Measurement of outcome in lumbar spine surgery

Validity and interpretability of frequently used outcome measures in the Swespine register

Catharina Parai

Department of Orthopaedics, Institute of Clinical Sciences
Sahlgrenska Academy, University of Gothenburg

Gothenburg 2020

Measurement of outcome in lumbar spine surgery
Validity and interpretability of frequently used outcome measures in the Swespine register

catharina.parai@spinecenter.se

Cover illustration by Ylva Nelson
Layout by Nikolaos Vryniotis

Printed in Borås, Sweden, 2020
Printed by Stema AB

Trycksak
3041 0234

SBN 978-91-7833-796-5 (PRINT)
ISBN 978-91-7833-797-2 (PDF)
http://hdl.handle.net/2077/63237

To my family,
whom I love beyond measure

TABLE OF CONTENTS

ABSTRACT

Measurement of outcome in lumbar spine surgery

Validity and interpretability of frequently used outcome measures in the Swespine register

Catharina Parai

Department of Orthopaedics, Institute of Clinical Sciences,
Sahlgrenska Academy, University of Gothenburg,
Gothenburg, Sweden

ABSTRACT

BACKGROUND
The purpose of elective lumbar spine surgery is mainly to reduce pain and to improve physical function and quality of life. The quality and results of the interventions are monitored in the Swedish spine register, Swespine. The large quantities of data offer unique opportunities to improve quality of care, decrease costs and enable benchmarking. For register-based data to be useful, however, the quality must be high, the variables must be carefully selected to ensure relevant data collection, and the logistics of data collection should be workable.

AIM
The overall purpose of the thesis was to find ways to simplify the assessment of patient-reported outcome without a loss in scientific credibility.

STUDY POPULATION
The main study population was obtained from the Swespine register and included patients operated in the lumbar spine in the period 1998-2015 for either disc herniation (n: 30,102), spinal stenosis (n: 50,194), isthmic spondylolysis/spondylolisthesis, or degenerative disc disorder. The two latter diagnoses

were treated as a single entity (n: 13,836). A test-retest study was performed on 182 individuals obtained from two spine-care hospitals (2017-2019). Analyses on non-respondents were computed using Swespine data from 2008-2012 that were linked to hospital data, Statistics Sweden, the National Patient Register, and the Social Insurance Agency (n: 21,961).

METHODS
The usefulness of the single-item retrospective outcome measure GA as an overall PROM (Patient-Reported Outcome Measure) was tested in correlation analyses with symptom-specific (i.e VAS), disease-specific (i.e ODI), and generic PROMs (i.e EQ-5D, SF-36). The capability of GA as a discriminator of treatment success was explored in ROC curve analyses. The level of treatment success was defined for each of the Swespine PROMs with different lumbar conditions. The proportion that achieved these scores one year after the operation was compared with the proportion at two years. PROM retest reliability was tested on a symptom stable population. The SDC at the 95% confidence level was computed. The retrospective measurements were tested using weighted kappa. Regression analyses were conducted to identify variables associated with non-response. The output was used to predict outcomes for patients with the characteristics of the non-respondent population.

RESULTS AND CONCLUSIONS
High correlations were seen between GA and VAS, and also the ODI, indicating that GA can replace these tools in effectiveness studies. The correlations were better for final scores than for changes in score, indicating present-state bias and/or recall bias. Correlations with EQ-5D were lower, indicating that GA works less well as a discriminator of quality of life. The ROC curve analyses support the use of GA as a reference criterion in the interpretation of VAS and ODI scores. A tough cut-off signifying a considerable improvement is encouraged. The change in a PROM score needed to achieve treatment success (i.e. the MIC value) varied somewhat between the degenerative conditions tested; thus, the ODI MICs were 14-22 points, the VAS_{BACK} MICs were 20-29 mm; the VAS_{LEG} MICs were 23-39 mm; and the EQ-5D MICs were 0.10-0.18. The proportion of patients who reached these levels at the one-year follow-up was similar to the proportion at the two-year follow-up. Thus, collection of PROM data in Swespine on the latter occasion is not necessary. The retest reliability for the PROMs tested was similar or lower than previously reported. In general, the SDC estimates exceeded the MIC values, thereby complicating the interpretation of score changes, as the PROMs were not sensitive enough to detect score

changes considered important. Being lost to follow-up was associated with male sex, younger age, smoking, lower disposable income, and lower education, and with being born outside the EU. Non-respondents were predicted to have a somewhat worse outcome than respondents.

Keywords: spine register, disc herniation, spinal stenosis, degenerative disc disorder, patient-reported outcome measure, Global Assessment, minimal important change, smallest detectable change, retest reliability, non-response to follow-up, attrition, measurement of change.

SAMMANFATTNING

SAMMANFATTNING SUMMARY IN SWEDISH

AVHANDLINGENS BAKGRUND.

De senaste decennierna har antalet elektiva ländryggsoperationer ökat påtagligt. Syftet med kirurgin är huvudsakligen att minska smärta och förbättra fysisk funktion och livskvalitet. Kvaliteten på och resultaten av operationerna dokumenteras sedan 1998 i det svenska ryggregistret, Swespine och idag är mer än 100 000 ländryggsoperationer registrerade. Information från patienter samlas in före operationen, samt efter 1, 2, 5 och 10 år. Patient-rapporterade utfallsmått förkortas PROMs. Det finns två typer av PROMs, dels de som mäts före och efter ett ingrepp, dels de som enbart mäts efteråt – båda har sina för- och nackdelar. Registrets data erbjuder unika möjligheter att förbättra vården, minska kostnader och kan fungera som en måttstock vid jämförelser. För att registerdata ska vara användbara behöver kvaliteten vara hög. Detta kan uppnås genom ett högt deltagande, noggrant utvalda bakgrundsvariabler, process- och utfallsmått med god validitet, samt en uppföljning som fungerar för såväl patienter som administratörer.

MÅLSÄTTNINGAR

Att undersöka hur det tillbakablickande enfrågemåttet Global Assessment (GA) fungerar som utfallsmått efter degenerativ ländryggskirurgi. Att ta reda på om det finns kliniskt relevanta skillnader i PROM-data mellan ett- och tvåårsuppföljningen, som berättigar datainsamling vid båda tillfällena. Att mäta den minsta statistiskt upptäckbara skillnaden mellan två mättillfällen för vart och ett av de PROMs som används i Swespine. Att jämföra dessa med den minsta skillnaden i PROM-värde som uppfattas som en viktig förbättring. Att undersöka skillnader i bakgrundsvariabler och i utfall mellan de individer som har registrerade uppföljningsformulär i Swespine med dem som inte har det.

UNDERSÖKTA INDIVIDER

Den huvudsakligen undersökta populationen inhämtades från Swespines databas och innehöll patienter som opererats mellan åren 1998–2015 för diskbråck, spinal stenos (ryggkanalsförträngning) eller kronisk ländryggssmärta. En så kallad retest-studie utfördes på en mindre grupp individer med

stabil symtombild. Undersökningen av individer utan uppföljningsformulär byggde på data från Swespine 2008–2012, som länkats samman med landstingens patientadministrativa system, Statistiska Centralbyråns register, Socialstyrelsens patientregister samt Försäkringskassans register.

METODER
Användbarheten hos GA undersöktes genom att detta utfallsmått korrelerades med etablerade utfallsmått som mäter smärta (VAS), fysisk funktion i relation till ryggsmärta (ODI), samt livskvalitet (EQ-5D). Förmågan hos GA att skilja ut patienter med ett eftersträvat resultat undersöktes med ROC-metoden. Den grad av förändring, mätt med respektive PROM, som kan tolkas som en klar förbättring definierades, också den med ROC-metoden. Andelen patienter som uppnådde den definierade förbättringen efter ett år jämfördes med andelen patienter som rapporterade samma grad av förbättring året därpå. McNemars statistiska test användes för att jämföra hur patienter svarade på de retrospektiva enfrågemåtten vid ett- respektive tvåårsuppföljningen. Ett PROMs pålitlighet vid upprepade mätningar testades på en patientgrupp vars symtom antogs vara oförändrade under tiden studien pågick. Mätfelet för respektive PROM räknades ut. Prediktionsmodeller baserade på en stor mängd variabler skapades för att identifiera faktorer som i högre utsträckning förekommer hos den grupp för vilken uppföljningsdata saknas i Swespine. Det predicerade resultatet beräknades för denna grupp.

RESULTAT OCH SLUTSATSER
GA korrelerade till VAS och ODI på ett sådant sätt att det skulle kunna ersätta dem vid rutinmässig uppföljning av erkänd kirurgisk behandling av degenerativ ländryggssjukdom. Analyserna talade dock för att patienternas nuvarande hälsotillstånd kan påverka hur GA besvaras. GA föreföll fungera sämre för att beskriva förändring i livskvalitet. Om GA ska användas som en referens för att tolka en förändring i ett PROM-värde bör svarsalternativen ”smärtfri” och ”mycket förbättrad” användas för att definiera en förbättring. Den förändring i PROM-värde som krävdes för att uppnå denna definition av förbättring varierade beroende på diagnosgrupp. För ODI låg förändringen på 14 – 22 poäng, för VAS_{RYGG} 20–29 mm, för VAS_{BEN}: 23–29 mm och för EQ-5D på 0.10-0.18. Storleken på mätfelet var oftast större än dessa förbättringsvärden. Detta är problematiskt eftersom en patientrapporterad förbättring i ett sådant fall inte kan särskiljas från PROM-instrumentets mätfel, eller med andra ord från slumpen. Andelen patienter som rapporterade förbättring efter sin operation vid ettårsuppföljningen var likvärdig med den andel som

uppgav förbättring efter två år. En tvåårsuppföljning är därför inte nödvändig att ha med i Swespines uppföljningsrutin. Gruppen av patienter som saknar uppföljningsdata består i högre utsträckning av yngre, män, samt rökare. Gruppen har också jämförelsevis lägre utbildning, en lägre inkomst, samt är född utanför EU. Den här gruppen predicerades att ha ett lite sämre resultat. Resultaten kan tolkas som att utfallet mätt med PROMs i Swespine är något överskattat.

Resultaten i avhandlingen talar för att man kan förenkla uppföljningsrutinen i Swespine genom att minska antalet PROMs och ta bort ett uppföljningstillfälle. Detta kan leda till en ökad svarsfrekvens vilket i sin tur ökar datakvaliteten när den insamlade informationen ska analyseras. Instrument som mäter subjektiva tillstånd som smärta hos en befolkning som behandlas för degenerativa åkommor, där en förändring i symtombilden inte enbart orsakas av operationen utan kanske också av den degenerativa processen i sig, eller av andra sjukdomar, är en utmaning. Svårigheten med att tolka resultaten är uppenbar. Avhandlingen bidrar till att underlätta denna tolkning.

LIST OF PAPERS

LIST OF PAPERS

This thesis is based on the following studies, which are referred to in the text by their Roman numerals.

I. Catharina Parai, Olle Hägg, Bengt Lind, Helena Brisby. The value of patient global assessment in lumbar spine surgery, an evaluation based on more than 90,000 patients.
Eur Spine J (2018) 27:554–563.

II. Catharina Parai, Olle Hägg, Bengt Lind, Helena Brisby. Follow-up of degenerative lumbar spine surgery - PROMS stabilize after 1 year: an equivalence study based on Swespine data.
Eur Spine J. 2019 Sep;28(9):2187-2197.

III. Catharina Parai, Olle Hägg, Bengt Lind, Helena Brisby. ISSLS prize in clinical science 2020: the reliability and interpretability of score change in lumbar spine research.
Eur Spine J. 2019 Nov 23.

IV. Catharina Parai, Olle Hägg, Carl Willers, Bengt Lind, Helena Brisby. Characteristics and predicted outcome of patients lost to follow-up after degenerative lumbar spine surgery.
Submitted

ABBREVIATIONS

ABBREVIATIONS

AUC	Area Under the Curve
COSMIN	COnsensus-based Standards for the selection of health Measurement INstruments
DDD	Degenerative Disc Disorder
EQ-5D	European Quality of Life 5-dimension questionnaire
FU	Follow-Up
GA	Global Assessment
ICC	Intra-class Correlation Coefficient
ICHOM	International Consortium and Health Outcomes Measurement
LDH	Lumbar Disc Herniation
LoA	Limits of Agreement
LSS	Lumbar Spinal Stenosis
MAR	Missing At Random
MCAR	Missing Completely At Random
MCS	Mental Component Summary
MIC	Minimal Important Change
MNAR	Missing Not At Random
NRS	Numeric Rating Scale
ODI	Oswestry Disability Index
PASS	Patient Acceptable Symptom State
PCS	Physical Component Summary
PREM	Patient-Reported Experience Measure
PROM	Patient-Reported Outcome Measure
PROMIS	Patient-Reported Outcome Measurement Information System
RCT	Randomized Controlled Trial
ROC	Receiver Operating Characteristic
SEM	Standard Error of Measurement
SF-36	Short Form-36
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology
TQ	Transition Question
VAS	Visual Analogue Scale

INTRODUCTION

1. INTRODUCTION

1.1 The registration of outcome

1.1.1 The historical background of a national quality register

Amory Codman lowered the newspaper and leaned back in his armchair. Absent-minded, he stroked the head of one of his dogs. The Germans seemed unstoppable in their incessant bombing of Britain. This was in the autumn of 1940 and the 69-year-old surgeon was fighting a war of his own, one that he could not win, against malignant melanoma. He came to think of another lost battle he had fought against his own peers and hospital administrators. For many years, he had urged them to study the outcome – the end result – of their treatments, but to no avail. In fact, it had cost him his career and reputation. Still, he was certain he was right. When publishing his paper on hospital efficiency, his statement that "the common sense notion that every hospital should follow every patient it treats, long enough to determine whether or not the treatment has been successful, and then to inquire, 'if not, why not?' with a view to preventing similar failures in the future"[1] was true. Getting ready for a slow walk with the dogs, he thought: "Honours, except those I have thrust upon myself, are conspicuously absent..., but I am able to enjoy the hypothesis that I may receive some more from a more receptive generation."[2]

Indeed, Codman's hypothesis proved correct, and today he is acknowledged as a pioneer in outcome assessment [3].

The contemporary giant of healthcare quality assessment was Avedis Donabedian (1919–2000), a professor of medical care organization at the University of Michigan School of Public Health. In the article "Evaluating the Quality of Medical Care"[4], he introduced what have become the three pillars in the evaluation of the quality of healthcare: structure, process, and outcome. Donabedian emphasized that – despite having many limitations – outcome measures remained the "ultimate validators of the effectiveness and quality of medical care". In a presentation summarizing 20 years of developments within the field, held in Portland, Oregon in

1984, Donabedian called for a clinically relevant quality assessment based on individual and social valuations [5].

Sweden is regarded as being a modern country with strong democratic values and a high degree of confidence in authorities [6], with traditionally positive attitudes towards registers. Since the introduction of the civic registration numbers in 1947, a large number of national registers such as the Cancer Register (1958) and the Patient Register (1987) have been developed. The first Swedish quality registers were the Knee Arthroplasty Register (1975) and the Hip Arthroplasty Register (1979) [7]. The Spine Register – Swespine – was founded in 1993, and was launched nationally in 1998 [8].

1.1.2 The framework of a national spine register

An outcome register is described as an organized system that uses observational study methods based on STROBE (Strengthening of the Reporting of Observational Studies in Epidemiology) recommendations [9]. The purpose of a national spine register is to ensure and improve the quality of the care provided, to enable benchmarking, to detect rare or late complications, and to make visible changes in surgical techniques, in the use of implants, in indications, and in outcomes [9].
Guidelines, such as the STROBE statement, that aim to ensure the accuracy and generalizability of a study [10], and recommendations from the ICHOM (International Consortium for Health Outcomes Measurement) collaboration [11] are used to ensure that the requirements for achievement of the purpose of a register are fulfilled. Such concerns include a standardized approach to data collection at baseline and follow-up of at least one year. Other recommendations concern the prevention of selection bias by providing accurate patient characteristics at baseline to enable adjustment for covariates. A national register has the advantage of having a lower risk of selection bias compared to institutional and sponsored registers because it covers the whole country.

The target groups must be properly defined and the patient-reported outcome measures should show good measurement properties. There is no consensus on a minimal patient response rate [12]. Postal, e-mail, or telephone reminders are encouraged. Finally, data analyzed should be presented to the participating spinal units and also to the public.

The internal validation of a national register is a continuous process. Logical checks are put into the software to avoid input of obviously incorrect data at the start-up of a register. Checks for inconsistent or unlikely

data are also run on a regular basis. A time consuming but important validation method is the comparison of register data with patient records. Thanks to the Swedish system of having personal identity numbers, there is the possibility of validating data against other registers such as the National Patient Register (NPR), Statistics Sweden (SCB), and the Social Insurance Authority (FK). Validation by an adjudication committee may be used for estimation of the degree of correct diagnoses [13].

There are a few basic concepts that have a profound effect on the external validity of the data in a national spine register. Coverage is the number of spinal units that report their operations to the register divided by the total number of spinal units. Completeness is the number of operated patients in the register divided by the total size of the operated population [13]. Hence, if fewer perioperative forms are registered a decrease in completeness will occur. Patients who do not return the follow-up questionnaires to the register for any reason are called non-respondents. Partially missing data, i.e. loss of one or several variables or items, do not affect register completeness, but they may cause less robustness of results as the data are analyzed.

1.1.3 The value and scientific character of a national spine register

The efficacy of an intervention, i.e. the ability of an intervention to produce a desired or intended result, is usually tested under ideal conditions in a randomized controlled trial (RCT). The effectiveness of that intervention, i.e. the ability of an intervention to produce a desired or intended result in practical clinical work, can be examined in register-based studies that allow for the variable conditions of real life to be included[14]. The efficiency, i.e. the ability of the intervention to produce a desired or intended result in practical clinical work with an optimum use of resources, may need data from additional sources, for instance the National Board of Health and Welfare and/or the Social Insurance Authority. Recent high-quality studies have indicated that a register-based study and a randomized controlled trial can produce equally valid results [15,16].

An RCT enables hypothesis testing. Any statistically significant differences are immediately interpretable. However, the conclusion applies only to the study sample that was considered eligible for the study after informed consent was given. A register-based study reflects reality. But in this case, the causality behind a statistically significant difference

in outcome between, for example, two spine-care units is not explicable before confounding factors have been considered. The Swedish Association of Local Authorities and Regions and the National Board of Health and Welfare display case-mix adjusted outcomes on the web page Open Comparisons, with the aim of more accurately reflecting the quality of care received [17]. It may be an unachievable task to account for all possible bias; thus, register-based studies may be regarded as hypothesis generators. Statistical models such as propensity score matching, which aim to overcome the biasing obstacles and mimic a randomized experiment, require the skills of an experienced statistician and a researcher with a vast knowledge of the register population and the quality of the register data.

Register data that are being collected on consecutive patients, including patient-reported outcome measures before the surgery and at specific time points after surgery, are considered to be prospectively collected data - even though the study question is not designed at the start of data collection [18].

1.1.4 Swespine: the Swedish spine register

The rapid increase in the number of surgical interventions in the spine led to the foundation of a spine register in Lund in 1993. It was launched nationally in 1998 as a patient-based protocol and a comprehensive computer application was introduced, and since 1999 all data except the surgical report have been patient-based [19]. There has been a gradual increase in coverage. Around the millennium, approximately 80% of the spinal units registered their operations in Swespine and in 2018, 97% did [20].

The preoperative data registered are age, sex, smoking habits, working conditions, sick listing time, pain duration, walking distance and consumption of analgesics. Several Patient-Reported Outcome Measures (PROMs) are collected. Pain severity was reported on the Visual Analogue Scale (VAS) until 2016, at which time it was replaced by a Numeric Rating Scale (NRS). Disability has been measured by the Oswestry Disability Index (ODI) since 2003. Quality of life has been registered with the Rand Short Form-36 (SF-36) and the European quality of life instrument EQ-5D-3L (EQ-5D), but since 2016 solely using the EQ-5D. In the 2016 revision, specific questions on opioid use and physiotherapy were also added.

The surgical data include diagnosis for surgery, type of intervention, implants, and adverse events.

Postoperative data are collected at 1, 2, 5, and 10 years, with the preoperative

Figure 1. The Swespine logo.
With permission from the Board of the Swedish Society of Spinal Surgeons

protocol and an additional question labelled Global Assessment (GA) about the patient´s opinion on back and leg pain as compared to before the surgery, and a question on patient satisfaction with outcome of surgery (Satisfaction).

Completeness nationwide has been approximately 80%. However, there is a large variation between spinal units (30-100%). Practice in spine registries was reviewed by van Hooff et al. in 2015 [9]. The authors concluded that Swespine is a spine register of high quality. Although non-respondents in Swespine are considered to be treated patients, the completeness of the register is completely dependent on the treating surgeon. The completeness in Swespine was 78.4% in 2018 [20], which means that 21.6% of the patients who were surgically treated for a spinal condition (other than trauma) were not registered. This means that missing data, in terms of completeness, are related to the ability or willingness of the surgeon or hospital administrators to register the perioperative data in Swespine.

As with any survey, Swespine does not achieve a complete response from the patients on the follow-up occasions, which may affect the external validity by introducing the risk of selection bias. When the registered patients are no longer representative of the target population, the value of the results is weakened. Follow-up rates can be improved by increasing the number of reminders, but these efforts are costly and at some point, ineffectual. A systematic loss to follow-up occurs if the characteristics of the non-respondents differ in a substantial way. A random loss to follow-up is less serious and results in a smaller number of patients on which calculations can be based, and wider confidence intervals.

1.2 Patient-Reported Outcome Measures (PROMs)

To make a clinical decision relevant, the priorities and preferences of the patient as well as the clinician must be considered. This interaction is a cornerstone in evidence-based medicine, and during the last four decades there has been a steep rise in the use of PROMs [18,21]. Other reasons for the use of PROMs are that objective

measures for subjective traits such as chronic pain are inconclusive, that the assessment made by the treating surgeon is not always consistent with that of the patient, and that a specific purpose of health services is to increase gain in health for patients in terms of patient self-assessment of health [22]. PROMs may also be useful in areas other than the monitoring of interventions, for instance in facilitating communication and shared clinical decision-making [22,23].

From a degenerative spine surgery point of view, PROMs are standardized and validated questionnaires or questions that are completed/answered by patients in adherence to the intervention to determine their opinion of their general health quality, function, and pain[24]. PROMs that measure quality of life in general and permit comparisons between different disease entities are called generic (for example, the quality of life questionnaires SF-36 and EQ-5D) whereas measures focusing on certain conditions are called disease-specific (such as the low back pain questionnaire ODI). Scales measuring a single construct (for example, the pain-specific VAS) are called symptom-specific PROMs.

Retrospective single-item measures concerning a globally perceived effect of the outcome or of a health state are called transition questions (TQs). Questions such as "How is your pain now as compared to before your treatment?" have been used by clinicians in daily practice for many years. However, when they are asked by the patient's own physician, bias is introduced. By inclusion in follow-up questionnaires completed by patients at home, the TQs are used in a scientifically more correct manner. Although readily understandable and easy to use, factors such as recall bias, present-state bias, response shift [25,26], and the risk of not covering all important aspects of the trait to be measured have called the validity of TQs into question [27-30]. Multiple-item PROMs measuring a health state or a disease-specific condition before and after an intervention have been developed in an attempt to overcome these obstacles. However, these PROMs are not protected from response shift. Furthermore, they have other problems - such as the difficulty of handling incomplete responses, and floor and ceiling effects. Also, a large amount of questions may contribute to lower response rates, greater administrative costs, and difficulty in interpretation [31].

The theoretical framework behind the development and use of PROMs in the form of questionnaire scales originates from the social sciences, and it was introduced to the health sciences via the psychological research sector in the 1960s [21]. In the field of

psychometry, measurement theories such as classical test theory and item response theory were developed, giving physicians and researchers the opportunity to evaluate "unmeasurable" traits like feelings and pain by asking questions in a systematic and scientifically sound way. The epidemiologist Alvan Feinstein was a major critic of the questionnaires developed by psychometricians because of the difficulty in using the measures in clinical practice, and in the 1980s his work gave rise to an alternative branch, called clinimetrics [32]. A clinimetric scale does not need an internal validation. Arguments were later put forward that psychometry and clinimetry are two sides of the same coin and that further development of this kind of outcome measure had everything to gain from cooperation between the two camps [33,34]. In recently published guides on health measurement, a division is avoided [35-37].

1.2.1 Measurement properties

Like any other measurement tool, PROMs need to be validated - that is, do they measure what we want them to measure, and how well [38]? There is an abundance of names and definitions for the same measurement property [37]. In an international Delphi survey (COSMIN), consensus was reached on quality criteria for the measurement properties and also on a common terminology and classification [39,40]. Guidelines from the COSMIN group were recently updated[41]. The COSMIN taxonomy will be adhered to in this thesis. A summary of how the measurement properties are related is given in Figure 2.

1.2.2 Reliability

Reliability is defined as "the degree to which the measurement is free from measurement error" [39]. Measurement error is expressed in the units of the measurement tool in question and it is affected by the instrument's ability to distinguish between patients (inter-individual variation) and also the size in score variation between repeated measures on the same patient (intra-individual variation). A reliability parameter tells us how well patients can be distinguished from measurement error.

The Limits of Agreement (LoA) described by Bland and Altman is a central concept in the measurement of agreement in method comparison studies [42]. A Bland-Altman plot can visualize the inter-rater repeatability of a method through the limits of agreement [43]. A frequently used reliability parameter is the Standard Error of Measurement (SEM). There are no parameters of measurement error for categorical variables since there are no units of measurement. Instead, percentage of agreement is calculated. Examples of reliability parameters

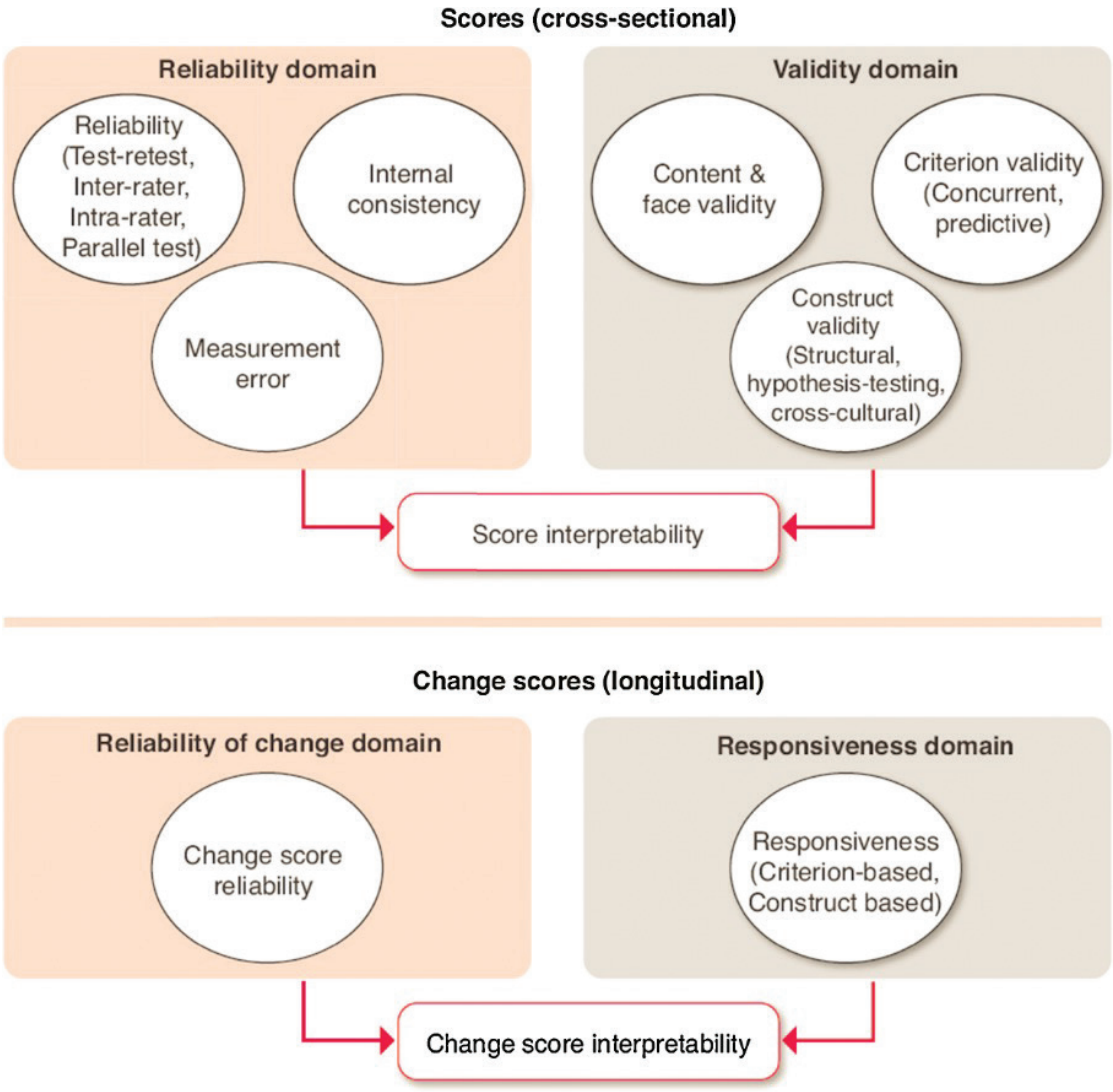


Figure 2. A taxonomy of measurement properties for an instrument's scores and change scores Reproduced from Measurement and the Measurement of Health with permission from Wolters Kluwer Health.

for continuous variables are the intra-class correlation coefficient (ICC) for absolute agreement, ranging from 0 to 1, and for categorical variables, Cohen's kappa (nominal variables), and weighted kappa (ordinal variables), ranging from –1 to 1.

When a measure is tested on stable patients (i.e. there are no symptom fluctuations) on two or several occasions, the scores can be expected to be more or less the same. This test situation is called a test-retest.

1.2.3 Validity

A criterion validity may be determined if there is a gold standard to compare with. There is, however, no such standard for PROMs. Instead, agreement with other measurements that concern the same concept is estimated. This is termed construct validity. One limitation is the potential inaccuracy of the outcome tool that is used as reference criterion. Face validity describes whether the purpose of the instrument is logical and readily understandable or not.

The focus in a validation process lies on the score obtained by the measurement tool, which means that the instrument should be validated each time it is applied to a new setting, for instance a new target population defined by age, culture, diagnosis, and so on. As in any other research situation, hypotheses should be formulated if possible. The degree of validity of a PROM is usually based on results from a number of validity studies[37].

Validation parameters are, among others, correlation coefficients, and the area under the curve (AUC) in receiver operating characteristic curve (ROC) analyses. The appropriate statistic to be used depends on the level of measurement of the two measures i.e. dichotomous, ordinal or continuous, as described by Polit and Yang (chapter 12) [36].

1.2.4 Responsiveness

Our goal as surgeons is that the interventions we perform will lead to a noticeable decrease in symptoms in our patients. The founding father of clinical epidemiology, clinimetrics, Alvan Feinstein pointed out the importance of a measure's sensitivity to change[32]. The ability of a measure to detect change is called responsiveness. The term was introduced into the clinical literature by Kirshner and Guyatt in 1985[44].

Responsiveness, the validity of change scores, can be tested using the same statistical methods as exemplified in the validity section. A transition question about global perceived effect of the intervention is often used as gold standard, although the reliability and validity of such questions have been criticized [45].

Parameters of responsiveness vary with context as well as with population, and several approaches to assessing the validity of change scores are recommended [46-48]

1.2.5 Interpretation of score changes

One approach to interpret changes in PROM scores is to identify a plausible score change beyond which the patient considers the intervention worthwhile. Jaeschke and colleagues

were the first to introduce the concept of minimal clinically important difference in 1989 [49]. Since then, a number of similar concepts have emerged [46]. In this thesis, the choice was made to adhere to the terminology of the COSMIN guidelines [39]. Hence, the Minimal Important Change (MIC) is used to describe the smallest detectable change in score that is considered important to patients [39]. Another variable that should always accompany the MIC is the Smallest Detectable Change (SDC) [50], which is the smallest change in score that is not due to chance.

The SDC is based on population variability in change and does not say anything about the patients' opinion of the outcome, and therefore the MIC is the parameter of choice when it comes to interpreting score changes. The problem, however, is that the SDC is sometimes larger than the MIC, making it impossible to distinguish MIC from chance (Figure 3). Usually only one of the two is given in a scientific paper.

Parameters of interpretability such as the SDC and MIC can be assessed by several methods – either anchor-based or distribution-based, or by the Delphi method. There is no consensus as to which one is preferable to the other. It has been proposed that the SDC should be determined with a distribution-based method and the MIC with an anchor-based method in the same population, and that a combination of the two should be used in the interpretation of change scores [51,52,47,53,54]. Parameters in distribution-based methods include the effect size (ES), the standardized response mean (SRM), and the SEM. Anchor-based approaches include average change in score, change difference, minimum detectable change (95% CI), and ROC curve (area under the curve), and they usually involve the patient's self-assessment of change as the reference criterion or, less commonly, a clinical anchor. The anchor assigns patients into groups reflecting their degree of change [48].

One way of circumventing the difficulties in the interpretation of changes in PROM scores is simply to estimate the patient's current health state only at follow-up. Tubach et al. described the cut-point in a PROM score above which the current health state at the follow-up is satisfactory - the "patient-acceptable symptom state" (PASS) [55] . According to a study by van Hoof et al., the ODI equivalent to PASS was 22 [56].

1.3 PROMs in Swespine

1.3.1 EQ-5D-3L

The European Quality of Life 5-dimension questionnaire is a standardized instrument developed by

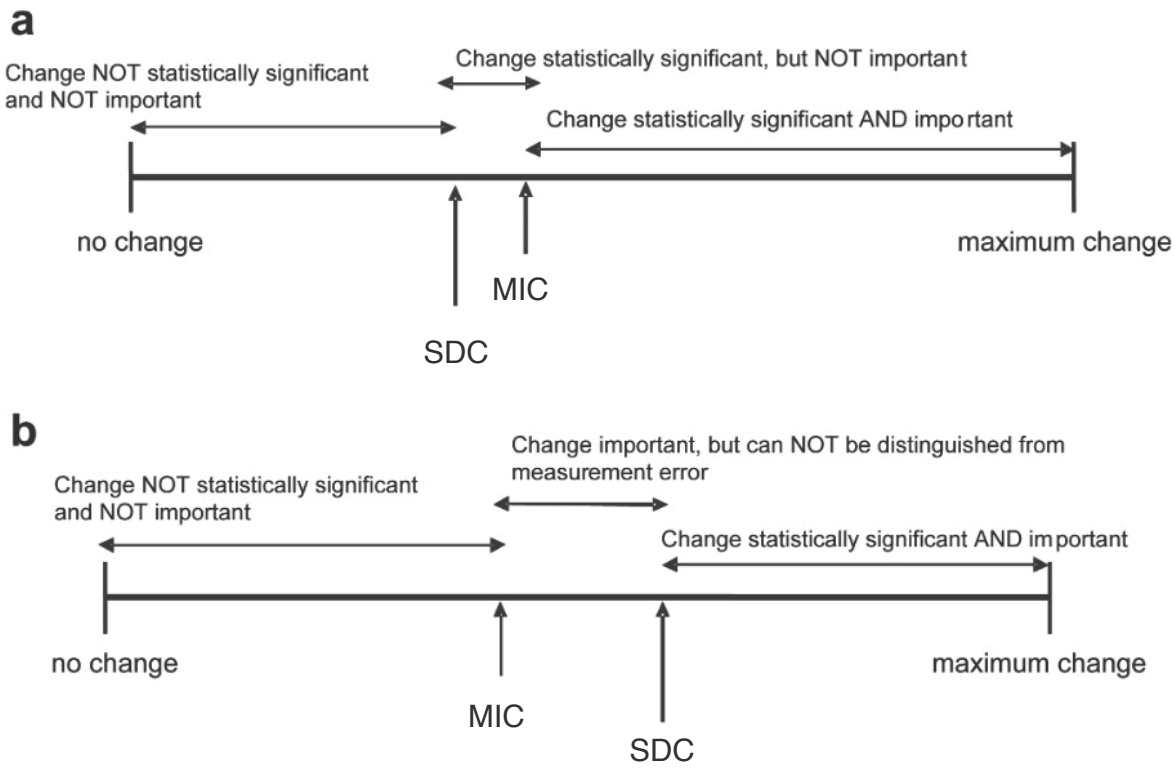


Figure 3. a. Interpretation of change when MIC is larger than SDC. b. Interpretation of change when MIC is smaller than SDC.
Reproduced from the Journal of Clinical Epidemiology with permission from Elsevier.
Reproduced from Measurement and the Measurement of Health with permission from Wolters Kluwer Health.

the EuroQol Group as a measure of health-related quality of life [57-59]. With the descriptive part of the instrument, the patient can classify his or her health in five dimensions - mobility, self-care, usual activities, pain/discomfort, and anxiety/depression - with three levels of severity: no problems, moderate problems, or extreme problems. The score can either be presented as a health profile or be converted to a single summary index number (utility) reflecting preferability compared to other health profiles. Value sets have been derived for EQ-5D in several countries, among others Denmark and Norway. These value sets have been obtained using the EuroQol Visual Analogue Scale (EQ VAS) or the Time Trade-Off (TTO) techniques, and reflect the opinion of the general population (see below).

The EQ VAS is a vertical scale ranging from 0 (representing worst imaginable health) to 100 (best imaginable health). As the respondent fills out the 5-dimension questionnaire, he/she is asked where on the scale his/her current state of health should be positioned. In the TTO task, the respondents are asked to imagine liv-

ing in a certain health state for ten years and then to specify how many years they are willing to give up living in full health instead. Swedish experience-based value sets were published in 2014 [60] but have so far not been implemented in Swespine. Instead, British value sets derived by the TTO technique are used [61]. The Swedish value sets were found to be more accurate in terms of representation of Swedish total hip replacement patients, than the UK TTO value sets, and are used in the Swedish hip arthroplasty register since 2017 [62,63].

The EQ-5D is a relatively short questionnaire that is considered easy to complete. However, the questions may appear irrelevant to patients with a low degree of impairment [64]. Its measurement properties have been tested in populations with low back pain, although with contradictory results [65]. It is notable that the distribution of the weighted (value-set based) means is bimodal. It appears that the index systematically divides the population in two: one with a less severe health state and one with a more severe health state. This has been considered to be a difficulty in group comparisons when presenting the EQ-5D index with central tendency and dispersion in clinical practice or in trials [66]. Furthermore, the EQ-5D lacks an algorithm that handles missing data. In 2009, another version with 5 response levels was introduced by the EuroQol Group, increasing the sensitivity and reducing the ceiling effect [67].

1.3.1.1 Measurement properties and interpretability of the EQ-5D

Parameters of reliability for the EQ-5D in populations with degenerative spinal conditions are scarce in the literature [68]. In a Norwegian retest study with a 2-week interval based on 200 patients with rheumatoid arthritis, the ICC was 0.79 (0.68-0.87) [69]. In a retest study by Mannion et al. including 63 patients with chronic low back pain, the ICC was also 0.79 [64]. The reliability as measured by 95% Limits of Agreement was 0 ± 0.27 [69]. The EQ-5D has been found to be responsive in populations undergoing lumbar surgery with an AUC of 0.75-0.97 [70-72].

According to a review by Coretti et al., the MIC for the EQ-5D in LDH and LSS populations varies from 0.15 to 0.43 [65]. In a population with chronic low back pain randomized to two programs of physiotherapy the 95% CI for SDC was 0.28 and that for MIC was 0.09 [73]. In a Norwegian population operated for disc herniation, the MIC was 0.3 [72]. In another study from Norway on 172 patients with DDD, the SEM was 0.16, the SDC 0.43, and the MIC 0.17 [74]. In a Swiss population with chronic low back pain, the SEM was 0.12 and the 95% CI for SDC was 0.33 [64]. In light of the large SDCs, the interpretation of a MIC in the EQ-5D is problematic.

1.3.2 SF-36

One of the most commonly used generic tools for measurement of health-related quality of life is the Short Form-36 (SF-36). The SF-36 was constructed to survey health status in the Rand Medical Outcomes Study during the 1980s [75]. It reflects the WHO definition of health as a state of ..."physical, mental, and social well-being, and not merely the absence of disease or infirmity" [76].

The SF-36 is a multiple-item scale that assesses eight health concepts: (1) limitations in physical activities because of health problems; (2) limitations in social activities because of physical or emotional problems; (3) limitations in usual role activities because of physical health problems; (4) bodily pain: (5) general mental health (psychological distress and well-being); (6) limitations in usual role activities because of emotional problems; (7) vitality (energy and fatigue); and (8) general health perceptions.

The respondent is given 36 questions about his or her health state during the previous 4 weeks, 35 of which (the one being left out is a separately reported transition question) are put in an algorithm to compute scores of the eight subscales, which can be transformed to scales ranging from 0 (worst) to 100 (best).

The SF-36 has been found to be valid, reliable, and responsive in populations with low back pain [77]. However, a recent review reported that studies assessing measurement properties of SF-36 in low back pain populations were of low quality [68]. The subscales can be merged into a physical dimension, called the Physical Component Summary (PCS), and a mental dimension, called the Mental Component Summary (MCS). The correct calculation of the summary measures requires the use of special algorithms, which can be purchased from the private company QualityMetric. The algorithms are constructed so that the highest score on PCS is obtained when the scores on the physical scales are high at the same time as the scores on the mental scales are low. This means that if there are very low scores on the mental subscales, a high score on PCS may reflect low mental health, instead of reflecting the true existence of good health. It is therefore recommended that the composite scales be presented and interpreted together with the eight subscales [78].

The SF-36 uses a norm-based scoring algorithm where each scale is scored to have a standardized mean of 50 and a standard deviation of 10, relative to the general population norms. Swedish norms are found in the Swedish Manual and Interpretation Guide [79]. The norm-based scores vary some-

what in range; they do not go as low as 0 and never above 70. This must be considered in comparisons of different studies.

The SF-36 provides an algorithm for the handling of missing data but has no score that reflects overall health-related quality of life. The Swedish version of the SF-36 has been psychometrically tested [80-82]. A decision to end the collection of SF-36 data in Swespine was made in 2016.

1.3.3 ODI

The Oswestry Disability Questionnaire (ODI) was initiated by John O'Brien in 1976 using interviews of patients with low back pain done by the orthopaedic surgeon Stephen Eisenstein, and the occupational therapist Judith Couper and the physiotherapist Jean Davies. The objective was to identify the disturbance of activities of daily living through chronic back pain [83]. It was published in 1980 [84] and subsequently became one of the most common PROMs used in the outcome assessment of lumbar spine surgery.

The Swedish version (ODI version 2.1a) used in Swespine is the one recommended for general use [85,86]. It consists of ten items that assess the difficulty in carrying out various activities of daily life (personal care, lifting, walking, sitting, standing, sleeping, sex life, social life, and travelling) in light of the patient's back pain. The questions are answered according to the patient's functional status "today". Each item is scored from 0 to 5. Higher values represent greater disability. The total score is divided by 50 (total possible score) and then multiplied by 100 to express the score as a percentage. If one or two sections are missed, the score may be summarized as follows: (total score/(5 × number of questions answered)) × 100% [85].

The ODI has been validated, modified, and improved - and also adapted to other cultures [85]. Its psychometric properties have been tested with modern techniques [87] supporting the use of the single summary score. However, there are concerns about large floor effects and small ceiling effects, and true unidimensionality (i.e. whether it appears to measure solely the one dimension of back disability) [88,89]. Gabel et al. concluded that the overwhelming influence of pain on the response options and the fact that estimates determining responsiveness and error are at approximately the same levels as those for numeric rating scales for back pain, suggests that the same response may be obtained by using no more than a single question [89].

1.3.3.1 Measurement properties and interpretability of the ODI

The ICC was 0.97 (0.94-0.98) in a retest situation with a time interval of 2-14 days in a population selected for lumbar surgery [90]. In another retest with 20 patients with a one-week interval, the ICC was 0.83 [91]. The ODI has been found to be responsive to change in populations undergoing lumbar surgery, yielding AUC values of 0.85-0.94 [70,92,71,72]

In studies on spine populations with chronic low back pain and/or sciatica, recruited as surgical candidates or treated with decompression and fusion, the SEM was within the range of 3.54 to 4.62 points. The SDC was 8.2-12.8 and the MIC was 9.0-20 points [93,74,90,92,94,95,71,72].

1.3.4.VAS and NRS for back or leg pain

A 10-cm horizontal line with no marked gradation – the Visual Analogue scale – has been a common pain assessment tool and outcome measure for decades. The way of graphically rating pain was borrowed from psychology, where it was used to measure traits such as personality, depression, and sleep. The VAS pain scale was introduced into medical research by Huskisson in 1974 [96]. The left end of the line is marked "no pain" and the right end is marked "worst imaginable pain". A mark on the line placed by the patient represents the current level of pain. The distance between the left end and the mark is reported in centimetres or millimetres.

An alternative to the VAS is the Numeric Rating Scale (NRS) which has the same anchors at both ends but is marked from 0 to 10. The scales are the most frequently used tools for measuring pain intensity in low back pain [97]. A recent review concluded that there is no evidence that either of the scales is superior to the other in terms of measurement properties [97] and the minimal important change has been reported to be of equal size [98], but the VAS has more practical difficulties than the NRS [99-101]. Thus, in 2016 Swespine switched to the NRS. According to Graves et al., simple single-item measures like the VAS may be less accurate than multiple-item questionnaires when a complex trait such as pain is to be measured [102].

1.3.4.1 Measurement properties and interpretability of the VAS_{BACK} or NRS_{BACK}

In a reliability study of the Swespine register, the ICC for VAS_{BACK} was found to be 0.78 (0.66-0.87) [103]. The VAS_{BACK} and NRS_{BACK} have been found to be responsive in populations undergoing lumbar surgery, showing AUC values of 0.93 (VAS_{BACK}) [104] and 0.78-0.88 (NRS_{BACK}) [70,92,72].

The SDC and MIC for VAS_{BACK} were 15 mm and 18 mm, respectively, in the study by Hägg and colleagues [95]. Parker et al. defined the SDC and MIC for VAS_{BACK} in two populations undergoing revision lumbar surgery. The authors found that the 95% CI for SDC was between 2.2 and 3.8 cm and that for MIC was between 4.0 and 6.0 cm [104,71]. In another study by the same authors, two different transition questions were tested as criterion standards (anchors) in patients with spondylolisthesis, operated with fusion. The conclusion was that the SDC in VAS_{BACK} was 2.1-2.4 cm (depending on which anchor was used) and the MIC was 2.0 cm for both anchors [105].

The SEM for NRS_{BACK} in a population undergoing lumbar surgery was found to be 0.42; the 95% CI for SDC was 1.19 and that for MIC was 2.5[93]. In a population with chronic low back pain randomized to either one of two physiotherapy programs, the 95% CI for SDC was 4.5 and that for MIC was 2.5[73]. In other studies on lumbar surgery populations – where the SEM and SDC were not presented - the MIC for NRS_{BACK} varied within the range of 1.2 to 2.5 [72,92].

1.3.4.2 Measurement properties and interpretability of the VAS_{LEG} or NRS_{LEG}

In a reliability study of the Swespine register, the ICC for VAS_{LEG} was 0.88 (0.81-0.93) [103]. The VAS_{LEG} has been found to be responsive in populations undergoing lumbar surgery, with AUC values of 0.93 [71] for VAS_{LEG} and 0.72– .84 [70,92,72] for NRS_{LEG}.

The SDC and MIC for VAS_{LEG} were 5.0 cm, and 6.0 cm respectively, for patients undergoing surgery for recurrent spinal stenosis, according to Parker et al. [104]. For patients operated for spondylolisthesis, the SDC varied between 2.5 and 2.8 cm and the MIC was 2.2 cm [105].

The SEM in a population undergoing lumbar surgery was 0.49; for NRS_{LEG} the 95% CI for SDC was 1.58 and that for MIC was 1.5[93]. In studies not presenting distribution-based estimates, the MIC varied between 1.6 and 3.5 [92,72].

1.3.5 Global Assessment

The Global Assessment (GA) is a so-called Transition Question (TQ). A TQ assesses patients' retrospective perception of treatment effect. In 1989, Jaeschke et al. reported on the use of patient retrospective rating of change by a global TQ 49, and it is now the most commonly used method for determining whether or not a score change is important to patients [46].

The question in GA is worded as "How is your back/leg pain today as compared to before you had your back surgery?" with six response options on a Likert format scale - (0) I had no back/leg pain, (1) Completely pain-free, (2) Much better, (3) Somewhat better, (4) Unchanged, and (5) Worse - and has been used as endpoint in several studies [94,106-108]. The scale is considered to be asymmetric, as it has an uneven number of response options on either side of the "unchanged" option. However, since there is a response option that no change has occurred, not forcing the patient to label herself or himself as being better or worse, one can argue that the scale is balanced [29].

The simple TQs have a high face validity [29]. Although the wordings and number of response options vary between TQs, the ability to differentiate between improved and unchanged patients does not appear to be significantly affected [109]. However, the TQ should have an adequate correlation to the outcome measure under validation [54,48]. Recall bias, for example because of influence of the current health state and also the risk of not covering all important aspects of the trait to be measured, has called the validity of the TQs into question [27-30].

1.3.6 Satisfaction

The question regarding Satisfaction is worded as "How would you describe your satisfaction with the surgical outcome?", with the response options (1) Satisfied, (2) Uncertain, and (3) Dissatisfied. As a question about content, the Satisfaction is regarded as a patient-reported experience measure (PREM) [22]. This kind of outcome measure, which is focused on patient evaluation of the hospital visit as a whole, especially the patient-provider interactions, has attracted a growing amount of attention. Communication with nurses, pain management, and timeliness of assistance have shown the highest degree of correlation with overall satisfaction; communication with doctors ranked fifth. Timeliness and the existence of a clear relation to the intervention of interest appear to be important factors in the explanation of inconsistent results in studies concerning PREMs. There is no common approach to defining satisfaction [110]. It is unclear whether the Satisfaction in Swespine should be considered a true PREM, since it specifically asks about the attitude to the surgical outcome and not to the hospital visit as a whole.

1. 4 Timing of follow-up with PROMs

A reasonably sufficient number of follow-ups to capture the main results of the surgery is important for the internal and external validity of a register, as are also the response rates at follow-up. Costs of distribution of follow-up questionnaires and data management, and also unwillingness of patients to respond at follow-up, are reasons to keep the number of follow-ups low. A follow-up period of at least one year is recommended, and several spine registers also collect outcome data at 2 years, and a few at 5 and 10 years after intervention [56]. The results, if measured with PROMs, appear to stabilize between 1 and 2 years [111,112], calling into question the need for a follow-up at both 1 and 2 years.

1.5 Missing data

Although recommended in guidelines [11,10], the reporting of missing data, management of missing data, and the possible impact that missing data might have on the outcome, are rarely reported in spine register research [9]. Statistically demanding models and also the complexity of the mechanisms behind missingness probably intimidate many clinically active researchers and the reporting of response rates has to suffice [12].

Data that were planned to be collected in a register, but never were, deserve attention - as the consequence might be that the internal validity (i.e. the robustness of the conclusions) or the external validity (i.e generalizability) is affected. An unwanted scenario in connection with a national quality register might be that routines or guidelines are implemented on false grounds [113].

1.5.1 Mechanisms of missing data

Two statisticians - Donald Rubin and Roderick Little - have had a particularly profound influence on missing data management. Although statistical models on handling of missing data in RCTs were described as early as in the 1930s [114], it was not until the work of Rubin and Little in 1987 [115] that this topic gained an obvious role in the broad scientific arena [113]. Rubin, who was aiming for a degree in psychology, ended up studying statistics since the Head of the psychology department found his undergraduate education to be scientifically deficient in statistics.

Rubin's and Little's classification system for missing data is the foundation for many of the missing data handling techniques. They defined missing data according to the statistical properties of the data: missing data are either missing completely at random (MCAR), conditionally at ran-

dom (MAR), or not at random (MNAR) [116]. Despite the acceptance and widespread use of these concepts, confusion easily arises around them. To facilitate the assessment of missing data, McNight and colleagues suggested an expansion of the system, as shown in Table 1.

1.5.2 Dimensions where data might be missing

In Swespine, sociodemographic data, transition questions, and multiple-item outcome questionnaires are collected on up to six occasions for each patient. Hence, data can be missing in a variety of different ways. Firstly, one or several responses can be left out in a PROM questionnaire, indicating missingness at the item level. Secondly, when the entire PROM or a single-item variable is missing, the variable level is affected. Thirdly, when all data are missing for a participant or for a subgroup on one or several occasions, the missingness is at the individual level and/or the occasion level.

1.5.3 Reasons for data being missing

The causes of missing data may be related to the characteristics of the study participants, or to the register design, or a combination of both. For instance, if data cannot be collected because of attitudes to sharing personal information, participant characteristics is the underlying cause. If the questionnaires are left unanswered because they are too time consuming to fill in, the cause is linked to the design of the register.

Many studies covering various populations have found that differences in gender, age, personality, economic and educational prerequisites, and way of living are common between people who accept participation in

Table 1. Merging of classification systems for missing data

	MCAR	MAR	MNAR
Variable (item)	Subjects randomly omit responses	Subjects omit responses that are traceable to other responses	Subject fails to respond to incriminating items
Individuals/ Subjects	Subject data missing at random	Subject data missing but related to available demographic data	Subject data missing and relate to unmeasured demographic data
Occasions	Subjects randomly fail to show up to data collection session	Subjects who perform poorly at previous session fail to show for subsequent session	Subjects who are doing poorly at the time of the session fail to show

MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random. Reprinted from "Missing Data, a Gentle Introduction" by McKnight P, McKnight K, Sidani S, and Aurelio JF, 2007. Editor: Kenny A. With permission from Guildford Press.

surveys and trials and those who do not. This difference remains when it comes to the probability of dropping out from study participation (i.e. attrition) [117-122].

An unwillingness to respond (because of lack of time, interest, or satisfaction with the results) or an inability to respond (due to linguistic difficulties or declining health status, or death) further illustrate that the drop-out group is heterogeneous [120].

Non-response due to refusal or failure to contact has been associated with sociodemographic and socio-economic factors, whereas non-response due to morbidity and mortality has been predicted by health-related variables [119,123].

A study of the Swedish fracture register found that the most common reasons for non-response were not having received the questionnaire, lack of time, lack of interest, and inability to respond because of illness. Only 1% of non-respondents reported that dissatisfaction was a reason for non-response [124]. Solberg et al. concluded that the most common reason for loss to follow-up in the Norwegian spine register was forgetfulness [125], while a study based on the Danish spine register found that lack of time, survey tiredness, and lack of improvement were the most common reasons. As in the fracture register study, only a few individuals reported that dissatisfaction with outcome was a reason for non-response.

Norqvist et al. [126] reported that there was a significant difference in physical function between study participants who were requested to assess their shoulder function via telephone as opposed to mail, highlighting the influence of study design. The authors also reported that study drop-outs had a considerably worse shoulder function than respondents, indicating that adverse events or treatment failure are factors that require more attention in registers; this was also the conclusion in a study on hip arthroplasty survival [127].

1.5.4 Example of missing data in Swespine

A simple analysis of data from Swespine on patients who were operated on for degenerative lumbar conditions during 2006, and who were monitored 1, 2, 5, and 10 years postoperatively with ODI, is shown in Figure 4. The largest loss to follow-up occurred on the first follow-up occasion (26.5%), and after that there was a drop in the response rate of about 10% at each follow-up. At the 10-year follow-up, nearly 57% of the operated patients were lost.

Depending on the research question, one has to decide on what level missing data might be important: on the individual level, or on the occasion level. If, for instance, the focus is on the outcome at follow-up, the missing data problem is at the individual level and one must try to find out whether the missing data are related to any covariates. If the focus, for example, is on differences in outcome between the first and second follow-up occasions and a subgroup of individuals have had a second surgery during this time and therefore fail to respond, there may be a missing data problem at the occasion level.

Figure 5 depicts that the ODI might have had a missing data issue at the item level, because item number 8 stands out as missing to a higher degree than the other items on all follow-up occasions. This item asks about sexual function in relation to back pain. If this question is of importance for the research question, then it must be dealt with. But if the ODI index algorithm - which allows for a loss of two items - is used, this problem may be disregarded.

1.5.5 Missing data handling techniques

There are numerous techniques for handling missing data. McKnight et al. offer non-statisticians an introduction to the topic and the following paragraphs are a very short summary [113]. In statistical software packages, listwise deletion or pairwise deletion is often the default procedure for handling missing data. The former method means that any observation (e.g. patient) with at least one missing value is excluded from the analysis, and the latter excludes data at the variable level. When pairwise deletion is used, the patients providing data for one variable could be different from those providing data for the other.

As indicated by the name, data augmentation procedures augment a dataset with extra information provided by an assumed underlying distribution or probability model. Maximum Likelihood and Markov Chain Monte Carlo are examples of augmentation procedures.

In single-imputation procedures, missing values are replaced with a single value, for instance, a value on the second follow-up occasion substitutes for a missing value on the third occasion. Multiple imputation produces multiple estimates for each parameter, which are combined to obtain the single best estimate for the parameter of interest. Multiple imputation has the ability to estimate what impact the missing data have on results and conclusions drawn from the results.

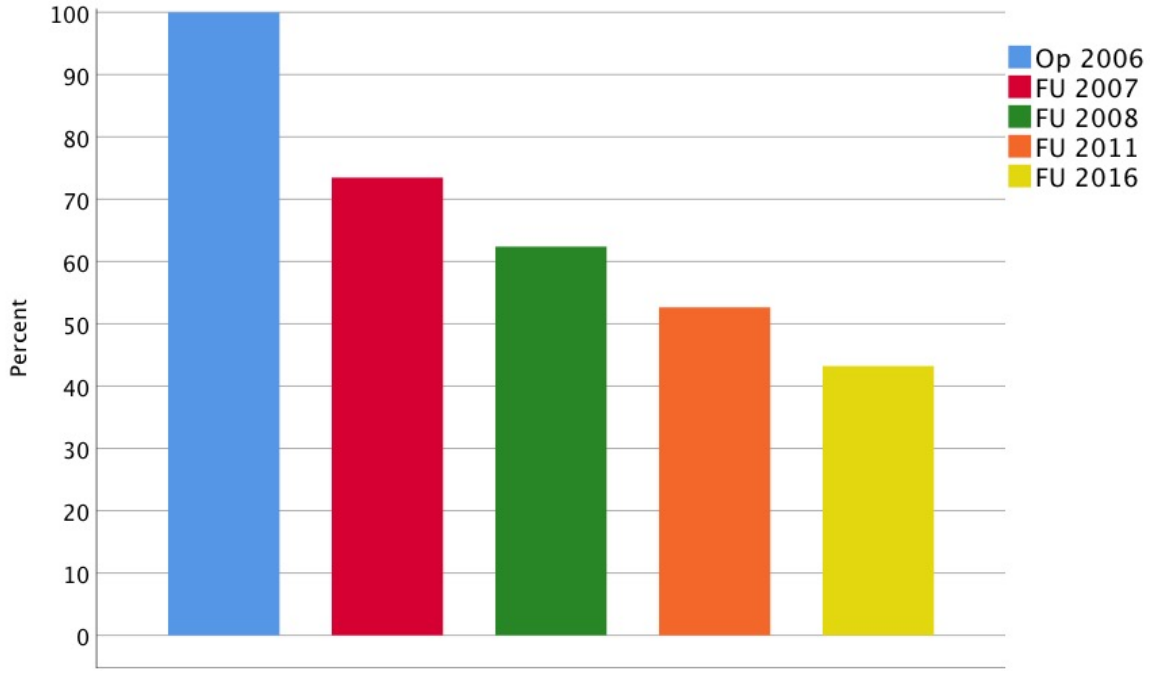


Figure 4.Percentage of patients registered in Swespine who were operated for degenerative lumbar spine conditions in 2006 and who returned the follow-up booklet at follow-up 1, 2, 5, and 10 years postoperatively. 100% = 4,732 cases.

Most missing data handling techniques assume that data are MCAR or MAR. It is therefore important to get as much information as possible about the missing data, in order to estimate whether a missing data handling procedure should be used or whether the risk of introducing bias is too high. Multiple imputation was used in a recently published paper based on Swespine data [128].

1.6 Degenerative conditions in the lumbar spine

The lower back is "a weak spot" in humans. The lifetime prevalence of low back pain is up to 80% [129]. Experimental studies have shown that low back pain can be initiated by noxious stimulation of selected structures such as muscles, facet joints, interspinous ligaments, the dura mater, and the posterior surface of the disc [130]. The stimulation of the same structures may also cause a somatic referred pain in the legs, without neurological signs or dermatomal pattern. It is described as dull, aching, and gnawing - in contrast to the radicular pain, which is sharp and radiates along the leg with a width of no more than 5-10 cm [130].

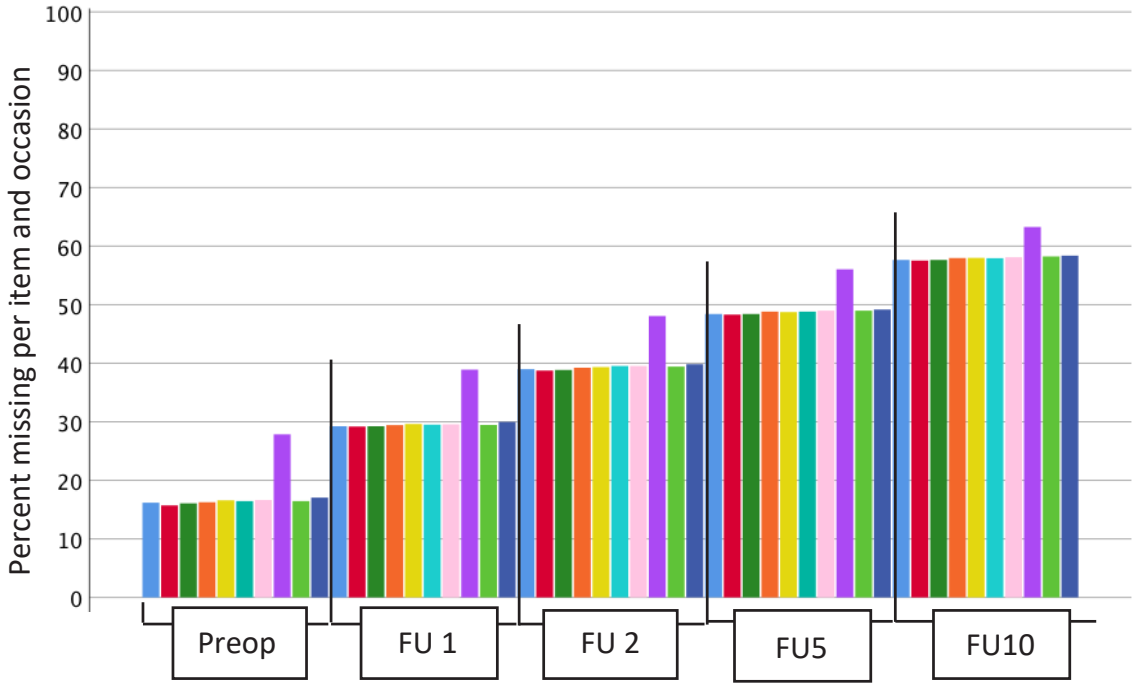


Figure 5.The 10-item Oswestry Disability Questionnaire at the time of surgery, and at follow-up (FU) 1, 2, 5, and 10 years postoperatively. Each coloured bar represents one item. The largest drop was seen between baseline and FU1. Thereafter, response rates dropped by approximately 10% on each occasion. Note that item number 8 was left out by the respondents to a considerably greater extent than the other items. N_{PREOP} varied between 3,414 and 3,966.

Degeneration of the spine has a complex aetiology involving an age-related process influenced by mechanical and genetic factors [131]. Degenerative changes, defined as the presence of intravertebral disc changes and, at a later stage, disc space narrowing, osteophytes, and sclerosis [132], do not have an obvious causal relationship to low back pain, although certain changes are more common in low back pain populations [133]. The most common degenerative conditions in the lumbar spine that may lead to a surgical intervention are described below.

1.6.1 Lumbar disc herniation

A disc herniation occurs when a portion of the nucleus pulposus is pushed out through a tear in the annulus fibrosus. When the radicular leg pain is resistant to conservative treatment, a surgical removal of the hernia with a conventional or microscopic technique is offered. A minority of the patients are treated with a supplementary fusion.

1.6.2 Lumbar spinal stenosis

Spinal stenosis is caused by narrowing of the spinal canal due to disc degeneration, facet joint arthritis,

and thickening of the ligamentum flavum. The location of the stenosis is described by the terms "central stenosis", "lateral/recess stenosis", and "foraminal stenosis". Degenerative slips usually occur at one of the lower levels of the lumbar spine as a result of long-standing disc and facet degeneration [134]. When conservative treatment has failed, there may be an indication for surgery, which includes central, lateral, or foraminal decompression with or without fusion [135]. Spinal stenosis is the most common reason for surgical intervention in the lumbar spine. The aim of the surgery is to relieve pain in the buttocks and the legs.

1.6.3 Degenerative changes and chronic low back pain

Low back pain without radicular pain may be caused by rare entities such as tumours, fractures, or infections. However, in most cases the pain usually decreases over time, or recurs intermittently, and the underlying cause is never found [136]. A subgroup of carefully selected patients – labelled degenerative disc disorder (DDD) - with MRI-verified degenerative changes in one or several of the lower levels of the lumbar spine, and severe chronic low back pain that is resistant to long-term conservative treatment, may be subjected to surgical intervention [137-139]. It is debated whether or not the so-called high-intensity zones, and Modic changes type 1 and 2, seen on MRI are associated with an increased rate of low back pain [140]. In a recent study, Modic changes were not found to be associated with long-term pain, disability, or sick leave [141].

Surgical options include posterolateral fusion, interbody fusion, and total disc replacement with the ultimate goal of reducing low back pain[139].

Spondylolysis is defined as a disruption of a vertebral structure; and spondylolisthesis refers to any forward slipping of one vertebra onto the one below it [134]. The most common form of spondylolisthesis is isthmic spondylolisthesis, which includes a fibrous defect in the pars interarticularis. Surgical options are in situ posterolateral or interbody fusion with or without decompression.

1.6.4 Reporting of Swespine data

Each year, the Swespine register presents a summary of the collected data in the Annual Register Report. As an example, Figure 6 shows that spinal stenosis is by far the most common reason for surgery. In 2019 a total of 2,554 patients were operated for lumbar disc herniation, 5,512 for spinal stenosis, 334 for isthmic spondylolysis or spondylolisthesis, and 711 for DDD [20]. As shown in Figure 7, there is a difference in PROM score between diagnostic groups.

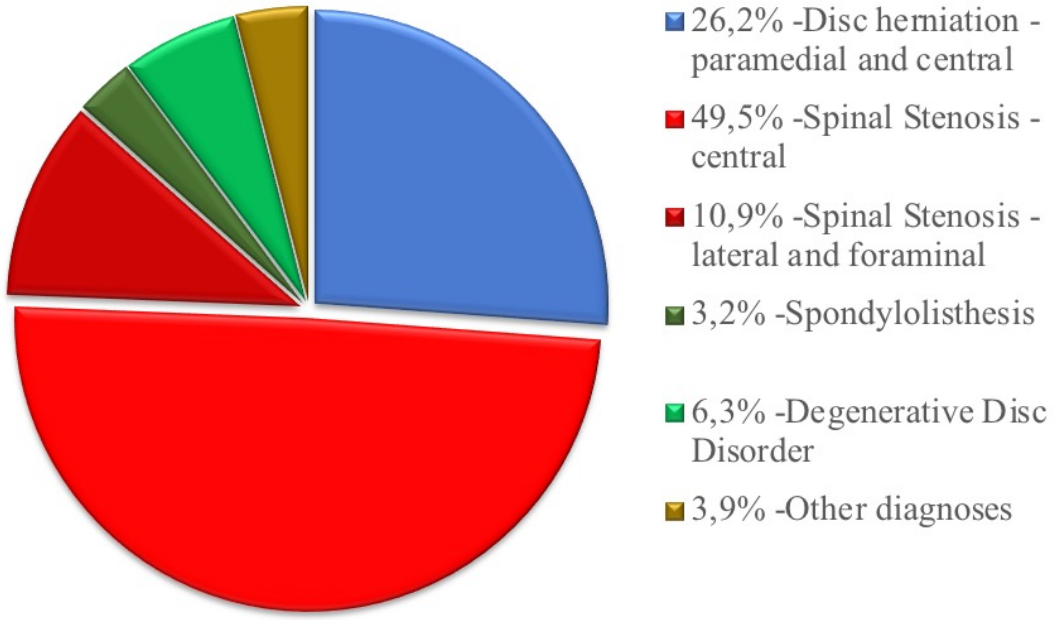


Figure 6. Breakdown of treated conditions of the lumbar spine registered in Swespine during 2017 (9,484 patients). Reproduced from the Swespine Annual Report (2017) with permission from the Board of the Swedish Society of Spinal Surgeons.

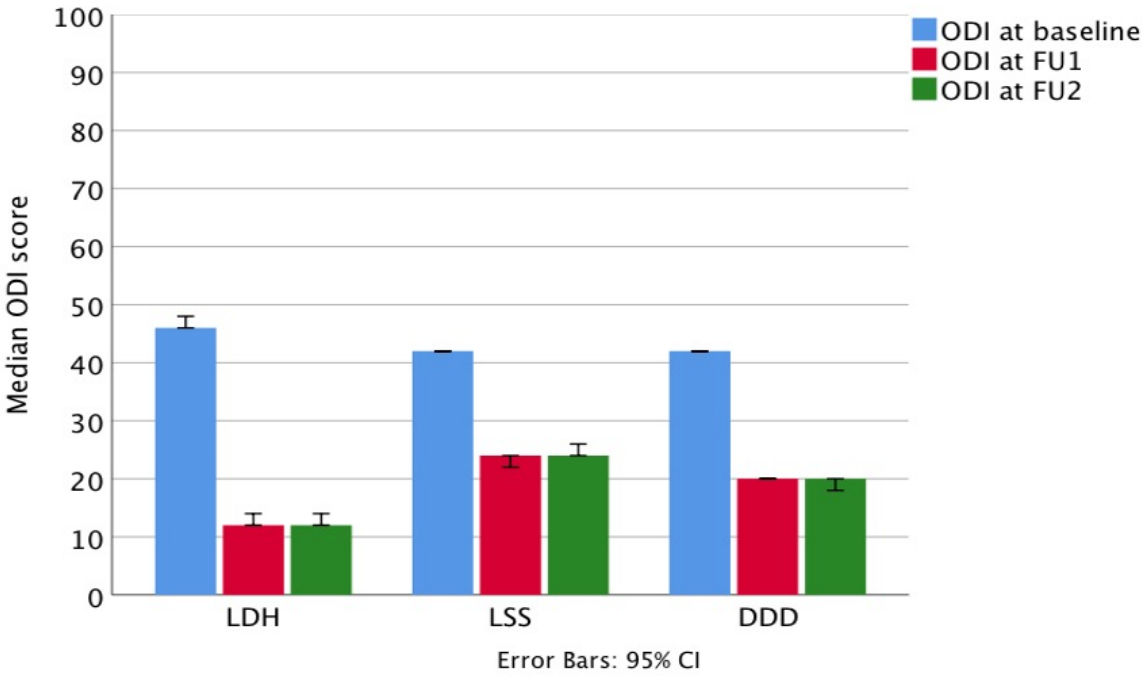


Figure 7. ODI at baseline and at one and two years after lumbar spine surgery, in relation to diagnosis, for patients registered in Swespine during the period 1998-2015. LDH, lumbar disc herniation; LSS, lumbar spinal stenosis; DDD, degenerative disc disorder. Only patients who responded on all three occasions are presented. LDH: n = 9,844; LSS: n = 17,325; DDD: n = 5,339

Painted by Ingrid Johansson, a dear friend suffering from chronic back and leg pain arising from a spondylodiscitis.

AIM

2. AIM

The overall aim of the work included in this thesis was to find ways to simplify the assessment of patient-reported outcome without any loss in scientific credibility. Specifically, the thesis aimed to find answers to the following questions:

1. Would it be possible to use only one retrospective question in the assessment of effectiveness in spine surgery? (Study I).

2. Could one year of follow-up be sufficient – as opposed to two years – when measuring the results of spine surgery if PROMs are used as outcome tools? (Study II).

3. What is the smallest clinically relevant change in outcome of each PROM? Does the size of that change vary between populations with different degenerative spinal disorders? Is the PROM sensitive enough to detect a clinically important change? (Study III).

4. Are PROM data affected by loss to follow-up? If so, how? (Study IV).

PATIENTS AND METHODS

3. PATIENTS AND METHODS

3.1 Ethical approval

Patients are informed in writing about their voluntary participation in Swespine, with an opt-out procedure. To use Swespine data in research, approval from the owners of the register (the Society of Swedish Spinal Surgeons) and also the Ethics Committee is required. The current studies were approved by the Ethics Committees in Gothenburg and Stockholm (Dnr 1039-15 and Dnr 2014/5:1, respectively). In study III, the retest participants gave their written consent.

3.2 Patient recruitment

In all studies, patient data were retrieved from the Swespine database. In addition, in study III participants in the retest study were recruited from GHP spine centres in Gothenburg and Stockholm. In study IV, Swespine data were linked to data from Statistics Sweden, the Social Insurance Authority, patient administrative systems, and the National Patient Register.

3.3 Inclusion criteria and exclusion criteria

3.3.1 Studies I, II, and III

Patients operated for degenerative lumbar spine disorders who were registered in Swespine and who were diagnosed with disc herniation, spinal stenosis, chronic low back pain, spondylolysis, or spondylolisthesis were all eligible. There were no limitations regarding age, comorbidity, previous history of back surgery, or surgical method.

The patients were divided into three groups: those operated for disc herniation (LDH), those operated for spinal stenosis (LSS), and those operated for degenerative disc disorder (DDD). The stenosis group included central as well as lateral and foraminal stenosis. The third group consisted of patients diagnosed with chronic low back pain, spondylolisthesis, and/or spondylolysis. The rationale for merging different diagnoses in the third group was low back pain as an indication for surgery, and similar levels in patient-reported

outcomes, as shown in the Swespine Annual Register Reports [20].

The numbers of patients included in studies I-III are shown in the flow chart in Figure 8. Between 1997 and 2015, 104,661 patients who underwent degenerative lumbar spine surgery were registered in Swepine. In study I, a dataset with registrations from 1997 to 2016 was used, but it was not noted that patients operated in 2016 were naturally excluded from analyses because they had not yet received the follow-up questionnaires. For study II, an updated set of data with registrations up to mid 2017 was used, where this particular lack of data was not accounted for. Also, the four (!) registrations from 1997 were excluded. For studies II and III, paired data were used, resulting in the exclusion of cases without complete responses to all PROMs at baseline and also at follow-up.

Study III also contained a retest study, where participants were obtained from GHP Stockholm Spine Center and GHP Spine Center Göteborg. Efforts were made to mimic a real-life setting by incorporating the retest study into the ordinary Swespine logistics, while at the same time keeping the retest situation as consistent as possible. In order to get a representative sample, participants from each of the diagnoses LDH, LSS, and DDD were included. The diagnosis itself was not, however, believed to play a decisive role in the potential variation in PROM score, since distribution-based methods - which are used to determine measurement error and repeatability - are assumed to be fairly sample-independent [142]. The calculations for repeatability and reliability were therefore performed on the whole study population without stratification regarding diagnosis.

The Swespine booklet for patient-reported data was used. To cover as much of the range of each PROM scale as possible, patients were included at two different stages in relation to the operation. Thus, one group was enrolled from the preoperative waiting list (the pre-op group), and the other was enrolled at the one-year follow-up (the post-op group). The participants in the former group filled out the first booklet (T1) at the clinic on the day they were listed for surgery. The second booklet (T2) was sent by post one week later, with a request to return the questionnaire within 5 days. One reminder was sent. The latter group was asked to participate at the time of the Swespine one-year follow-up (T1). One week after the booklet was registered at the Swespine office, the second booklet (T2) was sent out.

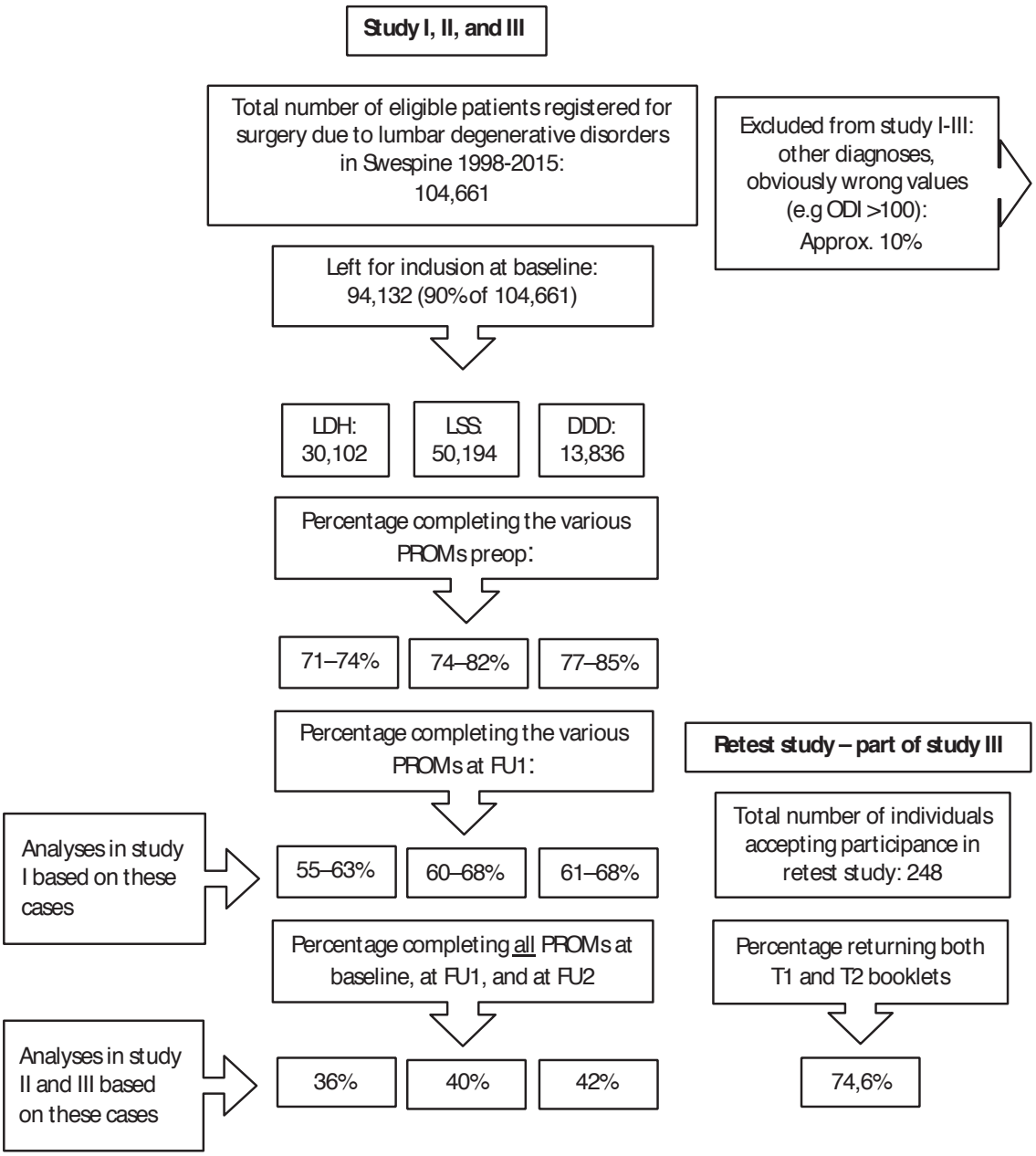


Figure 8. Flow chart of studies I-III.

One major concern was avoidance of the possibility that a true change in the attribute being measured occurred during the test period. Since the patients in the first group were on the waiting list for surgery, they were considered to be symptomatically stable regarding the spine condition for which they were awaiting treatment. The same assumption was made for the second group, regarding having a stable condition between the two measurements, because a full year had passed since the operation.

The number of days between the test and the retest should neither be too small (to avoid carry-over effects) nor be too large (to reduce fluctuations in symptoms) [143]. The optimal timing of measurements is not known, although a time interval of 1-2 weeks has been recommended [143].

Inclusion stopped when the total number of participants exceeded 30 in each of the three diagnosis groups. Based on an estimated ICC of 0.80, a 95% CI of ±0.10, and two measurement replications, a sample of at least 50 has been suggested by de Vet et al, chapter 5 [37]. A greater precision, for example a 95% CI of ±0.05 for an ICC of 0.80, which has been recommended by Polit [143], would require a sample size of 200. A sample size closer to 200 was therefore aimed for. It was rather time consuming to acquire study participants for the LDH group, as they were frequently operated before the time of T2.

3.3.2 Study IV

A flow chart of study IV is given in Figure 9. The data were taken from a database created in the project "Swedish national collaboration for value-based reimbursement and monitoring of healthcare" [144]. In this database, Swespine data from the period 2008–2012 were linked at the patient level with data from seven patient-administrative systems (covering 65% of the Swedish population), Statistics Sweden, the National Patient Register, and the Social Insurance Authority. All individuals who fitted into one of the diagnosis groups (LDH, LSS, or DDD) and who answered the baseline questionnaire were eligible. However, this time the inclusion criteria were more stringent. Cases diagnosed as spondylolysis or spondylolisthesis were excluded, as were patients with pseudarthrosis as the main diagnosis. Also, the subgroups were defined by a combination of diagnosis code and procedure code. In total, 21,961 patients were included. Non-respondents were defined as those who did not return the Swespine FU1 booklet.

3.4 Outcome variables

The same outcome variables were used throughout the four papers, with a few disparities (Table 2). In study I, the EQ-5D and ODI were analyzed at the item or dimension level, and at the algorithm-based indices level. In study II, Satisfaction was added. In study III, the first question of the Short Form-36 questionnaire ($SF\text{-}36_{GH}$) was added to reveal changes in global health during the retest period. The question is worded: "In general, would you say

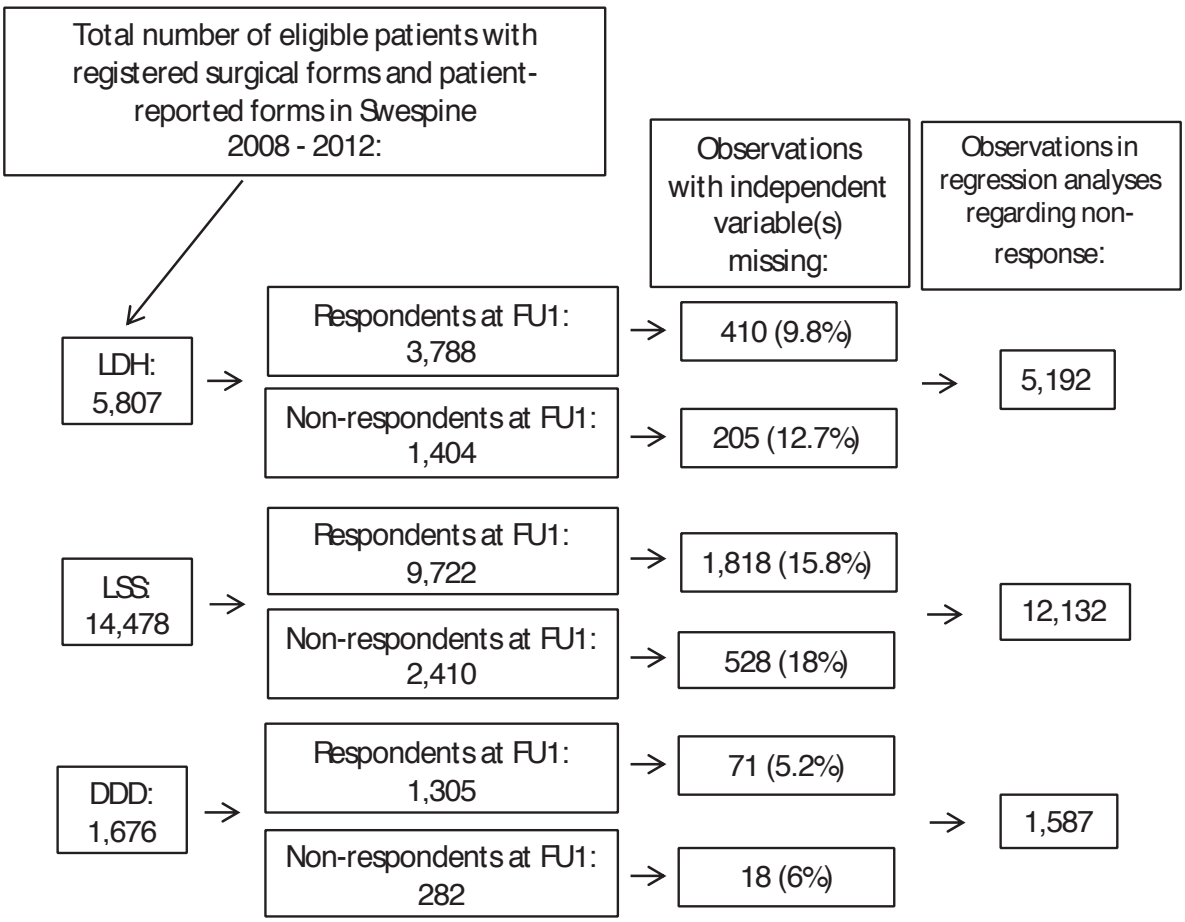


Figure 9. Flow chart of study IV.

Table 2. Outcome measures

	Studies			
Outcome Measure	**I**	**II**	**III**	**IV**
EQ-5D$_{INDEX}$	x	x	x	x
EQ-5D (items)	x			
SF-36 (domains)	x			
SF-36$_{GH}$			x	
ODI	x	x	x	x
ODI (items)	x			
VAS$_{BACK/LEG}$	x	x	x	x
NRS$_{BACK/LEG}$			x	
GA$_{BACK/LEG}$	x	x	x	x
Satisfaction		x		

EQ-5D, EuroQol 5 dimensions; SF-36, Short Form-36; $SF\text{-}36_{GH}$, Short Form-36 global health single item; ODI, Oswestry Disability Index; $VAS_{BACK/LEG}$, Visual Analogue Scale for back and leg pain, respectively; $GA_{BACK/LEG}$: Global Assessment for back and leg pain, respectively

your health is" with response options: Excellent/VeryGood/Good/Fair/Poor. Furthermore, in the same retest study NRS was used instead of VAS because the latter had replaced the former in the Swespine booklet from 2016 onwards. As the MIC calculations were done on VAS scores, these estimates were adapted to the NRS scale by simply adding a comma.

3.5 Statistical methods

3.5.1 Spearman rank correlations (Study I)

A "correlation" can be defined as "a measure of the strength of the linear relationship between two random variables" [145]. But as the concept of linearity assumes that the data have interval-scale properties, which the GA does not have, it is better described as a monotonically increasing or decreasing bivariate pattern, or a correlation of the ranks. Hence the use of the non-parametric Spearman's correlation. Confidence intervals were calculated using the method for Pearson´s correlation as the distribution of the correlations for larges samples are similar [146].

3.5.2 McNemar's test (Study II)

The McNemar test is a non-parametric test for paired nominal data and is used when analyzing changes in proportions for the paired data [146].

3.5.3 Receiver Operating Characteristic curve (ROC) analysis (Studies I, III, and IV)

Deyo and Centor proposed scales to be viewed as diagnostic tests and that they could be evaluated using sensitivity and specificity statistics. Thus, the ability of an outcome measure to detect an improvement could be measured through a ROC curve analysis [147].

The Area Under the ROC Curve (AUC) is a way of measuring the accuracy of a diagnostic test (in this thesis, the "diagnostic test" is a PROM). The measure of interest is compared with a gold standard considered to be a perfect test, i.e. 100% sensitivity and specificity equals an AUC of 1.0. Thus, the AUC can be interpreted as the probability of accurately discriminating between a successful outcome and an unsuccessful outcome. Sensitivity is interpreted as the ability of the cut-point to define all patients who assessed themselves to be "improved" on the global perceived effect scale of choice. Specificity refers to the ability of the cut-point to exclude patients who did not assess themselves as being improved on the external criterion [92].

The external criterion (anchor) that is chosen as gold standard is usually a transition question of some kind, for example the health transition question of the SF-36, the NASS Satisfaction questionnaire, the Satisfaction with results scale, the global perceived effect scale, or the Global Assessment [93,70,73,94].

3.5.4 MIC, Minimal Important Change (Studies II and III)

The ROC curve provides a visual illustration of the location of the optimal cut-point (MIC) in a PROM score for improved and unimproved patients according to the definition set by the transition question that was chosen as external criterion. The choice of transition question should be based on the degree of correlation to the outcome measure being validated. Also, a critical point lies in the anchor´s definition of important change [92,54]. The MIC varies by context, population, the choice of reference criterion, the method, and according to the strength of the relationship between the PROM and the anchor [37].

3.5.5 SDC, Smallest Detectable Change (Study III)

When a PROM is used repeatedly on the same patient, there will be measurement error because of natural fluctuations in symptoms, variation in the measurement process, or both. The repeatability measure SDC is a useful way of presenting the measurement error in the context of score change. The SDC is described by Polit and Yang as a change in score of sufficient magnitude that the probability of it being the result of random error is low [36]. The SDC $=1.96 \times \sqrt{2} \times \mathrm{SEM}$ (Standard Error of Measurement).

3.5.6 Measurement Error (Study III)

The Standard Error of Measurement (SEM) is a standard error in an observed score that obscures the true score and is given in the units of the PROM [142].

By obtaining the standard deviation of repeated measurements on the same patient the size of the measurement error can be measured. The SEM $=\sqrt{\text{intra individual variance}}$ The difference between a subject's measurement and the true value would be expected to be within ±1.96 × SEM for 95% of the observations.

The assumption that the standard deviation is unrelated to the magnitude of the measurement (heteroscedasticity) is checked by plotting of the individual patient's standard deviations against his or her means [146].

3.5.7 ICC (Study III)

The reliability parameter used in a test-retest reliability situation is called the intra-class correlation coefficient (ICC). The ICC reflects the variation

in measurements taken by a PROM on the same patient under the same conditions [148]. The ICC is defined as $SD^2_{subject}/SD^2_{total}$. Based on the 95% CI of the ICC estimate values, less than 0.40 indicates poor reliability, while estimates of 0.4-0.59 indicate fair reliability, 0.6–0.74 good reliability, and 0.75-1.00 excellent reliability [149]. The relation between the ICC and the SEM has been described by de Vet et al., Chapter 5, as $SEM = SD\sqrt{(1 - ICC)}$[37] where the SD is the pooled standard deviation of the sample and ICC, the reliability parameter of the PROM, is calculated with an absolute agreement, two-way random-effects single-measures model [150].

3.5.8 Kappa (Study III)

Another test-retest reliability parameter, Kappa, is used to statistically describe test-retest reliability of categorical variables. In study III, kappa was calculated for $GA_{BACK/LEG}$ and $SF\text{-}36_{GH}$. A PROM can be considered reliable when the kappa is above 0.75, signifying a substantial agreement [151]. Measurements with several response options require the use of a weighting scheme. In study III, the scheme of quadratic weights - which is identical to an ICC of absolute agreement according to Polit and Yang, Chapter 8 [36] - was computed. An overall agreement between the two test occasions, T1 and T2, and the proportion of respondents who reported having a better outcome at T1 than at T2, or vice versa, were also calculated.

3.5.9 Logistic regression and ordinary least-squares regression (Study IV)

Statistical analyses were computed to recognize any statistically significant systematic differences between respondents and non-respondents (two-sided tests at the 5% significance level). χ^2 tests were used for dichotomous and ordinal variables, the Kruskal-Wallis test was used for count variables, and the t-test was used for continuous variables.

Explanatory variables were included based on clinical relevance and to assess the possible impact of any sociodemographic variables. Previous sick leave and disability pension were included as explanatory factors for the disc herniation subgroup and the DDD subgroup. Neither of these two explanatory factors were used in the regression analysis for the LSS group, as only 45% of the sample was less than 65 years of age at surgery and therefore younger than the general retirement age in Sweden at the time.

Two different sets of logistic regression analyses were performed: (1) logistic regression analyses, with non-response as outcome, showing the degree of association between baseline variables and non-response; (2) regression analyses with PROM values reported at FU1 as dependent variable. Output in the regression analyses was expressed as odds ratios (ORs), reflecting the interplay of the explanatory variables and the probability of non-response. Furthermore, the output from these regression analyses was used to predict outcome in the non-response group. The levels predicted were compared with the actual outcome in the response group for all three diagnosis groups. The same set of explanatory variables was used in all regression analyses.

The results were presented as proportions of successful $GA_{BACK/LEG}$ (logistic regression) and levels of ODI, VAS, and EQ-5D (ordinary least-squares regression), respectively.

Based on the second regression analysis, all explanatory variables included were also used to predict the chance of having a successful outcome of surgery, for the individuals who answered the follow-up questionnaire as well as those who did not, to enable comparison of predicted GA values for these two groups.

Receiver operating characteristic (ROC) values were calculated for the regression analysis of successful outcome, to evaluate the predictive ability of the models.

SUMMARY OF RESULTS

4. SUMMARY OF RESULTS

The baseline characteristics of each diagnosis group are given in Table 3.

Table 3. Baseline characteristics of patients operated for disc herniation, spinal stenosis, or degenerative disc disorder in the lumbar spine

	LDH (N = 31,314*)	LSS (N = 53,043*)	DDD (N = 14,375*)
Age mean (SD)	45 (14)	67 (11)	47 (13)
Female %	45	54	53
Smoker %	19	12	14
Previous spine surgery %	13	20	26
Unemployed %	11	10	13
Employed %	81	37	75
Back pain > 1 year, %	36	75	89
Leg pain > 1 year, %	29	69	70

* The numbers below do not always correspond to the group numbers because of missing data

4.1 Study I

In the first paper, the usefulness of the simple transition question GA as an overall PROM was explored.

The study population was tested in a non-respondent analysis; comparison of sex, age, and baseline PROM mean scores was done using Fisher's exact test and the independent-samples t-test. Although highly significant statistically, the differences in absolute numbers were minor.

GA showed the highest correlations to the VAS_{LEG} in the LDH and LSS cohorts, and to the VAS_{BACK} in the DDD cohort. Further, GA was correlated to the scores at the 1-year follow-up to a greater extent than to the score changes (Figure 10). The boxplots in Figure 11 depict patient assessments of VAS_{LEG} scores at the one-year follow-up and of VAS_{LEG} score changes (i.e. scores at FU1 minus scores at baseline) according to patient responses to GA_{LEG} in the LDH group. Figure 12 shows the same data for the LSS group, and in Figure

13 data for the DDD group can be seen. In the latter, VAS_{BACK} scores according to GA_{BACK} are shown. Two subgroups emerged in the plots of the final VAS scores, one consisting of patients who considered that there had been a considerable improvement, and the other with patients who considered that there had been little or no improvement. The subgroups were not as evident in the plots illustrating the score changes. The boxplots also highlight the large spread of VAS scores within each GA response category, including outliers 1.5 times and three times the interquartile range. Note that another set of boxplots were presented in the published version of study I, showing VAS_{BACK}, ODI, and EQ-5D for the whole study population (i.e. the LDH, LSS, and DDD groups were merged).

The pain-specific domain of SF-36 as well as the pain-specific items in the EQ-5D, and ODI, all showed a higher correlation to $GA_{BACK/LEG}$ than the remaining domains/items (Table 4).

The discriminative ability of PROMs with $GA_{BACK/LEG}$ as reference criterion, defining success and lack of success, was examined with ROC curve analyses, as shown in Figures 14-16. In all three groups and for all PROMs, the

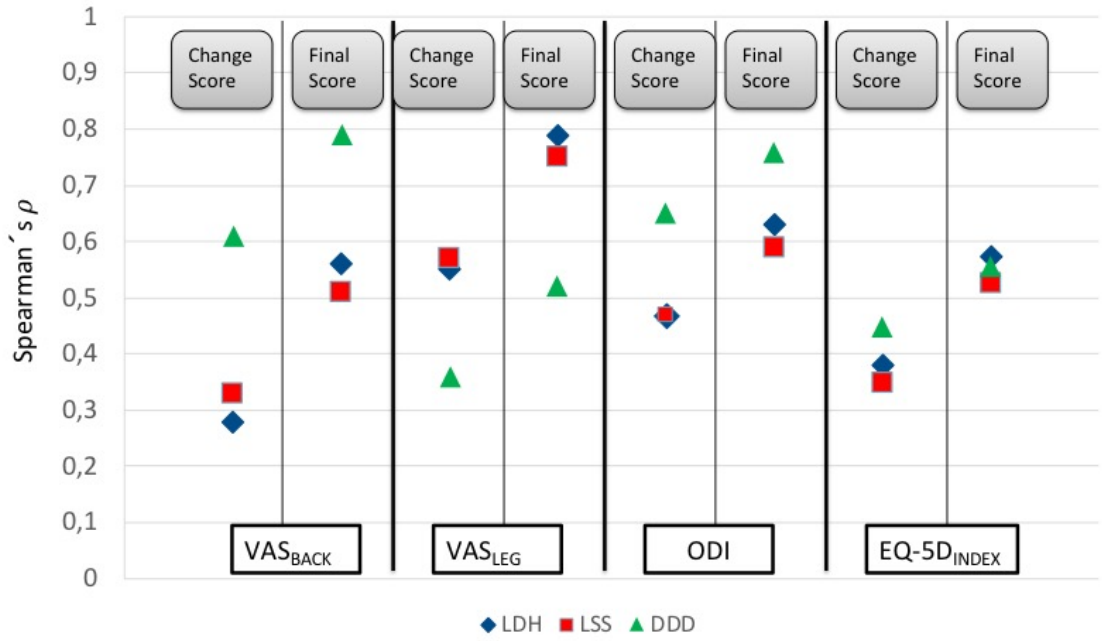


Figure 10. Spearman rank correlation analyses where GA was correlated to score changes (between baseline and FU1), and final scores (FU1) of VAS_{BACK}, VAS_{LEG}, ODI, and EQ-5D. GA_{LEG} was used in the LDH and LSS cohorts, and GA_{BACK} in the DDD cohort. Correlation coefficients were generally higher when GA was correlated to the final scores rather than to score changes. The correlations were higher with pain-specific and disease-specific PROMs (VAS and ODI) than with the generic quality of life PROM (EQ-5D).

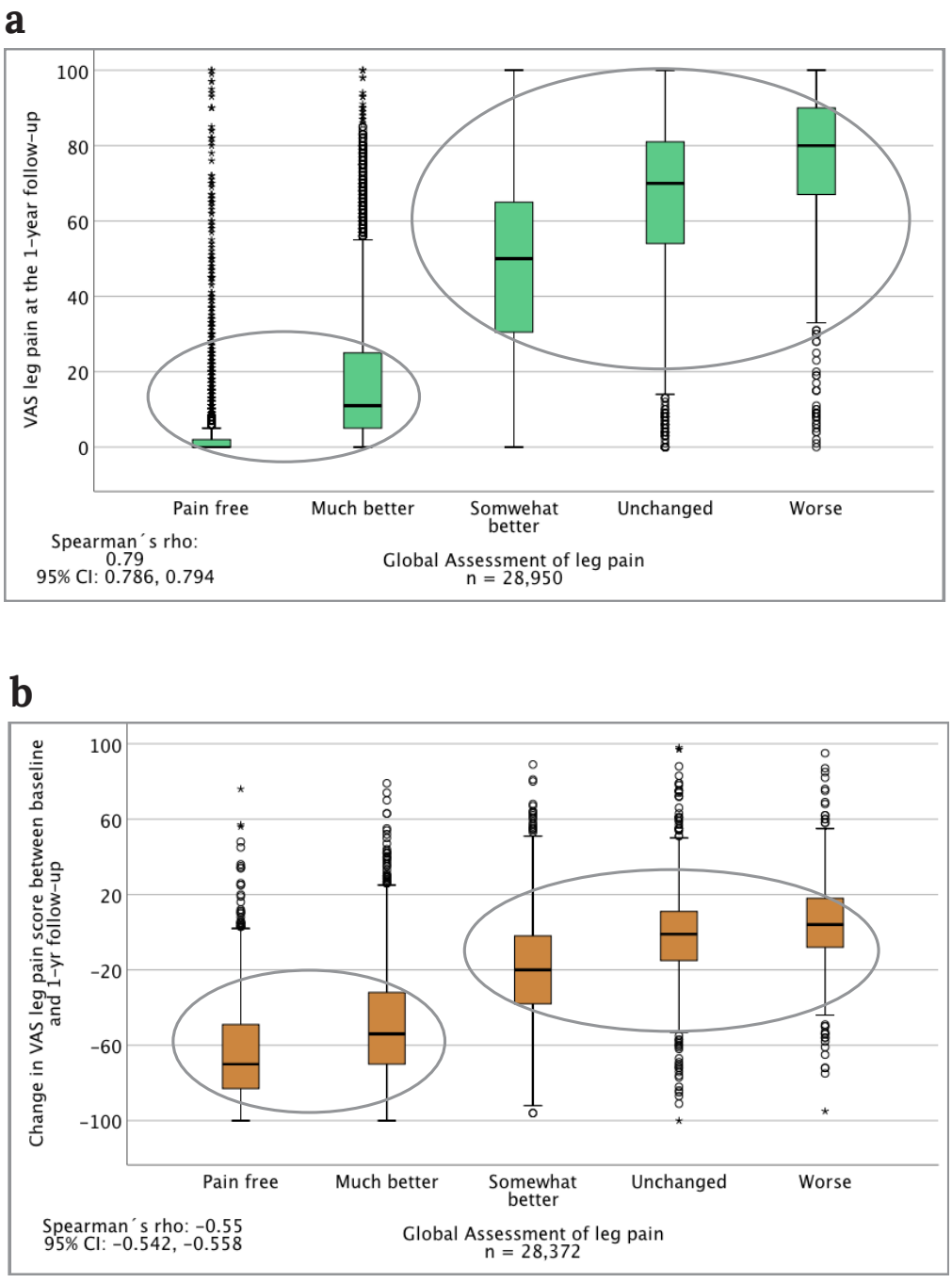


Figure 11. Distribution of absolute scores at FU1 (a), and score changes (b) in VAS_{LEG} according to GA_{LEG} in the LDH cohort. Self-assessments such as "pain-free" and "much better" were considered to be a successful outcome. This cut-off was more evident for final scores than for score changes.

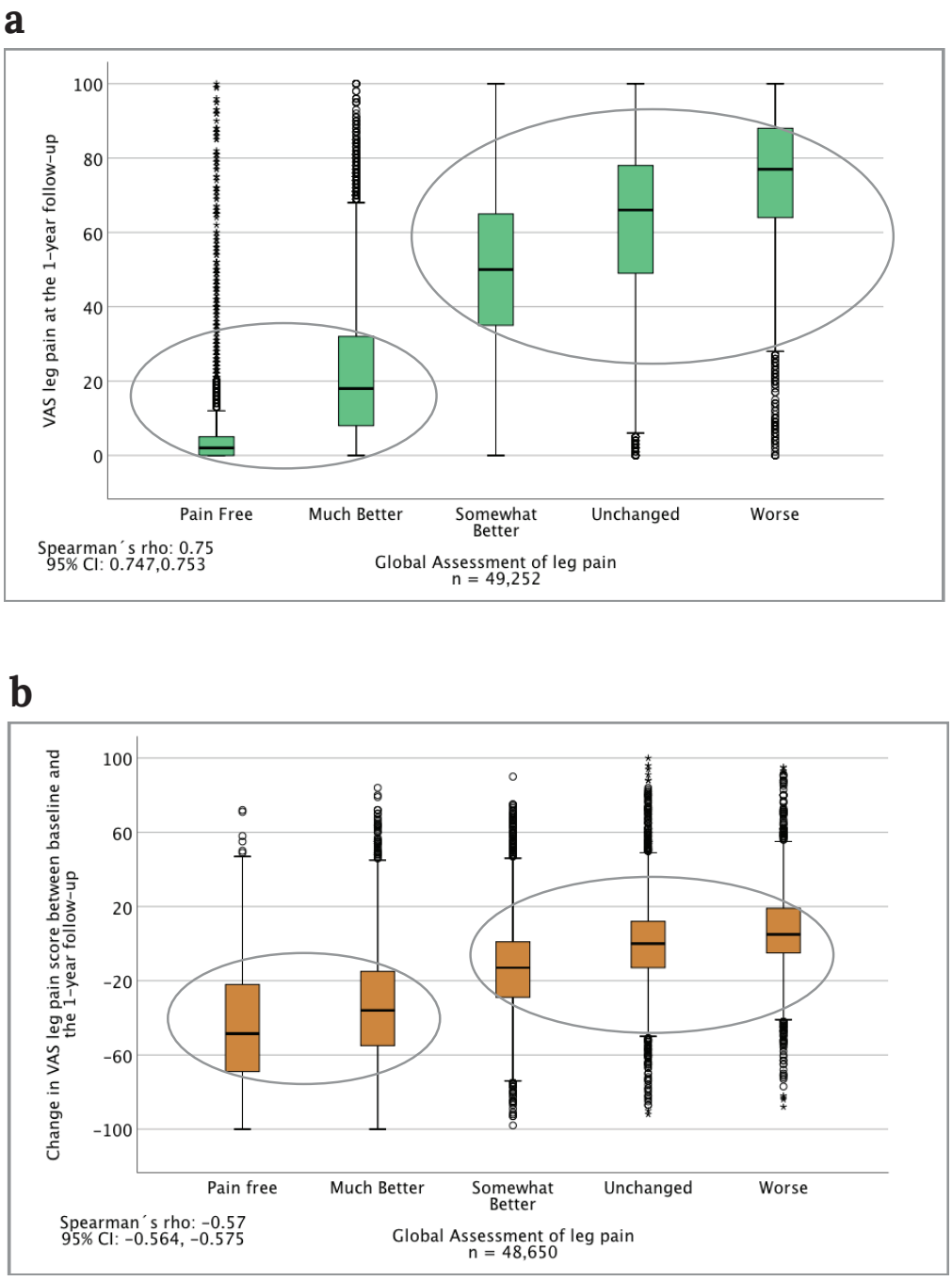
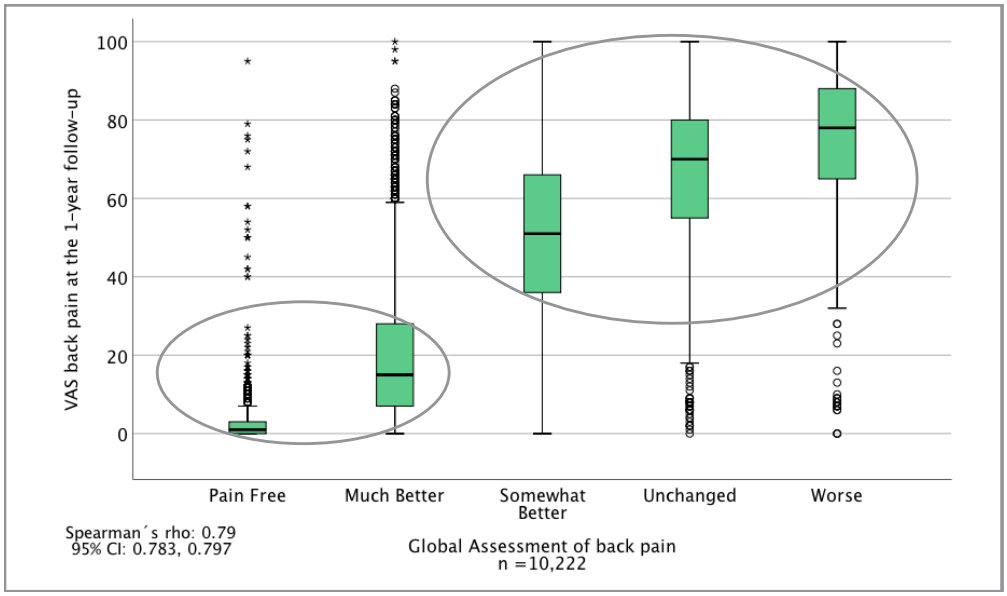


Figure 12. Distribution of absolute scores at FU1 (a), and score changes (b) in VAS_{LEG} according to GA_{LEG} in the LSS cohort. Self-assessments such as "pain-free" and "much better" were considered to be a successful outcome. This cut-off was more evident for final scores than for score changes.

a



b

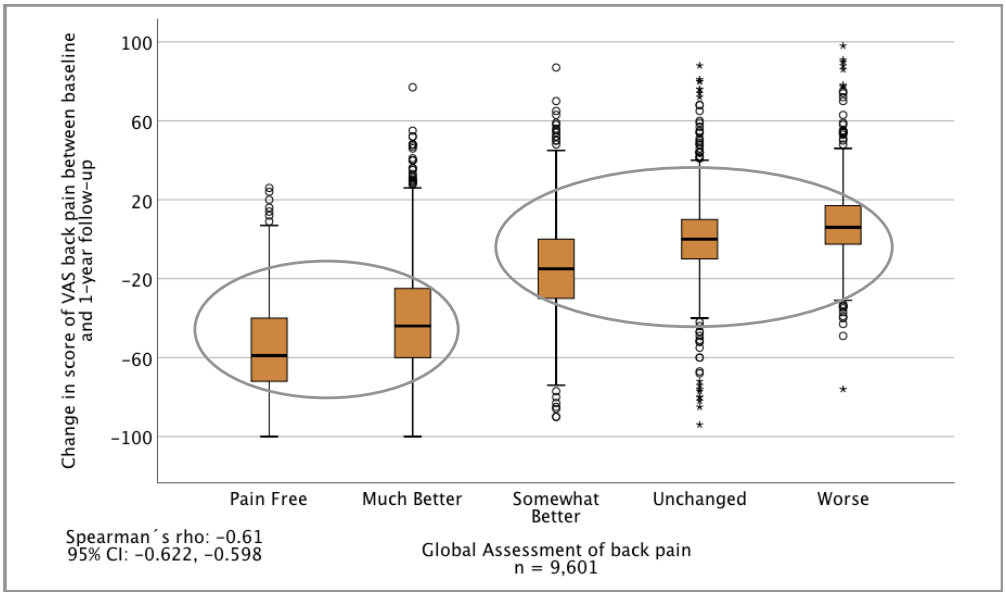


Figure 13. Distribution of absolute scores at FU1 (a), and score changes (b) in VAS_{BACK} according to GA_{BACK} in the DDD cohort. Self-assessments such as "pain-free" and "much better" were considered to be a successful outcome. This cut-off was more evident for final scores than for score changes.

Table 4. Spearman rank correlations between Global Assessment and specific items/domains in ODI, SF36, and EQ-5D at one year postoperatively

ODI vs	GA		SF-36 vs	GA		EQ-5D vs	GA	
	GA$_{BACK}$	GA$_{LEG}$		GA$_{BACK}$	GA$_{LEG}$		GA$_{BACK}$	GA$_{LEG}$
Pain intensity	**0.73**	**0.58**	PF	-0.56	-0.52	Mobility	-0.46	-0.47
Personal care	0.55	0.46	RP	-0.51	-0.46	Self-care	-0.27	-0.24
Lifting	0.52	0.43	RE	-0.40	-0.37	Ususal Activities	-0.48	-0.42
Walking	0.48	0.47	SF	-0.50	-0.45	Pain/Discomfort	**-0.61**	**-0.53**
Sitting	0.51	0.43	BP	**-0.69**	**-0.60**	Anxiety/Depression	-0.40	-0.35
Standing	0.56	0.48	MH	-0.43	-0.38			
Sleeping	0.53	0.49	VT	-0.54	-0.46			
Sex life	0.53	0.47	GH	-0.49	-0.44			
Social life	0.58	0.51						
Travelling	0.58	0.51						

The number of respondents in each item ranged from 50,212 to 58,879 in the ODI, from 63,597 to 65,823 in the SF36, and from 61,299 to 61,680 in the EQ-5D.
PF, physical functioning; RP, role limitations due to physical health; RE, role limitations due to emotional problems; SF, social functioning; BP, bodily pain; MH, general mental health; VT, vitality; GH, general health perceptions

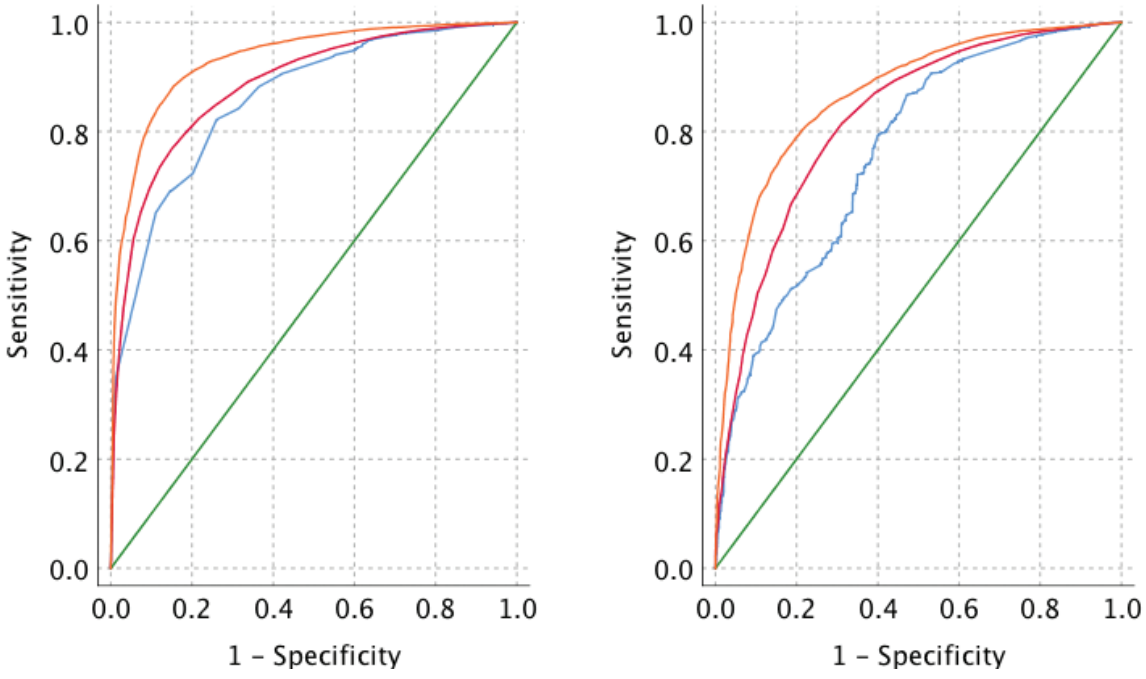


Figure 14. LDH cohort. ROC analyses based on individuals with complete responses to all PROMs at baseline and at the one-year follow-up. VAS$_{LEG}$ (AUC for final score: 0.93; AUC for score change: 0.87), ODI (AUC for final score: 0.89; AUC for score change: 0.82), and EQ-5D$_{INDEX}$ (AUC for final score: 0.86; AUC for score change: 0.86) The analysis on final scores is given on the left and the analysis on score changes is given on the right (n = 10,855).

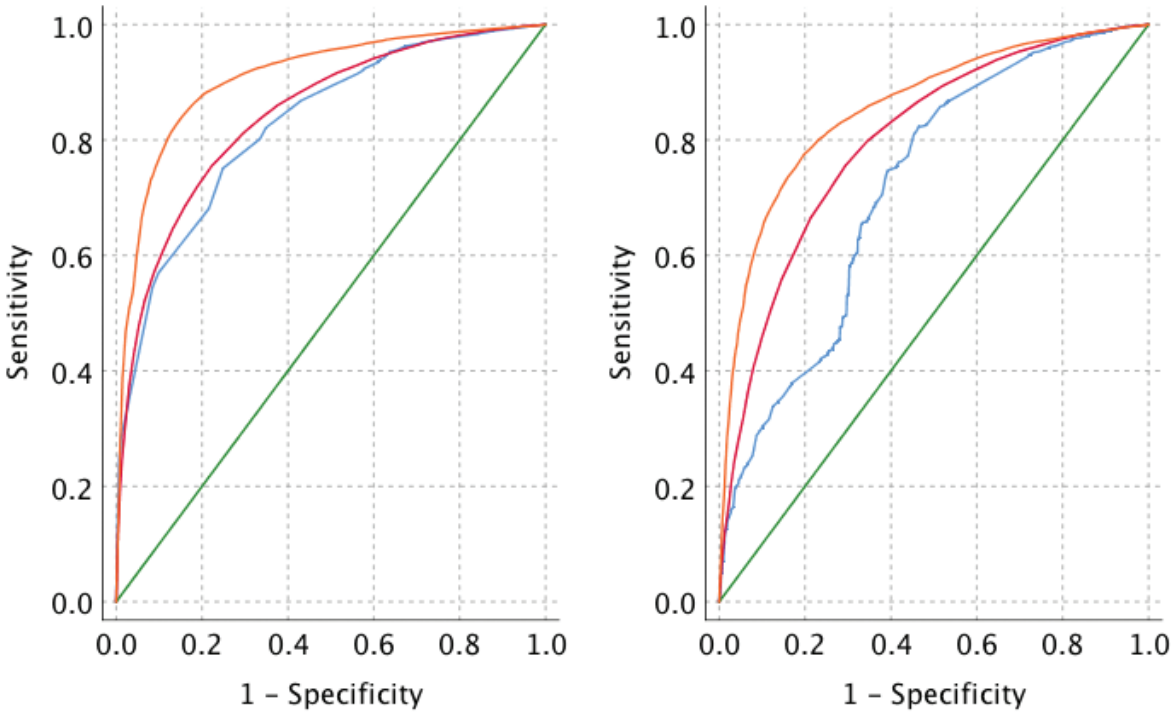


Figure 15. LSS cohort. ROC analyses based on individuals with complete responses to all PROMs at baseline and at the one-year follow-up. VAS_{LEG} (AUC for final score: 0.91; AUC for score change: 0.86), ODI (AUC for final score: 0.84; AUC for score change: 0.80), and $EQ\text{-}5D_{INDEX}$ (AUC for final score: 0.83; AUC for score change: 0.72) The analysis on final scores is given on the left and the analysis on score changes is given on the right (n = 19,805).

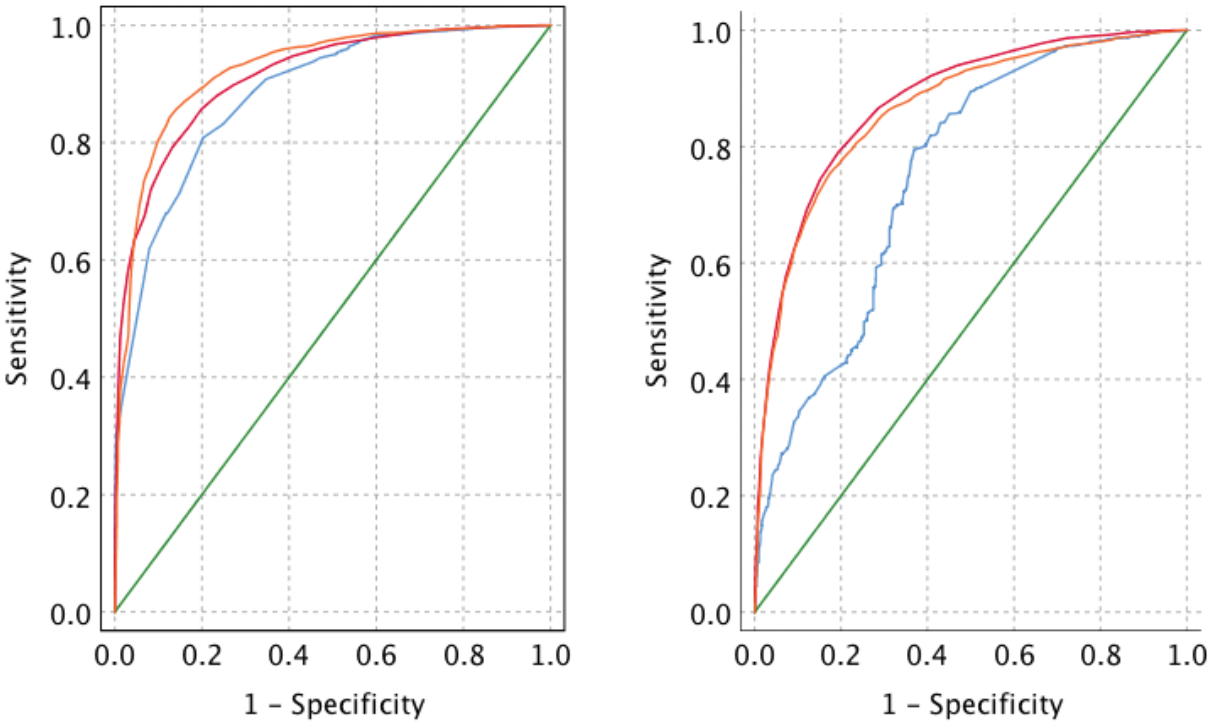


Figure 16. DDD cohort. ROC analyses based on individuals with complete responses to all PROMs at baseline and at the one-year follow-up. VAS_{LEG} (AUC for final score: 0.92; AUC for score change: 0.86), ODI (AUC for final score: 0.91; AUC for score change: 0.88), and $EQ\text{-}5D_{INDEX}$ (AUC for final score: 0.88; AUC for score change: 0.75) The analysis on final scores is given on the left and the analysis on score changes is given on the right(n = 6,522).

discriminative ability improved when final scores rather than score changes were used. The overall quality, expressed as AUC, was better for the VAS and ODI than for the EQ-5D, suggesting that the GA has the ability to measure outcomes in pain and back-pain related disability, but it measures health-related quality of life less well.

4.2 Study II

The aim of the second paper was to investigate possible clinically relevant differences in outcome between the first follow-up occasion and the second follow-up occasion.

A considerable change in PROM score was seen in all groups between baseline and FU1. Changes thereafter were

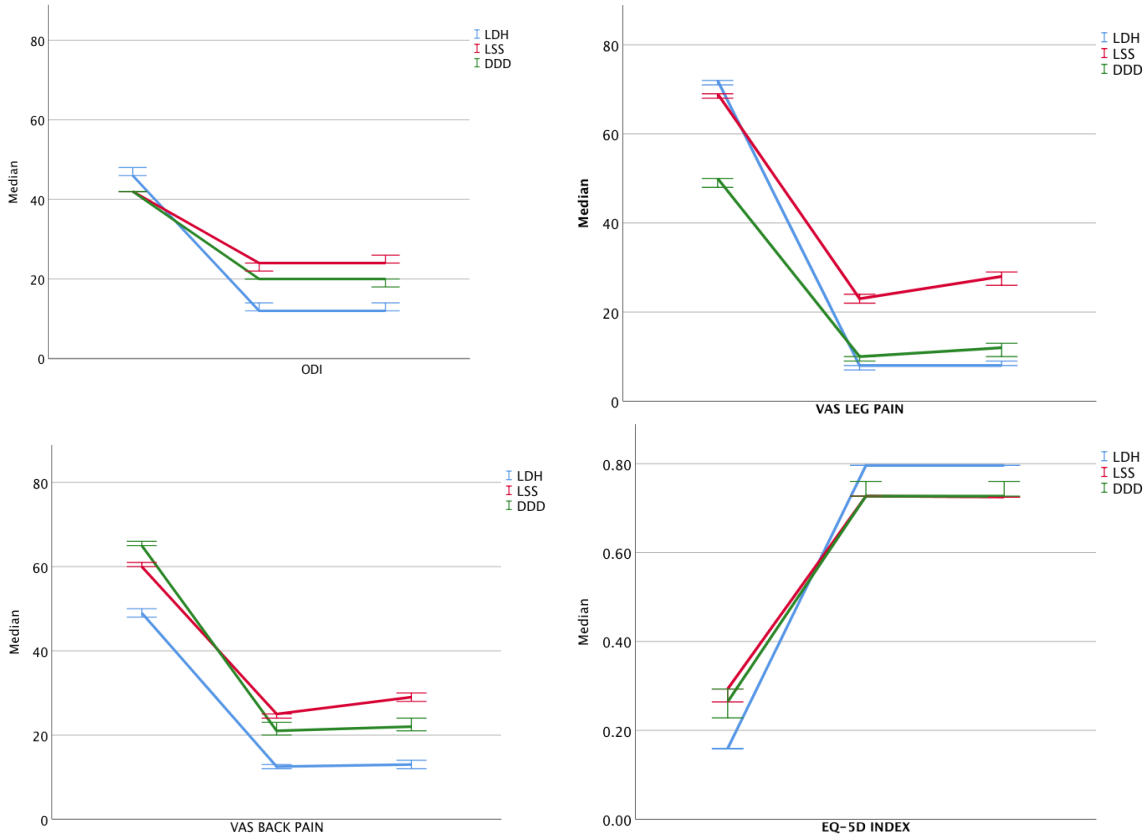


Figure 17. Median values (with 95% confidence intervals) of four PROMs at baseline, and at FU1 and FU2 in the diagnostic groups lumbar disc herniation (LDH; n = 31,314), lumbar spinal stenosis (LSS; n = 53,043), and degenerative disc disorder (DDD; n = 14,375). At baseline, 17-35% of the data were missing. At FU1, 33-45% of the data were missing, and at FU2 the corresponding proportion was 47-56%.

Table 5. Mean and median PROM estimates before surgery, at the one-year follow-up, and at the two-year follow-up in patients who were operated for disc herniation, spinal stenosis, or degenerative disc disorder in the lumbar spine

	Diagnostic group		
	LDH	LSS	DDD
	mean (SD)/median (IQ range)	mean (SD)/median (IQ range)	mean (SD)/median (IQ range)
Baseline			
ODI	48(18)/ 48(36-60)	43(16)/ 44(34-54)	44(15)/ 44(34-52)
VAS$_{BACK}$	48(28)/ 50(24-72)	57(26) 63(45-77)	67(22)/ 66(50-78)
VAS$_{LEG}$	67(26)/ 72(54-86)	64(26)/ 70(52-81)	47(30)/ 53(29-72)
EQ-5D	0.26(0.34)/ 0.16(-0.03-0.62)	0.35(0.32)/ 0.2(0.09-0.69)	0.33 (0.32)/ 0.23(0.09-0.69)
FU1			
ODI	19(17)/ 14(6-30)	26(19)/ 26(10-42)	24(19)/ 22(8-38)
VAS$_{BACK}$	24(25)/ 15(3-42)	32(29)/ 28(7-60)	31(28)/ 23(6-54)
VAS$_{LEG}$	21(26)/ 9(1-35)	33(31)/ 27(5-62)	24(28)/ 14(2-48)
EQ-5D	0.73 (0.30)/ 0.80(0.69-0.85)	0.64(0.30)/ 0.73(0.62-0.80)	0.65(0.32)/0.73(0.62-0.80)
FU2			
ODI	19(18)/ 14(4-30)	27(20)/ 28(12-44)	24(20)/ 22(8-40)
VAS$_{BACK}$	25(26)/ 15(3-46)	36(30)/ 33(9-64)	32(29)/ 23(6-54)
VAS$_{LEG}$	23(28)/ 10(1-41)	36(31)/32(6-66)	26(29)/ 16(2-52)
EQ-5D	0.72(0.30)/ 0.80(0.69-0.85)	0.62(0.30)/ 0.73(0.59-0.80)	0.65(0.32)/ 0.73(0.62-0.80)

Estimates are based on participants who responded to all PROMs at all assessments, as presented in Figure 8

minor. As depicted in Figure 17 and detailed in Table 5, the LDH cohort showed the greatest improvement in all PROMs except VAS_{BACK}. The DDD group reported less severe leg pain compared to the other groups. Large variations in score change were seen between baseline and FU1, but between the two follow-up occasions, changes were close to zero.

To detect any clinically important differences in outcome between FU1 and FU2, the proportion of patients who reached the MIC of treatment success on the two follow-up occasions was determined (Table 6). Depending on the diagnosis group and PROM, 0-4% fewer patients reached the MIC at FU2 compared to FU1. In a similar analysis, the MIC was replaced by thresholds of treatment success based not on score changes, but on absolute scores at the one-year follow-up. In general, a higher proportion of patients reached the thresholds than when MIC values were used as estimates of treatment success, but the differences between FU1 and FU2 remained at 0-5%.

Depending on the diagnosis, 81-85% of the participants made the same assessment about their outcome at FU1 as at FU2 according to $GA_{BACK/LEG}$, and 85-89% according to Satisfaction. There was a statistically significant

Table 6. Proportion of patients in three diagnosis groups who reached minimal important change (MIC) estimates for treatment success at the one-year follow-up, and at the two-year follow-up

PROM	Diagnostic group	N	MIC for success	reaching MIC at FU1 (%)	reaching MIC at FU2 (%)	Cases where baseline scores cause inability to reach MIC (%)
ODI						
	LDH	8,359	−22	64	64	5
	LSS	17,549	−14	57	54	1.5
	DDD	5,493	−16	58	59	1.5
VAS$_{BACK}$						
	LDH	9,193	−20	53	51	21
	LSS	16,038	−28	45	41	14
	DDD	5,510	−29	55	52	8.5
VAS$_{LEG}$						
	LDH	9,778	−39	62	60	13
	LSS	16,083	−27	54	51	8.5
	DDD	4,081	−23	50	49	19.5
EQ-5D						
	LDH	9,214	0.18	71	69	0.5
	LSS	19,252	0.10	61	59	0.1
	DDD	5,930	0.10	66	66	0.1

PROM, patient-reported outcome measure; MIC, minimal important change for treatment success; FU, follow-up; ODI, Oswestry Disability Index; VAS_{BACK}, visual analogue scale for back pain; VAS_{LEG}, visual analogue scale for leg pain; EQ-5D, EuroQol 5-dimension index score; LDH, lumbar disc herniation; LSS, lumbar spinal stenosis; DDD, degenerative disc disorder

deterioration in outcome from FU1 to FU2 of 1-3%. In summary, there was no clinically important change in PROMs at the group level between FU1 and FU2.

4.3 Study III

Paper III was designed to define the SDC at the 95% confidence level for each PROM in a symptom-stable back pain population, and to compare the SDCs to the corresponding opinion-based MICs.
In total, 248 participants filled out the Swespine baseline form on the first test occasion. After the second test, 74.6% had returned both questionnaires. The reliability and measurement error calculations were based on this group. The MIC calculations were based on the Swespine population (n = 98,732).

The time between T1 and T2 was 20 ± 8 days. There was no correlation between the time interval and any of the PROM scores; nor could any statistically significant systematic differences in PROM score between T1 and T2 be detected.

Thus, the influence of random error resulted in the SDCs presented in Table 7, in which the MIC values of the four prospective PROMs, stratified by diagnosis group, can also be seen.

It is notable that there was a considerable gap between the SDC and MIC of the EQ-5D$_{INDEX}$ in all groups. The interpretation of change according to the SDCs and MIC estimates of each PROM and diagnosis group can be seen in Figure 18. The reliability and percentage of agreement between T1 and T2 for the retrospective PROMs are given in Table 8. The exact agreement for GA$_{BACK}$ was 74%, for GA$_{LEG}$ it was 65.5%, and for SF-36$_{GH}$ it was 69%. Overall, an imprecision was seen for all the instruments tested, mainly due to the random error.

4.4 Study IV

The purpose of paper IV was to determine characteristics of patients who

Table 7. Measurement of change parameters (SDC and MIC) for PROMs in three lumbar spine conditions

Parameter of change	NRS$_{BACK}$	NRS$_{LEG}$	ODI	EQ-5D$_{INDEX}$
SDC LDH group	3.6	3.7	18	0.49
MIC LSS group	2.0	3.9	22	0.18
MIC DDD group	2.8	2.7	14	0.10
MIC	2.9	2.3	16	0.10

The MIC calculations were based on the Swespine population operated for LDH (ODI, n = 8,359; NRS$_{BACK}$, n = 9,193; NRS$_{LEG}$, n = 9,778; EQ-5D, n = 9,214), LSS (ODI, n = 17,549; NRS$_{BACK}$, n = 16,038; NRS$_{LEG}$, n = 16,083; EQ-5D, n = 19,252), and DDD (ODI, n = 5,493; NRS$_{BACK}$, n = 5,519; NRS$_{LEG}$, n = 4,081; EQ-5D, n = 5,930) in the period 1998-2016, using the anchor-based ROC curve method.

Table 8. Reliability of retrospective single-item questions

PROM	Exact Agreement %	T1>T2 %	T1<T2 %	Weighted Kappa
GA$_{BACK}$(n =96)	74	12.5	13.5	0.86
GA$_{LEG}$ (n =96)	65.5	18	16.5	0.75
SF-36$_{GH}$ (n =94)	69	19	12	0.81

Exact agreement: the proportion who gave the same response at T1 as at T2; T1 > T2: the proportion who responded that they had a better outcome at T1 than at T2; T1 < T2: the proportion who responded that they had a worse outcome at T1 than at T2.
GA$_{BACK}$, Global Assessment for back pain; GA$_{LEG}$, Global Assessment for leg pain; SF-36$_{GH}$, Short Form-36 single-item question on global health; T1, first test occasion; T2, retest occasion.

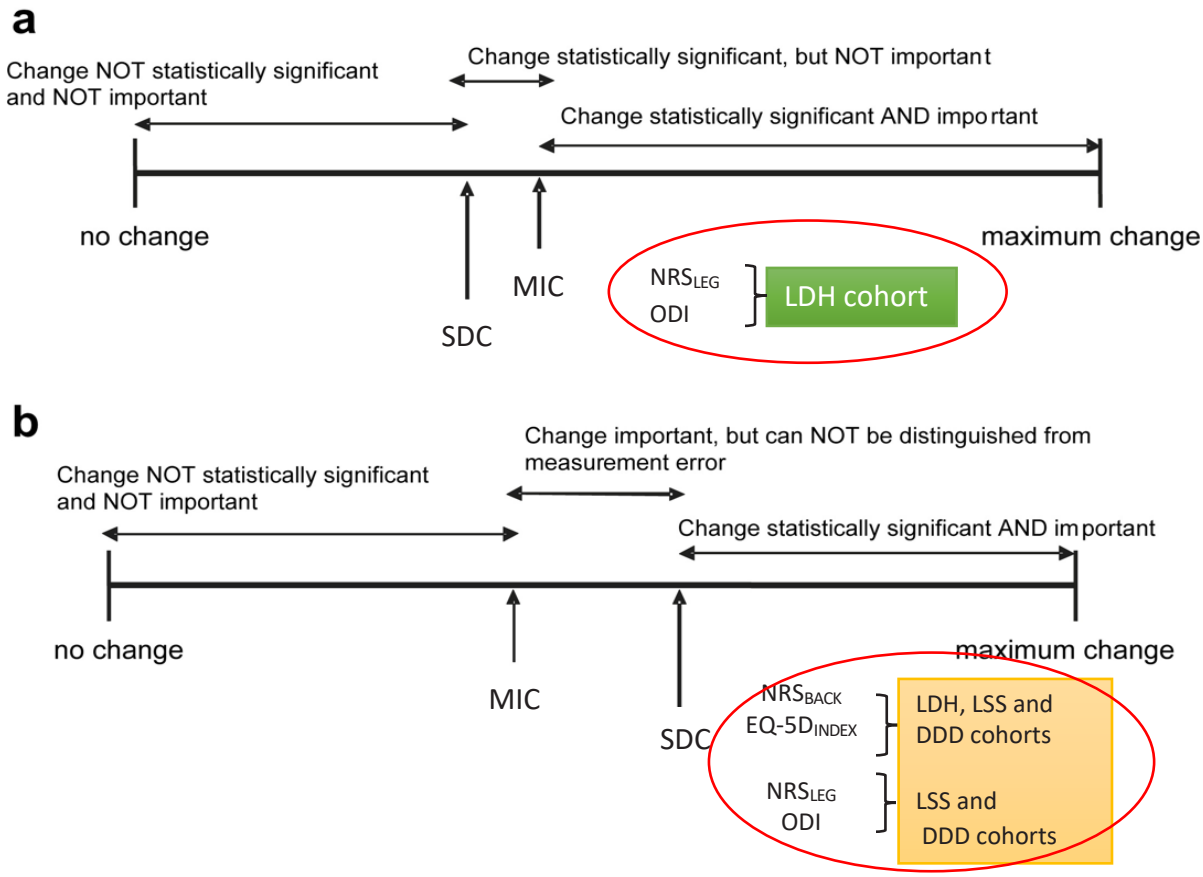


Figure 18. Relation of minimal important change (MIC) and smallest detectable change (SDC) for the outcome measures $NRS_{BACK/LEG}$, ODI, and $EQ\text{-}5D_{INDEX}$ in the diagnosis groups disc herniation (LDH), lumbar spinal stenosis (LSS), and degenerative disc disorder (DDD). Only in the LDH cohort did the MIC exceed the measurement errors (SDCs) estimated for NRS_{LEG} and ODI. All other MIC estimates could not be distinguished from measurement error.

were lost to follow-up, and also their predicted outcome.

In all three diagnosis groups, there was a significant difference between respondents and non-respondents in many of the variables. However, in absolute numbers, the differences were small. Non-respondents consumed a somewhat higher amount of postoperative healthcare.

Younger age, male sex, and being born outside of the EU were predictors of non-response in all of the three groups. In the LSS and LDH groups, non-response was also predicted by a lower disposable income, living alone, smoking, and a higher degree of comorbidity. A low level of education was predictive of non-response in the LDH and DDD groups. Furthermore, previous spine surgery was found to be predictive of non-response in the

Table 9. Percentage of patients self-assessing as having a successful outcome at the one-year follow up compared to the outcome as predicted by the regression models.

			GA$_{SUCCESS}$			
Diagn. group	Observed outcome	95% Conf. Interv.	Predicted outcome, respondents	95% Conf. Interv.	Predicted outcome non-respondents	95% Conf. Interv.
LDH	78.8	77.4-80.1	78.7	78.2-79.2	75.4	74.6-76.3
LSS	58.2	55.6-57.6	58.7	57.9-58.6	53.9	53.2-54.5
DDD	67.4	64.9-70.0	67.5	66.6-68.3	62.7	60.6-64.8

Successful outcome: self-assessments such as "pain-free" and "much better" at the one-year follow-up. Differences between respondents and non-respondents were significant ($p < 0.001$). LDH, lumbar disc herniation (n = 5,192); LSS, lumbar spinal stenosis (n = 12,132); DDD, degenerative disc disorder (n = 1,587)

LSS and DDD cohorts and, finally, having an unexpected event during the first 12 postoperative months predicted non-response in the LSS group. Apart from a low EQ-5D$_{INDEX}$ score in the LSS cohort, baseline PROM scores were not predictive of non-response.

Measured with the GA, the predicted successful outcome was significantly lower for non-respondents (Table 9). In the LDH cohort, 78.7% self-assessed themselves as being "pain-free" or "much better", whereas the corresponding percentage for non-respondents was 75.4%. The corresponding figures for the LSS cohort were 58.2% and 53.9%; and in the DDD cohort 67.5% and 62.7%. The secondary outcome measures showed the same pattern, suggesting a somewhat worse outcome for patients with the characteristics of non-respondents. The accuracy of the models in predicting outcome was estimated in ROC curve analyses. The AUC for the LDH model was 0.73; for the LSS model 0.69; and for the DDD model 0.72.

DISCUSSION

5. DISCUSSION

5.1 Patient values as indicators of outcome

Each scientific field is like a clock with a blank face. A slowly turning hand points to research questions that are important for the present time. Some points in time are cardinal and the clock strikes a new hour, a paradigm shift. Amory Codman was ahead of his time and did not live to see his End-result idea become deeply rooted and a given a part in the evidence-based medicine paradigm of today [152,3].

The starting point for the development of PROMs directly related to end results in spine surgery might have been a paper written by Lee et al., presenting an evaluation system for patients with spinal stenosis [153]. The evaluation form relied on a combination of the patient's own perceptions and the physician's examination. Since the 1980s, the patient's point of view has been increasingly important [154], and there has been a veritable explosion in the development of questionnaires measuring treatment success in patients with chronic low back pain [155]

5.2 Challenges in the interpretation of change

Participants in the Swespine register are requested to fill out the prospective PROMs VAS/NRS, ODI, and EQ-5D (and until recently, the SF-36) before the operation and 1, 2, 5, and 10 years afterwards. The retrospective PROM GA is added to the follow-up questionnaires.

It is not easy to define and quantify treatment success. As an example, imagine two patients with spinal stenosis. Patient A cannot play golf any more because his leg pain prevents him from walking more than 500 metres. He considers "being able to walk for two hours" to be a clinically relevant change. Patient B cannot walk 50 metres to get his newspaper in the morning. He considers "being able to walk 100 metres" to be a clinically relevant change. Both have spinal stenosis, but they have different goals for their treatments.

The same could be applied to differences between surgeons/researchers aiming to interpret whether or not a change in PROM score is clinically important. Some argue that the smallest detectable change beyond the measurement error is sufficiently important, while others claim that the improvement has to be considerable in order to be important [46]. King et al. recently proposed that patients reporting getting "a little better" constitute the minimal change group (i.e. MIC), and ratings of "much better" or "very much better" might be used to define responders [48]. As long as there is no consensus on where to draw the line, there will be a variety of recommended MIC values for the same instrument [156].

Efforts have been made to reach consensus on which PROMs are the most appropriate. Prospective PROMs and also TQs are recommended [157]. However, the problem of how to interpret the results given by the PROMs still remains. How to interpret change is particularly cumbersome – for prospective PROMs as well as for TQs - because cognitive processes will affect a person's appraisal of the outcome over time [154].

Some of the disadvantages of retrospective questions are that they cannot be psychometrically tested, that the recall bias is too much of a millstone, and that too little information can be extracted. Such criticism is important. However, recent reviews have pointed out that prospective PROMs are used without a thorough consideration of their psychometric properties [97,158,68].

When there is no ultimate surgical technique for a certain condition, there are usually a number of surgical options available. This also applies to calculation of the MIC, and a recent systematic review reported 11 different methods [159]. The anchor-based method most often uses a transition question as a substitute for the absent gold standard, and is considered the method of choice by influential researchers despite having well documented shortcomings [160].

Sometimes the PROM is not sensitive enough to distinguish a true change from measurement error (i.e. the MIC value lies below the limit of what is statistically detectable). This suggests that the PROM is either not good enough or that it was used for measuring a condition that it was not designed for [161].

5.3 Lessons from the current studies

Big datasets, like the ones provided by Swespine, containing tens to a hundred thousand of observations and hundreds of different variables, might give researchers or others the impression that all the necessary information is included, so any study could be safely performed. However, data lost in a systematic manner may affect the output variables and inferences based on the results may be wrong, or applicable only to the respondents (paper IV). Furthermore, it is important to evaluate how well different PROMs reflect changes in aspects of health in the actual patient group (papers I and III) and also when after an intervention the changes obtained are apparent and should be measured (paper II).

The missing data mechanism behind attrition at the one-year follow-up appear to be MNAR (i.e non-ignorable) (paper IV). Other mechanisms may apply to data that are missing at the item level or the variable level. The effect of the lost data appears to be that the outcome presented by Swespine data would be a little over-rated.

The results from study IV can be interpreted in different ways. A recently published paper found that patients who were lost to follow-up reported a statistically significant worse outcome according to GA_{BACK} than patients who remained in the register, but it concluded that data could be treated as missing at random (MAR) [162]. When data are missing at random, adjustments can be made using various statistical models. When data are MNAR, however, there is a risk that the adjustments will lead to an increased degree of bias instead. The use of modern missing data methods, such as multiple imputation, is encouraged - even when the missing data mechanism is MNAR. But, in addition to requiring the help of an experienced statistician, they require careful planning and a deep knowledge of which variables might have an effect on the outcome if they are missing [116].

As documented in previous studies, typical socio-economic features signifying non-respondents emerged in the analyses [116,164,118], which might threaten the generalizability of the register. For instance, patients who were born outside the European Union were associated with non-response, suggesting that this subgroup may have a worse outcome after sur-

gery for any of the diagnoses disc herniation, spinal stenosis, or DDD, than what is reported in Swespine (paper IV). Younger age was associated with non-response, but at the same time it was a predictor of a successful outcome, indicating that the outcome in young individuals may be underrated. Also, the results suggest that when conducting a clinical trial, one should include a higher number of patients with the characteristics of non-respondents to prevent attrition bias.

Overall, Swespine participants who were lost to the follow-up at one year were predicted to have a somewhat worse outcome than patients who completed the follow-up questionnaires requested by Swespine at the one-year follow-up. This was seen in all three diagnosis cohorts and for all the PROMs tested.

Bearing this in mind, the registration should be made as simple and the least time consuming as possible for the patients, to bring up the response rate. One such improvement might be to use less and shorter questionnaires.

The results from paper I suggest that GA has the capacity to detect patients who gain from lumbar surgery. If GA were to be used as a proxy for a gold standard, the cut-off should be between the response options "much better" and "somewhat better" (signifying a considerable change) rather than between "somewhat better" and "unchanged". This finding is supported by a study on the individual conceptions of a good outcome among surgeons and patients [163]. The GA might be used as a gold standard proxy in the determination of MIC estimates of prospective PROMs measuring back and leg pain, and disability related to back pain.

GA might replace prospective PROMs estimating pain and function and perhaps also quality of life in routine follow-ups. The index score or composite score of prospective PROMS is frequently presented in scientific work. As questionnaire responses are transformed into a single total score, a figure is created that is no more detailed than a transition question. In fact, additional information might be lost, since the questions in the questionnaire might not cover all relevant aspects of clinical improvement. A TQ has the advantage of leaving the matter of relevance to the respondent.

Prospective PROMs do not appear to capture the outcome of degenerative spine surgery in a better fashion than

a TQ does. One might argue that using a transition question is as bad as using a prospective PROM, but for different reasons. If the two ways of measurement are of equally low quality – although for different reasons, as described above – why not use the least complicated one, without taking the detour of the MIC?

One of the shortcomings of GA is that while the ODI, VAS/NRS, EQ-5D, and SF-36 appear in studies all over the world, GA is not used on any other population than the Swedish one, which reduces the possibility of comparison. On the other hand, the prospective PROMs come in different versions, which might easily be overlooked when comparisons are being made.

Another important point is that multiple-item questionnaires are valuable instruments when a more detailed answer about the outcome is desirable. Research questions like "how is a person's physical function affected by lumbar surgery?" or "apart from leg pain, which other symptoms may be influenced by lumbar surgery and, if so, how?" or "in what ways does symptomatic spinal stenosis affect a person's quality of life?" demand a psychometrically sound battery of questions. These instruments, however, need to be interpreted in the light of how well they can measure clinical important differences in different patient groups.

The results in paper III indicated that possible differences between groups may be difficult to distinguish from measurement error if the NRS, ODI, and EQ-$5D_{INDEX}$ are to be used as outcome variables.

The SDC values at the 95% confidence level between two points of estimation of NRS, ODI, and EQ-5D were higher than the corresponding MIC estimates for patients operated for spinal stenosis or degenerative disc disorder. The SDCs of NRS_{LEG} and ODI were above the MIC levels for patients who were operated for disc herniation. There was a considerable gap between the SDC and the MIC for the EQ-5D in all three groups. The SDC for EQ-$5D_{INDEX}$ was remarkably higher than the MIC value, suggesting that this index is not appropriate as a tool for measuring change in this way. The various items may be interpreted separately instead, in order to get a more nuanced picture of the results of the intervention.

The retest reliability of $GA_{BACK/LEG}$ was substantial, according to the weighted kappa estimates. However, the

exact agreement was no more than 65.5% for GA_{LEG} and 74% for GA_{BACK}. One way of circumventing some of the difficulties arising from the use of MIC estimates would be to measure the outcome exclusively by using the final scores (labelled threshold of treatment success), instead of by score changes, since GA showed a stronger correlation to the former. When only one point of estimation is used, the SDC estimates are replaced with ±1.96 SEM, which would be below the threshold of treatment success.

Another consideration when using PROM questionnaires in quality registers is when after the surgery, and/or how often, the surveys should be administered. The necessity of both a one-year and a two-year follow-up in effectiveness studies has been in question for some time [164,111,112,165]. Arguments put forward for a one-year follow-up only are the reduction in cost and patient burden. Arguments against bring up the risk of not capturing unfavourable outcomes or adverse events, and also that most scientific journals and authors consider a monitoring time shorter than two years to be inappropriate.

In paper II, a minor deterioration in PROMs from FU1 to FU2 was seen - as measured by the proportion reaching the MIC value of each PROM. The same pattern was seen for the proportion reaching a threshold of treatment success based on PROM scores at FU1 instead of on score changes. A significant deterioration of 0.5-3% was found when outcome was measured by GA or Satisfaction.

Although statistically significant, the result is not to be interpreted as any encouragement to monitor PROMs at both one and two years postoperatively. Depending on the diagnosis, there are more plausible explanations for this deterioration. For instance, the reoperation rate for recurrent disc herniation, which is the most common cause of recurrent back and leg pain after discectomy, was approximately 5% [166]. Another factor is that normative values for the EQ-5D and ODI would be expected to decrease with increasing age [167,168]. Comorbidity, such as hip osteoarthritis, affects back pain-related PROMs [169]. Symptoms from the hip joint probably occur in some patients, shortly before the time of the second follow-up. These symptoms might be mistaken for a recurrent stenosis, thus biasing the outcome. Many patients who are treated for spinal stenosis are also reoperated because of a recurrent stenosis. According to Försth et al.,

approximately one-fifth of LSS patients in Sweden underwent a new operation at the same level or at an adjacent level during a follow-up of 6.5 years [16]. It could be expected that some of these patients would report a deterioration already in their two-year follow-up questionnaire.

Instead of monitoring patients at FU2, it might be beneficial to introduce a follow-up at three months postoperatively to better capture early complications, such as infections.

5.4 A promising future

In 1978, Lee and colleagues wrote that "An accurate and objective evaluation of patients with chronic low back pain is very difficult, since most of their complaints are subjective in nature"[153]. It is a disturbing reality that after 40 years of outcomes research, the search for an "objective" outcome measure prevails. Neither a sophisticated multiple-item questionnaire nor a simple single-item question appears to be the ultimate outcome measurement. On the subject of effectiveness, the latter would suffice, as would a shorter follow-up period - not least to prevent patient drop-out. A new system of patient-reported outcomes measurement (PROMIS) was launched by the US National Institutes of Health in 2004. PROMIS provides so-called item banks, which have been calibrated and referenced to the general US population [170]. The item banks, developed using "modern" psychometric methods such as item response theory, enable computerized adaptive testing (i.e. the test adapts the choice of items to respondents' levels of attainment, as determined by previous responses), which may result in improved measurement precision and responsiveness [171]. The goal is to standardize measurements and thereby facilitate comparability of data across studies and settings at the international level [172].

The idea of PROMIS is appealing, but it is too early to tell if and when researchers and clinicians will be ready to exchange established outcome measures for this new concept.

STRENGHTS AND LIMITATIONS

6. STRENGHTS AND LIMITATIONS

The ultimate situation for a register-based study would be a complete set of data consisting of variables with the ability to provide clear answers to the research question.

If no data were missing, the information obtained from the register would be safely generalizable to the entire target population. But in every register, data are missing, and in Swespine, the external validity appears to be somewhat affected by attrition. Known confounding variables may be adjusted for, but there will always be uncertainty about missing confounders with a possible impact on the outcome. Furthermore, little is known about patients who are lost at the recruitment stage, i.e before the operation. According to the National Patient Register, approximately 15% of the patents who are operated in the spine are not reported in Swespine.

If data have been collected successfully, the internal validity may be threatened if the outcome variables are not simple to use, reliable, and sufficiently accurate to allow clear inferences.

The above arguments are limitations of both RCTs and register-based studies. Both types of studies are equally important, but during different stages of the scientific process. Another limitation of register-based studies is the difficulty in formulating a priori hypotheses, because the data have already been collected.

Can sound inferences be drawn from the studies in this thesis? The first study had an exploratory approach and further research on factors affecting the response pattern of GA to confirm the results would be beneficial. In all the studies, selection bias may have been present because of large numbers of missing data at the variable level as well as at the individual and occasion levels, which were not adjusted for. The use of an imperfect measurement instrument (GA) as a gold standard in the testing of the ability of other instruments to discriminate between treatment success and treatment failure, as was done in studies II and III, is debatable. The absence of a gold standard reduces the certainty of the MIC estimates. Finally, in study IV, the predictive models were based on data provided by patients who did respond to the Swespine questionnaires. Although the quality of the models was satisfactory, the true outcome in non-respondents remains unknown.

CONCLUSIONS

7. CONCLUSIONS

- The transition question Global Assessment can be used as the single patient-reported outcome measure in the assessment of effectiveness in routine follow-up of degenerative lumbar spine surgery (study I).

- For outcome assessment using one of the PROMs examined, a one-year follow-up is sufficient. The Swespine two-year follow-up assessment with PROMs can be excluded (study II).

- The SDCs in NRS_{BACK}, NRS_{LEG}, and in the ODI exceeded the corresponding MICs estimated in populations treated for spinal stenosis and DDD, suggesting that these PROMs are not sufficiently responsive to detect small but potentially important changes in outcome in these groups. The SDC in NRS_{LEG} and in the ODI were lower than the MIC that was defined for patients treated for disc herniation, indicating sufficient responsiveness when used in this population. Measurement of change in $EQ\text{-}5D_{INDEX}$ should not be expressed in terms of SDC and MIC (study III).

- PROM data appear to be somewhat affected by patients who are lost to follow-up, resulting in an overestimation of the outcome one year after surgery. Measures need to be taken to mitigate attrition, e.g. by reducing questionnaire burden (study IV).

FUTURE WORK

8. FUTURE WORK

Given the limitations of the PROMs studied, the following research options may be considered. There could be development of the GA, either by exploring the possibility of rephrasing the question with the goal of decreasing present-state bias or by expanding the question also to include a measure of experience. If GA was able to measure the present health status in relation to the past surgery, it might be suitable as a measure of a patient-acceptable symptom state (PASS). In this way, obstacles such as recall bias and present-state bias could be partly overcome. A first step would be to compare GA with PASS measures. Furthermore, it might be better to rephrase the question in Satisfaction, (a measure regarding the overall experience of the surgery) and to separate it in time from the GA. The questions in GA and Satisfaction have similar wording - and when asked at the same time, the Satisfaction may be interpreted by the respondent as just another way of asking the same thing as GA does. The general increase in computer capacity and also the goal expressed by healthcare authorities of making patient records and registers digital, are arguments supporting the implementation of PROMIS. This would require close cooperation between existing outcome registers, but it might also enable comparisons with other patient populations that would be more accurate than can be achieved today, and allow outcome assessment on a more individual basis.

ACKNOWLEDGEMENTS

9. ACKNOWLEDGEMENTS

I would like to thank everyone who has supported me throughout my work behind this thesis. In particular, I would like to express my sincere gratitude to:

Professor **Helena Brisby**, for her guidance in every part of my research and writing of this thesis. I could not have imagined having a better advisor and mentor for my PhD study,

Olle Hägg MD PhD, one of few with expert knowledge in spine surgery as well as in spine surgery outcome research. Thank you for your patience, your rational and calm reasoning, and for prompt and wise responses to my many e-mails,

Assoc. Professor **Bengt Lind**, for taking on the responsibility as co-supervisor despite being retired,

Aldina Pivodic and **Carl Willers**, PhD, for their statistical expertise,

the University of Gothenburg, the Dept. of Orthopedics and professor **Ola Rolfson**, Chairman, Dept. of Orthopaedics,

GHP Spine Center Gothenburg, represented by former Head of Spine Center Gothenburg, **Åke Blixt** MD PhD and current Head of Spine Center Gothenburg, **Christian Hagelberg** MD,

Cina Holmer, for taking care of administrative matters surrounding the thesis,

the secretaries at the Swespine register, **Carina Blom**, **Lena Mellgren** and **Linda Brun**,

Jannis Ioannidis MD, dear friend, colleague and former employer at the orthopedic department in Karlstad, who supported the start-up of this thesis,

Rolf Sandberg MD PhD, the late **Leif Måwe** MD, **Åke Blixt** MD PhD, **Hans Laestander** MD, **Leif Anderberg** MD PhD, who have taken me under their wings and mentored me about life in general and patients and spine surgery in particular,

my good friend **Olof Thoreson** MD PhD, for his continuous encouragement and good advice,

the staff at Spine Center Göteborg, a professional, smiling, cheering team. I am lucky to have co-workers like you,

all patients participating in Swespine,

my beloved sister **Ylva**, supportive in all dimensions of life and cover illustrator,

my dear mother **Anna-Catharina** (Kim), constantly encouraging and helping me to correct my English,

my husband **Stamatis**, the love of my life and my best friend.

This thesis project was financially supported by FoU Värmland, government grants under the LUA agreement, GHP Spine Center Göteborg, and dr Félix Neubergh Foundation.

REFERENCES

10. REFERENCES

1. Codman EA. The classic: A study in hospital efficiency: as demonstrated by the case report of first five years of private hospital. Clin Orthop Relat Res. 2013;471(6):1778-83.

2. Reverby S. Stealing the Golden Eggs: Ernest Amory Codman and the Science and Management of Medicine. Bulletin of the History of Medicine. 1981;55(2):156.

3. Kaska SC, Weinstein JN. Historical perspective. Ernest Amory Codman, 1869-1940. A pioneer of evidence-based medicine: the end result idea. Spine. 1998;23(5):629-33.

4. Donabedian A. Evaluating the quality of medical care. 1966. The Milbank quarterly. 2005;83(4):691-729.

5. Donabedian A. Twenty Years of Research on the Quality of Medical Care: 1 964-1984. Evaluation & the Health Professions. 1985;8(3):243-65.

6. Tipple F. Den nationella SOM-undersökningen 2017. Göteborgs universitet Göteborg; 2018. p. 409-17.

7. Jacobsson Ekman G, Lindahl B, Nordin A. National quality registries in Swedish health care: Stockholm : Karolinska Institutet University Press; 2016.

8. Stromqvist B, Fritzell P, Hagg O, Jonsson B. The Swedish Spine Register: development, design and utility. European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society. 2009;18 Suppl 3:294-304.

9. van Hooff ML, Jacobs WC, Willems PC, Wouters MW, de Kleuver M, Peul WC, et al. Evidence and practice in spine registries. Acta orthopaedica. 2015;86(5):534-44.

10. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet (London, England). 2007;370(9596):1453-7.

11. Clement RC, Welander A, Stowell C, Cha TD, Chen JL, Davies M, et al. A proposed set of metrics for standardized outcome reporting in the management of low back pain. Acta orthopaedica. 2015;86(5):523-33.

12. Johnson TP, Wislar JS. Response Rates and Nonresponse Errors in Surveys. Jama. 2012;307(17):1805-6.

13. Kvalitetsregister N. Valideringshandbok [Internet]: Sveriges Kommuner och Regioner; 2016 [Available from: http://kvalitetsregister.se/drivaregister/valideringshandbok.1911.html.

14. Barreto ML. Efficacy, effectiveness, and the evaluation of public health interventions. Journal of epidemiology and community health. 2005;59(5):345-6.

15. Forsth P, Michaelsson K, Sanden B. Does fusion improve the outcome after decompressive surgery for lumbar spinal stenosis?: A two-year follow-up study involving 5390 patients. The bone & joint journal. 2013;95-b(7):960-5.

16. Forsth P, Olafsson G, Carlsson T, Frost A, Borgstrom F, Fritzell P, et al. A Randomized, Controlled Trial of Fusion Surgery for Lumbar Spinal Stenosis. The New England journal of medicine. 2016;374(15):1413-23.

17. Vården i Siffror: Sveriges Kommuner och Regioner; [Available from: https://vardenisiffror.se/dashboard?relatedmeasuresbyentry=registry&relatedmeasuresbyid=svenska-ryggregistret-swespine&units=13&units=03&units=23&units=14&units=04&units=24.]

18. Karlsson J, Marx RG, Nakamura N, Bhandari M. A Practical Guide to Research: Design, Execution, and Publication. Arthroscopy: The Journal of Arthroscopic and Related Surgery. 2011;27(4):S1-S112.

19. Stromqvist B, Jonsson B, Fritzell P, Hagg O, Larsson BE, Lind B. The Swedish National Register for lumbar spine surgery: Swedish Society for Spinal Surgery. Acta orthopaedica Scandinavica. 2001;72(2):99-106.

20. Svenska Ryggregistret Swespine: Swedish Society of Spinal Surgeons; [Available from: http://www.swespine.se.]

21. Streiner DL, Norman GR. Health Measurement Scales : A practical guide to their development and use: Oxford University Press, Incorporated; 2008.

22. Nilsson E, Orwelius L, Kristenson M. Patient-reported outcomes in the Swedish National Quality Registers. Journal of internal medicine. 2016;279(2):141-53.

23. Iderberg H, Willers C, Borgström F, Hedlund R, Hägg O, Möller H, et al. Predicting clinical outcome and length of sick leave after surgery for lumbar spinal stenosis in Sweden: a multi-register evaluation. European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society. 2019;28(6):1423-32.

24. McCormick JD, Werner BC, Shimer AL. Patient-reported outcome measures in spine surgery. The Journal of the American Academy of Orthopaedic Surgeons. 2013;21(2):99-107.

25. Ross M. Relation of Implicit Theories to the Construction of Personal Histories. Psychological Review. 1989;96(2):341-57.

26. Schwartz CE, Sprangers MAG. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. Social Science & Medicine. 1999;48(11):1531-48.

27. Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. Journal of clinical epidemiology. 2002;55(9):900-8.

28. Grovle 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. Journal of clinical epidemiology. 2014;67(5):508-15.

29. Kamper SJ, Maher CG, Mackay G. Global rating of change scales: a review of strengths and weaknesses and considerations for design. The Journal of manual & manipulative therapy. 2009;17(3):163-70.

30. 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. Journal of clinical epidemiology. 2010;63(7):760-6.e1.

31. Bowling A. Just one question: If one question works, why ask several? Journal of epidemiology and community health. 2005;59(5):342-5.

32. Feinstein AR. Clinimetrics. New Haven: Yale University Press; 1987. xi, 272 p. p.

33. Streiner DL. Clinimetrics vs. psychometrics: an unnecessary distinction. Journal of clinical epidemiology. 2003;56(12):1142-5; discussion 6-9.

34. de Vet HCW, Terwee CB, Bouter LM. Clinimetrics and psychometrics: two sides of the same coin. Journal of clinical epidemiology. 2003;56(12):1146-7.

35. Di Fabio RP. Essentials of rehabilitation research : a statistical guide to clinical practice. Philadelphia: F.A. Davis; 2013.

36. Polit DF, Yang FM. Measurement and the measurement of change : a primer for the health professions. Philadelphia: Wolters Kluwer; 2016. x, 350 pages p.

37. Vet HCWd. Measurement in medicine : a practical guide. Cambridge: Cambridge University Press; 2011. x, 338 p. p.

38. Bland JM, Altman DG. Measurement error.(Statistics Notes). British Medical Journal. 1996;312(7047):1654.

39. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. Journal of clinical epidemiology. 2010;63(7):737-45.

40. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires.(Author abstract). Journal of clinical epidemiology. 2007;60(1):34.

41. Mokkink LB, de Vet HCW, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation. 2018;27(5):1171-9.

42. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet (London, England). 1986;1(8476):307-10.

43. Bland JM, Altman DG. Measuring agreement in method comparison studies. Statistical methods in medical research. 1999;8(2):135-60.

44. Kirshner B, Guyatt G. A methodological framework for assessing health indices. Journal of chronic diseases. 1985;38(1):27-36.

45. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. Journal of clinical epidemiology. 1997;50(8):869-79.

46. King MT. A point of minimal important difference (MID): a critique of terminology and methods. Expert review of pharmacoeconomics & outcomes research. 2011;11(2):171-84.

47. Wells G, Beaton D, Shea B, Boers M, Simon L, Strand V, et al. Minimal clinically important differences: review of methods. The Journal of rheumatology. 2001;28(2):406-12.

48. King MT, Dueck AC, Revicki DA. Can Methods Developed for Interpreting Group-level Patient-reported Outcome Data be Applied to Individual Patient Management? Medical care. 2019;57 Suppl 5 Suppl 1(Suppl 5 1):S38-S45.

49. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Controlled clinical trials. 1989;10(4):407-15.

50. Terwee CB, Roorda LD, Knol DL, De Boer MR, De Vet HC. Linking measurement error to minimal important change of patient-reported outcomes. Journal of clinical epidemiology. 2009;62(10):1062-7.

51. Schwind J, Learman K, O'Halloran B, Showalter C, Cook C. Different minimally important clinical difference (MCID) scores lead to different clinical prediction rules for the Oswestry disability index for the same sample of patients. The Journal of manual & manipulative therapy. 2013;21(2):71-8.

52. Terwee CB, Roorda LD, Dekker J, Bierma-Zeinstra SM, Peat G, Jordan KP, et al. Mind the MIC: large variation among populations and methods. Journal of clinical epidemiology. 2010;63(5):524-34.

53. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. Journal of clinical epidemiology. 2003;56(5):395-407.

54. de Vet HC, Ostelo RW, Terwee CB, van der Roer N, Knol DL, Beckerman H, et al. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation. 2007;16(1):131-42.

55. Tubach F, Ravaud P, Baron G, Falissard B, Logeart I, Bellamy N, et al. Evaluation of clinically relevant states in patient reported outcomes in knee and hip osteoarthritis: the patient acceptable symptom state. Annals of the rheumatic diseases. 2005;64(1):34-7.

56. van Hooff ML, Mannion AF, Staub LP, Ostelo RW, Fairbank JC. 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. The spine journal : official journal of the North American Spine Society. 2016;16(10):1221-30.

57. Brooks R. EuroQol: the current state of play. Health policy (Amsterdam, Netherlands). 1996;37(1):53-72.

58. EuroQol--a new facility for the measurement of health-related quality of life. Health policy (Amsterdam, Netherlands). 1990;16(3):199-208.

59. Rabin R, Gudex C, Selai C, Herdman M. From translation to version management: a history and review of methods for the cultural adaptation of the EuroQol five-dimensional questionnaire. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research. 2014;17(1):70-6.

60. Burström K, Sun S, Gerdtham U-G, Henriksson M, Johannesson M, Levin L-Å, et al. Swedish experience-based value sets for EQ-5D health states. Quality of Life Research. 2014;23(2):431-42.

61. Dolan P. Modeling valuations for EuroQol health states. Medical care. 1997;35(11):1095-108.

62. Svenska höftprotesregistret Årsrapport 2018.

63. Nemes S, Burström K, Zethraeus N, Eneqvist T, Garellick G, Rolfson O. Assessment of the Swedish EQ-5D experience-based value sets in a total hip replacement population. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation. 2015;24(12):2963-70.

64. Mannion AF, Boneschi M, Teli M, Luca A, Zaina F, Negrini S, et al. Reliability and validity of the cross-culturally adapted Italian version of the Core Outcome Measures Index. European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society. 2012;21 Suppl 6:S737-49.

65. Coretti S, Ruggeri M, McNamee P. The minimum clinically important difference for EQ-5D index: a critical review. Expert review of pharmacoeconomics & outcomes research. 2014;14(2):221-33.

66. Ranstam J, Robertsson O, A WD, Lofvendahl S, Lidgren L. [EQ-5D--a difficult-to-interpret tool for clinical improvement work]. Lakartidningen. 2011;108(36):1707-8.

67. Janssen MF, Pickard AS, Golicki D, Gudex C, Niewada M, Scalone L, et al. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation. 2013;22(7):1717-27.

68. Chiarotto A, Terwee CB, Kamper SJ, Boers M, Ostelo RW. Evidence on the measurement properties of health-related quality of life instruments is largely missing in patients with low back pain: A systematic review. Journal of clinical epidemiology. 2018;102:23-37.

69. Linde L, Sørensen J, Ostergaard M, Hørslev-Petersen K, Hetland ML. Health-related quality of life: validity, reliability, and responsiveness of SF-36, 15D, EQ-5D corrected RAQoL, and HAQ in patients with rheumatoid arthritis. The Journal of rheumatology. 2008;35(8):1528.

70. Godil SS, Parker SL, Zuckerman SL, Mendenhall SK, Glassman SD, McGirt MJ. Accurately measuring the quality and effectiveness of lumbar surgery in registry efforts: determining the most valid and responsive instruments. The spine journal : official journal of the North American Spine Society. 2014;14(12):2885-91.

71. Parker SL, Mendenhall SK, Shau DN, Adogwa O, Anderson WN, Devin CJ, et al. Minimum clinically important difference in pain, disability, and quality of life after neural decompression and fusion for same-level recurrent lumbar stenosis: understanding clinical versus statistical significance. Journal of neurosurgery Spine. 2012;16(5):471-8.

72. Solberg T, Johnsen LG, Nygaard OP, Grotle M. Can we define success criteria for lumbar disc surgery? : estimates for a substantial amount of improvement in core outcome measures. Acta orthopaedica. 2013;84(2):196-201.

73. van der Roer N, Ostelo RW, Bekkering GE, van Tulder MW, de Vet HC. Minimal clinically important change for pain intensity, functional status, and general health status in patients with nonspecific low back pain. Spine. 2006;31(5):578-82.

74. Johnsen LG, Hellum C, Nygaard OP, Storheim K, Brox JI, Rossvoll I, et al. Comparison of the SF6D, the EQ5D, and the oswestry disability index in patients with chronic low back pain and degenerative disc disease. BMC Musculoskelet Disord. 2013;14:148.

75. Ware JE, Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Medical care. 1992;30(6):473-83.

76. The WHO Health Promotion Glossary Geneva: https://www.who.int/healthpromotion/about/HPG/en/; 1998 [updated 2006.

77. Chapman JR, Norvell DC, Hermsmeyer JT, Bransford RJ, DeVine J, McGirt MJ, et al. Evaluating common outcomes for measuring treatment success for chronic low back pain. Spine. 2011;36(21 Suppl):S54-68.

78. Mätning med RAND-36/SF-36 och EQ-5D RegisterCentrumSydOst; 2012 [Available from: http://rcso.se/wp-content/uploads/2018/02/M%C3%A4tning-med-RAND-36_SF-36-och-EQ-5D.pdf.

79. Sullivan M, Karlsson J, Taft C. SF-36 Hälsoenkät. Svensk Manual och Tolkningsguide : SF-36 Health Survey: Swedish Manual and Interpretation Guide: Sahlgrenska University Hospital Göteborg; 2002.

80. Persson LO, Karlsson J, Bengtsson C, Steen B, Sullivan M. The Swedish SF-36 Health Survey II. Evaluation of clinical validity: results from population studies of elderly and women in Gothenborg. Journal of clinical epidemiology. 1998;51(11):1095-103.

81. Sullivan M, Karlsson J. The Swedish SF-36 Health Survey III. Evaluation of criterion-based validity: results from normative population. Journal of clinical epidemiology. 1998;51(11):1105-13.

82. Sullivan M, Karlsson J, Ware JE, Jr. The Swedish SF-36 Health Survey--I. Evaluation of data quality, scaling assumptions, reliability and construct validity across general populations in Sweden. Social science & medicine (1982). 1995;41(10):1349-58.

83. Roland M, Fairbank J. The Roland-Morris Disability Questionnaire and the Oswestry Disability Questionnaire. Spine. 2000;25(24):3115-24.

84. Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. Physiotherapy. 1980;66(8):271-3.

85. Fairbank JC, Pynsent PB. The Oswestry Disability Index. Spine. 2000;25(22):2940-52; discussion 52.

86. Mannion AF, Junge A, Fairbank JCT, Dvorak J, Grob D. Development of a German version of the Oswestry Disability Index. Part 1: cross-cultural adaptation, reliability, and validity. European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society. 2006;15(1):55-65.

87. Saltychev M, Mattie R, McCormick Z, Barlund E, Laimi K. Psychometric properties of the Oswestry Disability Index. International journal of rehabilitation research Internationale Zeitschrift fur Rehabilitationsforschung Revue internationale de recherches de readaptation. 2017;40(3):202-8.

88. Brodke DS, Goz V, Lawrence BD, Spiker WR, Neese A, Hung M. Oswestry Disability Index: a psychometric analysis with 1,610 patients. The spine journal: official journal of the North American Spine Society. 2017;17(3):321-7.

89. Gabel CP, Cuesta-Vargas A, Qian M, Vengust R, Berlemann U, Aghayev E, et al. The Oswestry Disability Index, confirmatory factor analysis in a sample of 35,263 verifies a one-factor structure but practicality issues remain. European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society. 2017;26(8):2007-13.

90. Miekisiak G, Kollataj M, Dobrogowski J, Kloc W, Libionka W, Banach M, et al. Validation and cross-cultural adaptation of the Polish version of the Oswestry Disability Index. Spine. 2013;38(4):E237-43.

91. Gronblad M, Hupli M, Wennerstrand P, Jarvinen E, Lukinmaa A, Kouri JP, et al. Intercorrelation and test-retest reliability of the Pain Disability Index (PDI) and the Oswestry Disability Questionnaire (ODQ) and their correlation with pain intensity in low back pain patients. The Clinical journal of pain. 1993;9(3):189-95.

92. Glassman SD, Copay AG, Berven SH, Polly DW, Subach BR, Carreon LY. Defining substantial clinical benefit following lumbar spine arthrodesis. The Journal of bone and joint surgery American volume. 2008;90(9):1839-47.

93. Copay AG, Glassman SD, Subach BR, Berven S, Schuler TC, Carreon LY. 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. The spine journal : official journal of the North American Spine Society. 2008;8(6):968-74.

94. Hagg O, Fritzell P, Oden A, Nordwall A. Simplifying outcome measurement: evaluation of instruments for measuring outcome after fusion surgery for chronic low back pain. Spine. 2002;27(11):1213-22.

95. Hagg O, Fritzell P, Nordwall A. The clinical importance of changes in outcome scores after treatment for chronic low back pain. European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society. 2003;12(1):12-20.

96. Huskisson EC. Measurement of pain. Lancet (London, England). 1974;2(7889):1127-31.

97. Chiarotto A, Maxwell LJ, Ostelo RW, Boers M, Tugwell P, Terwee CB. Measurement Properties of Visual Analogue Scale, Numeric Rating Scale, and Pain Severity Subscale of the Brief Pain Inventory in Patients With Low Back Pain: A Systematic Review. The journal of pain : official journal of the American Pain Society. 2018.

98. Ostelo RW, de Vet HC. Clinically important outcomes in low back pain. Best practice & research Clinical rheumatology. 2005;19(4):593-607.

99. Williamson A, Hoggart B. Pain: a review of three commonly used pain rating scales. Journal of clinical nursing. 2005;14(7):798-804.

100. Karcioglu O, Topacoglu H, Dikme O, Dikme O. A systematic review of the pain scales in adults: Which to use? The American journal of emergency medicine. 2018;36(4):707-14.

101. Hawker GA, Mian S, Kendzerska T, French M. Measures of adult pain: Visual Analog Scale for Pain (VAS Pain), Numeric Rating Scale for Pain (NRS Pain), McGill Pain Questionnaire (MPQ), Short-Form McGill Pain Questionnaire (SF-MPQ), Chronic Pain Grade Scale (CPGS), Short Form-36 Bodily Pain Scale (SF-36 BPS), and Measure of Intermittent and Constant Osteoarthritis Pain (ICOAP). Arthritis care & research. 2011;63 Suppl 11:S240-52.

102. Graves C, Meyer S, Knightly J, Glassman S. Quality in Spine Surgery. Neurosurgery. 2018;82(2):136-41.

103. Zanoli G, Nilsson LT, Stromqvist B. Reliability of the prospective data collection protocol of the Swedish Spine Register: test-retest analysis of 119 patients. Acta orthopaedica. 2006;77(4):662-9.

104. Parker SL, Adogwa O, Mendenhall SK, Shau DN, Anderson WN, Cheng JS, et al. Determination of minimum clinically important difference (MCID) in pain, disability, and quality of life after revision fusion for symptomatic pseudoarthrosis. The spine journal : official journal of the North American Spine Society. 2012;12(12):1122-8.

105. Parker SL, Adogwa O, Paul AR, Anderson WN, Aaronson O, Cheng JS, et al. Utility of minimum clinically important difference in assessing pain, disability, and health state after transforaminal lumbar interbody fusion for degenerative lumbar spondylolisthesis. Journal of neurosurgery Spine. 2011;14(5):598-604.

106. Cheng T, Gerdhem P. Outcome of surgery for degenerative lumbar scoliosis: an observational study using the Swedish Spine register. European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society. 2018;27(3):622-9.

107. Endler P, Ekman P, Moller H, Gerdhem P. Outcomes of Posterolateral Fusion with and without Instrumentation and of Interbody Fusion for Isthmic Spondylolisthesis: A Prospective Study. The Journal of bone and joint surgery American volume. 2017;99(9):743-52.

108. Fritzell P, Knutsson B, Sanden B, Stromqvist B, Hagg O. Recurrent Versus Primary Lumbar Disc Herniation Surgery: Patient-reported Outcomes in the Swedish Spine Register Swespine. Clinical orthopaedics and related research. 2015;473(6):1978-84.

109. Lauridsen HH, Hartvigsen J, Korsholm L, Grunnet-Nilsson N, Manniche C. Choice of external criteria in back pain research: Does it matter? Recommendations based on analysis of responsiveness. Pain. 2007;131(1-2):112-20.

110. Manary MP, Boulding W, Staelin R, Glickman SW. The patient experience and health outcomes. The New England journal of medicine. 2013;368(3):201-3.

111. Adogwa O, Elsamadicy AA, Han JL, Cheng J, Karikari I, Bagley CA. Do measures of surgical effectiveness at 1 year after lumbar spine surgery accurately predict 2-year outcomes? Journal of neurosurgery Spine. 2016;25(6):689-96.

112. Fekete T, Loibl M, Jeszenszky D, Haschtmann D, Banczerowski P, Kleinstück F, et al. How does patient-rated outcome change over time following the surgical treatment of degenerative disorders of the thoracolumbar spine? European Spine Journal. 2018;27(3):700-8.

113. McKnight PE, McKnight KM, Sidani S, Figueredo AJ. Missing data: A gentle introduction: Guilford Press; 2007.

114. Li F, Mealli F, Rubin DB. A Conversation with Donald B. Rubin. Statistical Science. 2014;29(3):439-57.

115. Little RJA. Statistical analysis with missing data. Rubin DB, editor. New York: New York : Wiley; 1987.

116. Little RJ, Rubin DB. Statistical analysis with missing data: John Wiley & Sons; 2019.

117. Biering K, Hjollund NH, Frydenberg M. Using multiple imputation to deal with missing data and attrition in longitudinal studies with repeated measures of patient-reported outcomes. Clin Epidemiol. 2015;7:91-106.

118. Damen NL, Versteeg H, Serruys PW, van Geuns R-JM, van Domburg RT, Pedersen SS, et al. Cardiac patients who completed a longitudinal psychosocial study had a different clinical and psychosocial baseline profile than patients who dropped out prematurely. Eur J Prev Cardiol. 2015;22(2):196-9.

119. de Graaf R, Bijl RV, Smit F, Ravelli A, Vollebergh WA. Psychiatric and sociodemographic predictors of attrition in a longitudinal study: The Netherlands Mental Health Survey and Incidence Study (NEMESIS). Am J Epidemiol. 2000;152(11):1039-47.

120. Etter JF, Perneger TV. Analysis of non-response bias in a mailed health survey. Journal of clinical epidemiology. 1997;50(10):1123-8.

121. Powers J, Tavener M, Graves A, Loxton D. Loss to follow-up was used to estimate bias in a longitudinal study: a new approach. Journal of clinical epidemiology. 2015;68(8):870-6.

122. Tambs K, Rønning T, Prescott CA, Kendler KS, Reichborn-Kjennerud T, Torgersen S, et al. The Norwegian Institute of Public Health twin study of mental health: examining recruitment and attrition bias. Twin research and human genetics : the official journal of the International Society for Twin Studies. 2009;12(2):158-68.

123. Schmidt CO, Raspe H, Pfingsten M, Hasenbring M, Basler HD, Eich W, et al. Does attrition bias longitudinal population-based studies on back pain? European journal of pain (London, England). 2011;15(1):84-91.

124. Juto H, Gartner Nilsson M, Moller M, Wennergren D, Morberg P. Evaluating non-responders of a survey in the Swedish fracture register: no indication of different functional result. BMC Musculoskelet Disord. 2017;18(1):278.

125. Solberg TK, Sorlie A, Sjaavik K, Nygaard OP, Ingebrigtsen T. Would loss to follow-up bias the outcome evaluation of patients operated for degenerative disorders of the lumbar spine? Acta orthopaedica. 2011;82(1):56-63.

126. Norquist BM, Goldberg BA, Matsen FA, 3rd. Challenges in evaluating patients lost to follow-up in clinical studies of rotator cuff tears. The Journal of bone and joint surgery American volume. 2000;82(6):838-42.

127. Murray DW, Britton AR, Bulstrode CJ. Loss to follow-up matters. The Journal of bone and joint surgery British volume. 1997;79(2):254-7.

128. MacDowall A, Heary RF, Holy M, Lindhagen L, Olerud C. Posterior foraminotomy versus anterior decompression and fusion in patients with cervical degenerative disc disease with radiculopathy: up to 5 years of outcome from the national Swedish Spine Register. Journal of neurosurgery Spine. 2019:1-9.

129. Statens beredning för medicinsk u. Ont i ryggen, ont i nacken en evidensbaserad kunskapssammanställning : sammanfattning och slutsatser. Stockholm: Stockholm : SBU; 2000.

130. Bogduk N. On the definitions and physiology of back pain, referred pain, and radicular pain. Pain. 2009;147(1-3):17-9.

131. Hadjipavlou AG, Tzermiadianos MN, Bogduk N, Zindrick MR. The pathophysiology of disc degeneration: a critical review. The Journal of bone and joint surgery British volume. 2008;90(10):1261-70.

132. Niosi CA, Oxland TR. Degenerative mechanics of the lumbar spine. The spine journal : official journal of the North American Spine Society. 2004;4(6 Suppl):202s-8s.

133. van Tulder MW, Assendelft WJ, Koes BW, Bouter LM. Spinal radiographic findings and nonspecific low back pain. A systematic review of observational studies. Spine. 1997;22(4):427-34.

134. Herkowitz HN, Dvorák J, Bell GR, Nordin M, Grob D, Herkowitz HN, et al. Lumbar Spine : Official Publication of the International Society for the Study of the Lumbar Spine. Philadelphia: Philadelphia: Wolters Kluwer Health; 2004.

135. Machado GC, Ferreira PH, Yoo RI, Harris IA, Pinheiro MB, Koes BW, et al. Surgical options for lumbar spinal stenosis. The Cochrane database of systematic reviews. 2016;11:Cd012421.

136. Statens beredning för medicinsk och social u. Ont i ryggen, ont i nacken en evidensbaserad kunskapssammanställning. Vol. 1. Stockholm: Stockholm : Statens beredning för medicinsk utvärdering; 2000.

137. Airaksinen O, Brox JI, Cedraschi C, Hildebrandt J, Klaber-Moffett J, Kovacs F, et al. Chapter 4. European guidelines for the management of chronic nonspecific low back pain. European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society. 2006;15 Suppl 2:S192-300.

138. Hagg O, Fritzell P, Nordwall A. Characteristics of patients with chronic low back pain selected for surgery: a comparison with the general population reported from the Swedish lumbar spine study. Spine. 2002;27(11):1223-31.

139. Phillips FM, Slosar PJ, Youssef JA, Andersson G, Papatheofanis F. Lumbar spine fusion for chronic low back pain due to degenerative disc disease: a systematic review. Spine. 2013;38(7):E409-22.

140. Laustsen AF, Bech-Azeddine R. Do Modic changes have an impact on clinical outcome in lumbar spine surgery? A systematic literature review. European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society. 2016;25(11):3735-45.

141. Udby PM, Bendix T, Ohrt-Nissen S, Lassen MR, Sørensen JS, Brorson S, et al. Modic Changes Are Not Associated With Long-term Pain and Disability: A Cohort Study With 13-year Follow-up. Spine. 2019;44(17):1186-92.

142. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. Journal of clinical epidemiology. 1999;52(9):861-73.

143. Polit DF. Getting serious about test-retest reliability: a critique of retest research and some recommendations. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation. 2014;23(6):1713-20.

144. Värdebaserad uppföljning av kirurgisk behandling vid diskbråck och spinal stenos-analys från framtagande av nya uppföljningssystem: Sveus; 2015 [Available from: https://www.sveus.se/documents/files/Sveus-Rygg_Webb.pdf.]

145. Read CB, Banks DL, Kotz S. Encyclopedia of statistical sciences. [Update/edition] ed. New York: New York : John Wiley & sons; 1997.

146. Altman DG. Practical statistics for medical research. London: London : Chapman and Hall; 1991.

147. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. Journal of chronic diseases. 1986;39(11):897-906.

148. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of chiropractic medicine. 2016;15(2):155-63.

149. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological assessment. 1994;6(4):284.

150. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res. 2005;19(1):231-40.

151. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.

152. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. Lancet (London, England). 2017;390(10092):415-23.

153. Lee CK, Hansen HT, Weiss AB. Developmental lumbar spinal stenosis. Pathology and surgical treatment. Spine. 1978;3(3):246-55.

154. Finkelstein JA, Schwartz CE. Patient-reported outcomes in spine surgery: past, current, and future directions. Journal of neurosurgery Spine. 2019;31(2):155-64.

155. Hollenberg AM, Bernstein DN, Baldwin AL, Beltejar M-J, Rubery PT, Mesfin A. Trends and Characteristics of Spine Research From 2006 to 2015: A Review of Spine Articles in a High Impact General Orthopedic Journal. Spine. 2020;45(2):141-7.

156. Chung AS, Copay AG, Olmscheid N, Campbell D, Walker JB, Chutkan N. Minimum Clinically Important Difference: Current Trends in the Spine Literature. Spine. 2017;42(14):1096-105.

157. Dworkin RH, Turk DC, Wyrwich KW, Beaton D, Cleeland CS, Farrar JT, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. The journal of pain : official journal of the American Pain Society. 2008;9(2):105-21.

158. Chiarotto A, Ostelo RW, Boers M, Terwee CB. A systematic review highlights the need to investigate the content validity of patient-reported outcome measures for physical functioning in patients with low back pain. Journal of clinical epidemiology. 2018;95:73-93.

159. Copay AG, Eyberg B, Chung AS, Zurcher KS, Chutkan N, Spangehl MJ. Minimum Clinically Important Difference: Current Trends in the Orthopaedic Literature, Part II: Lower Extremity: A Systematic Review. JBJS Rev. 2018;6(9):e2-e.

160. Cook CE. Clinimetrics Corner: The Minimal Clinically Important Change Score (MCID): A Necessary Pretense. The Journal of manual & manipulative

161. Draak THP, de Greef BTA, Faber CG, Merkies ISJ, PeriNom Ssg. The minimum clinically important difference: which direction to take. Eur J Neurol. 2019;26(6):850-5.

162. Endler P, Ekman P, Hellström F, Möller H, Gerdhem P. Minor effect of loss to follow-up on outcome interpretation in the Swedish spine register. European Spine Journal. 2019.

163. Haefeli M, Elfering A, Aebi M, Freeman BJC, Fritzell P, Guimaraes Consciencia J, et al. What comprises a good outcome in spinal surgery? A preliminary survey among spine surgeons of the SSE and European spine patients. European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society. 2008;17(1):104-16.

164. Staartjes VE, Siccoli A, de Wispelaere MP, Schröder ML. Patient-reported outcomes unbiased by length of follow-up after lumbar degenerative spine surgery: Do we need 2 years of follow-up? The spine journal : official journal of the North American Spine Society. 2019;19(4):637-44.

165. Glassman SD, Schwab F, Bridwell KH, Shaffrey C, Horton W, Hu S. Do 1-year outcomes predict 2-year outcomes for adult deformity surgery? The spine journal : official journal of the North American Spine Society. 2009;9(4):317-22.

166. Rogerson A, Aidlen J, Jenis LG. Persistent radiculopathy after surgical treatment for lumbar disc herniation: causes and treatment options. International Orthopaedics. 2019;43(4):969-73.

167. Burström K, Johannesson M, Rehnberg C. Deteriorating health status in Stockholm 1998-2002: results from repeated population surveys using the EQ-5D. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation. 2007;16(9):1547-53.

168. Tonosu J, Takeshita K, Hara N, Matsudaira K, Kato S, Masuda K, et al. The normative score and the cut-off value of the Oswestry Disability Index (ODI). European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society. 2012;21(8):1596-602.

169. Eneqvist T, Bülow E, Nemes S, Brisby H, Garellick G, Fritzell P, et al. Patients with a previous total hip replacement experience less reduction of back pain following lumbar back surgery. J Orthop Res. 2018;36(9):2484-90.

170. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. Journal of clinical epidemiology. 2010;63(11):1179-94.

171. Cook KF, Cella D, Reeve BB. PRO-Bookmarking to Estimate Clinical Thresholds for Patient-reported Symptoms and Function. Medical care. 2019;57 Suppl 5 Suppl 1:S13-S7.

172. Alonso J, Bartlett SJ, Rose M, Aaronson NK, Chaplin JE, Efficace F, et al. The case for an international patient-reported outcomes measurement information system (PROMIS®) initiative. Health and quality of life outcomes. 2013;11:210-.