, Radcliff K (2015) Which Variables Are Associated With Patient-reported Outcomes After Discectomy? Review of SPORT Disc Herniation Studies. Clin Orthop Relat Res

473(6):2000–2006

19. Konstantinou K, Dunn KM (2008) Sciatica: review of epidemiological studies and prevalence estimates. Spine (Phila Pa 1976) 33(22):2464–2472

20. Lurie JD, Tosteson TD, Tosteson ANA, Zhao W, Morgan TS, Abdu WA, Herkowitz H, Weinstein JN (2014) Surgical Versus Nonoperative Treatment for Lumbar Disc Herniation. Spine (Phila Pa 1976) 39(1):3–16

21. Madsbu MA, Salvesen Ø, Carlsen SM, Westin S, Onarheim K, Nygaard ØP, Solberg TK, Gulati S (2020) Surgery for herniated lumbar disc in private vs public hospitals: A pragmatic comparative effectiveness study. Acta Neurochir (Wien) 162(3):703–711

22. Mannion AF, Elfering A (2006) Predictors of surgical outcome and their assessment. EurSpine J 15 Suppl 1(0940-6719 (Print)):S93-108

23. Mannion AF, Impellizzeri FM, Leunig M, Jeszenszy D, Becker HJ, Haschtmann D, Preiss S, Fekete TF (2018) EUROSPINE 2017 FULL PAPER AWARD: Time to remove our rose-tinted spectacles: a candid appraisal of the relative success of surgery in over 4500 patients with degenerative disorders of the lumbar spine, hip or knee. Eur Spine J 27(4):778–788

24. McGirt MJ, Bydon M, Archer KR, et al (2017) An analysis from the Quality Outcomes Database, Part 1. Disability, quality of life, and pain outcomes following lumbar spine surgery: Predicting likely individual patient outcomes for shared decision-making. J Neurosurg Spine 27(4):357–369

25. McGirt MJ, Sivaganesan A, Asher AL, Devin CJ (2015) Prediction model for outcome after low-back surgery: Individualized likelihood of complication, hospital readmission, return to work, and 12-month improvement in functional disability. Neurosurg Focus 39(6):1–10

26. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG (2009) Prognosis and prognostic research: What, why, and how? BMJ 338(7706):1317–1320

27. Nerland US, Jakola AS, Solheim O, et al (2015) Minimally invasive decompression versus open

laminectomy for central stenosis of the lumbar spine: pragmatic comparative effectiveness study. BMJ 350(apr01 1):h1603–h1603

28. Peul WC, van Houwelingen HC, van den Hout WB, Brand R, Eekhof JAH, Tans JTJ, Thomeer RTWM, Koes BW, Leiden-The Hague Spine Intervention Prognostic Study Group (2007) Surgery versus prolonged conservative treatment for sciatica. N Engl J Med 356(22):2245–56

29. Pieber K, Salomon N, Inschlag S, Amtmann G, Resch KL, Ebenbichler G (2016) Predictors of an unfavorable outcome 1.5 and 12 years after a first, uncomplicated lumbar disc surgery. Eur Spine J 25(11):3520–3527

30. Riley RD, Hayden JA, Steyerberg EW, Moons KGM, Abrams K, Briggs A, Schroter S, Altman DG, Kyzas PA (2013) Prognosis Research Strategy (PROGRESS) 2 : Prognostic Factor Research. doi: 10.1136/bmj.e5595

31. Royston P, Moons KGM, Altman DG, Vergouwe Y (2009) Prognosis and prognostic research: Developing a prognostic model. BMJ 338:b604

32. Solberg TK, Nygaard OP, Sjaavik K, Hofoss D, Ingebrigtsen T (2005) The risk of "getting worse" after lumbar microdiscectomy. EurSpine J 14(0940-6719 (Print)):49–54

33. Solberg TK, Sorlie A, Sjaavik K, Nygaard OP, Ingebrigtsen T (2011) Would loss to follow-up bias the outcome evaluation of patients operated for degenerative disorders of the lumbar spine?: A study of responding and non-responding cohort participants from a clinical spine surgery registry. Acta Orthop 82(1):56–63

34. Steyerberg E, Moons KGM, van der Windt D, Hayden J, Perel P, Schroter S, Riley R, Hemingway H, Altman RB (2013) Prognosis research strategy PROGRESS series 3: prognostic model research. PLoS Med 10(2):e1001381

35. Taylor RS, Taylor RJ (2012) The economic impact of failed back surgery syndrome. Br J Pain 6(4):174–181

36. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M, STROBE Initiative (2007) Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Epidemiology 18(6):805–35

37. Weber H (1983) Lumbar disc herniation. A controlled, prospective study with ten years of observation. Spine (Phila Pa 1976) 8(0362–2436):131–140

38. Werner DAT, Grotle M, Gulati S, Austevoll IM, Lønne G, Nygaard Øystein P., Solberg TK (2017) Criteria for failure and worsening after surgery for lumbar disc herniation: a multicenter observational study based on data from the Norwegian Registry for Spine Surgery. Eur Spine J. doi: 10.1007/s00586-017-5185-5

39. Werner DAT, Grotle M, Gulati S, Austevoll IM, Madsbu MA, Lønne G, Solberg TK (2019) Can a Successful Outcome After Surgery for Lumbar Disc Herniation Be Defined by the Oswestry Disability Index Raw Score? Glob Spine J 219256821985148

Table 1

Table 1. Baseline characteristics including patient reported outcome measures of respondents vs non-respondents (lost to follow-up)

	Respondent	Non-respondent	P-value
Characteristics	**n= 7397 (67%)**	**n=3621 (33 %)**	
Female	3097 (41.9)	1374 (38.1)	<0.001
Age > 60	1403 (19)	307 (8.6)	<0.001
Living alone	1642 (22.4)	1048 (29.3)	<0.001
Non-native speaker	416 (5.6)	240 (6.7)	0.031
Low education[1]	2870 (39.1)	1168 (32.8)	<0.001
Had leg pain	7156 (96.7)	3518 (97.7)	0.007
Leg pain > 12 months	1668 (23.8)	855 (25.5)	0.066
Back pain > 12 months	2441 (34.8)	1219 (36.0)	0.212
Operated for paresis	1542 (20.8)	651 (18.1)	0.001
Paresis < grade 4	529 (35.2)	195 (30.7)	0.046
Emergency surgery	757 (10.2)	350 (9.7)	0.417
Comorbidity[2]	1891 (29.1)	842 (26.9)	0.026
ASA[3] grade > 2	408 (5.6)	152 (4.3)	0.004
Smoker	1935 (26.4)	1317 (37.0)	<0.001
Obesity[4]	1236 (18.6)	735 (22.4)	<0.001
Diabetes Mellitus	236 (3.2)	95 (2.6)	0.123
Anxiety/Depression[5]	3062 (42.1)	1608 (45.8)	<0.001
Unresolved disability pension issue[6]	879 (12.3)	398 (11.3)	0.173
Unresolved insurance claim[7]	419 (5.8)	230 (6.5)	0.167
Previous surgery	1602 (21.7)	932 (25.9)	<0.001
Previously operated > 2 times	72 (1.0)	53 (1.5)	0.026
PROMs	**mean (SD)**	**mean (SD)**	
ODI[8]	46.3 (19.2)	45.7 (18.6)	0.166
EQ-5D	0.27 (0.36)	0.25 (0.36)	0.125
NRS[9] back pain	6.2 (2.5)	6.4 (2.4)	0.024
NRS leg pain	6.9 (2.2)	6.9 (2.2)	0.492

[1]Less than four years of college/university education. [2]Rheumatoid arthritis, Ankylosing spondylitis, Other rheumatic disorder, Hip arthrosis, Knee arthrosis, Chronic generalized musculoskeletal pain, Chronic neurologic disorder, Cerebrovascular disorder, Heart disease, Vascular disease, Chronic lung disease, Cancer, Osteoporosis, Hypertension, Diabetes mellitus, Other endocrine disorder. [3]American Society of Anesthesiologists grade. [4]Body mass index ≥30. [5]EQ-5D 3L questionnaire; 5th item, moderate to severe problems. [6]Pending medical claim/litigation against the Norwegian public welfare agency fund concerning disability

pension. [7]Pending medical compensation claim/litigation against private insurance companies or the public Norwegian System of Compensation to Patients.
[8]Oswestry Disability Index, 0-100 (no-maximal disability). [9]Numeric rating scale (0-10).

Table 2. Results from the multiple regression model showing associations (Odds Ratio (OR) and 95% confidence intervals (CI)) between predictors and patient reported "failure" (unchanged or worse, yes/no) of lumbar disc surgery, as defined by validated cut offs on the Oswestry Disability Index (ODI), split on subgroups with low, medium and high baseline ODI scores (percentiles). For all predictors, except age and gender, NS indicates statistical insignificance, p value > 0.05.

	OR for failure by baseline ODI score[1]								
	Low ODI <25th percentile			Medium ODI 25th – 75th percentile			High ODI > 75th percentile		
Predictor	OR	95% CI[3]	P value	OR	95% CI	P value	OR	95% CI	P value
Female	1.3	0.9 – 1.7	0.146	1.2	1.0 – 1.5	0.092	1.3	0.9 – 1.7	0.175
Age > 60	1.0	0.7 – 1.5	0.941	1.2	0.9 –1.6	0.318	1.1	0.7 – 1.6	0.833
Low education[2]	1.5	1.1 – 2.0	0.011	1.8	1.4 – 2.3	<0.001	1.7	1.1 – 2.3	0.007
Non-native Norwegian speaker	NS	NS	NS	1.7	1.1 – 2.7	0.010	2.4	1.4 – 4.1	0.002
ASA[3] grade > 2	NS	NS	NS	NS	NS	NS	2.6	1.5 – 4.8	0.002
Obesity[4]	1.8	1.3 – 2.6	0.001	NS	NS	NS	1.5	1.1 – 2.3	0.025
Smoking	1.9	1.4 – 2.6	<0.001	1.6	1.3 – 2.1	0.001	1.6	1.1 – 2.3	0.008
Anxiety/Depression[5]	1.5	1.1 – 2.1	0.009	1.5	1.2 – 1.8	0.001	1.4	1.0 – 2.0	0.041
Back pain > NRS[6] 5	NS	NS	NS	1.5	1.1 – 2.0	0.015	3.0	1.3 – 2.7	0.009
Back pain > Leg pain	NS	NS	NS	1.7	1.3 – 2.2	<0.001	NS	NS	NS
Back pain > 12 months	2.3	1.8 – 3.1	<0.001	2.4	1.9 – 3.0	<0.001	2.8	2.0 – 3.9	<0.001
Previously operated	1.9	1.3 – 2.8	<0.001	2.3	1.8 – 3.0	<0.001	1.9	1.4 – 2.7	0.009
Unresolved disability pension issue[7]	2.8	1.7 – 4.9	<0.001	1.7	1.2 – 2.4	0.001	1.7	1.1 – 2.5	0.013
Unresolved insurance claim[8]	NS	NS	NS	1.6	1.0 – 2.5	0.048	1.7	1.0 – 3.0	0.048

[1]Range: 0-100 (no-maximal disability). The ODI score was <33, 33-58 and >58 in the subgroups with low, medium high baseline disability, respectively. [2]Less than four years of college/university education. [3]American Society of Anesthesiologists grade. [4]Body Mass Index ≥30. [5]EQ-5D 3L questionnaire; 5th item, moderate to severe problems. [6]Numeric Rating Scale (0-10). [7]Pending medical claim/ litigation the Norwegian public welfare agency fund concerning disability pension. [8]Pending medical compensation claim/litigation against private insurance companies or the public Norwegian System of Compensation to Patients.

Table 3

Table 3. Results from the multiple regression model showing associations (Odds Ratio (OR) and 95% confidence intervals (CI)) between predictors and patient reported worsening (yes/no) after lumbar disc surgery, as defined by validated cut offs on the Oswestry Disability Index (ODI), split on subgroups with low, medium and high baseline ODI scores (percentiles). For all predictors, except age and gender, NS indicates statistical insignificance, p value > 0.05.

	OR for worsening by baseline ODI score[1]								
	Low ODI <25th percentile			Medium ODI 25th – 75th percentile			High ODI > 75th percentile		
Predictor	OR[2]	95% CI	P value	OR	95% CI	P value	OR	95% CI	P value
Female	1.6	0.9 – 2.7	0.076	1.0	0.7 – 1.5	0.949	0.9	0.5 – 1.5	0.695
Age > 60	1.5	0.8 – 2.9	0.182	1.1	0.7 – 1.7	0.695	0.8	0.4 – 1.6	0.562
Low education[2]	2.7	1.5 – 5.1	0.002	1.8	1.1 – 2.7	0.010	2.0	1.1 – 3.7	0.022
Non-native Norwegian speaker	NS	NS	NS	2.8	1.6 – 4.9	0.001	3.8	1.9 – 7.6	<0.001
ASA[3] grade > 2	NS	NS	NS	NS	NS	NS	3.3	1.6 – 3.7	0.002
Obesity[4]	NS	NS	NS	NS	NS	NS	NS	NS	NS
Smoking	2.1	1.2 – 3.5	0.008	2.2	1.5 – 3.1	<0.001	2.3	1.4 – 3.8	0.001
Anxiety/Depression[5]	1.9	1.1 – 3.2	0.021	NS	NS	NS	NS	NS	NS
Back pain > NRS[6] 5	NS	NS	NS	2.2	1.2 – 4.1	<0.011	NS	NS	NS
Back pain > 12 months	2.7	1.6 – 4.5	<0.001	2.9	2.0 – 4.2	<0.001	3.4	2.1 – 5.6	<0.001
Previously operated	2.6	1.4 – 4.6	0.002	3.3	2.3 – 4.8	<0.001	NS	NS	NS
Unresolved insurance claim[7]	NS	NS	NS	NS	NS	NS	2.9	1.8 – 4.9	0.002

[1]Range: 0-100 (no-maximal disability)) The ODI score was <33, 33-58 and >58 in the subgroups with low, medium high baseline disability, respectively. [2]Less than four years of college/university education. [3]American Society of Anesthesiologists grade. [4]Body Mass Index ≥30. [5]EQ-5D 3L questionnaire; 5th item, moderate to severe problems. [6]Numeric Rating Scale (0-10). [7]Pending medical compensation claim/litigation against private insurance companies or the public Norwegian System of Compensation to Patients.

Table 4

Table 4. Example cases from the validation set (patients 1-3) with different predicted probability (6 risk matrices) for failure and worsening based on baseline ODI score and presence (yes) or absence (no) of predictors. An open cell indicates that predictor was not relevant for the risk matrix the patient was assigned to.

	Patient 1	Patient 2	Patient 3
Preoperative ODI score[1]	32	53	68
Female	No	Yes	No
Age > 60	Yes	Yes	Yes
Low education[2]	No	Yes	Yes
Non-native Norwegian speaker		No	No
ASA[3] grade > 2			Yes
Obesity[4]	No		No
Smoking	No		Yes
Anxiety/Depression[5]	Yes	Yes	Yes
Back pain > NRS[6] 5		Yes	Yes
Back pain > Leg pain		No	
Back pain > 12 months	No	Yes	Yes
Previously operated	No	No	Yes
Unresolved disability pension issue[7]	No	Yes	Yes
Unresolved insurance claim[8]		No	Yes
Predicted risk for failure	13%	50%	94%
Predicted risk for worsening	2%	6%	55%

[1]Range: 0-100 (no-maximal disability). [2]Less than four years of college/university education. [3]American Society of Anesthesiologists grade [4]Body Mass Index ≥30. [5]EQ-5D 3L questionnaire; 5th item, moderate to severe problems. [6]Numeric Rating Scale (0-10). [7]Pending medical claim/ litigation the Norwegian public welfare agency fund concerning disability pension. [8]Pending medical compensation claim/litigation against private insurance companies or the public Norwegian System of Compensation to Patients.

Figure 1

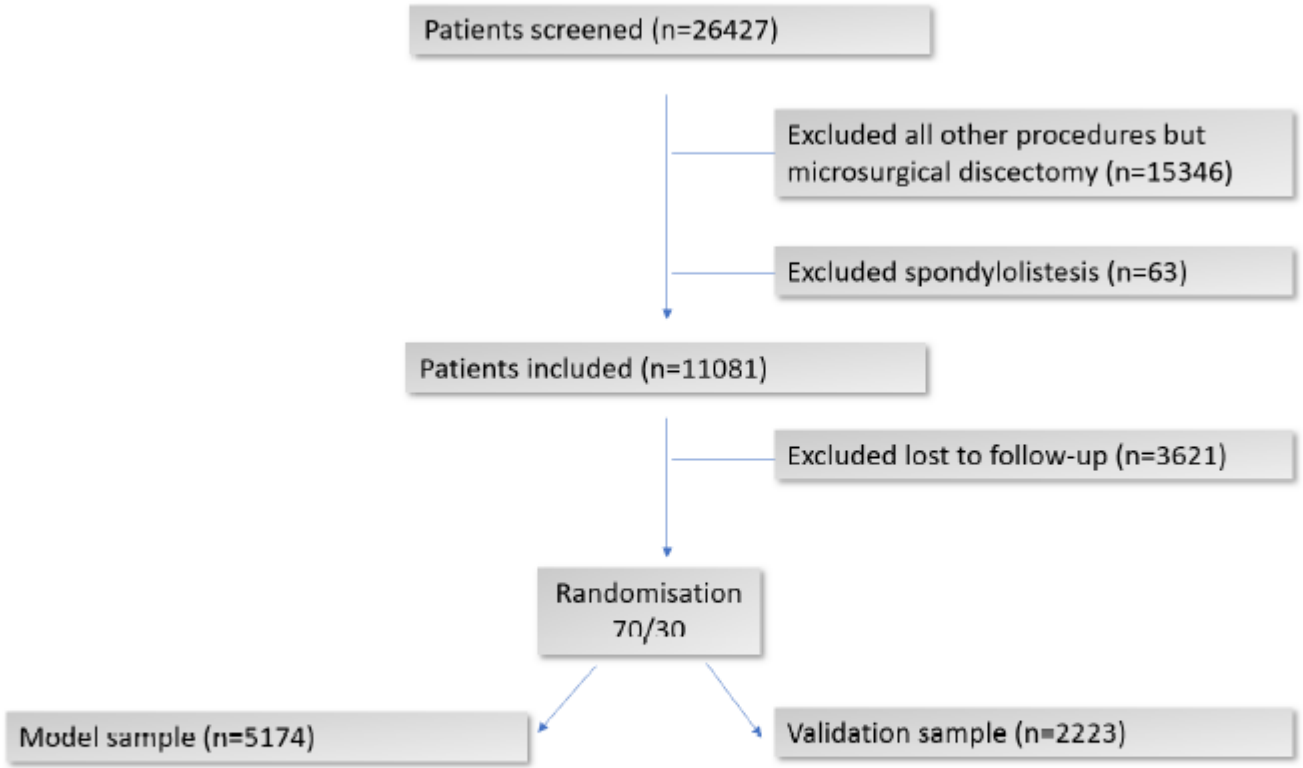


Figure 1. Flow diagram of patient enrollment, exclusion and allocation.

Figure 2

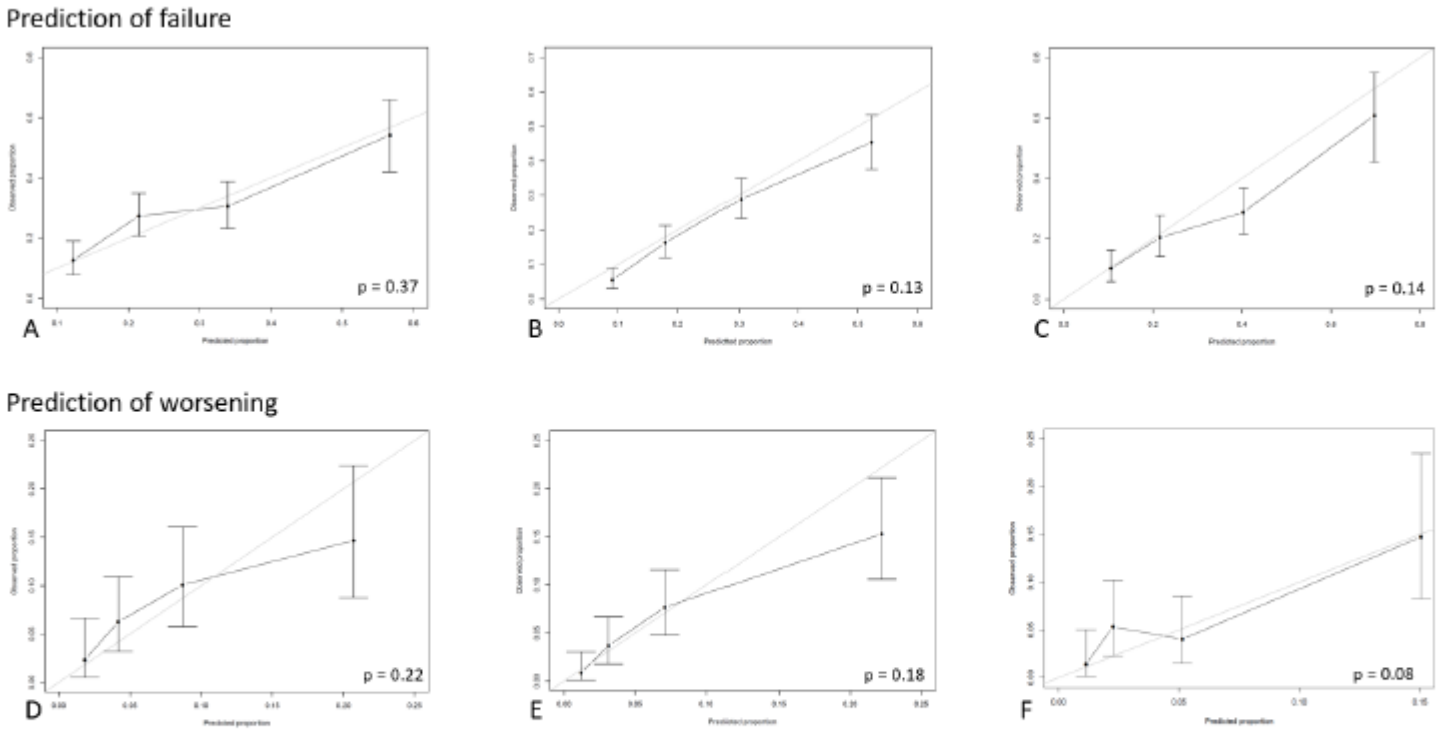
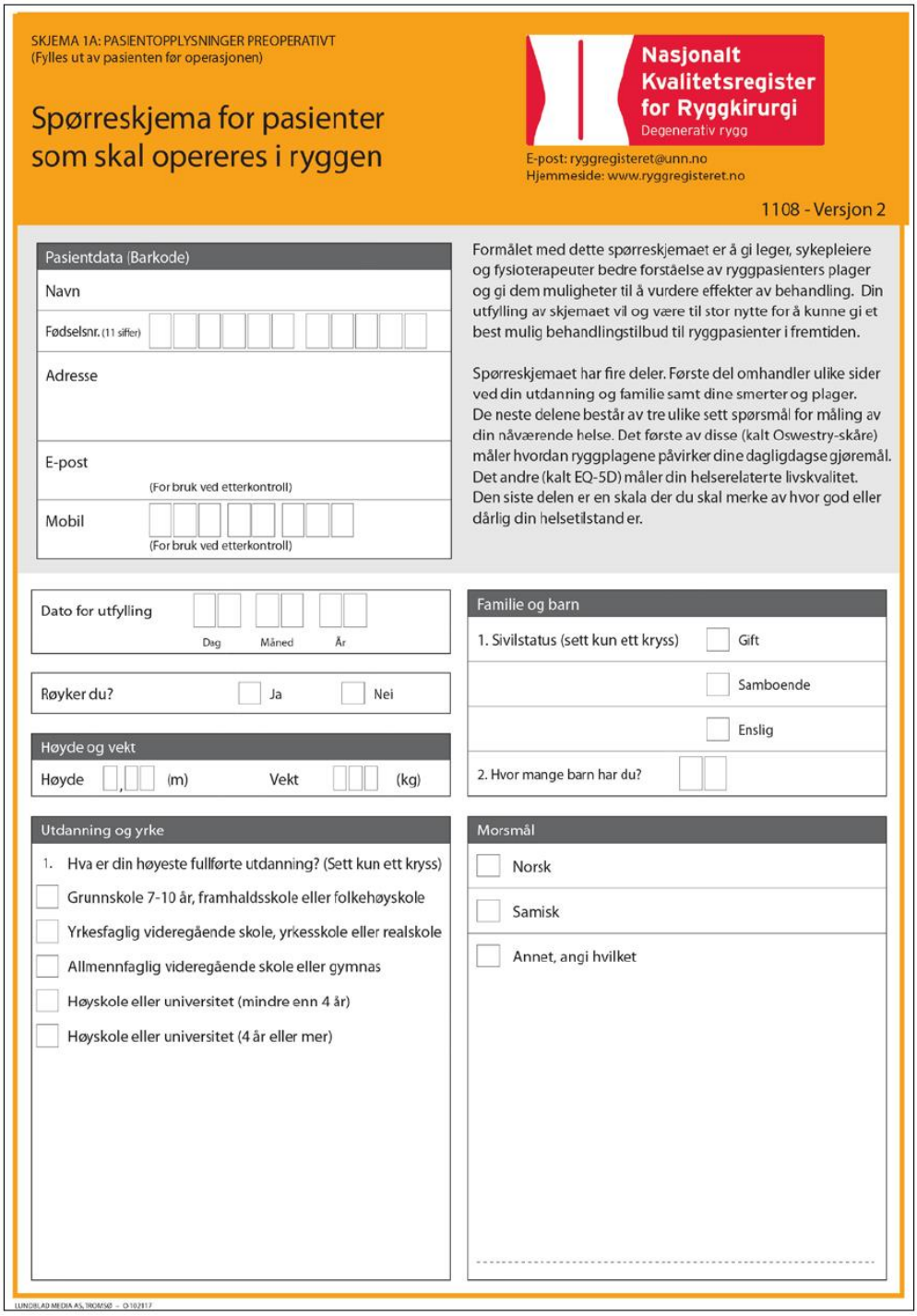


Figure 2. Model validation. Observed proportion of the outcome (with confidence interval) on the vertical axis against average predicted probability of the outcome on the horizontal axis. Each coordinate with whiskers represents one quartile of estimated probability and its 95% confidence interval, compared to the observed proportion of the predicted outcome. The p-value from the chi square test for the coordinates vs the optimal prediction line is indicated in the lower right corner. A p-value < 0.1 indicates significant deviation from the average predicted probability. A-C show prediction of failure for the three baseline invalidity groups (A: Baseline ODI <25th percentile, B 25-75th percentile, C >75th percentile). D-F show prediction of worsening for the three baseline invalidity groups (D: Baseline ODI <25th percentile, E: 25th – 75th percentile, F: >75th percentile).

16 Appendix

16.1 NORspine questionnaires (in Norwegian)



SKJEMA 1A: PASIENTOPPLYSNINGER PREOPERATIVT
(Fylles ut av pasienten før operasjonen)

Spørreskjema for pasienter som skal opereres i ryggen

Nasjonalt Kvalitetsregister for Ryggkirurgi
Degenerativ rygg

E-post: ryggregisteret@unn.no
Hjemmeside: www.ryggregisteret.no

1108 - Versjon 2

Pasientdata (Barkode)

Navn

Fødselsnr. (11 siffer)

Adresse

E-post
(For bruk ved etterkontroll)

Mobil
(For bruk ved etterkontroll)

Formålet med dette spørreskjemaet er å gi leger, sykepleiere og fysioterapeuter bedre forståelse av ryggpasienters plager og gi dem muligheter til å vurdere effekter av behandling. Din utfylling av skjemaet vil og være til stor nytte for å kunne gi et best mulig behandlingstilbud til ryggpasienter i fremtiden.

Spørreskjemaet har fire deler. Første del omhandler ulike sider ved din utdanning og familie samt dine smerter og plager. De neste delene består av tre ulike sett spørsmål for måling av din nåværende helse. Det første av disse (kalt Oswestry-skåre) måler hvordan ryggplagene påvirker dine dagligdagse gjøremål. Det andre (kalt EQ-5D) måler din helserelaterte livskvalitet. Den siste delen er en skala der du skal merke av hvor god eller dårlig din helsetilstand er.

Dato for utfylling
Dag Måned År

Røyker du? Ja Nei

Høyde og vekt

Høyde , (m) Vekt (kg)

Familie og barn

1. Sivilstatus (sett kun ett kryss) Gift
Samboende
Enslig

2. Hvor mange barn har du?

Utdanning og yrke

1. Hva er din høyeste fullførte utdanning? (Sett kun ett kryss)
- Grunnskole 7-10 år, framhaldsskole eller folkehøyskole
- Yrkesfaglig videregående skole, yrkesskole eller realskole
- Allmennfaglig videregående skole eller gymnas
- Høyskole eller universitet (mindre enn 4 år)
- Høyskole eller universitet (4 år eller mer)

Morsmål

- Norsk
- Samisk
- Annet, angi hvilket

Hvor sterke smerter har du hatt siste uke?

Hvordan vil du gradere smertene du har hatt i rygg/hofte i løpet av den siste uken? Sett ring rundt ett tall.

0 1 2 3 4 5 6 7 8 9 10

Ingen smerter — Så vondt som det går an å ha

Hvordan vil du gradere de smertene du har hatt i benet (ett eller begge) i løpet av den siste uken? Sett ring rundt ett tall.

0 1 2 3 4 5 6 7 8 9 10

Ingen smerter — Så vondt som det går an å ha

Funksjonsscore (Oswestry)

Disse spørsmålene er utarbeidet for å gi oss informasjon om hvordan dine smerter har påvirket dine muligheter til å klare dagliglivet ditt. Vær snill å besvare spørsmålene ved å sette kryss (kun ett kryss for hvert avsnitt) i de rutene som passer best for deg.

1. Smerte

- ☐ Jeg har ingen smerter for øyeblikket
- ☐ Smertene er veldig svake for øyeblikket
- ☐ Smertene er moderate for øyeblikket
- ☐ Smertene er temmelig sterke for øyeblikket
- ☐ Smertene er veldig sterke for øyeblikket
- ☐ Smertene er de verste jeg kan tenke meg for øyeblikket

2. Personlig stell

- ☐ Jeg kan stelle meg selv på vanlig måte uten at det forårsaker ekstra smerter
- ☐ Jeg kan stelle meg selv på vanlig måte, men det er veldig smertefullt
- ☐ Det er smertefullt å stelle seg selv, og jeg gjør det langsomt og forsiktig
- ☐ Jeg trenger noe hjelp, men klarer det meste av mitt personlige stell
- ☐ Jeg trenger hjelp hver dag til det meste av eget stell
- ☐ Jeg kler ikke på meg, har vanskeligheter med å vaske meg og holder sengen

3. Å løfte

- ☐ Jeg kan løfte tunge ting uten å få mer smerter
- ☐ Jeg kan løfte tunge ting, men får mer smerter
- ☐ Smertene hindrer meg i å løfte tunge ting opp fra gulvet, men jeg greier det hvis det som skal løftes er gunstig plassert, for eksempel på et bord
- ☐ Smertene hindrer meg i å løfte tunge ting, men jeg klarer lette og middels tunge ting, hvis det er gunstig plassert
- ☐ Jeg kan bare løfte noe som er veldig lett
- ☐ Jeg kan ikke løfte eller bære noe i det hele tatt

4. Å gå

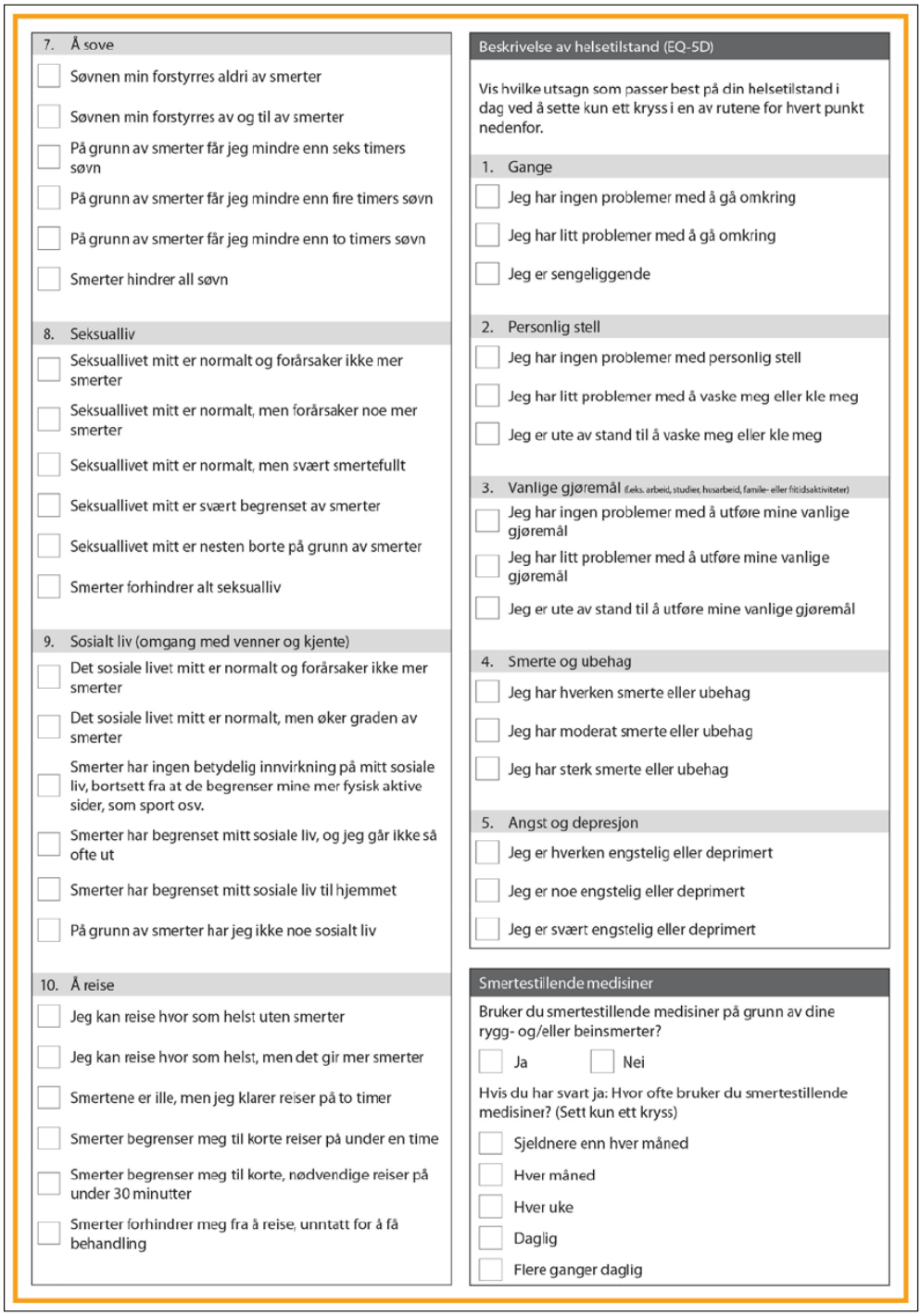
- ☐ Smerter hindrer meg ikke i å gå i det hele tatt
- ☐ Smerter hindrer meg i å gå mer enn 1 ½ km
- ☐ Smerter hindrer meg i å gå mer enn ¾ km
- ☐ Smerter hindrer meg i å gå mer enn 100 m
- ☐ Jeg kan bare gå med stokk eller krykker
- ☐ Jeg ligger for det meste i sengen, og jeg må krabbe til toalettet

5. Å sitte

- ☐ Jeg kan sitte så lenge jeg vil i en hvilken som helst stol
- ☐ Jeg kan sitte så lenge jeg vil i min favorittstol
- ☐ Smerter hindrer meg i å sitte i mer enn en time
- ☐ Smerter hindrer meg i å sitte i mer enn en halv time
- ☐ Smerter hindrer meg i å sitte i mer enn ti minutter
- ☐ Smerter hindrer meg i å sitte i det hele tatt

6. Å stå

- ☐ Jeg kan stå så lenge jeg vil uten å få mer smerter
- ☐ Jeg kan stå så lenge jeg vil, men får mer smerter
- ☐ Smerter hindrer meg i å stå i mer enn en time
- ☐ Smerter hindrer meg i å stå i mer enn en halv time
- ☐ Smerter hindrer meg i å stå i mer enn ti minutter
- ☐ Smerter hindrer meg i å stå i det hele tatt



7. Å sove

- Søvnen min forstyrres aldri av smerter
- Søvnen min forstyrres av og til av smerter
- På grunn av smerter får jeg mindre enn seks timers søvn
- På grunn av smerter får jeg mindre enn fire timers søvn
- På grunn av smerter får jeg mindre enn to timers søvn
- Smerter hindrer all søvn

8. Seksualliv

- Seksuallivet mitt er normalt og forårsaker ikke mer smerter
- Seksuallivet mitt er normalt, men forårsaker noe mer smerter
- Seksuallivet mitt er normalt, men svært smertefullt
- Seksuallivet mitt er svært begrenset av smerter
- Seksuallivet mitt er nesten borte på grunn av smerter
- Smerter forhindrer alt seksualliv

9. Sosialt liv (omgang med venner og kjente)

- Det sosiale livet mitt er normalt og forårsaker ikke mer smerter
- Det sosiale livet mitt er normalt, men øker graden av smerter
- Smerter har ingen betydelig innvirkning på mitt sosiale liv, bortsett fra at de begrenser mine mer fysisk aktive sider, som sport osv.
- Smerter har begrenset mitt sosiale liv, og jeg går ikke så ofte ut
- Smerter har begrenset mitt sosiale liv til hjemmet
- På grunn av smerter har jeg ikke noe sosialt liv

10. Å reise

- Jeg kan reise hvor som helst uten smerter
- Jeg kan reise hvor som helst, men det gir mer smerter
- Smertene er ille, men jeg klarer reiser på to timer
- Smerter begrenser meg til korte reiser på under en time
- Smerter begrenser meg til korte, nødvendige reiser på under 30 minutter
- Smerter forhindrer meg fra å reise, unntatt for å få behandling

Beskrivelse av helsetilstand (EQ-5D)

Vis hvilke utsagn som passer best på din helsetilstand i dag ved å sette kun ett kryss i en av rutene for hvert punkt nedenfor.

1. Gange

- Jeg har ingen problemer med å gå omkring
- Jeg har litt problemer med å gå omkring
- Jeg er sengeliggende

2. Personlig stell

- Jeg har ingen problemer med personlig stell
- Jeg har litt problemer med å vaske meg eller kle meg
- Jeg er ute av stand til å vaske meg eller kle meg

3. Vanlige gjøremål (f.eks. arbeid, studier, husarbeid, famile- eller fritidsaktiviteter)

- Jeg har ingen problemer med å utføre mine vanlige gjøremål
- Jeg har litt problemer med å utføre mine vanlige gjøremål
- Jeg er ute av stand til å utføre mine vanlige gjøremål

4. Smerte og ubehag

- Jeg har hverken smerte eller ubehag
- Jeg har moderat smerte eller ubehag
- Jeg har sterk smerte eller ubehag

5. Angst og depresjon

- Jeg er hverken engstelig eller deprimert
- Jeg er noe engstelig eller deprimert
- Jeg er svært engstelig eller deprimert

Smertestillende medisiner

Bruker du smertestillende medisiner på grunn av dine rygg- og/eller beinsmerter?

- Ja
- Nei

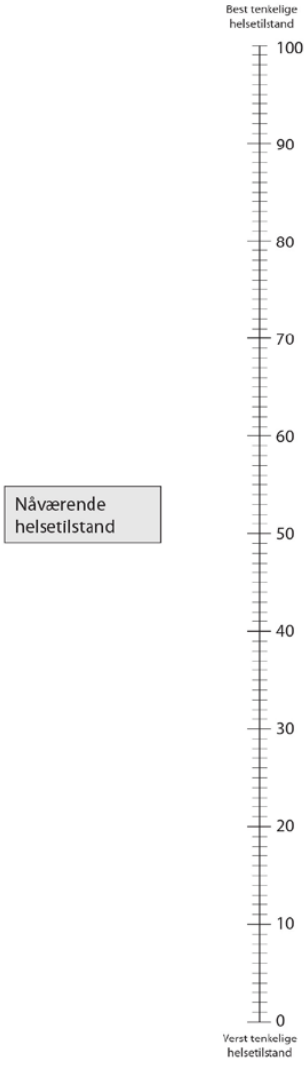
Hvis du har svart ja: Hvor ofte bruker du smertestillende medisiner? (Sett kun ett kryss)

- Sjeldnere enn hver måned
- Hver måned
- Hver uke
- Daglig
- Flere ganger daglig

Helsetilstand

For at du skal kunne vise oss hvor god eller dårlig din helsetilstand er, har vi laget en skala (nesten som et termometer), hvor den beste helsetilstanden du kan tenke deg er markert med 100 og den dårligste med 0.

Vi ber om at du viser din helsetilstand ved å trekke ei linje fra boksen nedenfor til det punkt på skalaen som passer best med din helsetilstand.



Symptomvarighet

Varighet av nåværende rygg-/hoftesmerter(sett kun ett kryss):

- ☐ Jeg har ingen rygg-/hoftesmerter
- ☐ Mindre enn 3 måneder
- ☐ 3 til 12 måneder
- ☐ 1 til 2 år
- ☐ Mer enn 2 år

Varighet av nåværende utstrålende smerter:

- ☐ Jeg har ingen utstrålende smerter
- ☐ Mindre enn 3 måneder
- ☐ 3 til 12 måneder
- ☐ 1 til 2 år
- ☐ Mer enn 2 år

Varighet sykemelding/attføring/rehabilitering pga aktuelle plager ☐☐☐ (uker)

Arbeidsstatus

- ☐ I arbeid
- ☐ Hjemmeværende, uiennet
- ☐ Student/skoleelev
- ☐ Alderspensjonist
- ☐ Arbeidsledig
- ☐ Sykemeldt
- ☐ Aktivt sykemeldt
- ☐ Delvis sykemeldt ..... % sykemeldt
- ☐ Attføring/rehabilitering
- ☐ Uføretrygdet evt ..... % uføretrygdet

Har du søkt om uføretrygd?

(Sett kun ett kryss)

- ☐ Ja
- ☐ Nei
- ☐ Planlegger å søke
- ☐ Er allerede innvilget

Har du søkt om erstatning fra forsikringsselskap eller folketrygden (eventuelt yrkesskadeerstatning)?

(Sett kun ett kryss)

- ☐ Ja
- ☐ Nei
- ☐ Planlegger å søke
- ☐ Er allerede innvilget

SKJEMA 2A:
SYKEPLEIER/LEGEOPPLYSNINGER PREOPERATIVT
(Fylles ut av lege samtidig med operasjonsbeskrivelsen
og suppleres evt. ved utstrivelse eller ved innrapportering)

Registreringsskjema for pasienter som opereres i ryggen

Nasjonalt Kvalitetsregister for Ryggkirurgi
Degenerativ rygg

E-post: ryggregisteret@unn.no
Hjemmeside: www.ryggregisteret.no

1108 - Versjon 2

Operasjonsdato (Må fylles ut) Dag Måned År

Dato for utfylling Dag Måned År

Pasientdata (Barkode)

Navn

Fødselsnr. (11 siffer)

Sykehistorie

Tidligere ryggoperert?

☐ Ja, samme nivå ☐ Ja, annet nivå ☐ Nei

- Pasienten har vært operert ☐☐ ganger tidligere i LS-kolumna

Andre relevante sykdommer, skader eller plager

☐ Nei

Ja, spesifiser:

- ☐ Reumatoid artritt
- ☐ Mb. Bechterew
- ☐ Annen reumatisk sykdom
- ☐ Hofte- eller kneartrose
- ☐ Depresjon / Angst
- ☐ Kroniske smerter i muskel-skjelettsystemet
- ☐ Kronisk nevrologisk sykdom
- ☐ Cerebrovaskulær sykdom
- ☐ Hjerte eller karsykdom
- ☐ Vaskulær Claudicatio
- ☐ Kronisk lungesykdom
- ☐ Kreftsykdom
- ☐ Osteoporose
- ☐ Hypertensjon
- ☐ Diabetes Mellitus
- ☐ Annen endokrin sykdom

Annet, spesifiser .....

Radiologisk vurdering (Sett evntuelt flere kryss)

1. Undersøkelse

- ☐ CT
- ☐ MR
- ☐ Radikulografi
- ☐ Diskografi
- ☐ Diagnostisk blokade
- ☐ Røntgen LS-columna
- ☐ Med fleksjon/ekstensjon

2. Funn

- ☐ Normal
- ☐ Skiveprolaps
- ☐ Sentral spinalstenose
- ☐ Lateral spinalstenose
- ☐ Foraminal stenose
- ☐ Degenerativ rygg/skivedegenerasjon
- ☐ Istmisk spondylolistese
- ☐ Degenerativ spondylolistese
- ☐ Degenerativ skoliose
- ☐ Synovial syste
- ☐ Pseudomeningocele

☐ Annet, spesifiser .....

Operasjonsindikasjon (Sett evntuelt flere kryss)

- ☐ Smerter
  - ☐ Rygg-/hoftesmerter
  - ☐ Bensmerter
  - ☐ Begge deler
- ☐ Parese, Grad (0-5): ..... Se eventuelt rettledning
- ☐ Cauda equina syndrom
- ☐ Annet, spesifiser .....

Ved tidlig reoperasjon (innen 90 dager), årsak: (Kun ett kryss)

- ☐ Recidiv prolaps
- ☐ Durarift
- ☐ Hematom
- ☐ Dyp infeksjon
- ☐ Overfladisk infeksjon
- ☐ Postoperativ spondylolisthese
- ☐ Løsning/feilplassering av osteosyntesemateriale

☐ Annet, spesifiser .....

Operasjonskategori

☐ Elektiv ☐ Øyeblikkelig hjelp ☐ ½ øyeblikkelig hjelp

Dagkirurgi (ingen døgnopphold på avdelingen)

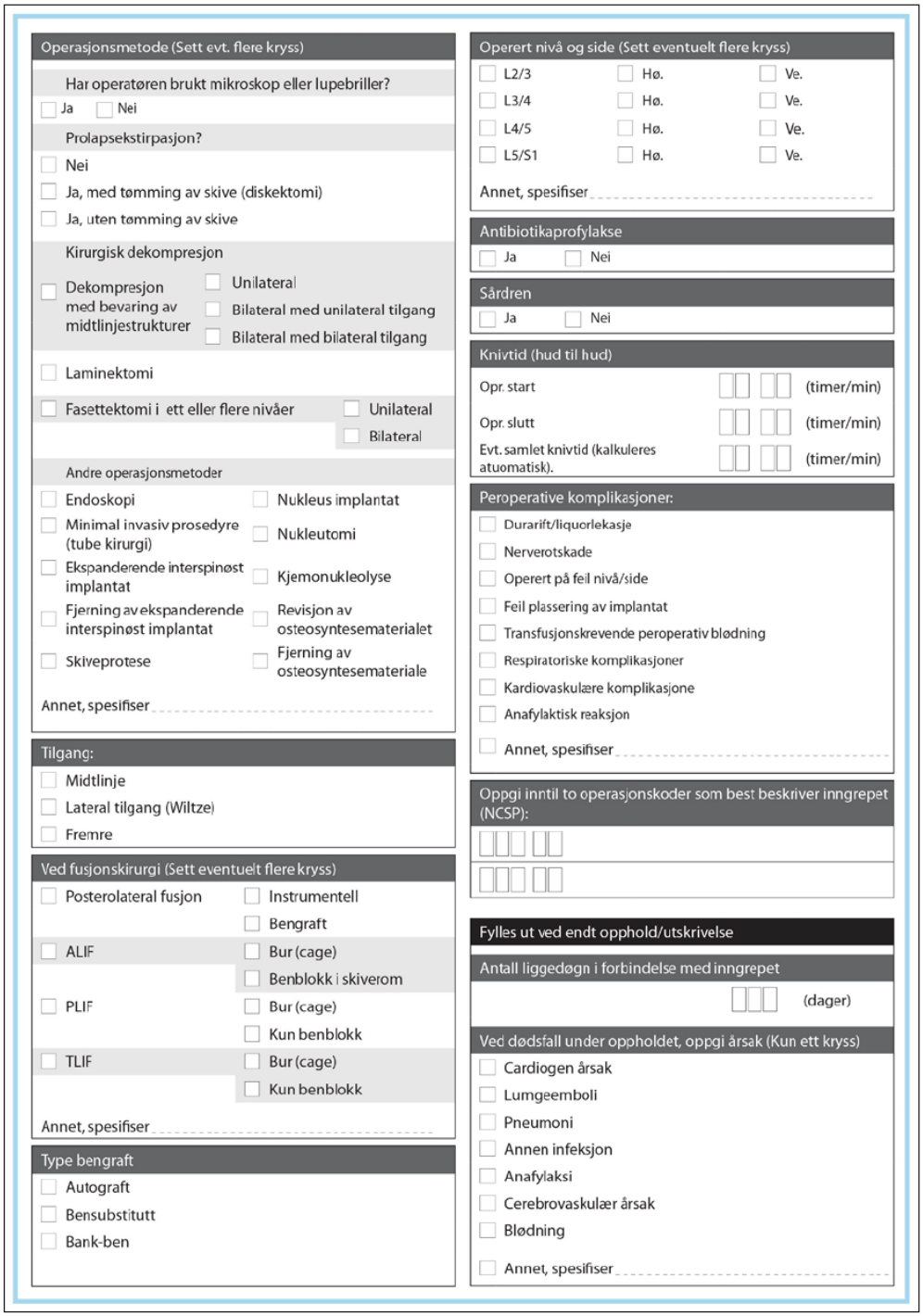
☐ Ja ☐ Nei

ASA-klassifisering

- ☐ I — Ingen organisk, fysiologisk, biokjemisk eller psykisk forstyrrelse. Den aktuelle lidelsen er lokalisert og gir ikke generelle systemforstyrrelser
- ☐ II — Moderat sykdom eller forstyrrelse som ikke forårsaker funksjonelle begrensninger
- ☐ III — Alvorlig sykdom eller forstyrrelse som gir definerte funksjonelle begrensninger
- ☐ IV — Livstruende organisk sykdom som ikke behøver å være knyttet til den aktuelle kirurgiske lidelse eller som ikke bedres ved det planlagte kirurgiske inngrepet
- ☐ V — Døende pasient som ikke forventes å overleve 24 timer uten kirurgi

LUNDBLAD MEDIA AS, TROMSØ – 0-102118

SNU



Operasjonsmetode (Sett evt. flere kryss)

Har operatøren brukt mikroskop eller lupebriller?
- ☐ Ja ☐ Nei

Prolapsekstirpasjon?
- ☐ Nei
- ☐ Ja, med tømming av skive (diskektomi)
- ☐ Ja, uten tømming av skive

Kirurgisk dekompresjon
- ☐ Dekompresjon med bevaring av midtlinjestrukturer
  - ☐ Unilateral
  - ☐ Bilateral med unilateral tilgang
  - ☐ Bilateral med bilateral tilgang
- ☐ Laminektomi
- ☐ Fasettektomi i ett eller flere nivåer
  - ☐ Unilateral
  - ☐ Bilateral

Andre operasjonsmetoder
- ☐ Endoskopi
- ☐ Minimal invasiv prosedyre (tube kirurgi)
- ☐ Ekspanderende interspinøst implantat
- ☐ Fjerning av ekspanderende interspinøst implantat
- ☐ Skiveprotese
- ☐ Nukleus implantat
- ☐ Nukleutomi
- ☐ Kjemonukleolyse
- ☐ Revisjon av osteosyntesematerialet
- ☐ Fjerning av osteosyntesemateriale

Annet, spesifiser .....

Tilgang:
- ☐ Midtlinje
- ☐ Lateral tilgang (Wiltze)
- ☐ Fremre

Ved fusjonskirurgi (Sett eventuelt flere kryss)
- ☐ Posterolateral fusjon
  - ☐ Instrumentell
  - ☐ Bengraft
- ☐ ALIF
  - ☐ Bur (cage)
  - ☐ Benblokk i skiverom
- ☐ PLIF
  - ☐ Bur (cage)
  - ☐ Kun benblokk
- ☐ TLIF
  - ☐ Bur (cage)
  - ☐ Kun benblokk

Annet, spesifiser .....

Type bengraft
- ☐ Autograft
- ☐ Bensubstitutt
- ☐ Bank-ben

Operert nivå og side (Sett eventuelt flere kryss)

☐ L2/3	☐ Hø.	☐ Ve.
☐ L3/4	☐ Hø.	☐ Ve.
☐ L4/5	☐ Hø.	☐ Ve.
☐ L5/S1	☐ Hø.	☐ Ve.

Annet, spesifiser .....

Antibiotikaprofylakse
- ☐ Ja ☐ Nei

Sårdren
- ☐ Ja ☐ Nei

Knivtid (hud til hud)

Opr. start □□ □□ (timer/min)

Opr. slutt □□ □□ (timer/min)

Evt. samlet knivtid (kalkuleres atuomatisk). □□ □□ (timer/min)

Peroperative komplikasjoner:
- ☐ Durarift/liquorlekasje
- ☐ Nerverotskade
- ☐ Operert på feil nivå/side
- ☐ Feil plassering av implantat
- ☐ Transfusjonskrevende peroperativ blødning
- ☐ Respiratoriske komplikasjoner
- ☐ Kardiovaskulære komplikasjone
- ☐ Anafylaktisk reaksjon
- ☐ Annet, spesifiser .....

Oppgi inntil to operasjonskoder som best beskriver inngrepet (NCSP):

□□□ □□

□□□ □□

Fylles ut ved endt opphold/utskrivelse

Antall liggedøgn i forbindelse med inngrepet

□□□ (dager)

Ved dødsfall under oppholdet, oppgi årsak (Kun ett kryss)
- ☐ Cardiogen årsak
- ☐ Lungeemboli
- ☐ Pneumoni
- ☐ Annen infeksjon
- ☐ Anafylaksi
- ☐ Cerebrovaskulær årsak
- ☐ Blødning
- ☐ Annet, spesifiser .....

16.2 Supplementary appendix to paper I

Table 1x. Cut-offs based on ROC analyses for all PROMs when splitting the patient population based on the baseline ODI

	ODI baseline < 25 percentile		ODI baseline >75 percentile	
	Failure	**Worsening**	**Failure**	**Worsening**
	Cut-off Sens, spec (AUC)	Cut-off Sens, spec (AUC)	Cut-off Sens, spec (AUC)	Cut-off Sens, spec (AUC)
ODI				
Mean change	**4** 0.84, 0.82 (0.91)		**30** 0.86, 0.90 (0.94)	
Mean % change	**21** 0.84, 0.83 (0.91)		**51** 0.91, 0.85 (0.95)	
12 months raw	**18** 0.88, 0.88 (0.92)	**33** 0.70, 0.66 (0.74)	**34** 0.91, 0.83 (0.95)	**58** 0.71, 0.65 (0.72)
NRS back-pain				
Mean change	**0.5** 0.79, 0.75 (0.84)		**2.5** 0.86, (0.87)	
Mean % change	**18** **0.**85, 0.77 (0.88)		**32** 0.90, 0.82 (0.90)	
12-month raw	**3.5** 0.78, 0.85 (0.89)	**6.5** 0.69, 0.64 (0.71)	**4.5** 0.86, 0.81 (0.90)	**7.5** 0.74, 0.56 (0.69)
NRS leg-pain				
Mean change	**1.5** 0.78, 0.76 (0.84)		**3.5** 0.84, 0.82 (0.89)	
Mean % change	**39** 0.84, 0.79 (0.86)		**32** 0.81, 0.88 (0.90)	
12-month raw	**3.5** 0.78, 0.85 (0.89)	**6.5** 0.69, 0.64 (0.71)	**4.5** 0.86, 0.81 (0.90)	**7.5** 0.74, 0.56 (0.69)
EQ-5D				
Mean change	**0.04** 0.76, 0.82 (0.84)		**0.60** 0.85, 0.84 (0.90)	
12-month raw	**0.73** 0.84, 0.81 (0.90)	**0.17** 0.70, 0.65 (0.68)	**0.60** 0.89, 0.81 (0.93)	**0.07** 0.73, 0.60 (0.70)

Table 2x. AUC and Cut-off values for all PROMs when identifying "failure".

	All		Operated with microsurgical technique		Operated with open discectomy		Operated for the first time		Previously Operated (on same level)	
	AUC 95% CI	Cut-off (sens., spec)	AUC 95% CI	Cut-off (sens., spec)	AUC 95% CI	Cut-off (sens., spec)	AUC 95% CI	Cut-off (sens., spec)	AUC 95% CI	Cut-off (sens., spec)
ODI										
Mean change	0.89 (0.88 - 0.91)	13 (0.82, 0.82)	0.89 (0.88 - 0.90)	13 (0.81, 0.82)	0.91 (0.88 - 0.94)	13 (0.89, 0.77)	0.90 (0.87 - 0.91)	13 (0.84, 0.82)	0.88 (0.85 - 0.91)	13 (0.80, 0.81)
Mean % change	0.93 (0.92 - 0.94)	33 (0.86, 0.86)	0.93 (0.92 - 0.94)	33 (0.86, 0.85)	0.94 (0.92 - 0.95)	33 (0.90, 0.82)	0.93 (0.92 - 0.94)	33 (0.86, 0.86)	0.91 (0.90 - 0.93)	33 (0.87, 0.82)
12-month raw	0.92 (0.91 - 0.93)	25 (0.89, 0.81)	0.92 (0.91 - 0.93)	25 (0.88, 0.82)	0.92 (0.90 - 0.94)	25 (0.90, 0.79)	0.92 (0.91 - 0.93)	25 (0.86, 0.83)	0.92 (0.91 - 0.94)	25 (0.94, 0.73)
NRS back-pain										
Mean change	0.85 (0.84 - 0.86)	1.5 (0.74, 0.86)	0.85 (0.83 - 0.86)	10.5 (0.80, 0.76)	0.85 (0.81 - 0.88)	1.5 (0.86, 0.72)	0.85 (0.83 - 0.87)	1.5 (0.81, 0.76)	0.84 (0.81 - 0.87)	1.5 (0.83, 0.74)
12-month raw	0.92 (0.91 - 0.93)	5.5 (0.81, 0.87)	0.92 (0.90 - 0.93)	5.5 (0.81, 0.87)	0.91 (0.86 - 0.94)	5.5 (0.82, 0.86)	0.91 (0.90 - 0.93)	5.5 (0.79, 0.88)	0.91 (0.89 - 0.94)	5.5 (0.87, 0.81)
Mean % change	0.87 (0.86 - 0.88)	24 (0.85 - 0.81)	0.87 (0.86 - 0.88)	24 (0.84, 0.81)	0.87 (0.85 - 0.90)	24 (0.90, 0.76)	0.87 (0.86 - 0.89)	24 (0.83, 0.82)	0.86 (0.83 - 0.88)	24 (0.88, 0.77)
NRS leg-pain										
Mean change	0.87 (0.86 - 0.88)	1.5 (0.81, 0.76)	0.87 (0.85 - 0.88)	1.5 (0.72, 0.86)	0.87 (0.84 - 0.91)	1.5 (0.79, 0.83)	0.88 (0.86 - 0.89)	1.5 (0.74, 0.86)	0.85 (0.82 - 0.88)	1.5 (0.69, 0.84)
12-month raw	0.90 (0.88 - 0.91)	4.5 (0.91, 0.85)	0.89 (0.88 - 0.91)	4.5 (0.80, 0.85)	0.91 (0.88 - 0.93)	4.5 (0.82, 0.82)	0.90 (0.88 - 0.91)	4.5 (0.81, 0.86)	0.88 (0.85 - 0.90)	4.5 (0.80, 0.80)
Mean % change	0.89 (0.88 - 0.90)	39 (0.86, 0.81)	0.89 (0.87 - 0.90)	39 (0.85, 0.82)	0.90 (0.87 - 0.92)	39 (0.88, 0.79)	0.90 (0.88 - 0.91)	39 (0.87, 82)	0.86 (0.83 - 0.89)	39 (0.84, 0.76)
EQ-5D										
Mean change	0.85 (0.84 - 0.87)	0.10 (0.76, 0.83)	0.85 (0.84 - 0.87)	0.10 (0.76, 0.84)	0.83 (0.79 - 0.87)	0.10 (0.76, 0.79)	0.85 (0.83 - 0.87)	0.10 (0.73, 0.85)	0.83 (0.80 - 0.87)	0.10 (0.76, 0.78)
12-month raw	0.91 (0.90 - 0.92)	0.63 (0.81 - 0.85)	0.91 (0.90 - 0.92)	0.63 (0.81, 0.85)	0.90 (0.88 - 0.93)	0.64 (0.81, 0.81)	0.91 (0.90 - 0.92)	0.64 (0.78, 0.86)	0.90 (0.87 - 0.92)	0.64 (0.86, 0.77)

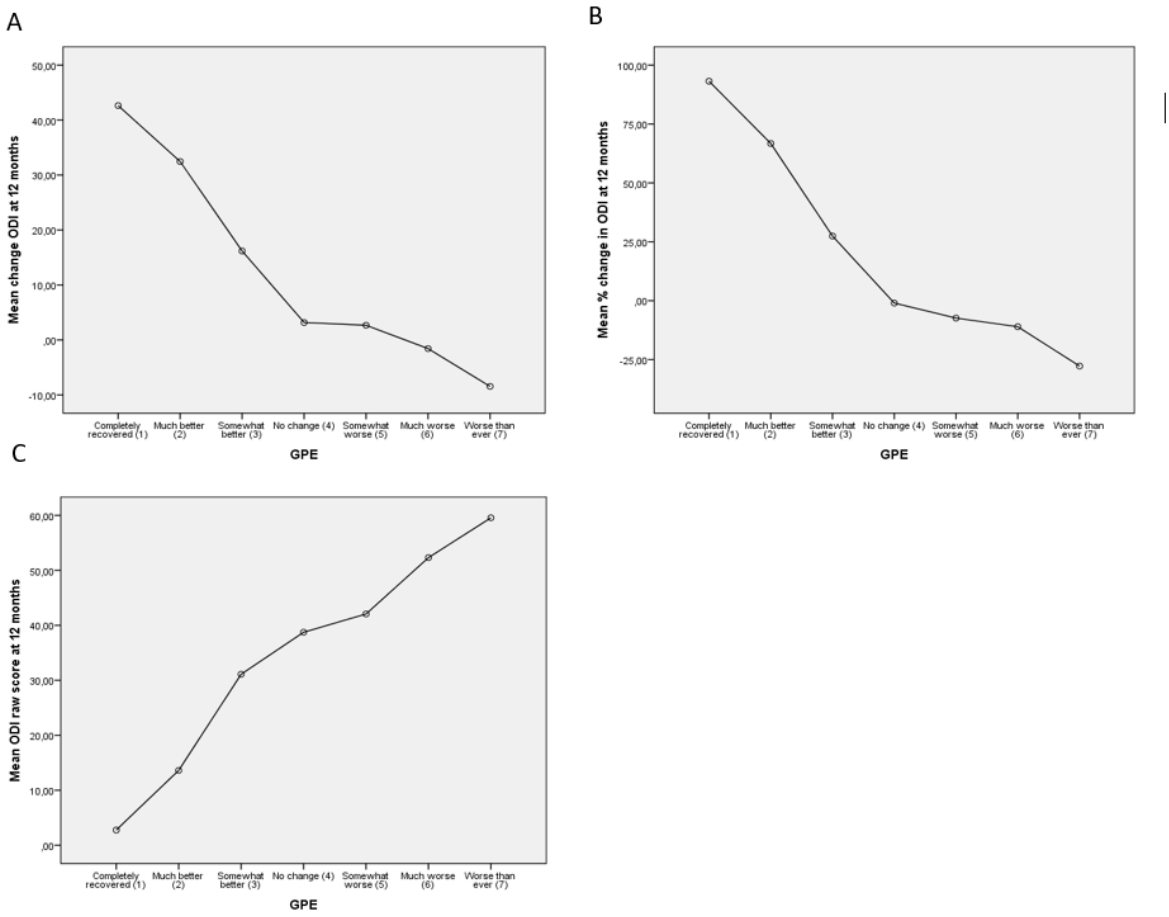
Table 3x. AUC and Cut-off values for all final raw scores when identifying "worsening".

	All		Operated with microsurgical technique		Operated with open discectomy		Operated for the first time		Previously Operated (on same level)	
	AUC (95% CI)	Cut-off (sens. spec)	AUC	Cut-off (sens. spec)	AUC	Cut-off (sens. spec)	AUC	Cut-off (sens. spec)	AUC	Cut-off (sens. spec)
ODI 12-month Raw	0.76 (0.72 - 0.80)	48 (0.70, 0.70)	0.76 (0.72 - 0.80)	48 (0.70, 0.70)	0.73 (0.63 - 0.83)	48 (0.67, 0.70)	0.75 (0.70 - 0.80)	48 (0.67, 0.73)	0.74 (0.66 - 0.81)	48 (0.68, 0.64)
NRS leg-pain 12-month raw	0.70 (0.66 - 0.75)	7.5 (0.64, 0.68)	0.69 (0.65, 0.74)	7.5 (0.61, 0.67)	0.76 (0.66 - 0.86)	7.5 (0.79, 0.70)	0.69 (0.63 - 0.74)	7.5 (0.60, 0.70)	0.71 (0.63 - 0.80)	7.5 (0.68, 0.66)
NRS back-pain 12-month raw	0.77 (0.73 - 0.81)	7.5 (0.78, 0.64)	0.76 (0.72, 0.80)	7.5 (0.77, 0.64)	0.81 (0.72 - 0.90)	7.5 (0.88, 0.66)	0.77 (0.72 - 0.82)	7.5 (0.76, 0.67)	0.74 (0.67 - 0.82)	7.5 (0.77, 0.58)
EQ-5D 12-month raw	0.71 (0.67 - 0.75)	0.09 (0.76, 0.60)	0.59 (0.54, 0.64)	0.09 (0.83, 0.68)	0.68 (0.56 - 0.79)	0.09 (0.66, 0.65)	0.69 (0.64 - 0.74)	0.09 (0.69, 0.63)	0.70 (0.62 - 0.78)	0.07 (0.70, 0.58)

Table 4x. Proportion of the study population classified into the categories "failure" or "worsening" based on all PROM cut-offs.

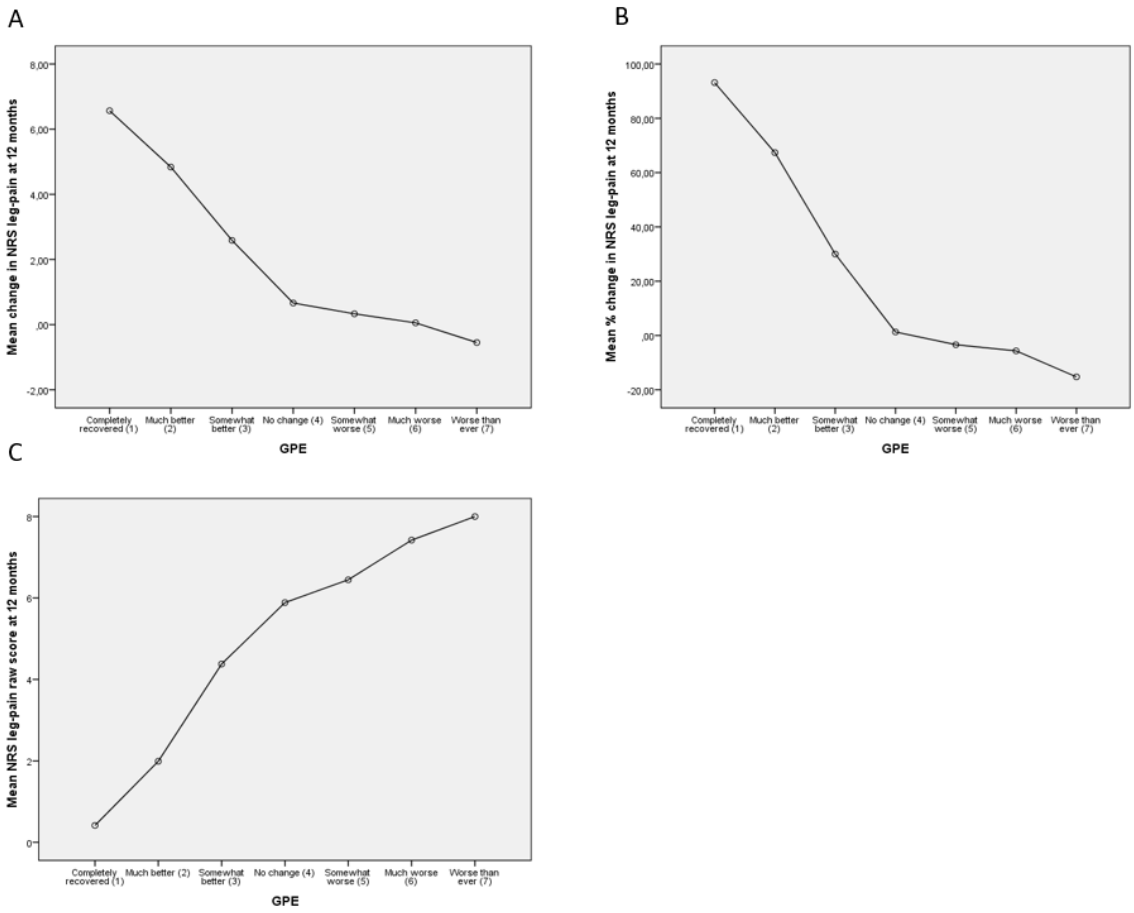
Instrument	% classified into cut-off category
	Failure
GPE	12
ODI	
Mean change	26
Mean % change	23
12-month raw	27
NRS back-pain	
Mean change	30
Mean % change	18
12-month raw	21
NRS leg-pain	
Mean change	20
Mean % change	20
12-month raw	23
EQ-5D	
Mean change	24
12-month raw	22
	Worsening
GPE	3
ODI 12-month raw	7
NRS back-pain 12-month raw	8
NRS leg-pain 12-month raw	7
EQ-5D 12 month raw	8

Figure 1x.



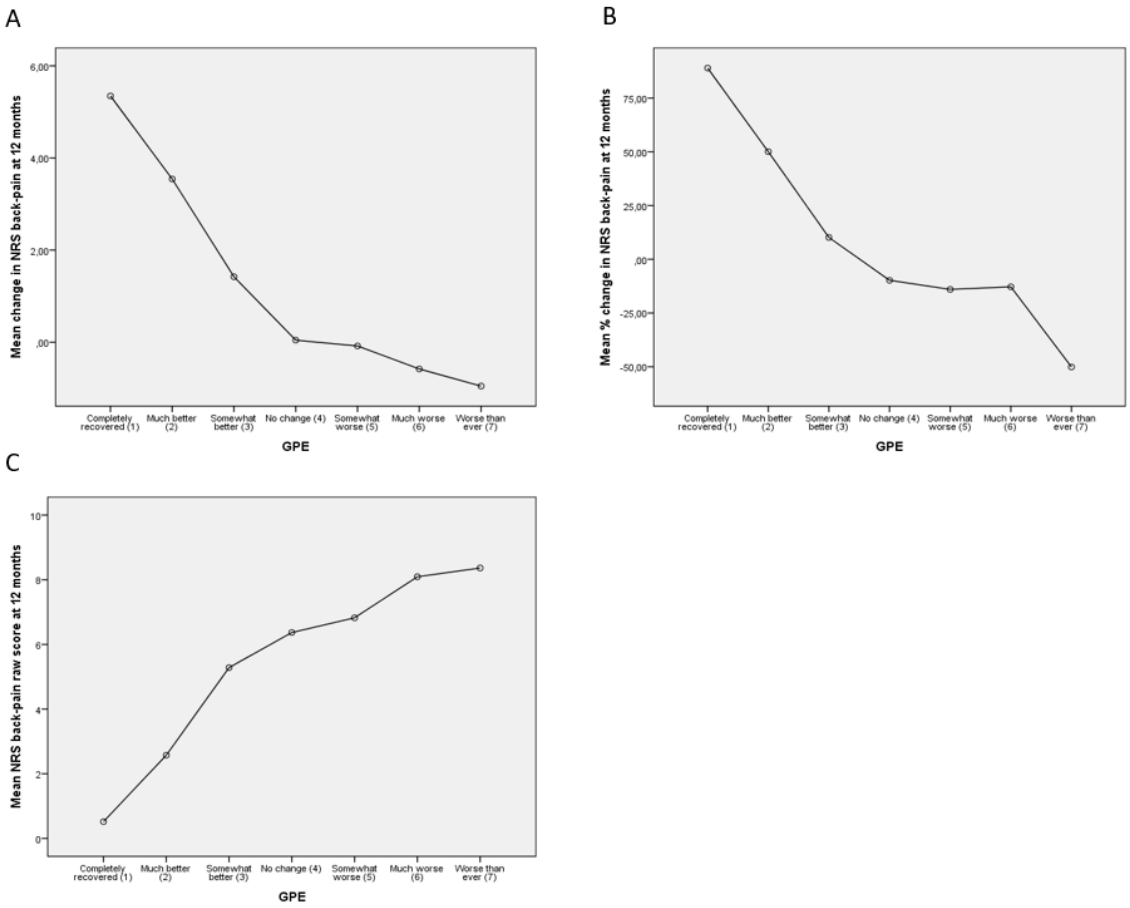
ANOVA of mean change in ODI (A), mean % change in ODI (B) and the final ODI raw score after 12 months (C)

Figure 2x.



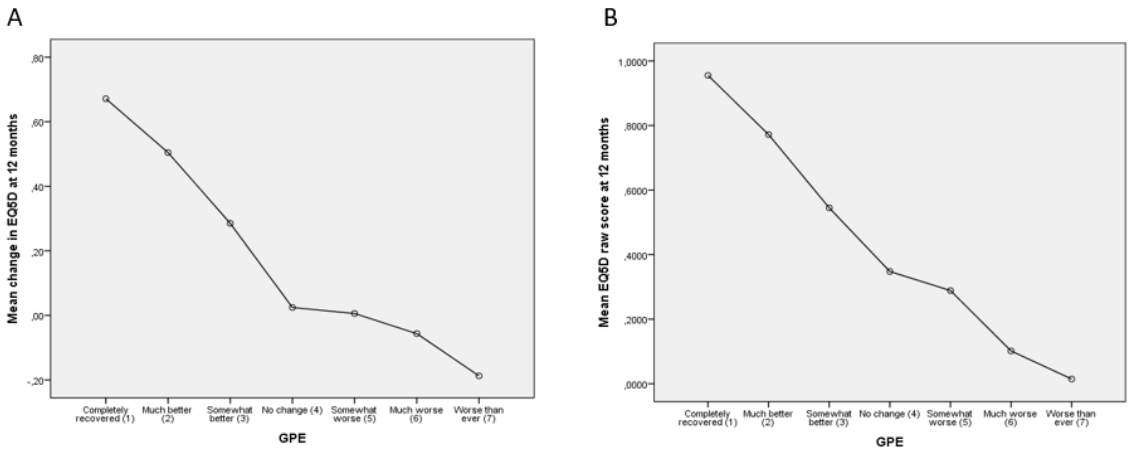
ANOVA of mean change in NRS leg-pain (A), mean % change in NRS leg-pain (B) and the final NRS leg-pain raw score after 12 months (C)

Figure 3x.



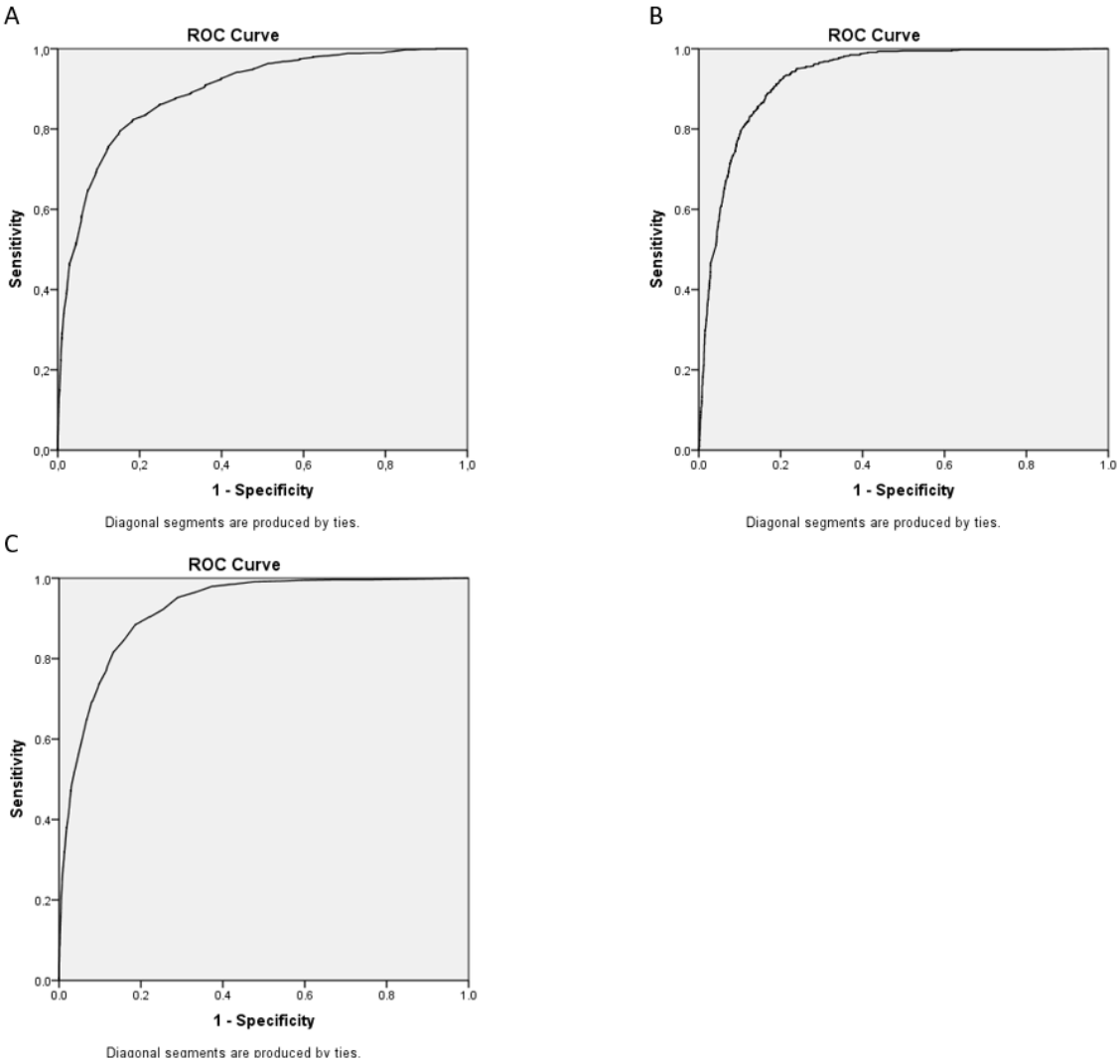
ANOVA of mean change in NRS back-pain (A), mean % change in NRS back-pain (B) and the final NRS back-pain raw score after 12 months (C)

Figure 4x.



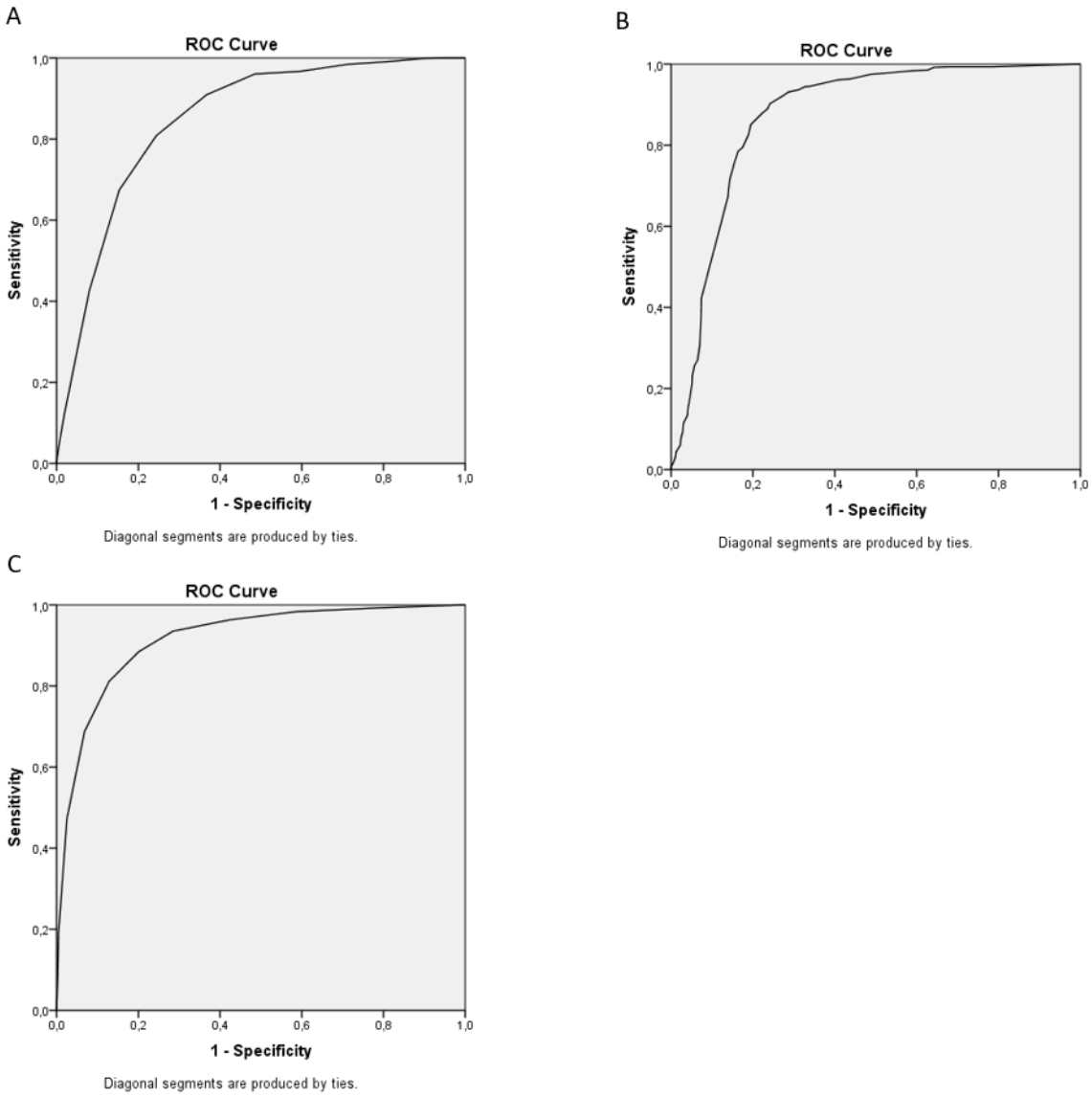
ANOVA of mean change in EQ-5D (A) and the final EQ-5D raw score after 12 months (B)

Figure 5x.



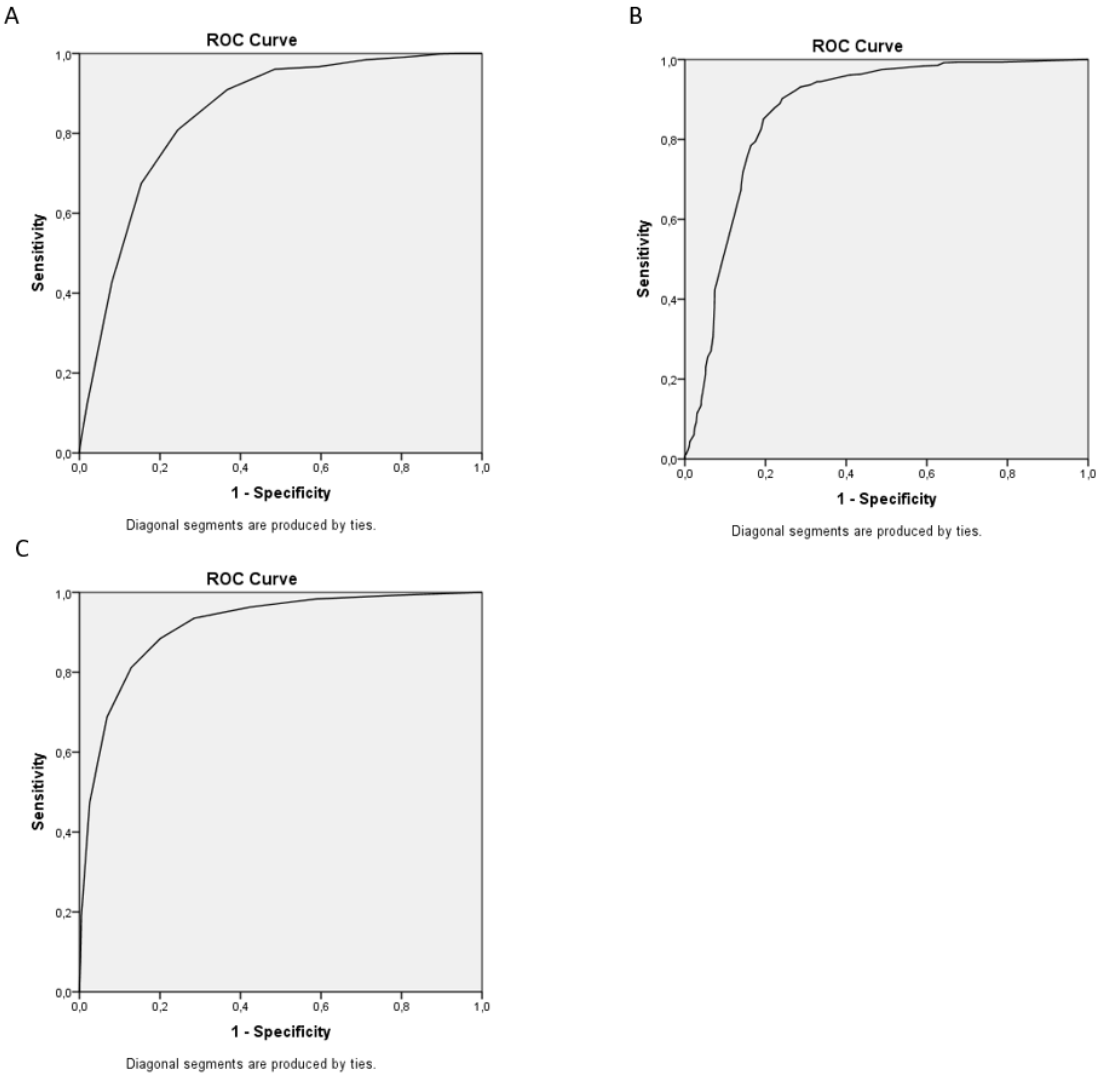
ROC analyses of ODI for mean change (A), mean % change (B) and final raw score after 12 months (C) when identifying failure

Figure 6x.



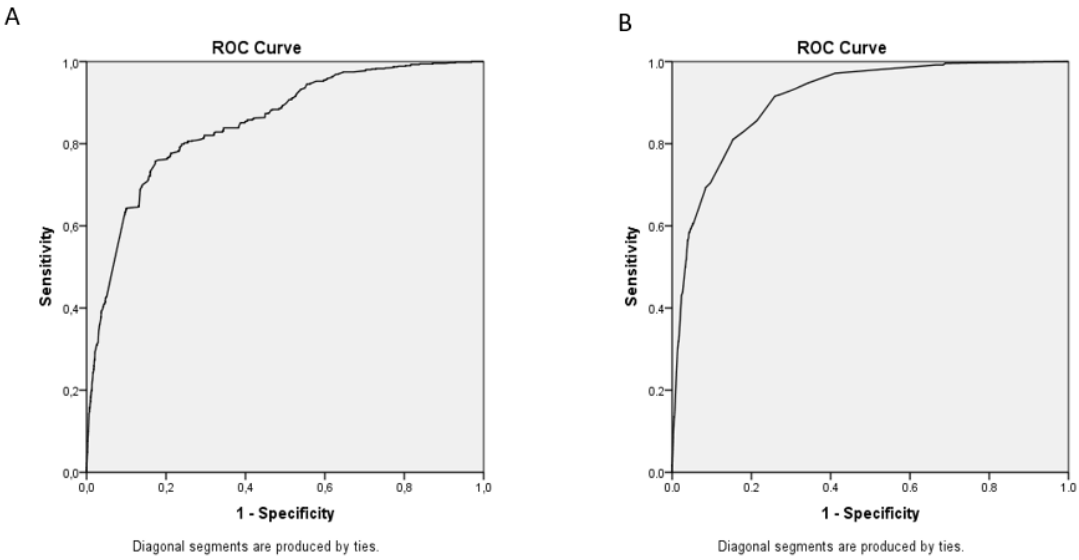
ROC analyses of NRS back-pain for mean change (A), mean % change (B) and final raw score after 12 months (C) when identifying failure.

Figure 7x.



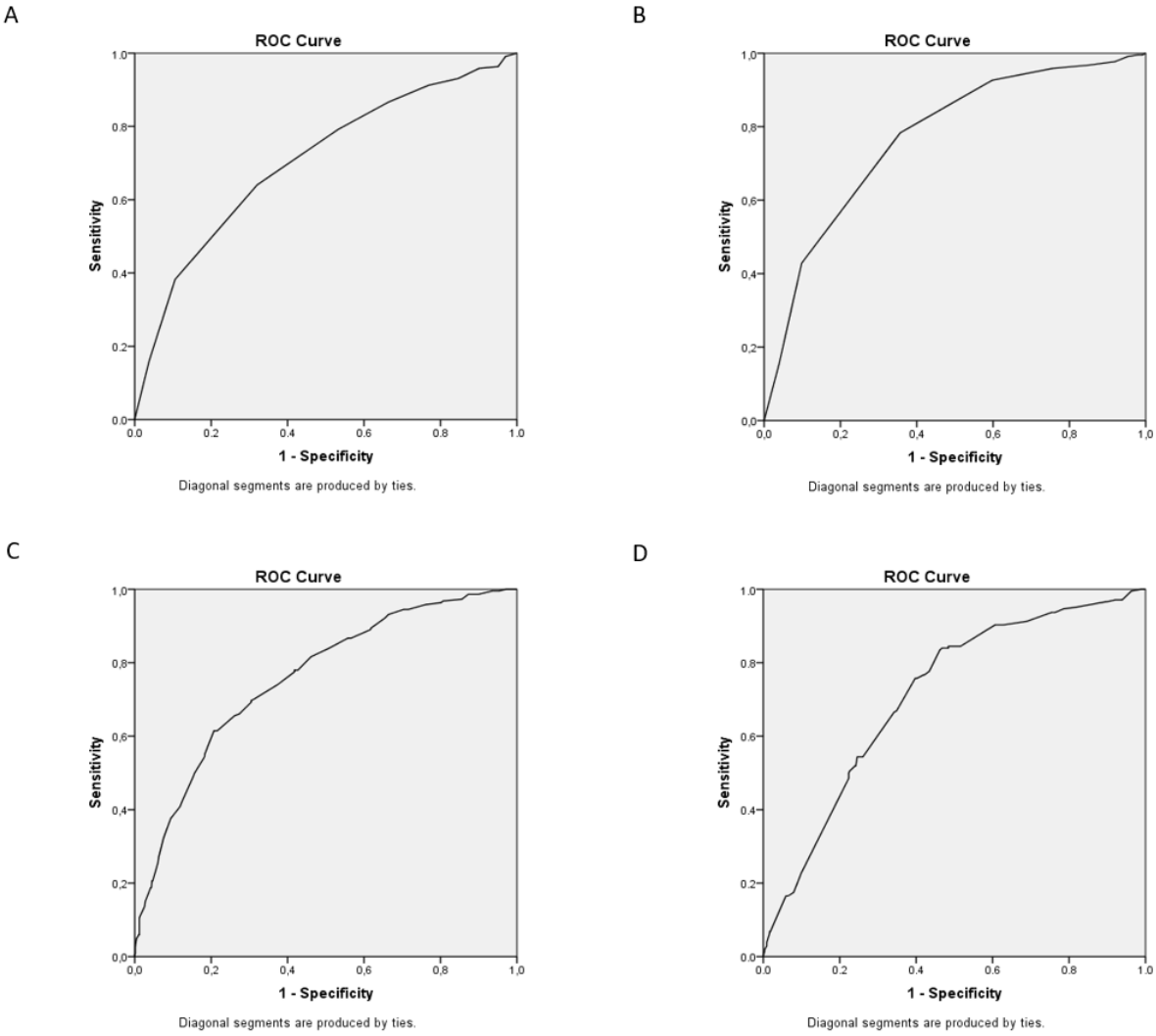
ROC analyses of NRS leg-pain for mean change (A), mean % change (B) and final raw score after 12 months (C) when identifying failure.

Figure 8x.



ROC analyses of EQ-5D mean change (A) and final raw score after 12 months (B) when identifying failure.

Figure 9x.



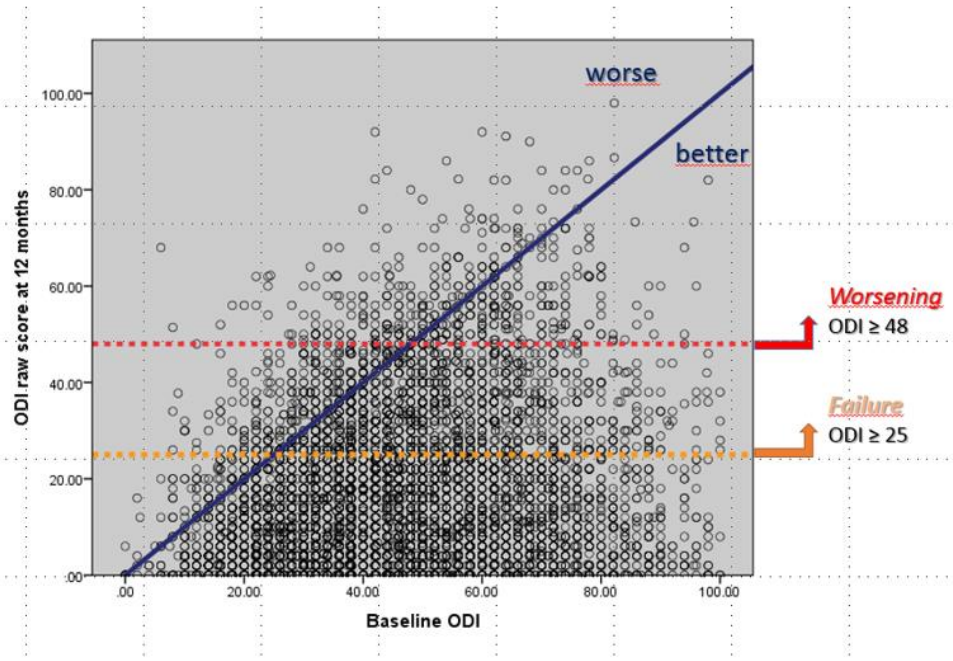
ROC analyses of 12-month final raw scores for NRS back-pain (A), NRS leg-pain (B), ODI (C), and EQ-5D (D) when identifying worsening

Table 5x. Mean differences in PROMs after 12 months for elective vs emergency operations

	Elective	Emergency	p-value
ODI			
Baseline score (SD)	43 (17)	59 (22)	<0.001
12-month raw score (SD)	18 (17)	16 (16)	0.06
NRS back-pain			
Baseline score (SD)	6.0 (2.4)	7.1 (2.7)	<0.001
12-month raw score (SD)	3.1 (2.7)	2.8 (2.5)	0.028
NRS leg-pain			
Baseline score (SD)	6.7 (2.2)	7.8 (2.2)	<0.001
12-month raw score (SD)	2.6 (2.8)	2.4 (2.6)	0.293
EQ-5D			
Baseline score (SD)	0.32 (0.34)	0.08 (0.37)	<0.001
12-month raw score (SD)	0.71 (0.29)	0.73 (0.26)	0.26
GPE			
Median	2	2	0.019

SD = Standard Deviation

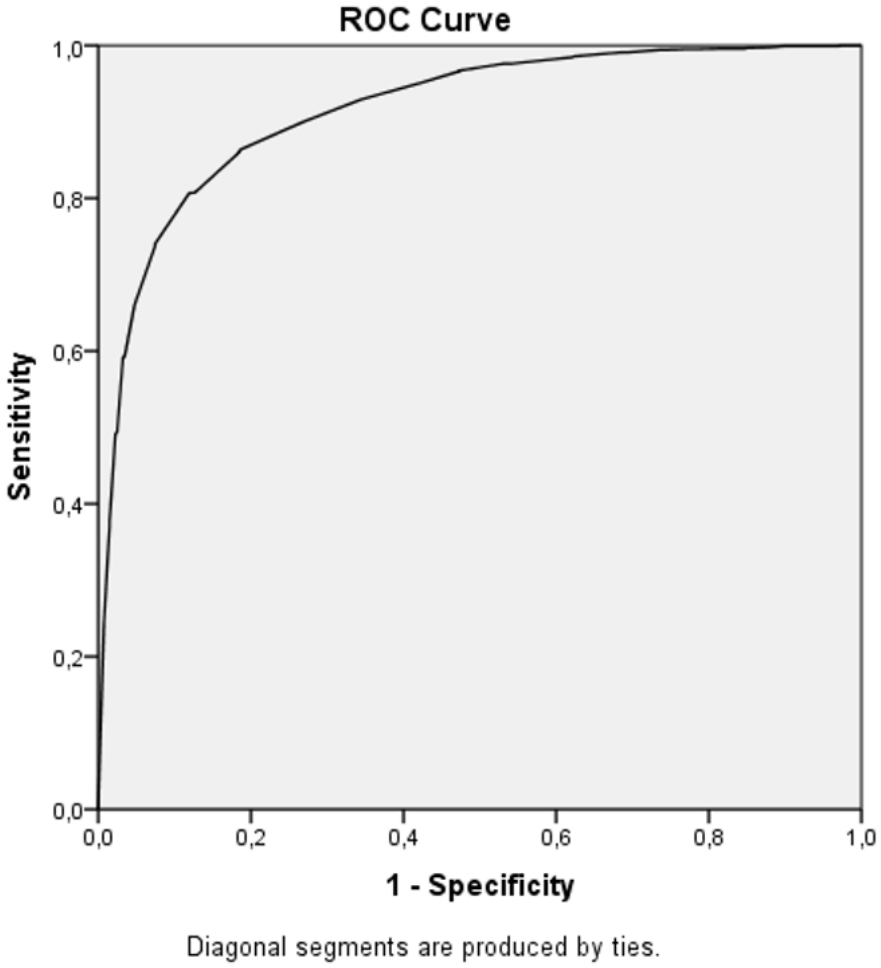
Figure 10x. Baseline ODI (preoperative) versus ODI 12 months after surgery



The diagonal line represents no change. Coordinates to the right of the diagonal line represent improvement from baseline, and coordinates to the left deterioration from baseline. The red line indicates the 12-month ODI raw cut-off above which patients consider themselves as worse, irrespective of the change experienced. The orange line represents the ODI raw cut-off for failure.

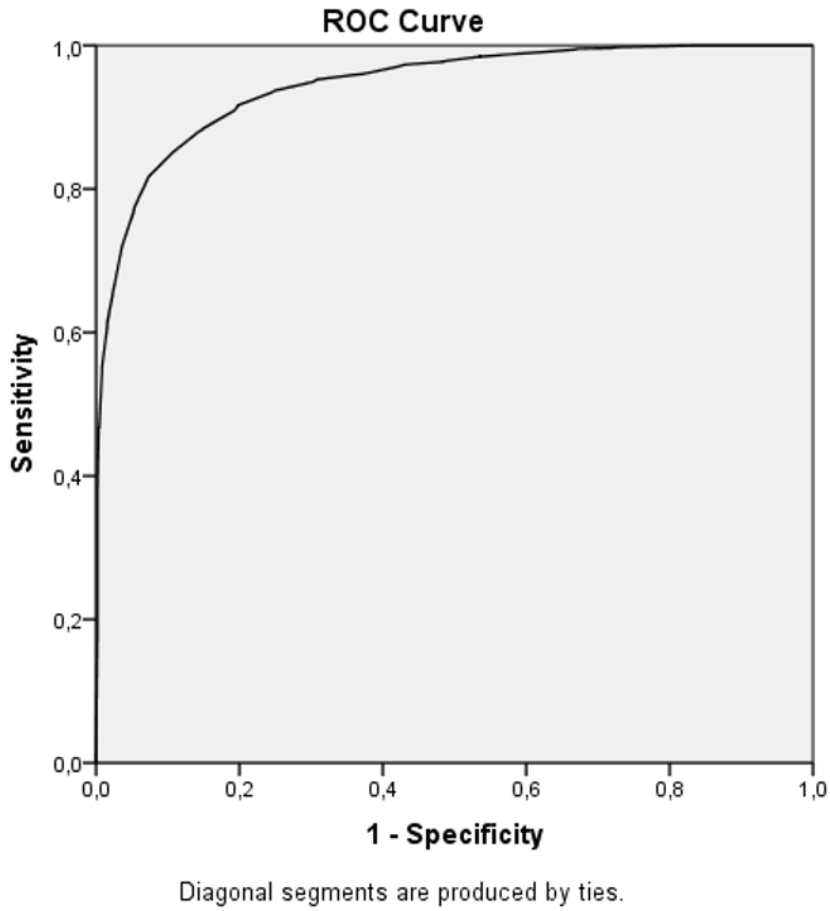
16.3 Supplementary appendix to paper II

Figure 1x.



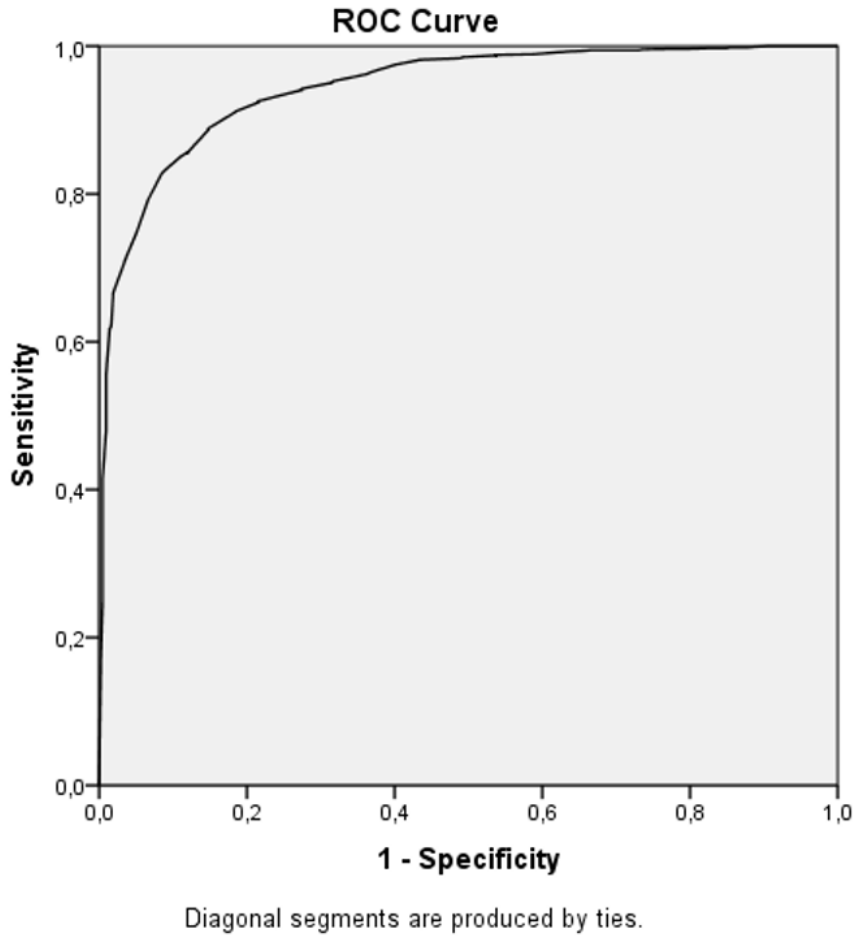
Receiver Operating Curve for the 12-months ODI raw cutoff for "success" in the subgroup "ODI prescore <25th percentile". AUC=0.92 (0.90-0.93).

Figure 2x.



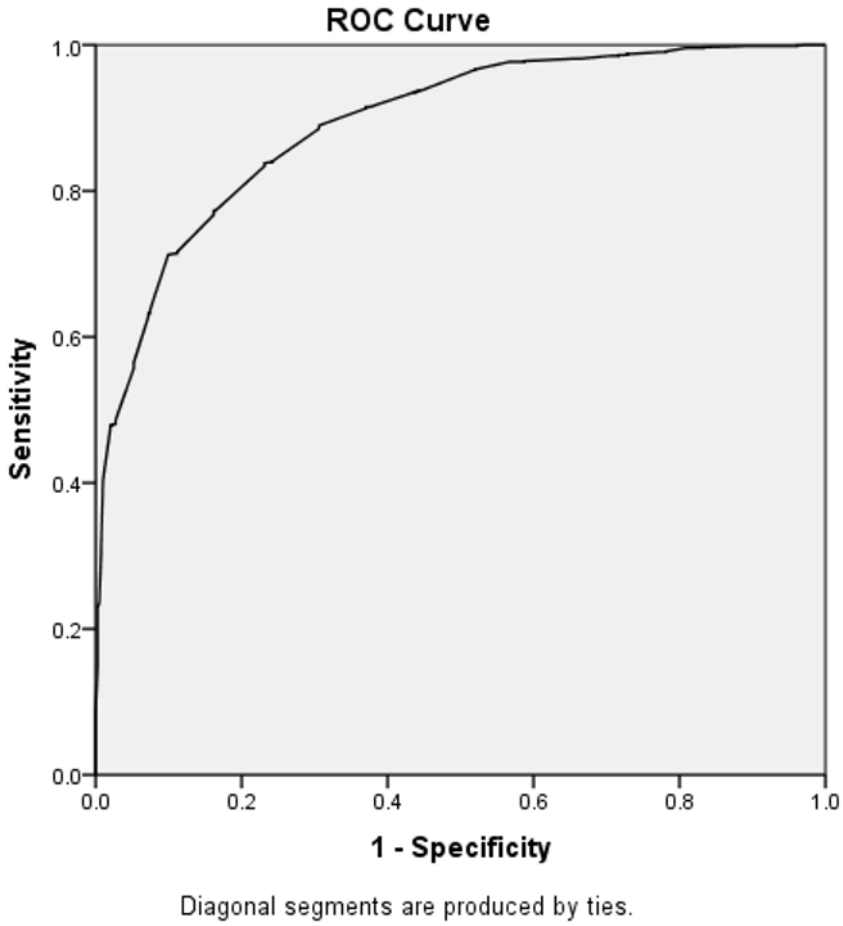
Receiver Operating Curve for the 12-months ODI raw cutoff for "success" in the subgroup "ODI prescore 25th – 75th percentile". AUC=0.95 (0.94-0.95).

Figure 3x.



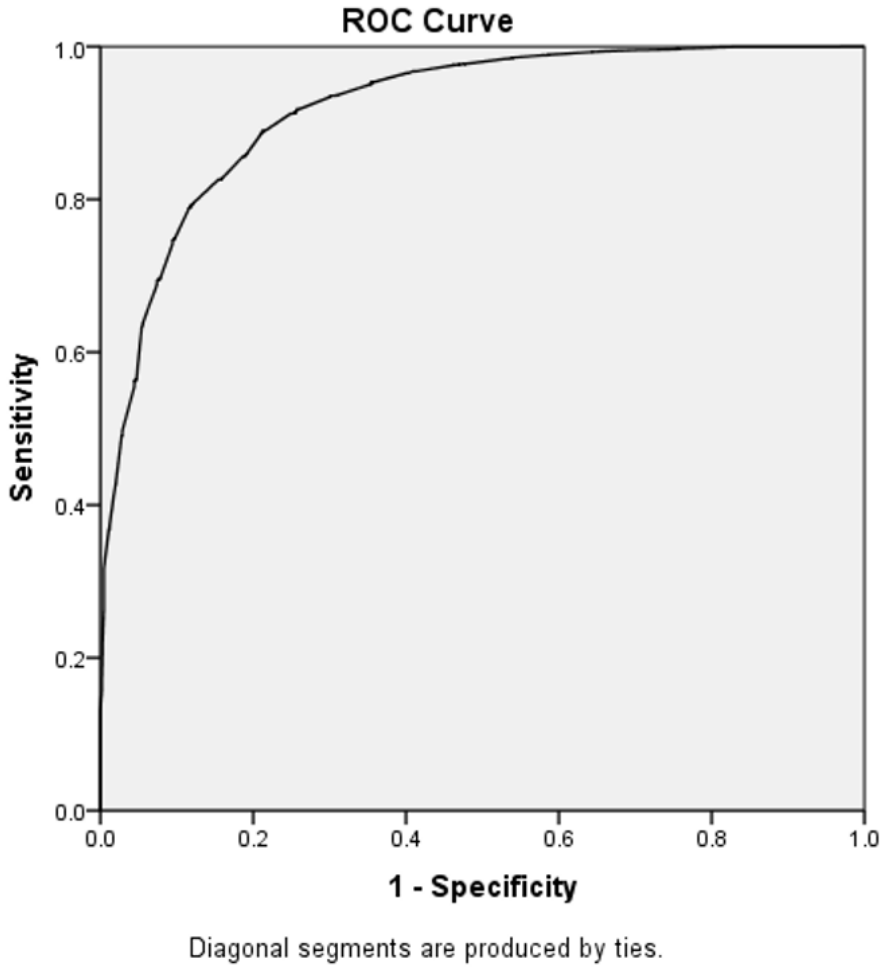
Receiver Operating Curve for the 12-months ODI raw cutoff for "success" in the subgroup "ODI prescore >75th percentile". AUC=0.94 (0.93-0.96).

Figure 4x.



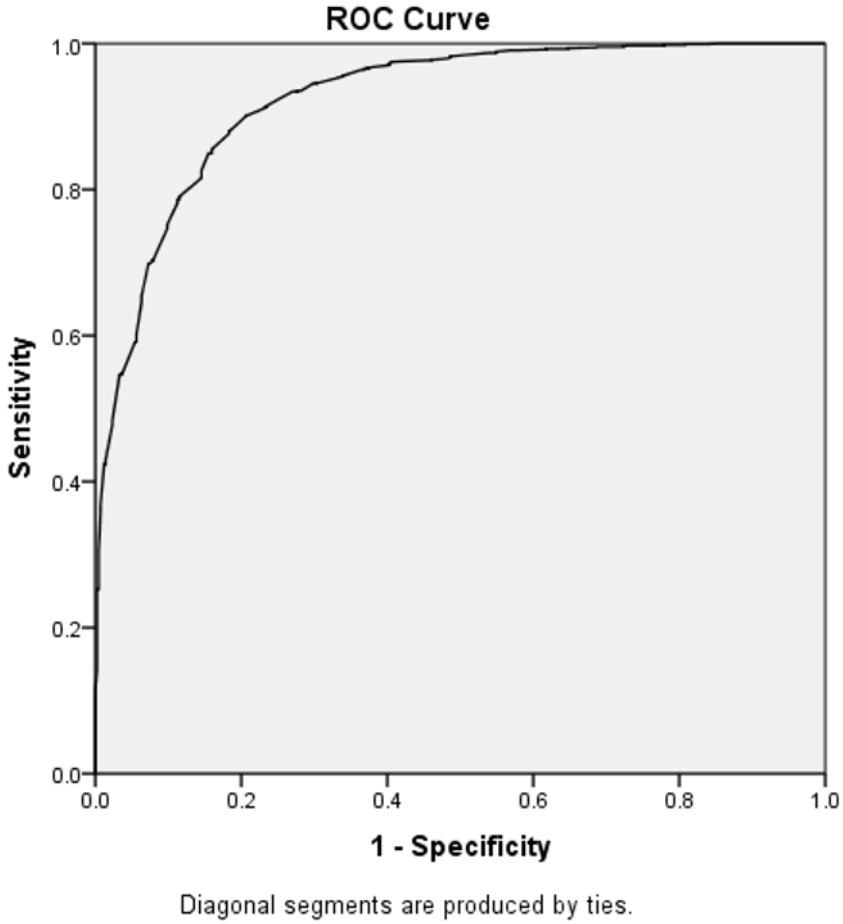
Receiver Operating Curve for the 12-months ODI change cutoff for "success" in the subgroup "ODI prescore <25th percentile". AUC=0.90 (0.88-0.91)

Figure 5x.



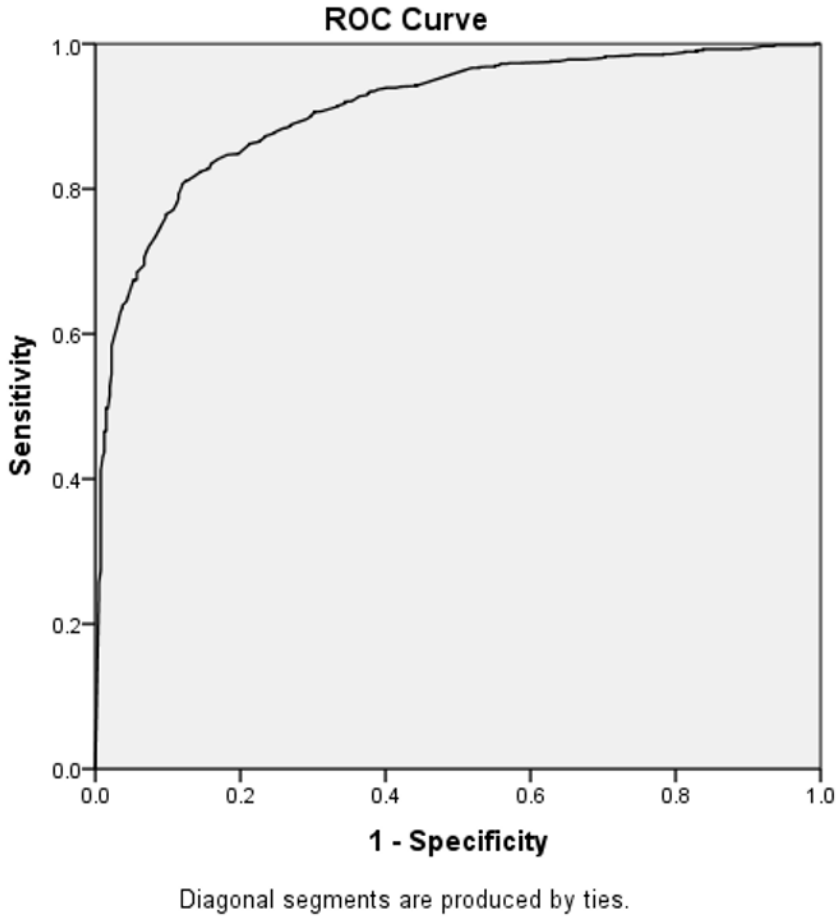
Receiver Operating Curve for the 12-months ODI change cutoff for "success" in the subgroup "ODI prescore 25th – 75th percentile". AUC=0.92 (0.91-0.93)

Figure 6x.



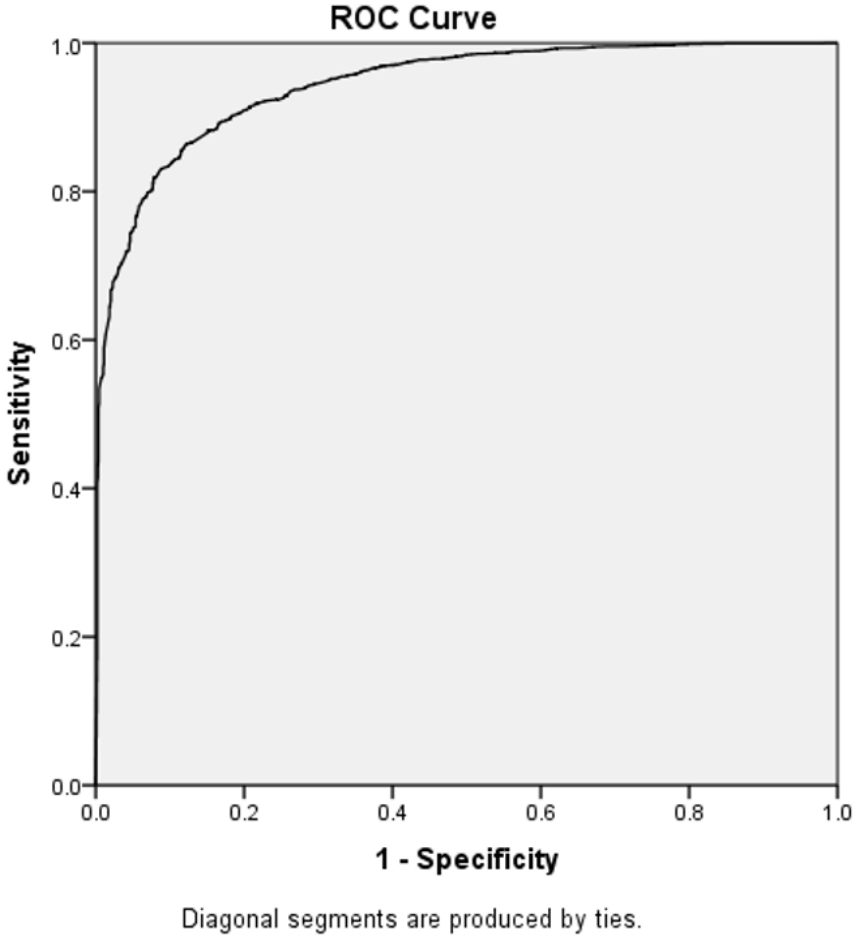
Receiver Operating Curve for the 12-months ODI change cutoff for "success" in the subgroup "ODI prescore >75th percentile". AUC=0.92 (0.91-0.94)

Figure 7x.



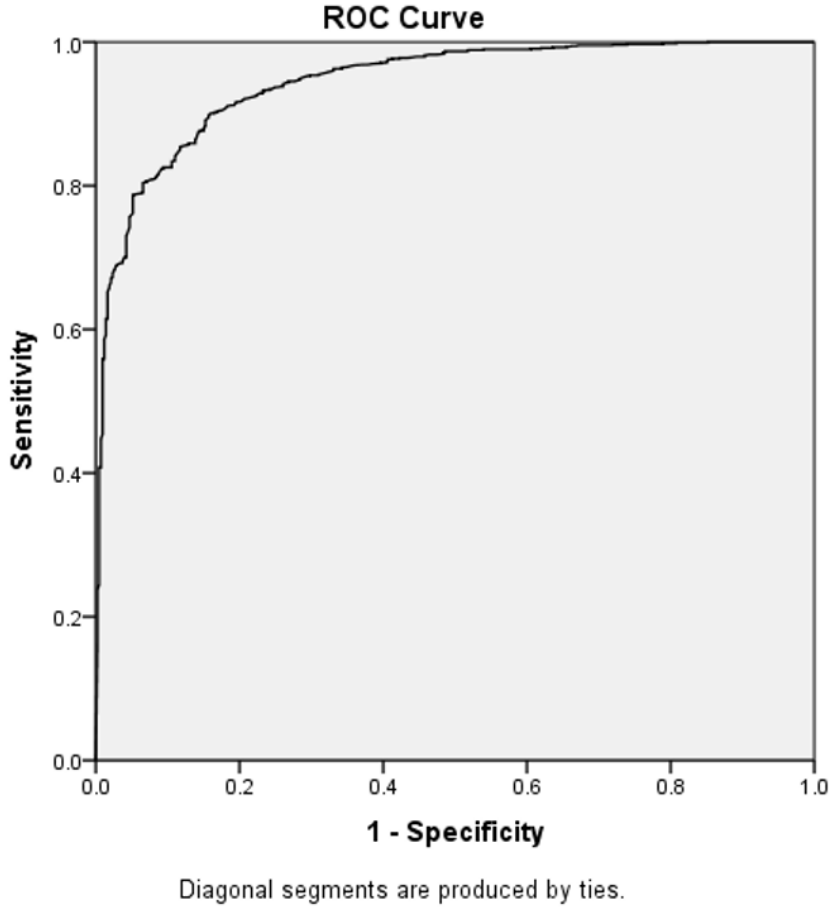
Receiver Operating Curve for the 12-months ODI percentage change cutoff for "success" in the subgroup "ODI prescore <25th percentile". AUC=0.91 (0.90-0.93)

Figure 8x.



Receiver Operating Curve for the 12-months ODI percentage change cutoff for "success" in the subgroup "ODI prescore 25th – 75th percentile". AUC=0.94 (0.94-0.95)

Figure 9x.



Receiver Operating Curve for the 12-months ODI percentage change cutoff for "success" in the subgroup "ODI prescore > 75th percentile". AUC=0.94 (0.93-0.96)

Table 1x. Proportion of cases classified as success after 12 months for different groups of baseline disability (based on the preoperative ODI)

ODI Prescore	12-months ODI raw n (%)	ODI percentage change n (%)	ODI change n (%)
<25th percentile	1025 (63.5)	1048 (65.3)	998 (61.9)
25-75th percentile	2355 (63.4)	2340 (63.2)	2328 (62.9)
>75th percentile	1058 (70.4)	986 (64.4)	981 (65.3)
Sum	**4438**	**4374**	**4307**

16.4 Supplementary appendix to paper III

Table 1x. Failure and worsening 12 months after surgery for subgroups of different baseline disability (low, medium and high percentiles of the ODI score) in the training (n= 5741, 70%) and validation (n= 2218, 30%) set.

	Training set			Validation set		
	ODI group[1] <25th	ODI group 25-75th	ODI group > 75th	ODI group <25th	ODI group 25-75th	ODI group > 75th
Cases n (%)	1243 (24)	2772 (54)	1159 (22)	608 (27)	1024 (46)	586 (26)
Failure n (%)	366 (26)	565 (23)	306 (23)	165 (27)	229 (22)	148 (25)
Worsening n (%)	76 (5)	157 (7)	85 (7)	50 (8)	65 (6)	36 (6)

[1]Baseline ODI group based on the baseline percentile of the ODI score – low (<25th percentile, <33 points), medium (25th – 75th percentile, 33-58 points), high (>75th percentile, >58 points). ODI range: 0-100 (no-maximal disability).

Table 2x. Results from the univariate binary logistic regression analyses of failure in both the training (n=5741) and validation (n=2218) cohort, showing associations (Odds Ratio (OR) and 95% confidence intervals (CI)) between predictors and patient reported "failure" (unchanged or worse, yes/no) of lumbar disc surgery, as defined by validated cut offs on the Oswestry Disability Index (ODI), split by subgroups with low, medium and high baseline ODI scores (percentiles). For all predictors, except age and gender, NS indicates statistical insignificance, p value >0.1.

	Training Cohort						Validation Cohort					
Baseline ODI[1] group	**<25th**		**25th-75th**		**>75th**		**<25th**		**25th-75th**		**>75th**	
Predictor	**OR[2]**	**P value**	**OR**	**P value**	**OR**	**P value**	**OR**	**P value**	**OR**	**P value**	**OR**	**P value**
Age >60	1.0 (0.7 – 1.4)	0.986	1.4 (1.2 – 1.8)	0.002	1.7 (1.2 – 2.2)	0.001	1.2 (0.7 – 1.9)	0.526	1.5 (1.1 – 2.1)	0.020	0.9 (0.6 – 1.4)	0.498
Living alone	NS[3]	0.488	NS	0.347	1.5 (1.1 – 2.0)	0.011	0.5 (0.3 – 0.9)	0.009	NS	0.213	1.7 (1.1 – 2.5)	0.014
Nonnative Norwegian speaker	1.9 (1.1- 3.5)	0.030	2.1 (1.5 – 3.0)	<0.001	2.3 (1.5 – 3.6)	<0.001	NS	0.415	1.9 (1.1 – 3.2)	0.027	2.2 (1.2 – 4.1)	0.015
Female	1.2 (0.9 – 1.6)	0.147	1.2 (1.0 – 1.4)	0.050	1.5 (1.2 – 2.0)	0.002	1.2 (0.8 – 1.7)	0.454	1.1 (0.8 – 1.5)	0.467	1.4 (0.9 – 2.0)	0.106
Smoking	2.3 (1.7 – 3.0)	<0.001	1.9 (1.6 – 2.3)	<0.001	2.0 (1.5 – 2.6)	<0.001	1.5 (1.0 – 2.3)	0.046	1.6 (1.2 – 2.2)	0.003	1.7 (1.1 – 2.6)	0.012
Low education[2]	2.0 (1.5 – 2.6)	<0.001	2.1 (1.7 – 2.5)	<0.001	2.3 (1.7 – 3.0)	<0.001	2.0 (1.4 – 3.0)	<0.001	2.3 (1.7 – 3.3)	<0.001	1.8 (1.2 – 2.8)	0.003
Obesity[3]	1.9 (1.4 – 2.6)	<0.001	1.2 (1.0 – 1.5)	0.096	2.0 (1.5 – 2.7)	<0.001	NS	0.211	1.7 (1.2 – 2.5)	0.003	2.5 (1.5 – 4.0)	<0.001
ASA[4] grade 2	NS	0.134	2.0 (1.5 – 2.8)	<0.001	4.3 (2.8 – 6.6)	<0.001	NS	0.264	2.5 (1.5 – 4.3)	0.001	2.5 (1.3 – 4.6)	0.005
Diabetes Mellitus	2.2 (1.1 – 4.7)	0.043	NS	0.149	4.0 (2.2 – 7.1)	<0.001	NS	0.446	2.2 (1.2 – 4.1)	<0.013	4.5 (1.7 – 12.0)	0.003
Anxiety/Depression	1.8 (1.5 – 2.5)	<0.001	1.8 (1.5 – 2.2)	<0.001	2.0 (1.4 – 2.5)	<0.001	1.8 (1.2 – 2.6)	0.004	1.9 (1.4 – 2.6)	<0.001	NS	0.161
Unresolved disability pension issue[6]	4.8 (3.0 – 7.6)	<0.001	3.1 (2.4 – 4.0)	<0.001	3.5 (2.6 – 4.7)	<0.001	2.8 (1.4 – 5.3)	0.002	3.3 (2.3 – 4.8)	<0.001	5.1 (3.2 – 8.1)	<0.001
Unresolved insurance claim[7]	1.9 (1.2 – 3.2)	0.011	2.0 (1.4 – 2.3)	<0.001	2.3 (1.5 – 3.5)	<0.001	2.2 (1.0 – 4.6)	0.042	2.0 (1.1 – 3.6)	0.019	3.3 (1.7 – 6.4)	<0.001
Back pain > 12 months	2.8 (2.2 – 3.6)	<0.001	2.6 (2.1 – 3.1)	<0.001	3.8 (2.9 – 5.1)	<0.001	2.2 (1.5 – 3.1)	<0.001	3.3 (2.4 – 4.5)	<0.001	2.0 (1.3 – 3.1)	0.001
Back pain worse than leg pain[8]	1.4 (1.0 – 1.8)	0.026	1.8 (1.5 – 2.3)	<0.001	1.5 (1.1 – 2.2)	0.015	NS	0.842	1.5 (1.0 – 2.1)	0.035	1.6 (1.0 – 2.7)	0.067
Paresis < grade 4	NS	0.160	NS	0.588	NS	0.565	NS	0.403	NS	0.782	NS	0.102
Previously operated	2.0 (1.4 – 2.7)	<0.001	2.2 (1.8 – 2.7)	<0.001	1.9 (1.4 – 2.5)	<0.001	2.5 (1.6 – 3.8)	<0.001	2.1 (1.5 – 2.9)	<0.001	2.3 (1.5 – 3.4)	0.001
> 2 previous surgeries	4.8 (1.2 – 20.3)	0.032	4.6 (1.8 – 11.6)	0.001	3.8 (1.5 – 9.9)	0.007	NS	0.169	2.9 (0.9 – 9.7)	0.078	7.0 (2.1 – 23.2)	0.001
Back pain > NRS[8] 5	1.6 (1.2 –	<0.001	2.2 (1.8 –	<0.001	4.4 (2.3 –	<0.001	2.0 (1.3 –	<0.001	2.7 (1.7 –	<0.001	2.6 (1.2 –	0.014

	2.0)		2.8)		8.5)		2.9)		4.3)		5.6)	
Leg pain > NRS 5	NS	0.911	NS	0.111	NS	0.387	NS	0.205	NS	0.757	NS	0.701
Daily use of analgesics	0.8 (0.6 – 1.0)	0.030	NS	0.932	NS	0.172	0.6 (0.4 – 1.9)	0.015	NS	0.165	NS	0.914

[1]Range: 0-100 (no-maximal disability). The ODI score was <33, 33-58, and >58 in the subgroups with low, medium high baseline disability.[2]Less than four years of college/university education. [3]Body Mass Index ≥30. [4]American Society of Anesthesiologists grade. [5]EQ-5D 3L questionnaire; 5th item, moderate to severe problems. [6]Pending medical claim/litigation with the Norwegian public welfare agency fund concerning disability pension. [7]Pending medical compensation claim/litigation against private insurance companies or the public Norwegian System of Compensation to Patients. [8]Numeric Rating Scale (0-10).

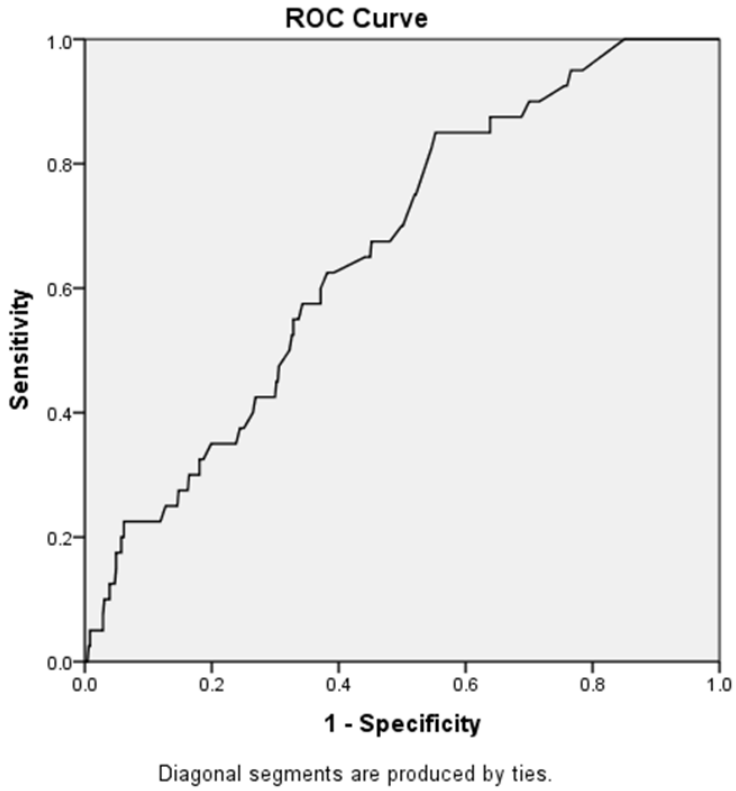
Table 3x. Results from the univariate binary logistic regression analyses of worsening in both the training (n=) and validation (n=) cohort, showing associations (Odds Ratio (OR) and 95% confidence intervals (CI)) between predictors and patient reported worsening (yes/no) of lumbar disc surgery, as defined by validated cut offs on the Oswestry Disability Index (ODI), split on subgroups with low, medium and high baseline ODI scores (percentiles). For all predictors, except age and gender, NS indicates statistical insignificance, p value > 0.1.

	Training Cohort						Validation Cohort					
Baseline ODI[1] group	**<25th**		**25th-75th**		**>75th**		**<25th**		**25th-75th**		**>75th**	
Predictor	**OR**	**P value**	**OR**	**P value**	**OR**	**P value**	**OR**	**P value**	**OR**	**P value**	**OR**	**P value**
Age >60	1.4 (0.8 – 2.5)	0.233	1.1 (0.8 – 1.7)	0.554	1.3 (0.8 – 2.2)	0.304	1.7 (0.8 – 3.3)	0.146	1.7 (1.0 – 2.9)	0.067	0.9 (0.4 – 2.0)	0.748
Living alone	1.8 (1.1 – 2.9)	0.028	NS[3]	0.389	NS	0.114	NS	0.506	NS	0.962	NS	0.264
Nonnative Norwegian speaker	3.2 (1.4 – 7.4)	0.006	2.6 (1.6 – 4.3)	<0.001	3.5 (1.9 – 6.3)	<0.001	NS	0.159	3.6 (1.8 – 7.4)	<0.001	NS	0.401
Female	1.5 (0.9 – 2.4)	0.097	1.0 (0.7 – 1.3)	0.754	1.1 (0.7 – 1.7)	0.640	1.3 (1 – 1.8)	0.104	1.3 (0.8 – 2.1)	0.360	0.6 (0.3 – 1.3)	0.185
Smoking	3.1 (1.9 – 5.0)	<0.001	2.5 (1.8 – 3.4)	<0.001	2.7 (1.7 – 4.2)	<0.001	2.4 (1.3 – 4.3)	0.006	1.8 (1.1 – 3.0)	0.029	NS	0.390
Low education[2]	3.9 (2.1 – 7.1)	<0.001	2.3 (1.5 – 3.3)	<0.001	2.7 (1.5 – 4.6)	<0.001	2.7 (1.3 – 5.5)	0.007	2.6 (1.4 – 4.9)	0.003	NS	0.550
Obesity[3]	NS	0.277	NS	0.184	1.8 (1.1 – 3.0)	0.024	NS	0.302	NS	0.832	NS	0.145
ASA[4] grade 2	2.5 (1.0 – 6.7)	0.058	NS	0.146	3.1 (1.7 – 5.7)	<0.001	NS	0.105	2.4 (1.1 – 5.2)	0.033	2.7 (1.0 – 6.8)	0.041
Diabetes Mellitus	NS	0.216	1.9 (0.9 – 3.9)	0.081	2.5 (1.1 – 5.8)	0.029	NS	0.592	NS	0.477	NS	0.954
Anxiety/Depression	2.4 (1.5 – 3.9)	<0.001	1.6 (1.2 – 2.2)	0.005	1.9 (1.2 – 3.1)	0.012	NS	0.849	3.1 (1.8 – 5.3)	<0.001	NS	0.253
Unresolved disability pension issue[6]	3.8 (2.0 – 7.3)	<0.001	2.1 (1.4 – 3.2)	<0.001	2.7 (1.7 – 4.3)	<0.001	3.3 (1.4 – 7.6)	0.005	3.3 (2.3 – 4.8)	<0.001	5.5 (2.7 – 11.3)	<0.001
Unresolved insurance claim[7]	NS	0.407	1.8 (1.2 – 2.6)	0.002	3.8 (2.1 – 6.9)	<0.001	3.8 (1.6 – 9.4)	0.004	NS	0.736	3.5 (1.4 – 9.2)	0.009
Back pain > 12 months	3.1 (1.8 – 5.1)	<0.001	3.3 (2.4 – 4.7)	<0.001	4.4 (2.8 – 7.0)	<0.001	NS	0.147	3.9 (2.3 – 6.6)	<0.001	5.1 (2.5 – 10.5)	<0.001
Back pain worse than leg pain[8]	NS	0.129	NS	0.121	NS	0.561	NS	0.490	NS	0.27	NS	0.519
Paresis < grade 4	NS	0.939	NS	0.588	NS	0.290	NS	0.143	NS	0.809	NS	0.657
Previously operated	2.4 (1.4 – 4.1)	0.001	3.0 (2.2 – 4.2)	<0.001	1.5 (0.9 – 2.4)	0.099	2.1 (1.1 – 4.0)	0.023	2.2 (1.3 – 3.8)	0.002	3.3 (1.7 – 6.6)	0.001
> 2 previous surgeries	11.0 (2.6 – 46.8)	0.001	3.693	0.021	3.2 (0.9 – 11.3)	0.074	NS	0.159	NS	0.127	11 (3.4 – 35.4)	<0.001
Back pain > NRS[8] 5	2.583	<0.001	3.044	<0.001	5.3 (1.3 –	0.021	2.9 (1.4 –	0.003	5.3 (1.7 –	0.005	NS	0.997

					21.8)		6.0)		17.1)			
Leg pain > NRS 5	1.544	0.098	NS	0.137	NS	0.998	1.8 (0.9 – 3.7)	0.870	NS	0.159	NS	0.999
Daily use of analgesics	NS	0.841	NS	0.678	NS	0.437	0.4 (0.2 – 0.7)	0.002	NS	0.595	NS	0.998

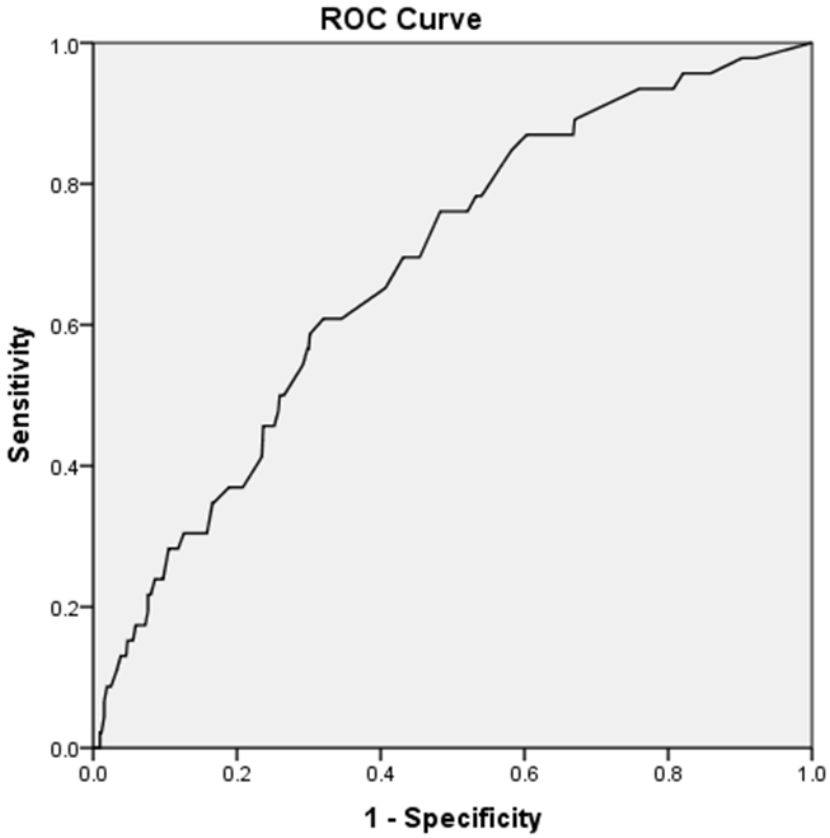
[1]Range: 0-100 (no-maximal disability). The ODI score was <33, 33-58, and >58 in the subgroups with low, medium high baseline disability.[2]Less than four years of college/university education. [3]Body Mass Index ≥30. [4]American Society of Anesthesiologists grade. [5]EQ-5D 3L questionnaire; 5th item, moderate to severe problems. [6]Pending medical claim/litigation with the Norwegian public welfare agency fund concerning disability pension. [7]Pending medical compensation claim/litigation against private insurance companies or the public Norwegian System of Compensation to Patients. [8]Numeric Rating Scale (0-10).

Figure 1x. ROC (Receiver operating characteristics) analysis of predicted risk for failure vs not in the low baseline group (ODI <25th percentile).



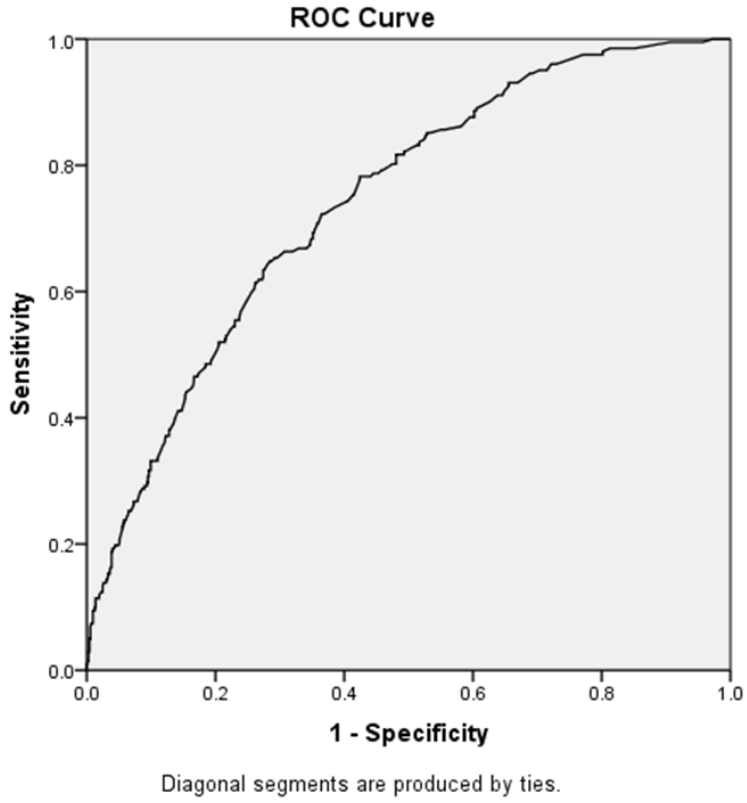
AUC (area under the curve, C-criterion) = 0.67, 95% CI = 0.58 - 0.74.

Figure 2x. ROC (Receiver operating characteristics) analysis of predicted risk for worsening vs not in the low baseline group (ODI <25th percentile).



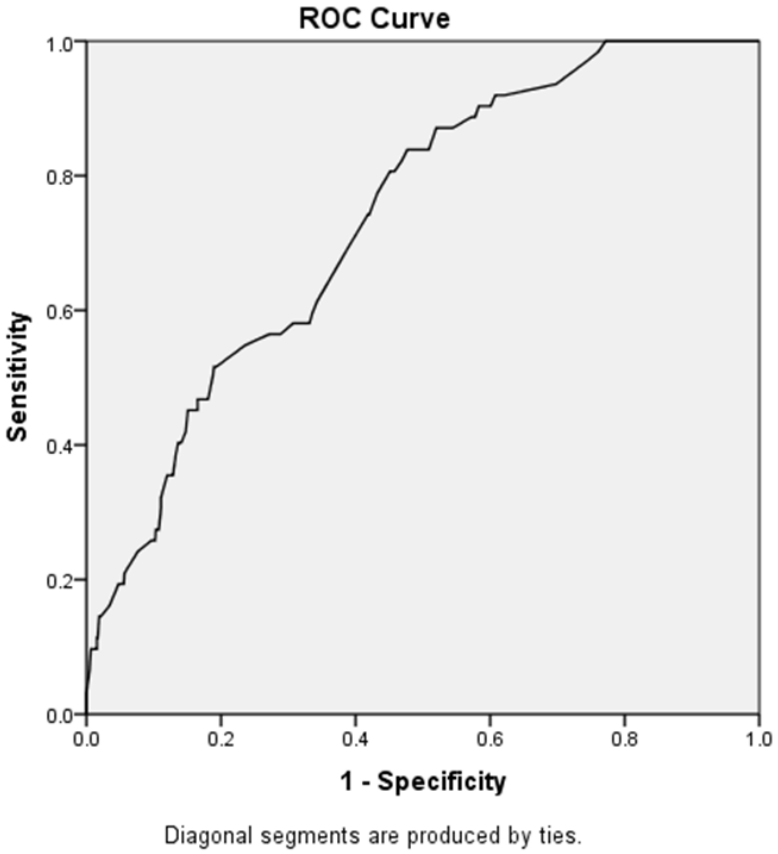
AUC (area under the curve, C-criterion) = 0.68, 95% CI = 0.60 - 0.76.

Figure 3x. ROC (Receiver operating characteristics) analysis of predicted risk for failure vs not in the medium baseline group (ODI 25th-75th percentile).



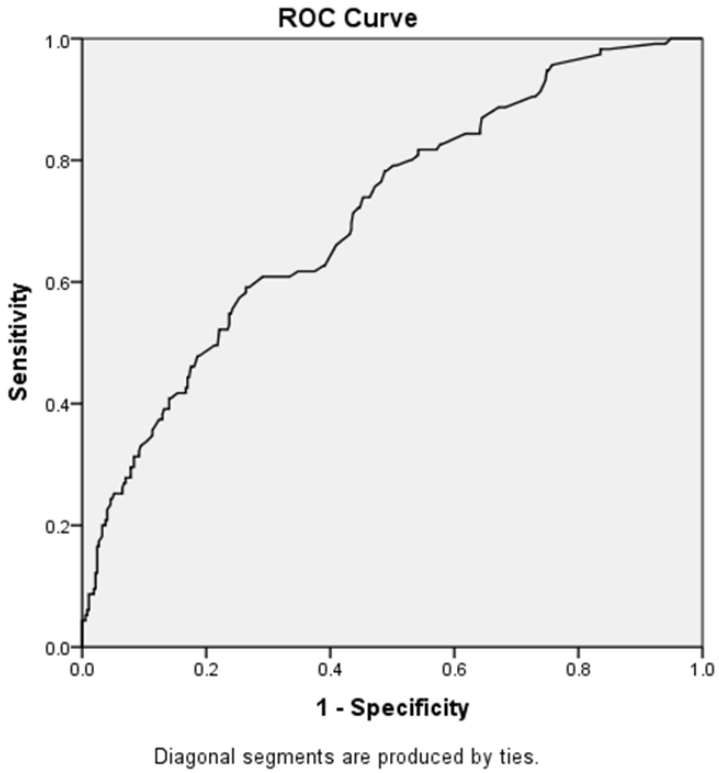
AUC (area under the curve, C-criterion) = 0.74, 95% CI = 0.70 - 0.78.

Figure 4x. ROC (Receiver operating characteristics) analysis of predicted risk for worsening vs not in the medium baseline group (ODI 25th-75th percentile).



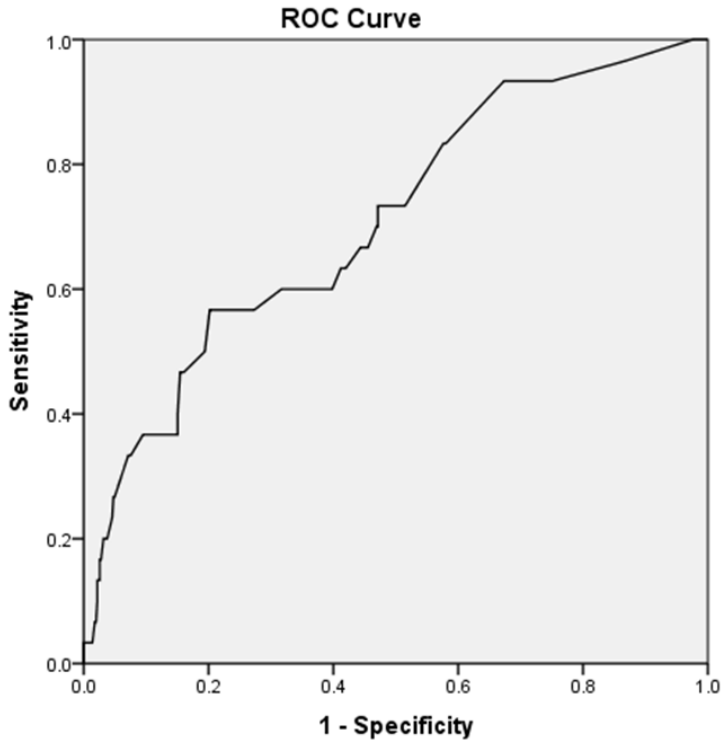
AUC (area under the curve, C-criterion) = 0.74, 95% CI = 0.68 - 0.79.

Figure 5x. ROC (Receiver operating characteristics) analysis of predicted risk for failure vs not in the high baseline group (ODI >75th percentile).



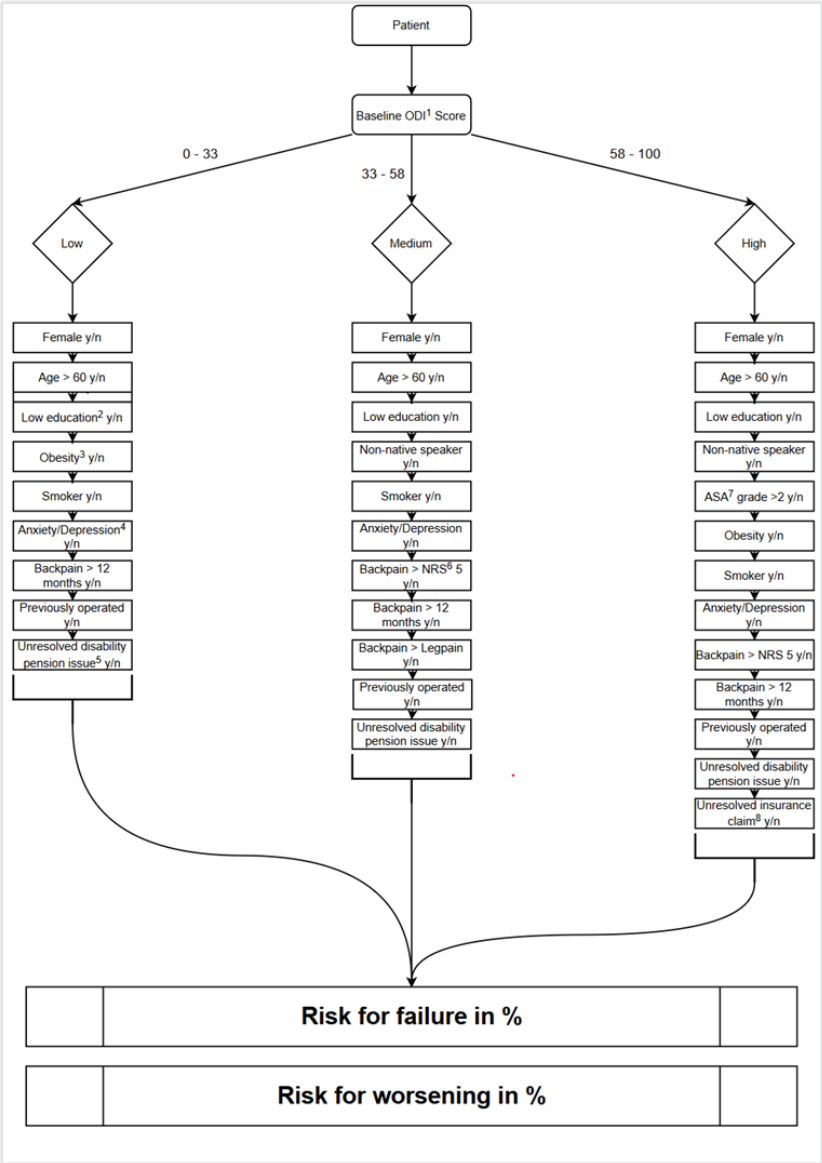
AUC (area under the curve, C-criterion) = 0.71, 95% CI = 0.66 - 0.76.

Figure 6x. ROC (Receiver operating characteristics) analysis of predicted risk for worsening vs not in the high baseline group (ODI >75th percentile).



AUC (area under the curve, C-criterion) = 0.71, 95% CI = 0.61 - 0.81.

Figure 7x. Model algorithm for the three ODI baseline groups. Based on the preoperative ODI the patient will be classified via one of the three pathways, calculating an overall risk for either failure or worsening. Risk is calculated from the odds of each risk factor. The risk factors are listed in random order, and their place in the sequence does not reflect their odds.



[1]Range: 0-100 (no-maximal disability). The ODI score was <33, 33-58 and >58 in the subgroups with low, medium high baseline disability, respectively.[2]Less than four years of college/university education. [3]Body Mass Index ≥30. [4]EQ-5D 3L questionnaire; 5th item, moderate to severe problems. [5]Pending medical claim/ litigation the Norwegian public welfare agency fund concerning disability pension. [6]Numeric Rating Scale (0-10). [7]American Society of Anesthesiologists grade [8]Pending medical compensation claim/litigation against private insurance companies or the public Norwegian System of Compensation to Patients