Cover Letter

Statistical Analysis Plan

Lumbar degenerative spondylolisthesis: Is only decompression good enough?
A prospective randomized clinical multi- institutional non-inferiority trial
The NORDSTEN-DS trial
SAP Version 3.0
NCT02051374

Participating authors

Ivar Magne Austevoll MD (Corresponding author) (1), E-mail: imau@helse-bergen.no

Erland Hermansen, PhD, E-mail: erland.hermansen@helse-bergen.no

Morten Fagerland PhD, (2), E-mail: morten.fagerland@medisin.uio.no

Frode Rekeland, MD, E-mail: frode.rekeland@helse-bergen.no

Tore Solberg MD, PhD, E-mail: tsdrugiss2@gmail.com

Kjersti Storheim PT, Professor, E-mail: UXSKJE@ous-hf.no

Oliver Grundnes MD PhD, E-mail: oliver.grundnes@ahus.no

Jens Ivar Brox MD, Professor, PhD, E-mail: j.i.brox@medisin.uio.no

Kari Indrekvam, MD, PhD, E-mail: kari.indrekvam@helse-bergen.no

Christian Hellum MD, PhD, (3) E-mail: christian.hellum@ous-hf.no

1. Kysthospitalet in Hagevik, Orthopedic Clinic, Haukeland
University Hospital, Bergen, Norway
Address: Hagaviksbakken 25, 5217 Hagevik

2. Oslo Centre for Biostatistics and Epidemiology, Research Support Services, Oslo University Hospital

3. Division of Orthopaedic Surgery, Oslo University Hospital, Norway

Statistical Analysis Plan

Lumbar degenerative spondylolisthesis: Is only decompression good enough?
A prospective randomized clinical multi- institutional non-inferiority trial
The NORDSTEN-DS trial
SAP Version 3.0
Date: January 15, 2020

1. Administrative Information

This document is a supplement to the NORDSTEN-DS protocol; "Decompression alone versus decompression with instrumental fusion; The NORDSTEN Degenerative Spondylolisthesis Trial (NORDSTEN-DS); Study protocol for a randomized controlled trial" [1].

Trial Registration Number

The trial is registered in ClinicalTrials.gov, first received January 10, 2014 (Identifier: NCT02051374).

Scientific board

Ivar Magne Austevoll MD (1,3), E-mail: imau@helse-bergen.no
Erland Hermansen MD (1,2,3), E-mail: erland.hermansen@helse-bergen.no
Morten Fagerland PhD (12), E-mail: morten.fagerland@medisin.uio.no
Frode Rekeland MD (2,3), E-mail: frode.rekeland@helse-bergen.no
Tore Solberg MD PhD (5,6,7), E-mail: tore.solberg@unn.no
Kjersti Storheim PT PhD (8), E-mail: kjersti.storheim@medisin.uio.no
Oliver Grundnes MD PhD (9), E-mail: oliver.grundnes@ahus.no
Jens Ivar Brox MD Professor (10), E-mail: j.i.brox@medisin.uio.no

Kari Indrekvam MD PhD (2,3), E-mail: kari.indrekvam@helse-bergen.no
Christian Hellum MD PhD (10), E-mail: christian.hellum@ous-hf.no

Sponsor/ Funding

Contributors to current SAP

Ivar Magne Austevoll MD
Erland Hermansen PhD
Morten Fagerland PhD
Frode Rekeland MD
Tore Solberg MD PhD
Kjersti Storheim PT Professor
Jens Ivar Brox MD Professor
Kari Indrekvam MD PhD
Christian Hellum MD PhD

Signature Page

I hereby declare that I have reviewed and approved the Statistical Analysis Plan:

Principal investigator

Signature _____ *Ivar M. Austevoll* _____

Ivar Magne Austevoll[1]

Date _____ JANUARY 15, 2020 _____

Study chair

Signature _____

Christian Hellum[2]

Date _____ January 14, 2020. _____

Responsible Biostatician

Signature _____

Morten Wang Fagerland[3]

Date _____

[1] Kysthospitalet in Hagevik, Orthopedic Clinic, Haukeland

University Hospital, Bergen, Norway

[2] Division of Orthopaedic Surgery, Oslo University Hospital, Norway

[3] Oslo Centre for Biostatistics and Epidemiology, Research Support Services, Oslo University

Hospital

Table of contents

Abbreviations and definitions

DA	Decompression alone
DF	Decompression with instrumental fusion
DS	Degenerative spondylolisthesis
EQ-5D	EuroQol 5-dimensional questionnaire utility index
GPE	Global perceived effect
ITT	Intention to treat
LSS	Lumbar spinal stenosis
NNT	Number needed to treat
NRS	Numeric rating scale
MI	Multiple imoutation
ODI	Oswestry disability index
PROM	Patient reported outcome
SAP	Statistical Analysis Plan
ZCQ	Zurich claudication questionnaire –score

Follow-up score = time-point value for the actual score

Change score = time point value – baseline value

Percentage change score = [(time-point value – baseline value) / baseline value] *100%

Justification for SAP revisions

The main context of the present SAP is in accordance with the primary SAP, the published study protocol, and the first registration in CinicalTrials.gov January 10, 2014 (IdentiferNCT02051374). The following clarifications are declared in updates to clincaltrials.gov:

1. Update September , 2014

Four exclusion criteria was recorded in the original protocol but were not recorded in first registration in Clinical trials. However, these criteria were used for enrolling patients from the start of the study. The criteria added to the update were:

1. Previous surgery in the level of spondylolisthesis; 2. Lumbosacral scoliosis of more than 20 degrees verified on AP-view; 3. Distinct symptoms in one or both legs due to other diseases, e.g. polynevropathy, vascular claudication or osteoarthtritis; 4. Radicular

pain due to a MRI-verified foraminal stenosis in the slipped level, with deformation of the nerve root because of a bony narrowing in the vertical direction.

2. Update January, 2016

From the start of inclusion (April 15, 2014) patients with ODI scores less than 25 were excluded. Due to experiences from participating surgeons that a considerable part of the patients were excluded due to ODI less than 25, even if their complaints from leg and back did justify an operation. To enhance the external validity of the study, the steering committee decided that from date 29th August 2015, the patients should not be excluded due to ODS- score lower than 25.

3. Update September, 2017

In accordance with the study "Follow-up score, change score or percentage change score for determining clinical important outcome following surgery?" the criteria for a clinical important outcome assessed by the primary outcome (ODI) was recorded.
We also changed the plan for handling missing data in the primary outcome. Instead of using Multiple imputation we planned different 'worse case – best case' imputation scenarios for sensitivity analysis.

Update January 2020

After thorough discussions in the study group, and review of current literature, we decided to reintroduce the original planned method for handling missing data due to 'lost to follow-up'. Patients without measurement available for dichotomizing into responder/non-responder (ODI, ZCQ, NRS leg pain, and NRS back pain) will receive two-year follow-up scores estimated with use of Multiple Imputation.
Statistical analyses:
We have modified the plan for the statistical analyses according to the final SAP. Relative efficacy will be evaluated by use of a Full Analysis Set (FAS-MI) and a Per Protocol Set (PPS). In the FAS-MI set missing scores necessary for the responder analyses will be imputed by use of Multiple imputation (MI). To recommend DA both the FAS-MI and the PPS analysis of the primary efficacy endpoint are required to show non-inferiority. Two sensitivity analysis will also be performed to evaluate relative efficacy; one using FAS including complete case analysis and one using FAS where missing values will be replaced with values at one year follow-up, if available.

To evaluate safety we will compare proportions of 'Substantially deteriorated' according to the GPE scale, 'Complications and side effects', 'Reoperations' and 'Volume of blood loss and blood transfusions'

In this update we have also recorded secondary outcome parameters described in the published original study protocol, but not previously recorded in Clinical trials. The last recorded outcomes are: 1) Duration of surgery; 2) Length of hospital stay; 3) Volume of blood loss and blood transfusions.

The study status has been changed during the study period. The inclusion period has been prolonged from the first anticipated time of December 2016 to the actual endpoint of patient inclusion due to December 31, 2017.

2 Introduction

Background and rationale

Degenerative spondylolisthesis (DS) is defined as a forward slippage of one vertebra over another without a disruption in the vertebral arch[2]. In most occasions the patients present symptoms related to a concomitant spinal stenosis, typically back and leg pain in supine position[3, 4].

Meta- analyses and systematic reviews have concluded with better clinical outcomes when decompression is combined with instrumented fusion [4-7]. A recently published randomized controlled trial (RCT) has supported this suggestion [8]. However, cohort studies [9-11] and another recently published RCT [12], have argued for the opposite conclusion; an additional fusion do not give superior clinical results when operating LDS.

The present document describes the planned statistical analysis plan for the NORDSTEN-DS study, a randomized controlled trial (RCT), comparing the efficacy for decompression alone (DA) and decompression with instrumented fusion (DF).

Objectives
Main Objective

6

The primary objective of the present study is to investigate whether the intervention-related difference in efficacy between decompression alone (DA) and decompression plus instrumented fusion (DF) 2 years after surgery, is large enough to justify the use of instrumentation. Our hypothesis is that decompression alone is non-inferior ("as good as") decompression with instrumented fusion, in treatment of spinal stenosis with degenerative spondylolisthesis.

Secondary objectives

Predictor analysis

To evaluate whether carefully selected radiological parameters and patient characteristics can be used to choose the most appropriate treatment; decompression alone or decompression with instrumented fusion. Detailed plan for the analysis is described in the published study protocol [1], and will not further be described in the present SAP.

Long- time follow up analysis

To investigate the clinical efficacy at long time follow-up. The statistical analyses plan for long-time follow-up data is similar to the analyses of two-year follow-up data.

Analysis of cost-effectiveness of the treatments at 2 years follow-up.

To investigate the cost-effectiveness of decompression alone relative to decompression with instrumented fusion. Details about this analysis will not be described in the present SAP.

3. Study Methods

Trial overview

The NORDSTEN-DS trial is a 1:1 block- randomized, controlled, multicenter, one-country, non- inferiority trial, with two parallel groups.

The study is one of three trials in the NORDSTEN- study, a Norwegian multicentre study on patients with lumbar spinal stenosis. In NORDSTEN-DS decompression without fusion will be compared to decompression with instrumented fusion.

Treatment groups

Decompression alone: The midline (i.e., the spinous process and the interspinous ligaments) will be preserved and one of the following techniques will be used: 1) unilateral laminotomy; 2) bilateral laminotomy; 3) unilateral laminotomy and crossover decompression. Magnifying devices (microscope or loupes) will be used.

Decompression and instrumented fusion: A decompression with or without preservation of the midline structures, and additional posterior pedicle screw instrumental fusion with or without an intervertebral cage. Magnifying devices (microscope or loupes) will be used.

The SPIRIT checklist [13] has been used as a template for the Study protocol [1].

The statistical analysis plan is prepared in accordance with guidelines for Statistical analysis plans in clinical trials [14].

The reporting of the trial will be based on an adapted Consolidated Standards of Reporting Trials (CONSORT) checklist for reporting non- inferiority trials[15].

The trial is monitored following the Helsinki Declaration, The International Conference on Harmonisation Guideline for Good Clinical Practice (ICH GCP) [16].

The protocol has been approved by the Norwegian Committee for Medical and Health Research Ethics Midt (2013/366).

Randomization

The computer generated randomization is block- permuted and centre- stratified. The system used for patient allocation was Medinsight - a registry tool developed for researchers and health professionals by Institute for Cancer Genetics and Informatics at Oslo University Hospital, Norway. The patients are randomized in a 1:1 ratio to one of two arms (randomly selected block size 4 and 6).

 After the patient has signed the informed consent form, the randomization is performed within 6 weeks before treatment. The computer generated randomisation procedure is concealed and administered by a central coordinator.

Details of block size, allocation sequence generation, and randomization, is unavailable to those who enrol patients or assign treatment.

Sample size

The sample size is computed by using the Blackwelder methodology for non- inferiority trials [17]. For the primary outcome, the study is designed to detect (β =0.2, α = 0.05) whether the

responder rate is less than 15 percentage lower for the decompression alone group compared to the instrumented fusion group. Considering these assumptions and adding 10% for possible dropouts, a total of 128 patients are required in each group.

Framework

Assessment of efficacy

For the main objective, the study is designed to establish the non-inferiority of decompression alone compared to decompression with an additional instrumented fusion with regard to the 'responder' rates. Test for non-inferiority will be performed for the primary efficacy outcome (ODI) and for responder rates assessed by the secondary outcomes NRS leg pain, NRS back pain and the Zürich Claudication Questionnaire.

*Rationale for non-inferiority testing*

Compared to decompression alone, the rate of decompression with instrumented fusion has increased significantly the last decades and has become the 'standard treatment' in many countries [18, 19]. Correspondingly, due to the more unfrequently use of decompression alone [18, 19] this technique is defined as the 'alternative method', although the method historically is the oldest method.

In generally terms a non-inferior trial intends to test whether an 'alternative treatment' is not inferior to the 'standard treatment' or 'control treatment'. In this study, we intend to test whether decompression alone is non-inferior to decompression with instrumented fusion. In other words we will investigate whether micro-decompression is 'as good as' or 'not unacceptably worse than' decompression plus instrumented fusion.

An important prerequisite for a non-inferiority trial is that the standard treatment, which the 'alternative' treatment is tested against, should be superior to placebo. It does not exist any papers that compare surgery and placebo (i.e., a sham procedure) in the treatment of DS. A modest grade of evidence exists for that decompression with instrumental fusion is better than non-surgical treatment [12, 41]. Another criterion for conducting a non-inferiority study is that the alternative treatment has some obvious advantages compared to the standard treatment. Decompression alone is associated with lower perioperative complications and lower hospital costs than decompression plus fusion [57, 67]. Decompression alone should therefore be advocated if clinical outcomes are not unacceptable worse than for decompression with instrumented fusion. Unfortunately, a clinically reliable margin for 'unacceptable worse' does not exist. Following thoroughly discussions in the study group, we

defined the margin of non-inferiority to be an absolute difference of 15% in the responder rate (proportion of patients with a clinically important improvement). This margin corresponds to a number needed to treat of seven patients (NNT = 100/15 = 6.67). That means that for accepting the null hypothesis, less than seven patients need instrumented fusion to achieve one extra responder. If more than seven patients are needed, we considered the advantages of decompression (e.g., less invasive and cheaper) to surpass the disadvantages with instrumented fusion (e.g., higher complication rate). This margin is in accordance with a prospective, randomized, multicenter Food and Drug Administration investigational trial comparing lumbar total disc replacement and lumbar fusion in patients [20]. The principles of testing non-inferior are illustrated in figure 2.

Figure 2. The 2 alternative results and conclusions for the primary outcome are illustrated. The bars indicate the difference in responder rate (DF minus DA) with 95% confidence interval limits.
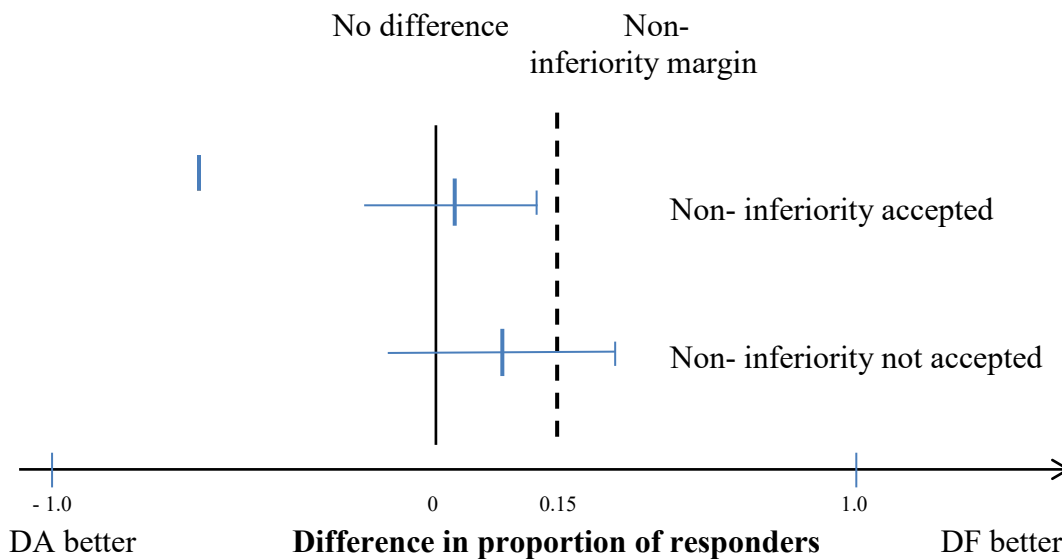
Blinding:

The treatment given is not blinded for the patients. For analysis and testing of the efficacy variables, the statistician will be blinded for treatment adherence.

Statistical interim analyses and stopping guidance

Due to ethical considerations in agreement with the Norwegian Committee for Medical and Health Research Ethics Midt, an interim analysis for safety was performed when 75 patients in

each group had completed the 12-month follow- up. If one of the proposed stop criteria were fulfilled the study would be terminated:

1. The proportion of patients needing reoperation due to any condition in the operated level(s) is statistically significantly ($p<0.05$) higher in one of the groups.
2. The proportion of responders in the DF group, assessed by the primary outcome measure, is higher than in the DA group by an amount of 0.20.

An independent statistician blinded for treatment adherence would perform the interim analysis. Only data on reoperations and on the primary outcome measure (ODI) would be available to the statistician. Following the analysis, the statistician would inform the steering committee, via the central coordinator whether the study can continue. Further information about the analysis would not be disclosed or available by anyone else than the independent statistician until the main analysis at 2-year follow-up.

Timing of final analyses

The data will be inaccessible to the research group until all available two-year follow-up participants has completed the 2-year questionnaire, and the study is declared adequate monitored according to principles of Good Clinical Practice (GCP). The Faculty of Research support, University of Oslo will declare this inaccessibleness and state which date the data has been accessible for the investigators. At that time, the primary and the secondary outcomes will be analysed.

Timing of outcome assessment

The study coordinators are responsible for the collection and administration of data at baseline and at 3-month follow-up. Data from 12-month 2-year, 5-year and 10-year follow-up is collected by the central coordinator at FORMI. All data will be stored at the Faculty of Research support, University of Oslo. Time schedule for assessment of data is presented in Table 2. For data assessed outside the time frame a "Note to file" will be recorded in the patients' Case Report Form (CRF).

Table 2. Time points at which the outcomes are measured for the NORDSTEN/DS trial

	Before operation	Hospital stay	3 months	12 months (±1 month)	2 years	5 years	10 years (±3 months)

		(±2 weeks)		(±2 months)	(±3 months)	
Demographics	X[1]					
Lifestyles	X[1]					
PROMs	X[1]	X	X	X	X	X
X-rays	X[2]			X	X	X
MRI scan	X[2]					
CT scan				X		
Operation data		X				
Data from hospital stay		X				
Complications, reoperations		X	X	X	X	X	X

Abbrevations: MRI= Magnetic resonance imaging, CT= Computed tomography, AP= anterior-posterior, PROMs= Patient reported outcome measurements,
[1] Maximum 6 weeks prior operation date
[2] Maximum 6 months prior operation date

4. STATISTICAL PRINCIPLES

A significance level of 5% will be used throughout. For all analyses a 95% CI will be estimated and reported.

Adherence and protocol deviations

The trial is monitored following the Helsinki Declaration, The International Conference on Harmonisation Guideline for Good Clinical Practice (ICH GCP) [50]. An independent monitor, without influence on the scientific work, will be responsible for the monitoring. Due to the non-regulated ICH GCP guideline for this trial (not including drug intervention) the risk and safety will be safeguarded at the same level as data quality. All informed consent forms will be checked and all registrations of serious events will be monitored. According to the monitoring plan selected variables will be checked. All hospitals will be visited regularly. Adapted versions of the 'Investigator's Site File (ISF)' and the 'Trial Master File (TMF)' will be checked for essential documents during the trial. Queries and deviations will be recorded and reported, and the coordinators at responsible hospitals have two months to send a written report with the required corrections to the monitor. All deviations from the protocol will subsequently be recorded at the 'Note to file form'. Recorded deviations will be presented in the final manuscript, tables or in Supplementary.

The patients have **major deviations from protocol if they**:

- have not received operative treatment in accordance with randomized allocation.

- have received operative treatment in accordance with randomized allocation and operated with a new operation at same level during the follow-up period.

- have not provided informed consent.

- have withdrew the informed consent and claimed their data withdrawn from analyses.

According to deviations to protocol the following analysis sets are defined:

Full Analysis Set (FAS): all randomised patients with primary operation according to the randomly assigned study treatment and with data on the primary outcome variable (ODI) at one or more time point.

Per Protocol Set (PPS): All randomized patients without major deviations from protocol and with data on the primary outcome variable at baseline and two-year follow-up.

5 Trial population

Sixteen Norwegian orthopedic and neurosurgical hospital departments participate in the study. Criteria for inclusion and exclusion are given in Table 2.

Table2

Inclusion criteria:	Exclusion criteria:
To be eligible for the study the participants must:	The participants will be excluded from the study if they:

|	
Be over 18 years of age.	Are not willing to give written consent.
Understand Norwegian language, spoken and written.	Are participating in another clinical trial that may interfere with this trial.
Have a spondylolisthesis, with a slip >=3 mm, verified on standing plain x-rays in lateral view.	Are ASA- grade > 3.
Have a spinal stenosis in the level of spondylolisthesis, shown on MRI, CT scan or myelogram.	Are older than 80 years.
Have clinical symptoms of spinal stenosis as neurogenic claudication or radiating pain into the lower limbs, not responding to at least 3 months of qualified conservative treatment.	Are not able to fully comply with the protocol, including treatment, follow-up or study procedures (psychosocially, mentally and physically).
Be able to give informed consent and to respond to the questionnaires.	Have cauda equina syndrome (bowel or bladder dysfunction) or fixed complete motor deficit.
	Have a slip >=3 mm in more than one level.
	Have an isthmic defect in pars interarticularis.
	Have a fracture or former fusion of the thoracolumbal region.
	Have had previous surgery in the level of spondylolisthesis.
	Have a lumbosacral scoliosis of more than 20 degrees verified on AP-view.
	Have distinct symptoms in one or both legs due to other diseases, e.g. polynevropathy, vascular claudication or osteoarthtritis.
	Have radicular pain due to a MRI-verified foraminal stenosis in the slipped level, with deformation of the nerve root because of a bony narrowing in the vertical direction.

MRI= Magnetic resonance imaging, CT= Computed tomography, AP= anterior- posterior, ASA = American Society of Anesthesiologists

All patients surgeons consider eligible at the participating hospitals will be recorded. The surgeon records the results of the screening at the Screening Form. The Screening Form consists of boxes where each box represents one particular inclusion and exclusion criteria. If all inclusion criteria and no exclusion criteria are checked, a patient will be randomised and included in the study. The check for eligibility and the result of the randomisation will be recorded in an Inclusion Form. Mandatory before inclusion is a signed consent form.

The result of the screening will be presented. A CONSORT flow chart is illustrated in Figure 1.

Figure 1
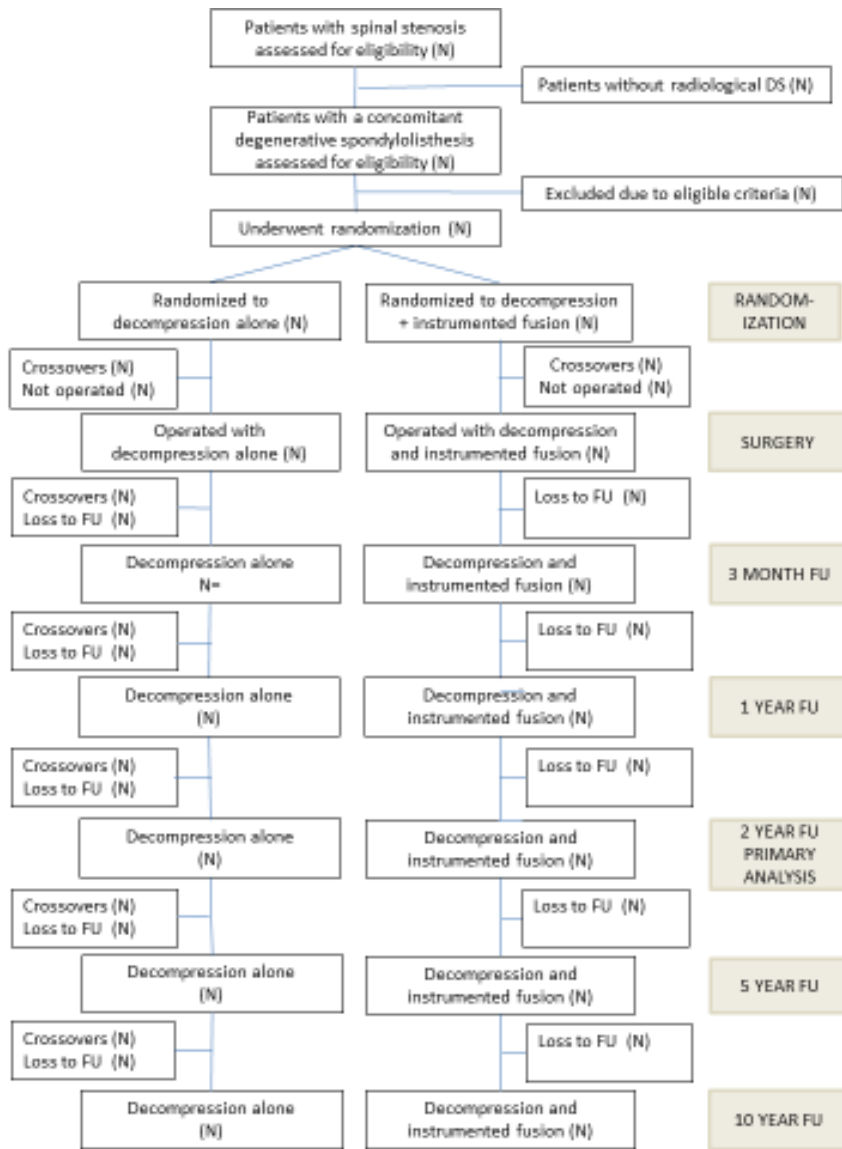
List of baseline data to be summarized
Demographics and lifestyles

- Age
- Gender
- Education (1-5), where '1' indicates primary/ junior high school and '5' indicate fulfilled degree of master.
- Marital status
- First language
- Smoking habitus
- Body Mass Index
- Former spinal surgery
- Duration of pain
- Use of analgesics

Comorbidities

American Society of Anesthesiologists (ASA) grade

Diagnosis assumed to be relevant for the operation: Rheumatoid arthritis, Morbus Bechterew, Other rheumatoid diseases, Cox- or Gonarthrosis, Neurological disease, Cerebrovascular disease, Coronary heart disease, Vascular claudication, Chronic lung disease, Cancer, Osteoporosis, Diabetes Mellitus, Other endocrine disease, Other.

Assessment of psychological distress

Hopkins symptom check list (HSCL-25) is a self- reports questionnaire for assessment of psychological variables, will be collected preoperatively solely. It includes emotional distress scores range from 1 to 4, with lower scores indicating less severe symptoms[21]. The scores will be used for descriptive presentation.

Baseline PROMs

- Oswestry disability index score

- Zürich Claudication Questionnaire (ZCQ)
- Numeric Rating Scale for leg pain

- Numeric Rating Scale for back pain

- EQ-5D score
- Hopkins symptom check list (HSCL-25)

Radiological parameters

- Degree of spondylolisthesis at standing x-rays in millimeter [22]
- Grading of spinal stenosis (Schizas A-D) [23]
- Segmental instability[22]
- Presence of foraminal stenosis [24]
- Amount of facet joint fluid [25]
- Orientation of the facet joint [26]
- Disc degeneration [27]
- Modic changes [28]
- Disc height in the level of olisthesis [29]
- Lumbal lordosis [30]

Surgical data

- Level with spondylolisthesis

- Number of level operated on
- Method used for decompression
- Method used for instrumented fusion

6. Analysis

Primary efficacy endpoint

The primary efficacy endpoint is the proportion of patients (the responder rate) with an improvement of Oswestry Disability Index[31] (ODI) V.2.0[32] of more than 30% from baseline to 2 year follow-up. Based on former studies [33, 34] and a study from The Norwegian Registry for Spine Surgery, an individual ODI improvement of 30% or more from baseline to follow-up has been chosen as the cut-off for being a responder [35]. The difference in the proportions of responders will be estimated with the Newcombe hybrid score CI [36].The null hypothesis (H0) is that the responder rate in the decompression alone group (n_{DA}) is at least 0.15 (15 percentage) lower (the non-inferiority margin) than for the instrumented fusion group (n_{DF}); **H0: $n_{DF} - n_{DA} \geq 0.15$** [1, 20]. H0 will be tested by forming a 95% confidence interval (CI) for the between-group difference in responder rate (responder rate DF minus responder rate DA) and will be rejected if the upper bound of the CI is below the non-inferiority margin of 0.15. The alternative hypothesis is that the decompression alone group is non-inferior, i.e., as good as, decompression with instrumented fusion; **H1: H0: $n_{DF} - n_{DA} < 0.15$**.

The responder rates, the difference in responder rates and the number needed to treat (NNT= 100 divided by the percentage difference in proportion of reponders) and the corresponding 95% CIs will be presented and interpreted.

Secondary key endpoints

All results from analysis of secondary endpoints will be will be supplements to analyses of the primary efficacy endpoint, and discussed and interpreted accordingly.

Mean scores at the Oswestry Disability Index

For the continuous scores we will use linear mixed models to estimate mean differences in change scores and follow-up scores between the treatment groups were all follow-up measurements from inclusion to 2-year follow-up will be included. Because most change from

baseline is expected to occur the first three months, the time development in the linear mixed models will be modelled as piecewise linear, with a knot at 3 months. The models will include fixed effects for treatment group, time, treatment group x time interaction, and center (stratification factor in the randomization). A random intercept will be used, and – if possible – a random effect for treatment group. Results of the analysis will be presented and interpreted.

Responder rates and mean scores for Zürich Claudication Questionnaire (ZCQ)

ZCQ is a self-completed disorder-specific functional score consisting of three domains: symptom severity, physical function and patient satisfaction. Non-inferiority will be tested with use of the Newcombe hybrid score CI for the difference in responder rate. Criteria for 'success' defined by Tulli et al. [37] will be used for dichotomizing patients into responders and non-responders. The hypothesis testing and the non-inferiority margin will be equivalent to testing of the primary efficacy endpoint (ODI).

We will use linear mixed models (see above) to estimate mean differences in change scores and follow-up scores between the treatment groups were all follow-up measurements from inclusion to 2-year follow-up will be included.

Results of the analysis will be presented and interpreted.

Responder rates and mean scores for Numeric Rating Scale (NRS) for back- and leg pain

NRS is a PRO that assesses self-reported pain the patients experienced in the last week from 0 (no pain) to 10 (the worst pain imaginable). Non-inferiority will be tested with use of the Newcombe hybrid score CI. The hypothesis testing and the non-inferiority margin will be equivalent to testing of the primary efficacy endpoint (ODI). Criteria for 'success', derived from the Norwegian registry for Spine Surgery, will be used for dichotomizing patients into responders and non-responders. The thresholds are defined as a ≥40% reduction in the NRS leg pain and a ≥33% reduction in NRS back pain.

We will use linear mixed models (see above) to estimate mean differences in change scores and follow-up scores between the treatment groups were all follow-up measurements from inclusion to 2-year follow-up will be included.

Results of the analysis will be presented and interpreted

Euroqol 5 dimensional descriptive system (EQ-5D)

EQ-5D is a generic PROM that is self-completed and comprises 5 questions relating to mobility, self-care, usual activity, pain/discomfort, and anxiety/depression [38]. Each question has a three-point descriptive scale where 3 is the worst possible health. The scale ranges from -0.59 to 1.0, with higher scores indicating better quality of life. We will use linear mixed models (see above) to estimate the mean differences in change scores and follow-up scores between the

18

treatment groups were all follow-up measurements from inclusion to 2-year follow-up will be included.

Results of the analysis will be presented and interpreted

Global Perceived Effect (GPE) scale

Patient- rated satisfaction with treatment outcome will be assessed using single question with seven-point descriptive scaling with the answers 'completely recovered', 'much improved', 'slightly improved', 'unchanged', 'slightly worse', 'much worse' and 'worse than ever' [39]. The GPE responses 'completely recovered' and 'much improved' will be categorized as 'substantially improved' and the responses 'much worse' and 'worse than ever' as 'substantially deteriorated'. The treatment groups will be compared with use of the Fisher Mid-P test, and the results will be presented and interpreted.

Duration of operation

Duration of surgery from skin incision to closure in minutes will be recorded. Student T-test with adjustment for unequal variance will estimate between-group differences, including 95% CI. The assumption of normal distribution will be checked by visual inspection of histograms and descriptive statistics. If major deviation from normally distributed data, median regression will be used for estimation of differences in medians, including 95% CI for the difference. Descriptive statistics and test statistics for between-group differences will be presented.

Length of hospital stay

Length of hospital stays after operation (days) will be analyzed and presented following similar statistical methods as for 'duration of surgery'.

Complications and side effects

An independent research coordinator will record complications and adverse effects (Table 3), consecutively during the hospital stay. Complications during the follow-up period will be recorded by study coordinators at the CRF1- form. Group differences in continuous variables will be tested following similar statistical methods as for 'duration of surgery'. Categorical secondary outcomes will be analysed with the Fisher mid-P test and the Newcombe hybrid score CI. All results regarding complications and adverse effects will be presented and interpreted.

Table 3. Complications and side effects to be recorded during the hospital stay

Perioperative	Postoperative

Dural tear	Liquor leakage
Nerve root lesion	Superficial infection
Operated on the wrong side	Neurological deterioration
Operated on the wrong level	Hematoma requiring reoperation
Cardiopulmonary complications	Infection requiring reoperation
Anaphylactic reaction	Thromboembolic episode
Blood transfusion	Cardiopulmonary complications
Other	Urological complication
	Wrong level/side revealed postoperatively
	Blood transfusion
	Other

Reoperations

Any new surgery in the lumbosacral column from the time of the index operation to follow-up will be recorded at the CRF-form. We will distinguish between an operation at the same level as the primary operation and an operation in a new segment. The reasons for a secondary operation will be recorded and presented. Group differences will be tested as for 'Complications and side effects' and presented and interpreted.

Analysis methods

All analyses described in this SAP will be applied and interpreted according to the description. All methods are considered *a priori* analyses in that they have been defined in ClinicalTrials.gov, the protocol and/or this SAP will be   If deviations between SAP, published protocol and ClinicalTrials.gov the SAP will be applied throughout. All *post hoc* analyses will be identified as such when interpreting results.

Descriptive statistics, including measures of centrality and variability, will be used to describe the baseline characteristics of the two treatment groups. The assumption of normal distribution will be checked by visual inspection of histograms.

Decision rules

Efficacy

The conclusion will be based on the efficacy analysis of the Full Analysis Set (FAS-MI), and the Per Protocol Set (PPS) at two-year follow-up [40]. In the FAS-MI set missing scores necessary for dichotomising patients into responders/non-responders will be imputed by use of Multiple imputation (MI). To recommend DA both the FAS-MI and the PPS analysis of the primary efficacy endpoint are required to show non-inferiority.

In addition we will perform two sensitivity analyses. One with responder analysis of FAS without imputation (a complete cases analysis (FAS-CC)) and one with responder analysis of FAS, where missing values will be replaced with values at one year follow-up, if available (FAS1-yearI).

Safety

Mean group-differences in 'Volume of blood loss and blood transfusions' and in proportions of 'Substantially deteriorated' according to the GPE scale, 'Complications and side effects' and 'Reoperations', will be evaluated in accordance to safety of the treatments.

Evaluation of costs

Although cost-utility analyses are not part of the efficacy study, the secondary outcomes 'Duration of operation' and 'Length of hospital stay' will be discussed in relation to costs of the treatments.

6.2 Managing of missing data

Although a strictly conducted study regarding routines for completing the follow-up questionnaires, some loss to follow-up is expected. Under the assumption of missingness at random (MAR), missing values necessary for estimating responder rates at two-year follow-up will be imputed by MI. The MAR assumption for patients not replying PROM-questionnaires is supported by a previous study from the Norwegian Registry for Spine Surgery [41]. Further, due to comprehensive set of available predictors for the imputation model, we consider the MI method robust regarding bias estimates. The imputation model, using linear regression, will include the following explanatory variables: Baseline patient characteristics (age; gender;

education; first language; smoking; body mass index; former spinal surgery; duration of pain; use of analgesics), radiological parameters at baseline (degree of the slip; segmental instability; Schizas grade; orientation of facet joint; disc height), operation time, length of hospital stay, baseline and follow-up scores for ODI, NRS leg pain, NRS back pain, Eq-5D, ZCQ, GPE, duration of surgery, length of hospital stay, complications, and reoperation. The imputation will be carried out stratified by treatment (i.e., by treatment group separately) [42]. The multiply imputing will be performed before dichotomising, as recommended [43]. It will be generated 50 data sets with complete two-year follow-up scores for ODI, ZCQ, NRS leg pain and NRS back pain. Before the responder analyses, which include the Newcombe hybrid score CI, the imputed scores will be estimated based on the 50 aggregated data sets.

Under the MAR assumption, the mean change and mean follow-up scores in continuous variables will be analysed by Linear mixed effect models, estimated with Full Information Maximum Likelihood.

<u>References</u>

1. Austevoll IM, Hermansen E, Fagerland M, Rekeland F, Solberg T, Storheim K, Brox JI, Lonne G, Indrekvam K, Aaen J, Grundnes O, Hellum C (2019) Decompression alone versus decompression with instrumental fusion the NORDSTEN degenerative spondylolisthesis trial (NORDSTEN-DS); study protocol for a randomized controlled trial. BMC musculoskeletal disorders 20:7. doi: 10.1186/s12891-018-2384-0
2. Farfan HF (2012) - The pathological anatomy of degenerative spondylolisthesis. A cadaver study. - Spine (Phila Pa 1976)1980 Sep-Oct;5(5):412-8:Oct
3. Fitzgerald JA, Newman PH (1976) Degenerative spondylolisthesis. The Journal of bone and joint surgery British volume 58:184-192
4. Watters WC, 3rd, Bono CM, Gilbert TJ, Kreiner DS, Mazanec DJ, Shaffer WO, Baisden J, Easa JE, Fernand R, Ghiselli G, Heggeness MH, Mendel RC, O'Neill C, Reitman CA, Resnick DK, Summers JT, Timmons RB, Toton JF, North American Spine S (2009) An evidence-based clinical guideline for the diagnosis and treatment of degenerative lumbar spondylolisthesis. The spine journal : official journal of the North American Spine Society 9:609-614. doi: 10.1016/j.spinee.2009.03.016
5. Martin CR, Gruszczynski AT, Braunsfurth HA, Fallatah SM, O'Neil J, Wai EK (2007) The surgical management of degenerative lumbar spondylolisthesis: a systematic review. Spine 32:1791-1798. doi: 10.1097/BRS.0b013e3180bc219e
6. Resnick DK, Watters WC, 3rd, Sharan A, Mummaneni PV, Dailey AT, Wang JC, Choudhri TF, Eck J, Ghogawala Z, Groff MW, Dhall SS, Kaiser MG (2014) Guideline update for the performance of fusion procedures for degenerative disease of the lumbar spine. Part 9: lumbar fusion for stenosis with spondylolisthesis. Journal of neurosurgery Spine 21:54-61. doi: 10.3171/2014.4.spine14274

7. Steiger F, Becker HJ, Standaert CJ, Balague F, Vader JP, Porchet F, Mannion AF (2014) Surgery in lumbar degenerative spondylolisthesis: indications, outcomes and complications. A systematic 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 23:945-973. doi: 10.1007/s00586-013-3144-3

8. Ghogawala Z, Dziura J, Butler WE, Dai F, Terrin N, Magge SN, Coumans JV, Harrington JF, Amin-Hanjani S, Schwartz JS, Sonntag VK, Barker FG, 2nd, Benzel EC (2016) Laminectomy plus Fusion versus Laminectomy Alone for Lumbar Spondylolisthesis. The New England journal of medicine 374:1424-1434. doi: 10.1056/NEJMoa1508788

9. Chang HS, Fujisawa N, Tsuchiya T, Oya S, Matsui T (2014) Degenerative spondylolisthesis does not affect the outcome of unilateral laminotomy with bilateral decompression in patients with lumbar stenosis. Spine 39:400-408. doi: 10.1097/BRS.0000000000000161

10. Forsth P, Michaelsson K, Sanden B (2013) Does fusion improve the outcome after decompressive surgery for lumbar spinal stenosis?: A two-year follow-up study involving 5390 patients. Bone Joint J 95-B:960-965. doi: 95-B/7/960 [pii];10.1302/0301-620X.95B7.30776 [doi]

11. Austevoll IM, Gjestad R, Brox JI, Solberg TK, Storheim K, Rekeland F, Hermansen E, Indrekvam K, Hellum C (2016) The effectiveness of decompression alone compared with additional fusion for lumbar spinal stenosis with degenerative spondylolisthesis: a pragmatic comparative non-inferiority observational study from the Norwegian Registry for Spine Surgery. European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society. doi: 10.1007/s00586-016-4683-1

12. Forsth P, Olafsson G, Carlsson T, Frost A, Borgstrom F, Fritzell P, Ohagen P, Michaelsson K, Sanden B (2016) A Randomized, Controlled Trial of Fusion Surgery for Lumbar Spinal Stenosis. The New England journal of medicine 374:1413-1423. doi: 10.1056/NEJMoa1513721

13. Agha RA, Altman DG, Rosin D (2015) The SPIRIT 2013 statement--defining standard protocol items for trials. International journal of surgery (London, England) 13:288-291. doi: 10.1016/j.ijsu.2014.12.007

14. Gamble C, Krishan A, Stocken D, Lewis S, Juszczak E, Dore C, Williamson PR, Altman DG, Montgomery A, Lim P, Berlin J, Senn S, Day S, Barbachano Y, Loder E (2017) Guidelines for the Content of Statistical Analysis Plans in Clinical Trials. Jama 318:2337-2343. doi: 10.1001/jama.2017.18556

15. Piaggio GF, - Elbourne DR FAU - Altman D, - Altman DG FAU - Pocock S, - Pocock SJ FAU - Evans S, Evans SJ (2010) - Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. - Jama2006 Mar 8;295(10):1152-60:60

16. ICH, Harmonised, Tripartite, Guideline (2014) Guideline for good clinical practice E6(R1). June 1996. http://wwwichorg/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6_R1/Step4/E6_R1__Guidelinepdf

17. Blackwelder WC, Chang MA (1984) Sample size graphs for "proving the null hypothesis". Controlled clinical trials 5:97-105

18. Deyo RA, Mirza SK, Martin BI, Kreuter W, Goodman DC, Jarvik JG (2010) Trends, major medical complications, and charges associated with surgery for lumbar spinal stenosis in older adults. Jama 303:1259-1265. doi: 10.1001/jama.2010.338

19. Kepler CK, Vaccaro AR, Hilibrand AS, Anderson DG, Rihn JA, Albert TJ, Radcliff KE (2014) National trends in the use of fusion techniques to treat degenerative spondylolisthesis. Spine 39:1584-1589. doi: 10.1097/brs.0000000000000486

20. Blumenthal S, McAfee PC, Guyer RD, Hochschuler SH, Geisler FH, Holt RT, Garcia R, Jr., Regan JJ, Ohnmeiss DD (2005) A prospective, randomized, multicenter Food and Drug Administration investigational device exemptions study of lumbar total disc replacement with the CHARITE artificial disc versus lumbar fusion: part I: evaluation of clinical outcomes. Spine 30:1565-1575; discussion E1387-1591

21. Derogatis LR, Lipman RS, Rickels K, Uhlenhuth EH, Covi L (1974) The Hopkins Symptom Checklist (HSCL): a self-report symptom inventory. Behavioral science 19:1-15

22. Dupuis PR, Yong-Hing K, Cassidy JD, Kirkaldy-Willis WH (1985) Radiologic diagnosis of degenerative lumbar spinal instability. Spine 10:262-276

23. Schizas C, Theumann N, Burn A, Tansey R, Wardlaw D, Smith FW, Kulik G (2010) Qualitative grading of severity of lumbar spinal stenosis based on the morphology of the dural sac on magnetic resonance images. Spine 35:1919-1924. doi: 10.1097/BRS.0b013e3181d359bd

24. Lee S, Lee JW, Yeom JS, Kim KJ, Kim HJ, Chung SK, Kang HS (2010) A practical MRI grading system for lumbar foraminal stenosis. AJR AmJRoentgenol 194:1095-1098. doi: 194/4/1095 [pii];10.2214/AJR.09.2772 [doi]

25. Cho IY, Park SY, Park JH, Suh SW, Lee SH (2017) MRI findings of lumbar spine instability in degenerative spondylolisthesis. Journal of orthopaedic surgery (Hong Kong) 25:2309499017718907. doi: 10.1177/2309499017718907

26. Berlemann UF, Jeszenszky DJ FAU, Buhler DW FAU, Harms J (2005) - Facet joint remodeling in degenerative spondylolisthesis: an investigation of joint orientation and tropism. - Eur Spine J1998;7(5):376-80:80

27. Pfirrmann CW, Metzdorf A, Zanetti M, Hodler J, Boos N (2001) Magnetic resonance classification of lumbar intervertebral disc degeneration. Spine (Phila Pa 1976) 26:1873-1878

28. Modic MT, Steinberg PM, Ross JS, Masaryk TJ, Carter JR (1988) Degenerative disk disease: assessment of changes in vertebral body marrow with MR imaging. Radiology 166:193-199. doi: 10.1148/radiology.166.1.3336678

29. Masharawi Y, Kjaer P, Bendix T, Manniche C, Wedderkopp N, Sorensen JS, Peled N, Jensen TS (2008) The reproducibility of quantitative measurements in lumbar magnetic resonance imaging of children from the general population. Spine 33:2094-2100. doi: 10.1097/BRS.0b013e31817f19f7

30. Schwab F, Lafage V, Patel A, Farcy JP (2009) Sagittal plane considerations and the pelvis in the adult patient. Spine 34:1828-1833. doi: 10.1097/BRS.0b013e3181a13c08

31. Fairbank JC, Couper J, Davies JB, O'Brien JP (1980) The Oswestry low back pain disability questionnaire. Physiotherapy 66:271-273

32. Baker DJ, Pynsent PB, J F (1989) The Oswestry disability index revisited: its reliability, repeatability and validity, and a comparison with the St Thomas's disability index. Roland MO, Jenner JR, eds New approaches to rehabilitation and education Manchester: Manchester University Press:174-186

33. Dworkin RH, Turk DC, Wyrwich KW, Beaton D, Cleeland CS, Farrar JT, Haythornthwaite JA, Jensen MP, Kerns RD, Ader DN, Brandenburg N, Burke LB, Cella D, Chandler J, Cowan P, Dimitrova R, Dionne R, Hertz S, Jadad AR, Katz NP, Kehlet H, Kramer LD, Manning DC, McCormick C, McDermott MP, McQuay HJ, Patel S, Porter L, Quessy S, Rappaport BA, Rauschkolb C, Revicki DA, Rothman M, Schmader KE, Stacey BR, Stauffer JW, von Stein T, White RE, Witter J, Zavisic S (2008) 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 9:105-121. doi: 10.1016/j.jpain.2007.09.005

34. Ostelo RW, Deyo RA, Stratford P, Waddell G, Croft P, Von Korff M, Bouter LM, de Vet HC (2008) Interpreting change scores for pain and functional status in low back pain: towards

international consensus regarding minimal important change. Spine 33:90-94. doi: 10.1097/BRS.0b013e31815e3a10

35. Austevoll IM, Gjestad R, Grotle M, Solberg T, Brox JI, Hermansen E, Rekeland F, Indrekvam K, Storheim K, Hellum C (2019) Follow-up score, change score or percentage change score for determining clinical important outcome following surgery? An observational study from the Norwegian registry for Spine surgery evaluating patient reported outcome measures in lumbar spinal stenosis and lumbar degenerative spondylolisthesis. BMC musculoskeletal disorders 20:31. doi: 10.1186/s12891-018-2386-y

36. Fagerland MW, Lydersen S, Laake P (2011) Recommended confidence intervals for two independent binomial proportions. StatMethods MedRes. doi: 0962280211415469 [pii];10.1177/0962280211415469 [doi]

37. Tuli SK, Yerby SA, Katz JN (2006) Methodological approaches to developing criteria for improvement in lumbar spinal stenosis surgery. Spine 31:1276-1280. doi: 10.1097/01.brs.0000217615.20018.6c

38. (1990) EuroQol--a new facility for the measurement of health-related quality of life. Health policy (Amsterdam, Netherlands) 16:199-208

39. Kamper SJ, Ostelo RW, Knol DL, Maher CG, de Vet HC, Hancock MJ (2010) Global Perceived Effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status. JClinEpidemiol 63:760-766. doi: S0895-4356(09)00304-7 [pii];10.1016/j.jclinepi.2009.09.009 [doi]

40. Rehal S, Morris TP, Fielding K, Carpenter JR, Phillips PP (2016) Non-inferiority trials: are they inferior? A systematic review of reporting in major medical journals. BMJ open 6:e012594. doi: 10.1136/bmjopen-2016-012594

41. Solberg TK, Sorlie A, Sjaavik K, Nygaard OP, Ingebrigtsen T (2011) Would loss to follow-up bias the outcome evaluation of patients operated for degenerative disorders of the lumbar spine? Acta Orthop 82:56-63. doi: 10.3109/17453674.2010.548024 [doi]

42. Yamaguchi Y, Ueno M, Maruo K, Gosho M (2020) Multiple imputation for longitudinal data in the presence of heteroscedasticity between treatment groups. Journal of biopharmaceutical statistics 30:178-196. doi: 10.1080/10543406.2019.1632878

43. Floden L, Bell ML (2019) Imputation strategies when a continuous outcome is to be dichotomized for responder analysis: a simulation study. BMC medical research methodology 19:161. doi: 10.1186/s12874-019-0793-x